UNIVERSITÀ
DEGLI STUDI
DI PADOVA

The way we were, the way we will be
Open Science: monitoring a change of scientific paradigm

Massimo Grassi
Department of General Psychology

Overview

- Why science is changing: the confidence/replication crisis
- The endogenous & exogenous origin of the crisis
- Responses to the endogenous & exogenous issues
- Open issues
- The science of the future

WHY SCIENCE IS CHANGING: THE CONFIDENCE/REPLICATION CRISIS

LET’S SUPPOSE YOU FELL ASLEEP IN 2011 AND OPEN YOUR EYES TODAY...

Empirical Article: Registered Report

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
2020, Vol. 3(1) 24–52

DOI: 10.1177/2515245919900809
www.psychologicalscience.org/AMPPS
SAGE

Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference

The ManyBabies Consortium*
*All consortium members are listed in the Transparency section at the end of the article.

Abstract

Psychological scientists have become increasingly concerned with issues related to methodology and replicability, and infancy researchers in particular face specific challenges related to replicability: For example, high-powered studies are difficult to conduct, testing conditions vary across labs, and different labs have access to different infant populations. Addressing these concerns, we report on a large-scale, multisite study aimed at (a) assessing the overall replicability of a single theoretically important phenomenon and (b) examining methodological, cultural, and developmental moderators. We focus on infants' preference for infant-directed speech (IDS) over adult-directed speech (ADS). Stimuli of mothers speaking to their infants and to an adult in North American English were created using seminaturalistic laboratory-based audio recordings. Infants' relative preference for IDS and ADS was assessed across 67 laboratories in North America, Europe, Australia, and Asia using the three common methods for measuring infants' discrimination (head-turn preference, central fixation, and eye tracking). The overall meta-analytic effect size (Cohen's *d*) was 0.35, 95% confidence interval = [0.29, 0.42], which was reliably above zero but smaller than the meta-analytic mean computed from previous literature (0.67). The IDS preference was significantly stronger in older children, in those children for whom the stimuli matched their native language and dialect, and in data from labs using the head-turn preference procedure. Together, these findings replicate the IDS preference but suggest that its magnitude is modulated by development, native-language experience, and testing procedure.

*The presumed preference of infants for high-pitched & sweet speech over standard speech

Open access publication: your library doesn't need to subscribe the journal!

Replication: not a novel finding!

Multilab: the joint effort of a research community 67 research units, more than 100 authors!

Digital materials available: you can re-run the experiment in 1 click!

Data and analysis scripts are available: you can re-run stats!

The hypothesis was publicly stated *before* collecting the data

0.0B Public 1

ManyBabies /

ManyBabies 1: Infant-Directed Speech Preference

Contributors: Christina Bergmann, Michael C. Frank, Nayeli Gonzalez, Elika Bergelson, Alejandrina Cristia, Brock Ferguson, Melissa Kline Struhl, Melanie Soderstrom, Daniel Yurovsky, Krista Byers-Heinlein, robin panneton, caroline floccia, Casey Lew-Williams, Kiley Hamlin, Mohinish Shukla, Heidi Baumgartner, Grace Zhou, Rodrigo Dal Ben

Date created: 2016-07-01 07:24 PM | Last Updated: 2022-09-12 09:11 PM

Category: Project

Description: *The first ManyBabies study*

Has supplemental materials for Quantifying sources of variability in infancy research using the infant-directed speech preference on PsyArXiv

Wiki

ManyBabies 1: Infant-Directed Speech Preference

This is the project page for the first study of the ManyBabies collaborative research project. It contains materials, details on the procedure, and instructions for participation / replication.

The main project was completed in 2019. The Stage 2 Registered Report of ManyBabies 1 has been accepted for publication at Advances in Methods and Practices ...

Read More

Files

Citation

Components

ManyBabies 1: START HERE
Kline Struhl, Bergmann, Frank & 11 more

ManyBabies 1: Project Admin
Kline Struhl, Bergmann, Frank & 11 more
Material for organizing MB1 as a whole

ManyBabies 1: Stimuli Creation and Norming

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

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract

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

2011: "Ordinary operations on the data enable to get significant results: always "

3839 citations (Scopus) for a method paper!

4362 citations (Scopus) for a replication study!

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.

Reproducibility is a core principle of scientific progress (1–6). Scientific claims should not gain credence because of the status or authority of their originator but ... replicability of their ... Scientists att... ... agree or dis... ...ence supports the claims, ... oretical or methodological reasons or by collecting new evidence. Such debates are meaningless, however, if the evidence being

results are false and therefo... Some empiric... ... ...inal results of ... only 11 and 25% of the ... cases, respectively (10, 11). These numbers are stunning but also difficult to interpret because no details are available about the studies, methodology, or results. With no transparency, the reasons for low reproducibility cannot be evaluated.

Other investigations point to practices and incentives that may inflate the likelihood of

facilitated each step of the process and maintained the protocol and project resources. Replication materials and data were required to be archived publicly in order to maximize transparency, accountability, and reproducibility of the project (https://osf.io/ezcuj).

In total, 100 replications were completed by 270 contributing authors. There were many different research designs and analysis strategies in the original research. Through consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs. Analyses converted results to a common effect size metric [correlation coefficient (r)] with confidence intervals (CIs). The units of analysis for inferences about reproducibility were the original and replication study effect sizes. The resulting open data set provides an initial estimate of the reproducibility of psychology and correlational data to support development of hypotheses about the causes of reproducibility.

Sampling frame ...

... tion ... and maxi- ... of the accumulated evi- ... simultaneously, to maintain high quality, within this sampling frame we matched individual replication projects with teams that had relevant interests and expertise. We pursued a quasi-random sample by defining the sampling frame as 2008 articles of three important psychology journals: *Psychological Science* (*PSCI*), *Journal of Personality and Social Psychology* (*JPSP*), and *Journal of Experimental Psychology: Learning, Memory, and Cognition* (*JEP: LMC*). The first is a premier outlet for all psy-

2015: "First, large empirical check of the literature

PLOS | BIOLOGY

META-RESEARCH ARTICLE

Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature

Denes Szucs[1]*, **John P. A. Ioannidis**[2]

1 Department of Psychology, University of Cambridge, Cambridge, United Kingdom, 2 Meta-Research Innovation Center at Stanford (METRICS) and Department of Medicine, Department of Health Research and Policy, and Department of Statistics, Stanford University, Stanford, California, United States of America

* ds377@cam.ac.uk

2017: numbers do not add up

We have empirically assessed the distribution of published effect sizes and estimated power by analyzing 26,841 statistical records from 3,801 cognitive neuroscience and psychology papers published recently. The reported median effect size was D = 0.93 (interquartile range: 0.64–1.46) for nominally statistically significant results and D = 0.24 (0.11–0.42) for nonsignificant results. Median power to detect small, medium, and large effects was 0.12, 0.44, and 0.73, reflecting no improvement through the past half-century. This is so because sample sizes have remained small. Assuming similar true effect sizes in both disciplines, power was lower in cognitive neuroscience than in psychology. Journal impact factors negatively correlated with power. Assuming a realistic range of prior probabilities for null hypotheses, false report probability is likely to exceed 50% for the whole literature. In light of our findings, the recently reported low replication success in psychology is realistic, and worse performance may be expected for cognitive neuroscience.

OPEN ACCESS

Citation: Szucs D, Ioannidis JPA (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol 15(3): e2000797. doi:10.1371/journal.pbio.2000797

Academic Editor: Eric-Jan Wagenmakers, University of Amsterdam, Netherlands

Received: August 10, 2016


How big is/was the iceberg? Nobody knows! Definitely, we now know how to do things better

THE ENDOGENOUS & EXOGENOUS ORIGIN OF THE CRISIS

Why it went so bad?

- The crisis was the result of two interacting processes [here described in series]:
  - Endogenous origins: problems due to the behaviour of the researcher
  - Exogenous origins: problems due to the environment where the researcher is working

Researcher-driven problems

- Questionable Research Practices (QRP):
  - p-hacking & HARKing
- Several malpractices [such as]:
  - Lack of transparency
  - Statistical issues
  - Assumptions never tested (eg replicability)

p-hacking

- p-hacking: look actively for a significant, positive result
- Examples:
  - Add/remove one or more data points
  - Add/remove one or more dependent variables
  - Add/remove one or more independent variables
  - Perform multiple analysis on the same data
  - …
  - Until “IT’S SIGNIFICANT!”

- Because degrees of freedom are many, the chances to observe a significant result increase, but this result it’s likely to be a Type I error*

*if you throw a dice multiple times, you’ll eventually score “6”

HARKing

- Hypothesizing After the Results are Known [write your hypothesis after you have seen your results]
- Fanelli (2010): “About 90% of published papers show results that are coherent with the hypothesis reported in the paper*”

*If this success rate was realistic, we should ask why we do empirical research altogether!

Empirical Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2, pp. 1–12

DOI: 10.1177/25152459211007467
www.psychologicalscience.org/AMPPS
SAGE

Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens
Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Selectively publishing results that support the tested hypotheses (“positive” results) distorts the available evidence for scientific claims. For the past decade, psychological scientists have been increasingly concerned about the degree of such distortion in their literature. A new publication format has been developed to prevent selective reporting: In Registered Reports (RRs), peer review and the decision to publish take place before results are known. We compared the results in published RRs ($N = 71$ as of November 2018) with a random sample of hypothesis-testing studies from the standard literature ($N = 152$) in psychology. Analyzing the first hypothesis of each article, we found 96% positive results in standard reports but only 44% positive results in RRs. We discuss possible explanations for this large difference and suggest that a plausible factor is the reduction of publication bias and/or Type I error inflation in the RR literature.

ROYAL SOCIETY OPEN SCIENCE
royalsocietypublishing.org/journal/rsos

Review

Cite this article: Haeffel GJ. 2022 Psychology needs to get tired of winning. *R. Soc. Open Sci.* **9**: 220099.
https://doi.org/10.1098/rsos.220099

Received: 28 March 2022
Accepted: 6 June 2022

Subject Category:
Psychology and cognitive neuroscience

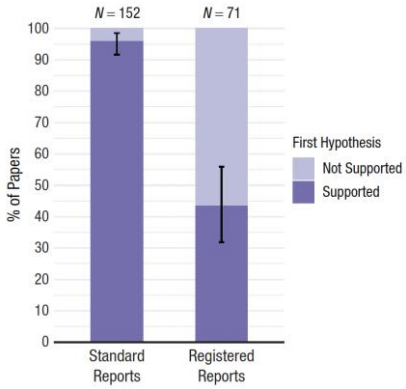
Subject Areas:
psychology

Keywords:
falsification, theory, predictions, philosophy of science, registered reports

Psychology needs to get tired of winning

Gerald J. Haeffel
Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, USA
GJH, 0000-0002-4029-1493

Psychological science is on an extraordinary winning streak. A review of the published literature shows that nearly all study hypotheses are supported. This means that either all the theories are correct, or the literature is biased towards positive findings. Results from large-scale replication projects and the prevalence of questionable research practices indicate the latter. This is a problem because science progresses from being wrong. For decades, there have been calls for better theories and the adoption of a strong inference approach to science. However, there is little reason to believe that psychological science is ready to change. Although recent developments like the open science movement have improved transparency and replicability, they have not addressed psychological science's method-oriented (rather than problem-oriented) mindset. Psychological science still does not embrace the scientific method of developing theories, conducting critical tests of those theories, detecting contradictory results, and revising (or disposing of) the theories accordingly. In this article, I review why psychologists must embrace being wrong and how the Registered Report format might be one strategy for stopping psychology's winning streak.



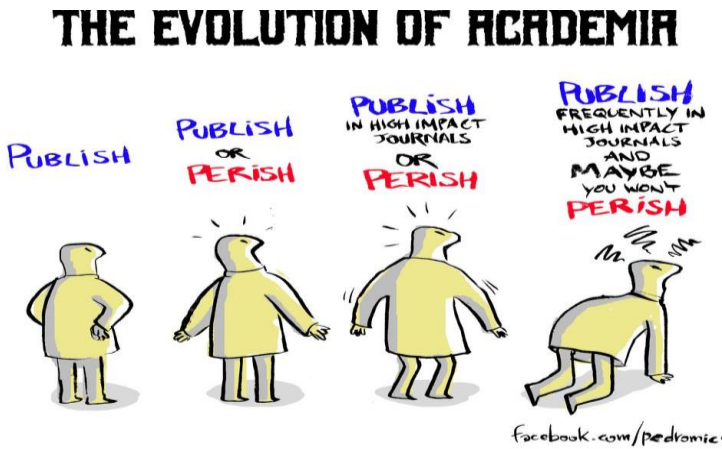
Other malpractices

- Lack of transparency:
  - studies came with no supporting materials: difficult to check the quality of papers
- Statistical issues:
  - Small N & low power (modal N=15, Balàzs et al. 2018)
  - Probabilistic results interpreted as true or false
- Assumptions never tested:
  - “psychology is replicable!” (but psychologists were not running direct replications to test it)

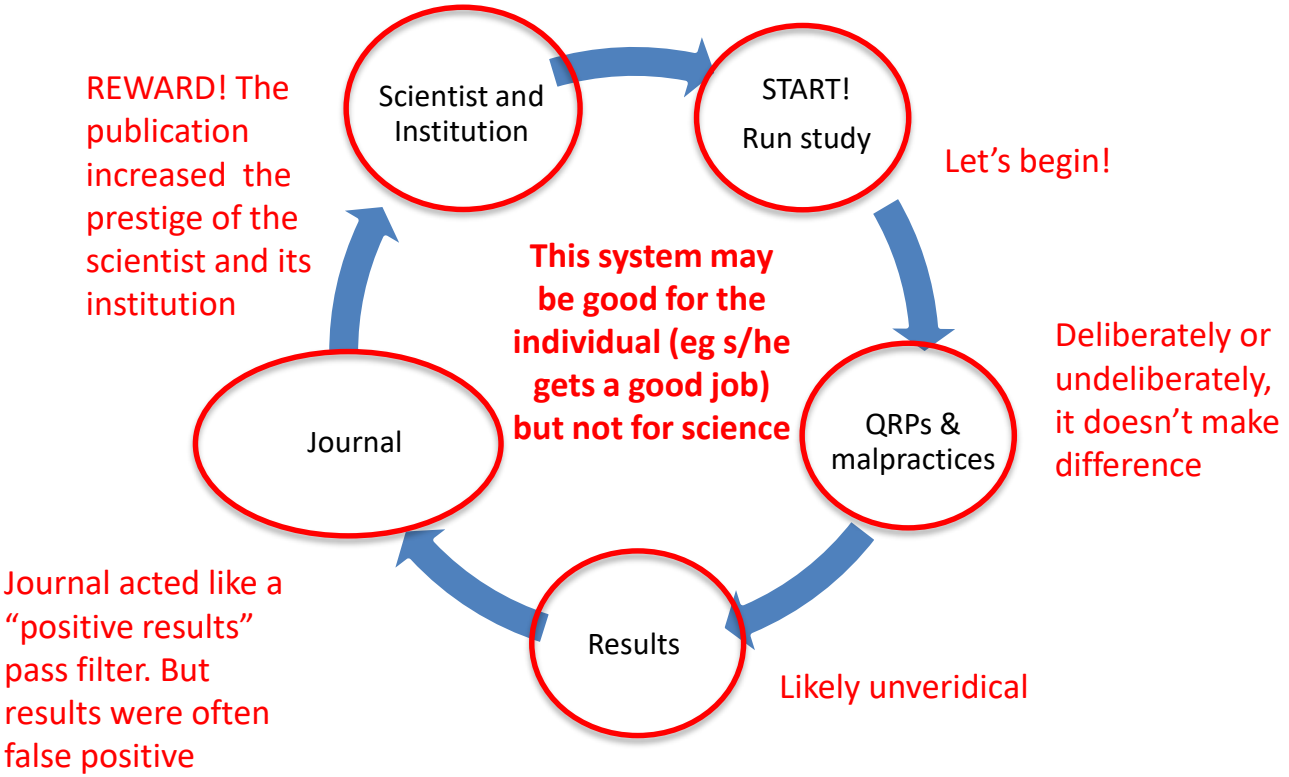
Environment-driven problems*

- Publication bias:
  - ie: journals were publishing mainly positive results
- Recruiting & funding system
  - Publication is the currency of academia: publish or perish!

*In reality, there is not such a thing as ”the environment”: scientists have a active role in this when they act as “editors”, “reviewers”, “member of evaluation panel”, “head of department/uni”



Recap: why old science was failing?



Open Science

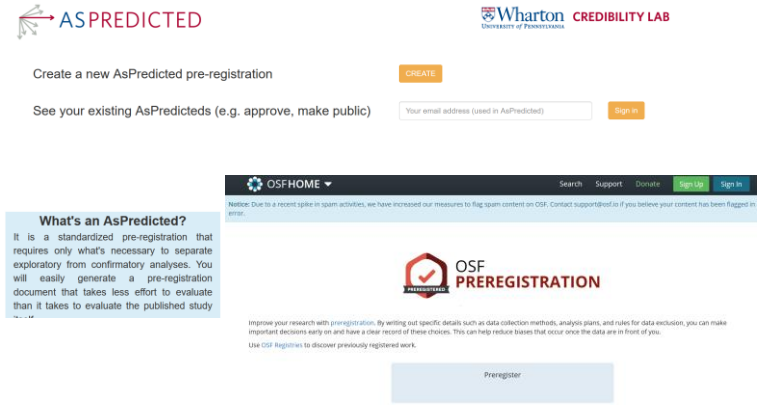
- In the last decade, several scientists worked with one goal in mind: improve quality and replicability of science
- In many cases, scientists acted by increasing the transparency of science. For this reason this movement falls within the large label “Open Science”

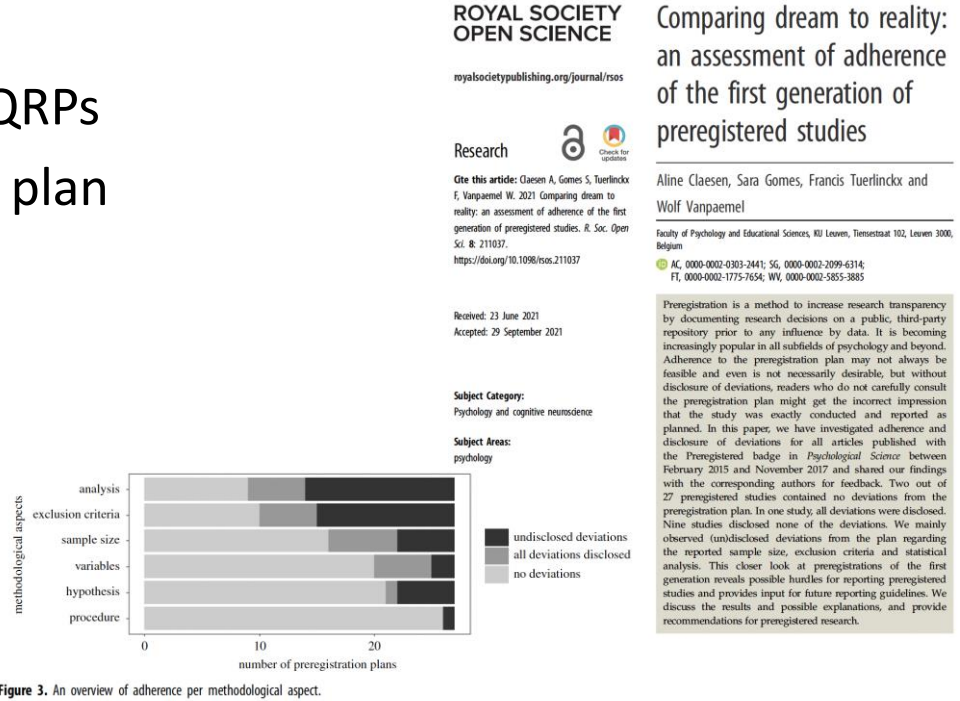
Are things changing?

RESPONSES TO THE ENDOGENOUS PROBLEMS

QRPs & HARKing -> Preregistration

- You state your hypothesis and analysis plan *before* collecting the data
  - Pro: increases awareness on QRPs
  - Con: easy to deviate from the plan





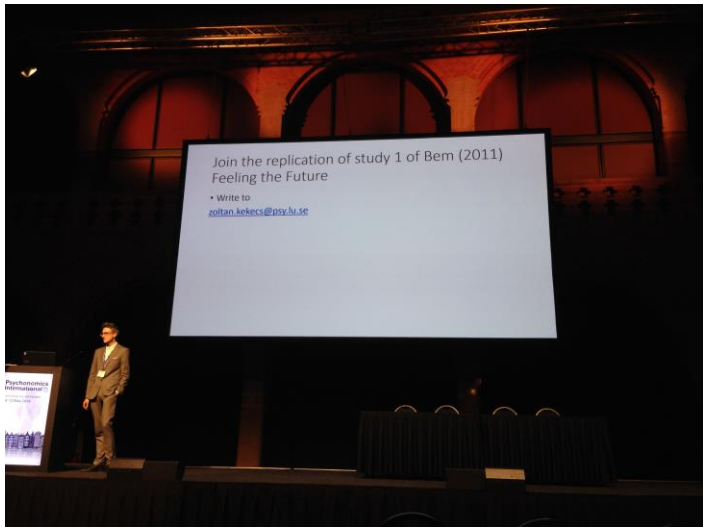
QRPs & HARKing -> Registered reports

- You discuss the study with reviewers and collect data after they say “GO!”
  - Pro: All
  - Con: None (except for publishers that may inflate journals with null findings)

Lack of transparency -> open data & materials

- Sharing data is becoming an ordinary practice, mandatory in many journals
  - However, journals often do not double check if data are available nor whether they are actually usable
- In addition, we are now asked to share (sometimes):
  - analysis script (eg, R-script)
  - scripts and digital materials that enabled to conduct the study

Statistical issues [small N] -> multilab*



Psychological Science Accelerator

a distributed laboratory network

HOME MAP PEOPLE PROJECT TRACKER STUDY SELECTION PROCESS TRANSLATION PROCESS MEDIA BLOG GET INVOLVED

The Psychological Science Accelerator is a globally distributed network of psychological science laboratories (currently over 300), representing over 45 countries on all six populated continents, that coordinates data collection for democratically selected studies.

Recent Posts

The Psychological Science Accelerator's First Year August

Article

Reproducible brain-wide association studies require thousands of individuals

https://doi.org/10.1038/s41586-022-04492-9
Received: 20 August 2020
Accepted: 31 January 2022
Published online: 16 March 2022
Check for updates

Scott Marek[1,39], Brenden Tervo-Clemmens[2,3,39], Finnegan J. Calabro[4,5], David F. Montez[6], Benjamin P. Kay[6], Alexander S. Hatoum[1], Meghan Rose Donohue[1], William Foran[4], Ryland L. Miller[1,5], Timothy J. Hendrickson[7], Stephen M. Malone[8], Sridhar Kandala[1], Eric Feczko[9,10], Oscar Miranda-Dominguez[9,10], Alice M. Graham[9], Eric A. Earl[6,11], Anders J. Perrone[6,11], Michaela Cordova[11], Olivia Doyle[11], Lucille A. Moore[12], Gregory M. Conan[6,11], Johnny Uriarte[13], Kathy Snider[11], Benjamin J. Lynch[6,11], James C. Wilgenbusch[6,42], Thomas Pengo[7], Angela Tam[13,14,15,16], Jianzhong Chen[13,14,15,16], Dillan J. Newbold[6], Annie Zheng[6], Nicole A. Seider[6], Andrew N. Van[6,17], Athanasia Metoki[6], Roselyne J. Chauvin[6], Timothy O. Laumann[1], Deanna J. Greene[18], Steven E. Petersen[6,17,19,20,21], Hugh Garavan[22], Wesley K. Thompson[23], Thomas E. Nichols[24], B. T. Thomas Yeo[13,14,15,16,25,26], Deanna M. Barch[1,27], Beatriz Luna[3,4], Damien A. Fair[6,10,27,28] ✉ & Nico U. F. Dosenbach[6,17,19,20,29,30] ✉

Magnetic resonance imaging (MRI) has transformed our understanding of the human brain through well-replicated mapping of abilities to specific structures (for example, lesion studies) and functions[1–3] (for example, task functional MRI (fMRI)). Mental health research and care have yet to realize similar advances from MRI. A primary challenge has been replicating associations between inter-individual differences in brain structure or function and complex cognitive or mental health phenotypes (brain-wide association studies (BWAS)). Such BWAS have typically relied on sample

***the topics investigated by multilab often emerge from a discussion within the relevant community**

Statistical issues [bad stat interpretation] -> effect size and & Bayes stats

- Journals now encourage to discuss the results in terms of effect size and drop the usual “it is significant/it’s not significant” binary descriptions
  - However, in many cases effect sizes are reported but not discussed
- Bayesian stats force to drop binary description of results

IMO: contemporary research groups need to include/hire a highly trained 'data analyst'

Statistical issues [bad method in general] -> internal methodological boards to filter out bad research at the source

nature

Explore content | About the journal | Publish with us | Subscribe

nature > world view > article

WORLD VIEW | 03 January 2023

Is my study useless? Why researchers need methodological review boards

Making researchers account for their methods before data collection is a long-overdue step.

Daniël Lakens

Should researchers have the freedom to perform research that is a waste of time? Currently, the answer is a resounding 'yes'. Or at least, no one stops to ask whether there are obvious methodological and statistical flaws in a proposed study that will make it useless from the get-go: a sample size that's simply too small to test a hypothesis, for example.

In my role as chair of the central ethical review board at Eindhoven University of Technology in the Netherlands, I've lost count of the number of times that a board member has remarked that, although we're not supposed to comment on non-ethical issues, the way a study has been designed means it won't yield any informative data. And yet we routinely wait until peer review — after the study has been done — to identify flaws that can't then be corrected.

In my own department at Eindhoven, we've been trialling a different approach. Five years ago, we instituted a local review board that also evaluates proposed methods. Although some colleagues found this extra hurdle frustrating at first, the improvements in study quality have led them to accept it. It's time to make dedicated methodological review boards a standard feature at universities and other research institutions, as institutional review boards are.

Lack of replications -> replications now exist!

BEHAVIORAL AND BRAIN SCIENCES (2018), Page 1 of 61
doi:10.1017/S0140525X17001972, e120

Making replication mainstream

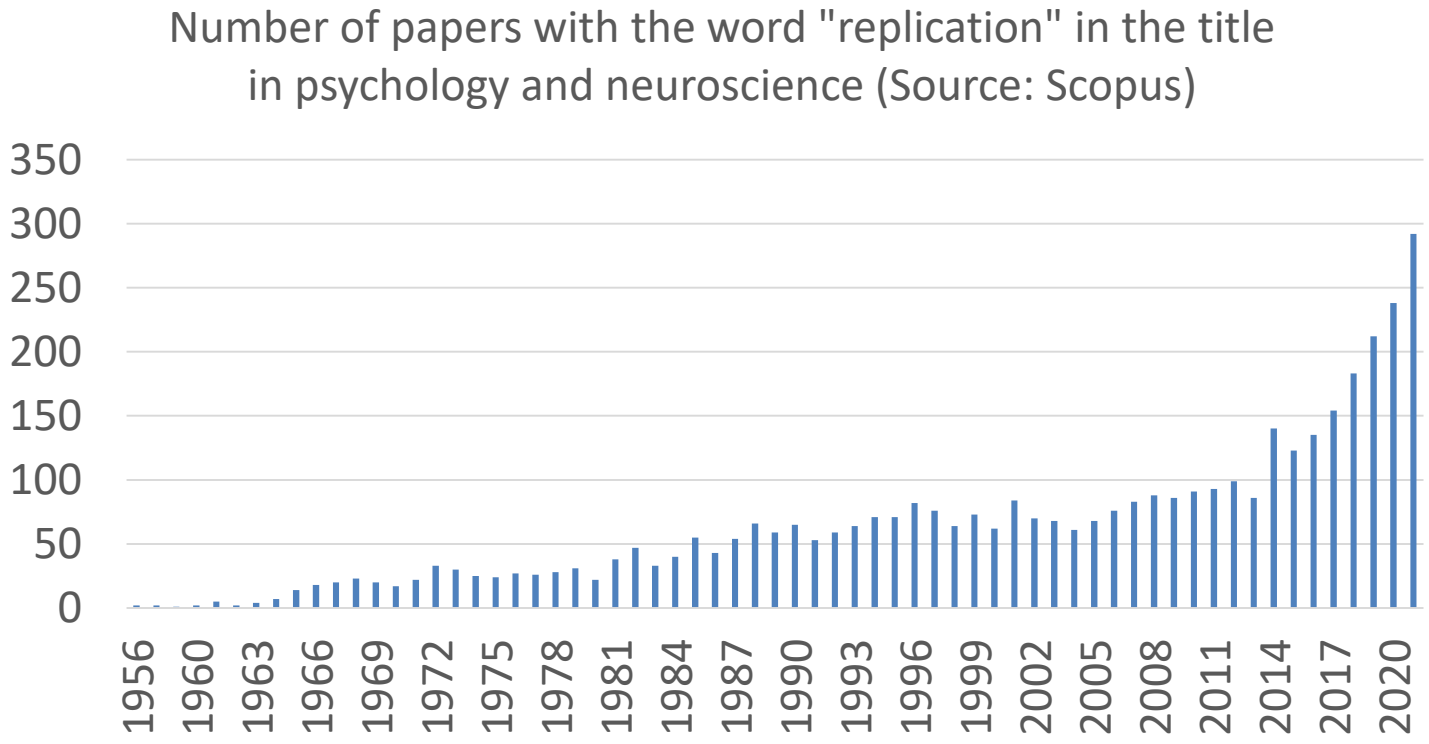
Rolf A. Zwaan
Department of Psychology, Education, and Child Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands
zwaan@essb.eur.nl
https://www.eur.nl/essb/people/rolf-zwaan

Alexander Etz
Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100.
etz.alexander@gmail.com
https://alexanderetz.com/

Richard E. Lucas
Department of Psychology, Michigan State University, East Lansing, MI 48824
lucasr@msu.edu
https://www.msu.edu/user/lucasr//

M. Brent Donnellan[1]
Department of Psychology, Texas A&M University, College Station, TX 77843
donnel59@msu.edu
https://psychology.msu.edu/people/faculty/donnel59

Abstract: Many philosophers of science and methodologists have argued that the ability to repeat studies and obtain similar results is an essential component of science. A finding is elevated from single observation to scientific evidence when the procedures that were used to obtain it can be reproduced and the finding itself can be replicated. Recent replication attempts show that some high profile results – most notably in psychology, but in many other disciplines as well – cannot be replicated consistently. These replication attempts have generated a considerable amount of controversy, and the issue of whether direct replications have value has, in particular, proven to be contentious. However, much of this discussion has occurred in published commentaries and social media outlets, resulting in a fragmented discourse. To address the need for an integrative summary, we review various types of replication studies and then discuss the most commonly voiced concerns about direct replication. We provide detailed responses to these concerns and consider different statistical ways to evaluate replications. We conclude there are no theoretical or statistical obstacles to making direct replication a routine aspect of psychological science.



But replications revealed a lack of “replication culture”

- We often “take it personally” and we still don’t know what “replication” means!

frontiers in Psychology | Quantitative Psychology and Measurement

SECTION ABOUT ARTICLES RESEARCH TOPICS FOR AUTHORS EDITORIAL BOARD ARTICLE ALERTS

‹ Articles

THIS ARTICLE IS PART OF THE RESEARCH TOPIC
Epistemological and Ethical Aspects of Research in the Social Sciences View all 12 Articles ›

CONCEPTUAL ANALYSIS ARTICLE

Front. Psychol., 04 September 2019 | https://doi.org/10.3389/fpsyg.2019.01884

Confounds in “Failed” Replications

Paola Bressan*

Dipartimento di Psicologia Generale, University of Padova, Padova, Italy

Reproducibility is essential to science, yet a distressingly large number of research findings do not seem to replicate. Here I discuss one underappreciated reason for this state of affairs. I make my case by noting that, due to artifacts, several of the replication failures of the vastly advertised Open Science Collaboration's *Reproducibility Project: Psychology* turned out to be invalid. Although these artifacts would have been obvious on perusal of the data, such perusal was deemed undesirable because of its *post hoc* nature and was left out. However, while data do not lie, unforeseen confounds can render them unable to speak to the question of interest. I look further into one unusual case in which a major artifact could be removed statistically—the nonreplication of the effect of fertility on partnered women's preference for single over attached men. I show that the “failed replication” datasets contain a gross bias in stimulus allocation which is absent in the original dataset; controlling for it replicates the original study's main finding. I conclude that, before being used to make a scientific point, all data should undergo a minimal quality control—a provision, it appears, not always required of those collected for purpose of replication. Because unexpected confounds and biases can be laid bare only after the fact, we must get over our understandable reluctance to engage in anything *post hoc*. The reproach attached to *p*-hacking cannot exempt us from the obligation to (openly) take a good look at our data.

Examine [the data] from every angle.

Daryl J. Bem (1987, p. 172)

Download Article

1,75 TOTAL VIEWS

View Article

Suggest a Research

SHARE ON

Replicator degrees of freedom allow publication of misleading failures to replicate

Christopher J. Bryan[a,1], David S. Yeager[b], and Joseph M. O’Brien[b]

[a]Booth School of Business, University of Chicago, Chicago, IL 60637; and [b]Department of Psychology, University of Texas at Austin, Austin, TX 78712

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved October 22, 2019 (received for review June 28, 2019)

In recent years, the field of psychology has begun to conduct replication tests on a large scale. Here, we show that “replicator degrees of freedom” make it far too easy to obtain and publish false-negative replication results, even while appearing to adhere to strict methodological standards. Specifically, using data from an ongoing debate, we show that commonly exercised flexibility at the experimental design and data analysis stages of replication testing can make it appear that a finding was not replicated when, in fact, it was. The debate that we focus on is representative, on key dimensions, of a large number of other replication tests in psychology that have been published in recent years, suggesting that the lessons of this analysis may be far reaching. The problems with current practice in replication science that we uncover here are particularly worrisome because they are not adequately addressed by the field’s standard remedies, including preregistration. Implications for how the field could develop more effective methodological standards for replication are discussed.

replication crisis | reproducibility | p-hacking | researcher degrees of freedom | null hacking

they could have an ironic and counterproductive effect: trading one sort of misleading research finding (false-positive original findings) for another (false-negative replication results). This is a bad trade because the latter sort of misleading finding undoes the field’s hard-won progress toward improved scientific understanding.

Others have already made versions of the 2 general methodological points that we make here: that empirical conclusions often hinge on analytic choices that competent investigators can disagree about and that replication tests that deviate from the design of the original study in material ways can create the misleading impression that the original finding was a false positive (19–25). Here, we provide an analysis of one prominent ongoing replication debate that demonstrates, concretely and directly, the implications of these 2 methodological principles for the field’s interpretation of the many ostensible failures to replicate that are already in the literature and for how replication tests should be conducted going forward.

The failure of many replication tests to adequately recreate important design elements of the original studies in question is perhaps the most widely discussed point of disagreement about

New Results

1 comment

Variability in the analysis of a single neuroimaging dataset by many teams

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, Alison Adcock, Paolo Avesani, Blazej Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland Benoit, Ruud Berkers, Jamil Bhanji, Bharat Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine Bottenhorn, Alexander Bowring, Senne Braem, Hayley Brooks, Emily Brudner, Cristian Calderon, Julia Camilleri, Jaime Castrellon, Luca Cecchetti, Edna Cieslik, Zachary Cole, Olivier Collignon, Robert Cox, William Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles Davis, Alberto De Luca, Mauricio Delgado, Lysia Demetriou, Jeffrey Dennison, Xin Di, Erin Dickie, Ekaterina Dobryakova, Claire Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Adriana Galvan, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle Goeman, Sergej Golowin, Carlos González-García, Krzysztof Gorgolewski, Cheryl Grady, Mikella Green, João Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott Huettel, Matthew Hughes, Vittorio Iacovella, Alexandru Iordan, Peder Isager, Ayse Ilkay Isik, Andrew Jahn, Matthew Johnson, Tom Johnstone, Michael Joseph, Anthony Juliano, Joseph Kable, Michalis Kassinopoulos, Cemal Koba, Xiang Kong, Nuri Erkut Kucukboyaci, Brice Kuhl, Sebastian Kupek, Angela Laird, Claus Lamm, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Flora Li, Monica Li, Phui Cheng Lim, Evan Linte, Geoffrey Lowe, Bradley Love, Michael Mack, Norberto Maggioni, Joseph McGuire, Helena Mehl, Georgios Mitsis, Adrian C. [illegible] Pischedda, [illegible] Richard Reynolds, [illegible] Romyn, Taylor Salo, [illegible] Schlichting, Douglas Schultz, Qiang Shen, [illegible], Kenny Skagerlund, Alec Smith, David Smith, [illegible], Sarah Tashjian, Bertrand Thirion, John Thorp, Gustav Tinghög, [illegible]son, Claudio Toro-Serey, Juan Torre, Leonardo Tozzi, Vuong Truong, [illegible] E. van't Veer, Tom Verguts, Jean Vettel, Sagana Vijayarajah, Khoi Vo, Matthew Wall, [illegible] D. Weeda, Susanne Weis, David White, David Wisniewski, Alba Xifra-Porxas, Emily Yearling, Sangsuk Yoon, Rui Yuan, Kenneth Yuen, Lei Zhang, Xu Zhang, Joshua Zosky, Thomas E. Nichols, Russell A. Poldrack, Tom Schonberg

doi: https://doi.org/10.1101/843193

Subject Area

Neuroscience

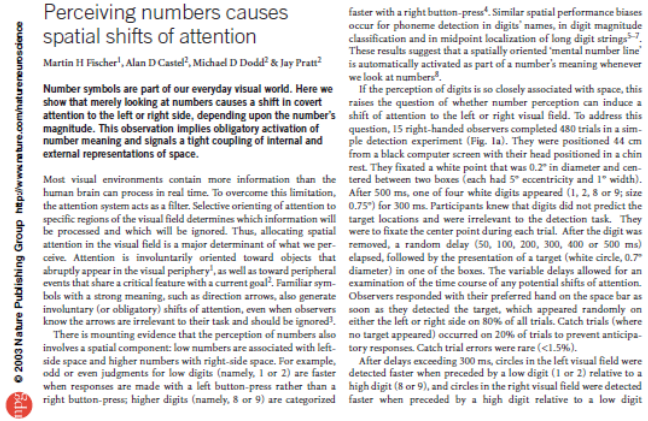
[illegible]vior and Cognition
Biochemistry
Bioengineering
Bioinformatics
Biophysics
Cancer Biology
Cell Biology
Clinical Trials*
Developmental Biology
Ecology
Epidemiology*
Evolutionary Biology
Genetics
Genomics
Immunology

Neuroimaging data: same dataset, 70 teams analysing with their own pipeline, different (in many cases sizeable) results!

NB: replications may cost incredible efforts!

Original study

- Fischer et al. (2003)
  - **15 participants**



Perceiving numbers causes spatial shifts of attention

Martin H Fischer[1], Alan D Castel[2], Michael D Dodd[2] & Jay Pratt[2]

Number symbols are part of our everyday visual world. Here we show that merely looking at numbers causes a shift in covert attention to the left or right side, depending upon the number's magnitude. This observation implies obligatory activation of number meaning and signals a tight coupling of internal and external representations of space.

Most visual environments contain more information than the human brain can process in real time. To overcome this limitation, the attention system acts as a filter. Selective orienting of attention to specific regions of the visual field determines which information will be processed and which will be ignored. Thus, allocating spatial attention in the visual field is a major determinant of what we perceive. Attention is involuntarily oriented toward objects that abruptly appear in the visual periphery[1], as well as toward peripheral events that share a critical feature with a current goal[2]. Familiar symbols with a strong meaning, such as direction arrows, also generate involuntary (or obligatory) shifts of attention, even when observers know the arrows are irrelevant to their task and should be ignored[3].

There is mounting evidence that the perception of numbers also involves a spatial component: low numbers are associated with left-side space and higher numbers with right-side space. For example, odd or even judgments for low digits (namely, 1 or 2) are faster when responses are made with a left button-press rather than a right button-press; higher digits (namely, 8 or 9) are categorized faster with a right button-press[4]. Similar spatial performance biases occur for phoneme detection in digits' names, in digit magnitude classification and in midpoint localization of long digit strings[5–7]. These results suggest that a spatially oriented 'mental number line' is automatically activated as part of a number's meaning whenever we look at numbers[8].

If the perception of digits is so closely associated with space, this raises the question of whether number perception can induce a shift of attention to the left or right visual field. To address this question, 15 right-handed observers completed 480 trials in a simple detection experiment (Fig. 1a). They were positioned 44 cm from a black computer screen with their head positioned in a chin rest. They fixated a white point that was 0.2° in diameter and centered between two boxes (each had 5° eccentricity and 1° width). After 500 ms, one of four white digits appeared (1, 2, 8 or 9; size 0.75°) for 300 ms. Participants knew that digits did not predict the target locations and were irrelevant to the detection task. They were to fixate the center point during each trial. After the digit was removed, a random delay (50, 100, 200, 300, 400 or 500 ms) elapsed, followed by the presentation of a target (white circle, 0.7° diameter) in one of the boxes. The variable delays allowed for an examination of the time course of any potential shifts of attention. Observers responded with their preferred hand on the space bar as soon as they detected the target, which appeared randomly on either the left or right side on 80% of all trials. Catch trials (where no target appeared) occurred on 20% of trials to prevent anticipatory responses. Catch trial errors were rare (<1.5%).

After delays exceeding 300 ms, circles in the left visual field were detected faster when preceded by a low digit (1 or 2) relative to a high digit (8 or 9), and circles in the right visual field were detected faster when preceded by a high digit relative to a low digit

Replication study

- Colling et al. (2020)
  - **17 labs & 1105 participants**

Registered Replication Report

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003)

Advances in Methods and Practices in Psychological Science
1–20
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920903079
www.psychologicalscience.org/AMPPS
SAGE

Lincoln J. Colling*, Dénes Szűcs*, Damiano De Marco, Krzysztof Cipora, Rolf Ulrich, Hans-Christoph Nuerk, Mojtaba Soltanlou, Donna Bryce, Sau-Chin Chen, Philipp Alexander Schroeder, Dion T. Henare, Christine K. Chrystall, Paul M. Corballis, Daniel Ansari, Celia Goffin, H. Moriah Sokolowski, Peter J. B. Hancock, Ailsa E. Millen, Stephen R. H. Langton, Kevin J. Holmes, Mark S. Saviano, Tia A. Tummino, Oliver Lindemann, Rolf A. Zwaan, Jiří Lukavský, Adéla Becková, Marek A. Vranka, Simone Cutini, Irene Cristina Mammarella, Claudio Mulatti, Raoul Bell, Axel Buchner, Laura Mieth, Jan Philipp Röer, Elise Klein, Stefan Huber, Korbinian Moeller, Brenda Ocampo, Juan Lupiáñez, Javier Ortiz-Tudela, Juanma de la Fuente, Julio Santiago, Marc Ouellet, Edward M. Hubbard, Elizabeth Y. Toomarian, Remo Job, Barbara Treccani, and Blakeley B. McShane*
*Lead authors

Multilab direct replication of: Experiment 2 from Fischer, M. H., Castel, A. D., Dodd, M. D., & Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. *Nature Neuroscience, 6*, 555–556. doi:10.1038/nn1066

Protocol vetted by: Martin H. Fischer

PLOS ONE

NB: and it is difficult to debunk a myth-result!

RESEARCH ARTICLE

Affirmative citation bias in scientific myth debunking: A three-in-one case study

Kåre Letrud[1]*, **Sigbjørn Hernes**[2]

1 Inland School of Business and Social Sciences, Inland Norway University of Applied Sciences, Lillehammer, Norway, **2** Lillehammer Campus Library, Inland Norway University of Applied Sciences, Lillehammer, Norway

* kare.letrud@inn.no

OPEN ACCESS

Abstract

Several uncorroborated, false, or misinterpreted conceptions have for years been widely distributed in academic publications, thus becoming scientific myths. How can such misconceptions persist and proliferate within the inimical environment of academic criticism? Examining 613 articles we demonstrate that the reception of three myth-exposing publications is skewed by an 'affirmative citation bias': The vast majority of articles citing the critical article will affirm the idea criticized. 468 affirmed the myth, 105 were neutral, while 40 took a negative stance. Once misconceptions proliferate wide and long enough, criticizing them not only becomes increasingly difficult, efforts may even contribute to the continued spreading of the myths.

Are things changing?

RESPONSES THE EXOGENOUS PROBLEMS

Journals: Open Access & publication bias

- In 2010 we switched from the traditional publishing scheme to OA publishing scheme
- The switch from traditional journals to OA has reduced the publication bias: OA journals often stress method over results

Traditional publication: stress on findings

Current Biology

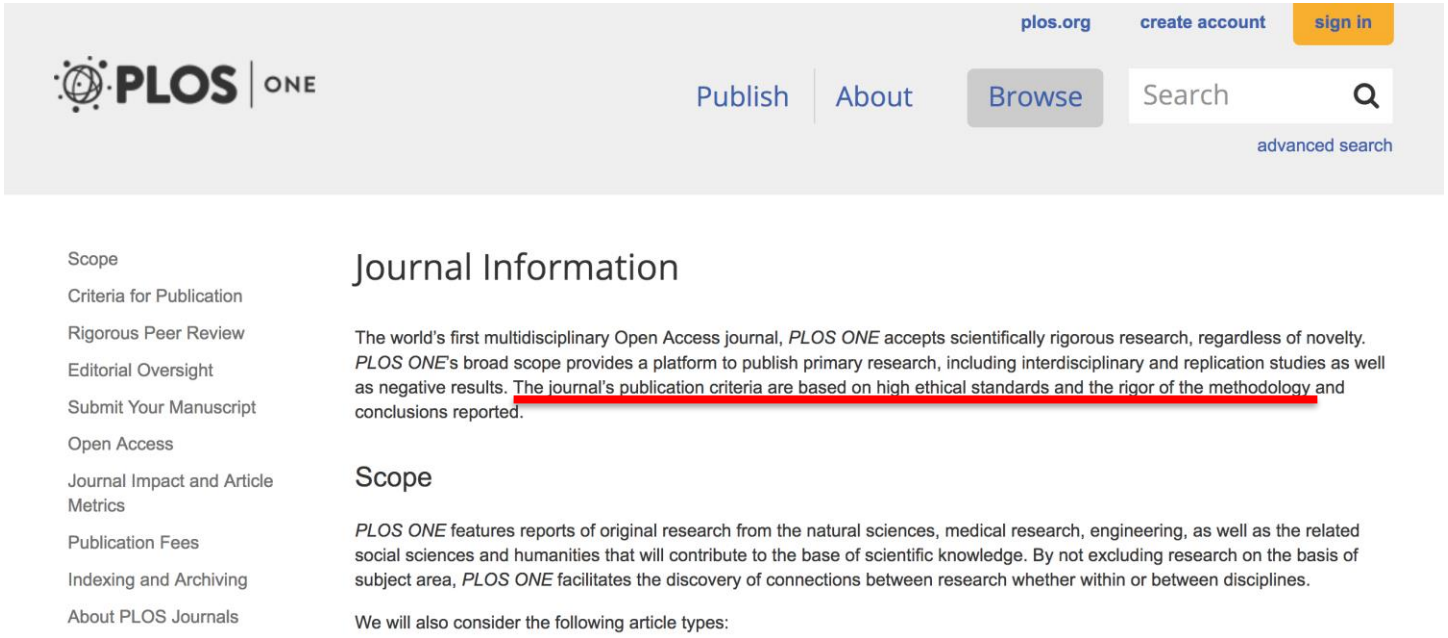
Aims and Scope

Important findings only

Current Biology is a general journal that publishes original research across all areas of biology together with an extensive and varied set of editorial sections. A primary aim of the journal is to foster communication across fields of biology, both by publishing important findings of general interest from diverse fields and through highly accessible editorial articles that explicitly aim to inform non-specialists.

Current Biology publishes papers reporting findings in any area of biology that have sufficient claim to be of general interest—this could be, for example, because the advance is important for a specific field, or because it is intrinsically of wide interest to biologists generally. We have several formats for publishing original research (Articles, Reports, and Correspondences); see our Information for Authors for details.

OA publication:
stress on method



Emphasis on method

Some journals now explicitly accept null results

NOW OPEN FOR SUBMISSIONS

Access: Full Open access

ISSN: 2516-712X (Online)

Editorial Board

Experimental Results is an open access journal providing a forum for experimental findings that disclose the small incremental steps vitally important to experimental research; experiments and findings which have so far remained hidden. Such results often go unpublished due to the traditional scholarly communication process, in which only a select group of experiments are chosen to make up the narrative of a single paper. Articles for consideration in *Experimental Results* include validation and reproducibility of existing findings, null results, supplementary findings, improvements or amendments to published results, as well as results that could be of importance, but for whatever reason, the researcher has not followed a particular line of questioning to produce a full narrative for a traditional paper. Where applicable, work published in *Experimental Results* will clearly link back to the related article. *Experimental Results* will publish short research papers from experimental disciplines across Science, Technology and Medicine, providing authors with an outlet for rapid publication of small chunks of research findings with maximum visibility. Articles will be accepted for publication if they are technically and methodologically sound and if the research reported answers a valid research question.

Open

MEET THE EDITO BOARD

Some journals have dropped the accept/reject decision

eLife

EDITORIAL

SCIENTIFIC PUBLISHING

Peer review without gatekeeping

eLife is changing its editorial process to emphasize public reviews and assessments of preprints by eliminating accept/reject decisions after peer review.

MICHAEL B EISEN, ANNA AKHMANOVA, TIMOTHY E BEHRENS, JÖRN DIEDRICHSEN, DIANE M HARPER, MIHAELA D IORDANOVA, DETLEF WEIGEL AND MONE ZAIDI

Last year eLife began exclusively reviewing papers already published as preprints and asking our reviewers to write public versions of their peer reviews containing observations useful to readers (***Eisen et al., 2020***). Over the past 18 months we have posted eLife reviews of more than 2,200 preprints to bioRxiv and medRxiv, along with a compact editorial assessment of the significance of the findings and the strength of the evidence for them.

We have found that these public preprint reviews and assessments are far more effective than binary accept or reject decisions ever could be at conveying the thinking of our reviewers and editors, and capturing the nuanced, multi-dimensional, and often ambiguous nature of peer review. eLife will now let them stand on their own by publishing every paper we review, along with our reviews and an assessment as a Reviewed Preprint, a new type of research output we hope will become the norm across science.

OA = equality?

A personal take on science and society

World view

By Tony Ross-Hellauer

Open science, done wrong, will compound inequities

Research-reform advocates must beware unintended consequences.

Ten years ago, as a new PhD graduate looking for my next position, I found myself in the academic cold. Nothing says "you are an outsider" more than a paywall asking US$38 for one article. That fuelled my advocacy of open science and, ultimately, drove me to research its implementation.

> "Open science could become just the extension of privilege."

a two-tier system, in which richer research teams publish more OA articles in the most prestigious journals. One analysis of 37,000 articles in hybrid 'parent' journals and their fully OA 'mirrors' (with the same editorial board and acceptance standards) found that the geographic diversity of authors was much greater for non-OA articles than for OA articles (A. C. Smith *et al. Quant. Sci. Stud.* **2**, 1123–1143; 2022). Another analysis found that authors of OA articles were more likely to be male, senior, federally funded and working at prestigious universities (A. J. Olejniczak and

«People that print OA come from richer countries!»

nature communications

View all journals Se

Explore content About the journal Publish with us Sign up for ale

nature › nature communications › article processing charges

Article processing charges

Nature Communications is an open access journal. To publish in *Nature Communications*, authors are required to pay an article processing charge (APC).

The APC for all published papers is as follows, plus VAT or local taxes where applicable:

OPEN ACCESS

£4,190 (UK)

Americas*, Greater China** and Japan)

€4,790 (Europe and rest of world)

i-Perception

Journal Home Browse Journal Journal Info Stay Connected **Submit Paper**

2. Article Processing Charge (APC)

If, after peer review, your manuscript is accepted for publication, a one-time article processing charge (APC) is payable. This APC is discounted for a limited time and covers the cost of publication and ensures that your article will be freely available online in perpetuity under a Creative Commons licence.

- The article processing charge (APC) for *Regular Articles* is 800 GBP (+ VAT where applicable*).
- The article processing charge (APC) for *Methods* is 385 GBP (+ VAT where applicable*).
- The article processing charge (APC) for *Registered Reports* is 800 GBP (+ VAT where applicable*).
- The article processing charge (APC) for *Journal Club and Short & Sweet Articles* is 285 GBP (+ VAT where applicable*).
- There is no charge for *i-reviews*.

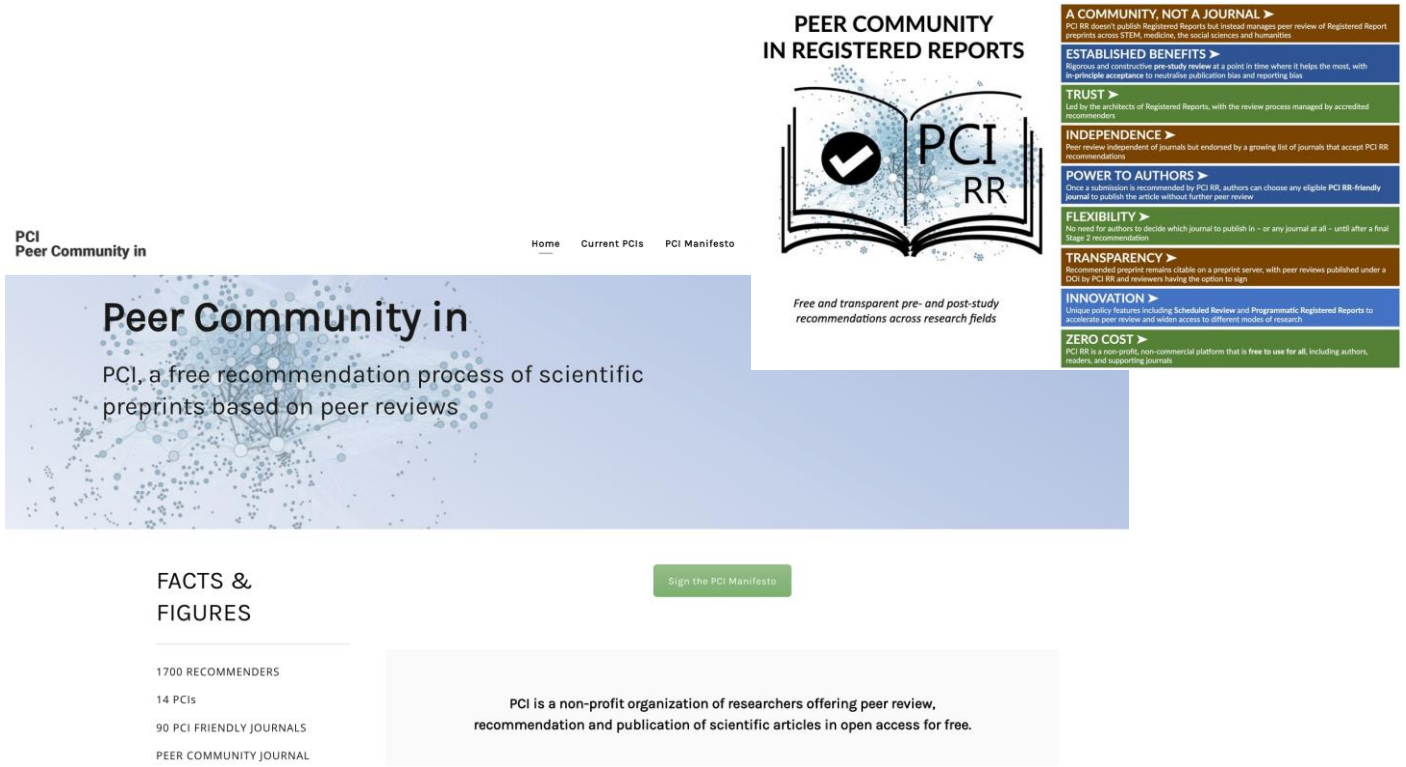
**The article processing charge (APC) is payable upon acceptance after peer review and is subject to value added tax (VAT) where applicable. If the paying author/institution is based in the European Union, to comply with European law, VAT must be added to the APC. Providing a VAT registration number will allow an institution to avoid paying this tax, except for UK institutions. Payments can be made in GBP or USD.*

Authors pay for brand in the same way we do for fashion

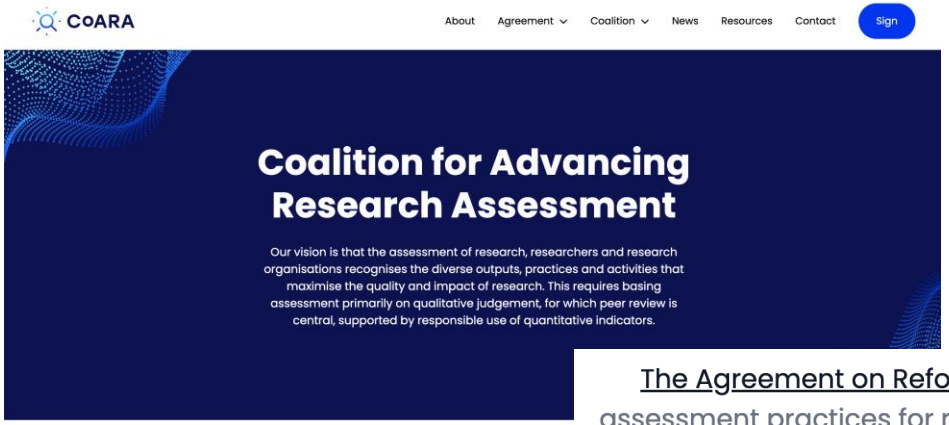
adidas

abibas

May we abandon scientific publishing altogether?



Research and researcher's assessment -> things are changing



unipd is among the 49 Italian academic institutions that signed the agreement

Who is behind CoARA?

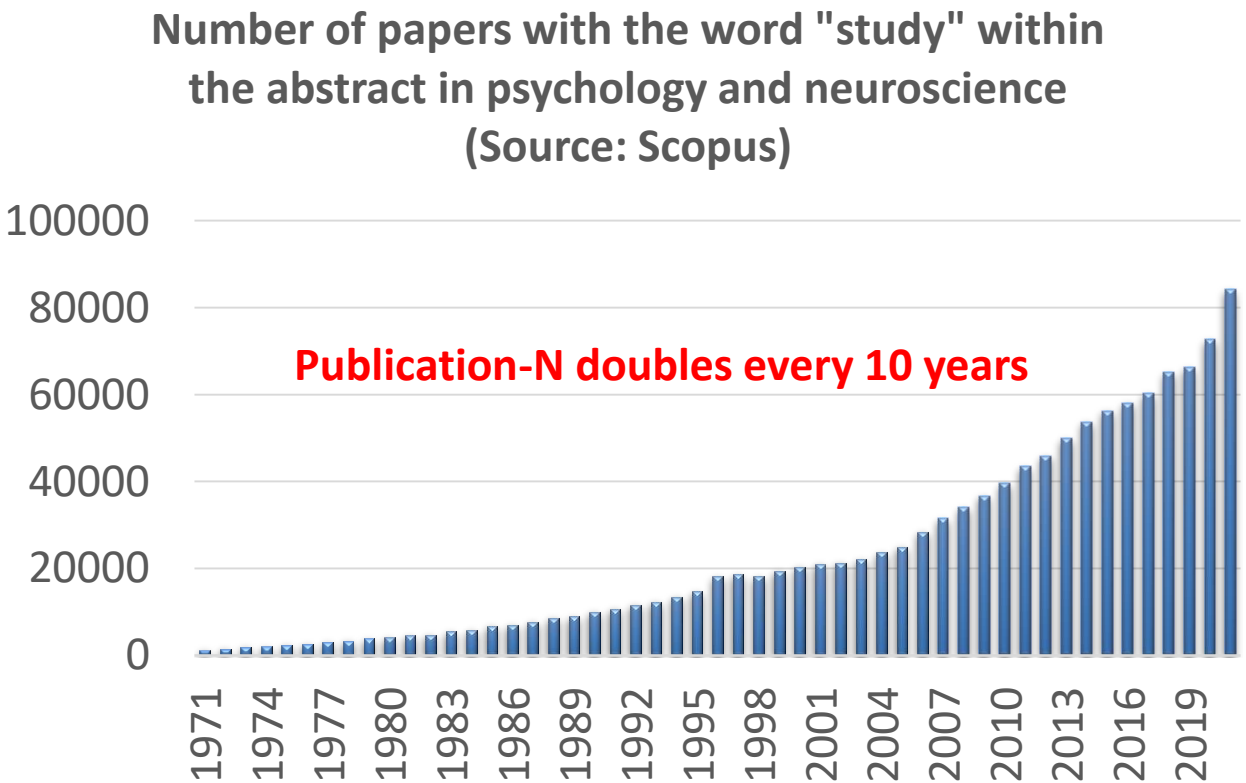
The process of drafting an Agreement on reforming research assessment was initiated in January 2022. More than 350 organisations from over 40 countries were involved. Organisations involved included public and private research funders, universities, research centres, institutes and infrastructures, associations and alliances thereof, national and regional authorities, accreditation and evaluation agencies, learned societies and associations of researchers, and other relevant organisations, representing a broad diversity of views and perspectives.

The Agreement on Reforming Research Assessment sets a shared direction for changes in assessment practices for research, researchers and research performing organisations, with the overarching goal to maximise the quality and impact of research. The Agreement includes the principles, commitments and timeframe for reforms and lays out the principles for a Coalition of organisations willing to work together in implementing the changes.

Signatories commit to a common vision, which is that the assessment of research, researchers and research organisations recognises the diverse outputs, practices and activities that maximise the quality and impact of research. This requires basing assessment primarily on qualitative judgement, for which peer-review is central, supported by responsible use of quantitative indicators.

OPEN ISSUES

Open issues (IMO): we are publishing too much



Trends in Cognitive Sciences

CellPress REVIEWS

Scientific Life

Fast Lane to Slow Science

Uta Frith[1,*]

Fast Science is bad for scientists and bad for science. Slow Science may actually help us to make faster progress, but how can we slow down? Here, I offer preliminary suggestions for how we can transition to a healthier and more sustainable research culture.

have become disillusioned by research bubbles and crashes that are reminiscent of free-market failures [2]. The relentless pace does not just mean that there is little chance to cultivate broader interests: it may be responsible for impairing the mental health and well-being of researchers. It also leads to a loss of talented people from the pool of researchers, inevitably resulting in decreased diversity. In addition, Fast Science leads to cutting corners and has almost certainly contributed to the reproducibility crisis. There are helpful recommendations to remedy the fail-

world. A farsighted vision is necessary to create and test big theories, regardless of obstacles. This perspective has consequences for how funders view the lengths of grant proposals and of intervals for evaluations. At present, early career researchers believe that they need to amass publications and citations to get grants. Established researchers need to continue to obtain grants to maintain their teams and facilities. Relentless expansion seems a rational strategy in these circumstances. With secure infrastructure including tenured key staff, there

More Speed, Mo
Stress, More Was

I vividly remember
Chaplin's film Mod
film, machine-like
to work more and
point of absurdity.
then that I would
strangely similar po
Over the past 50 ye
enced increasing p
up' and increasing
I used to be amaz
of a button you
Now I am amazed t
data by a click. P
gering though is
crease in the numbe
each year. Some
simply because th
searchers, but indi
have also dramatic
output. As implied
review papers ar
hardly anyone has ti
studies. When I wa
proudly published
years later I had ac
bunch of 28 publica
cited. I would have

Many of us ackn

Theory & Psychology

Article

Theory & Psychology
1–23
© The Author(s) 2020

Publish less, read more

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0959354319898250
journals.sagepub.com/home/tap

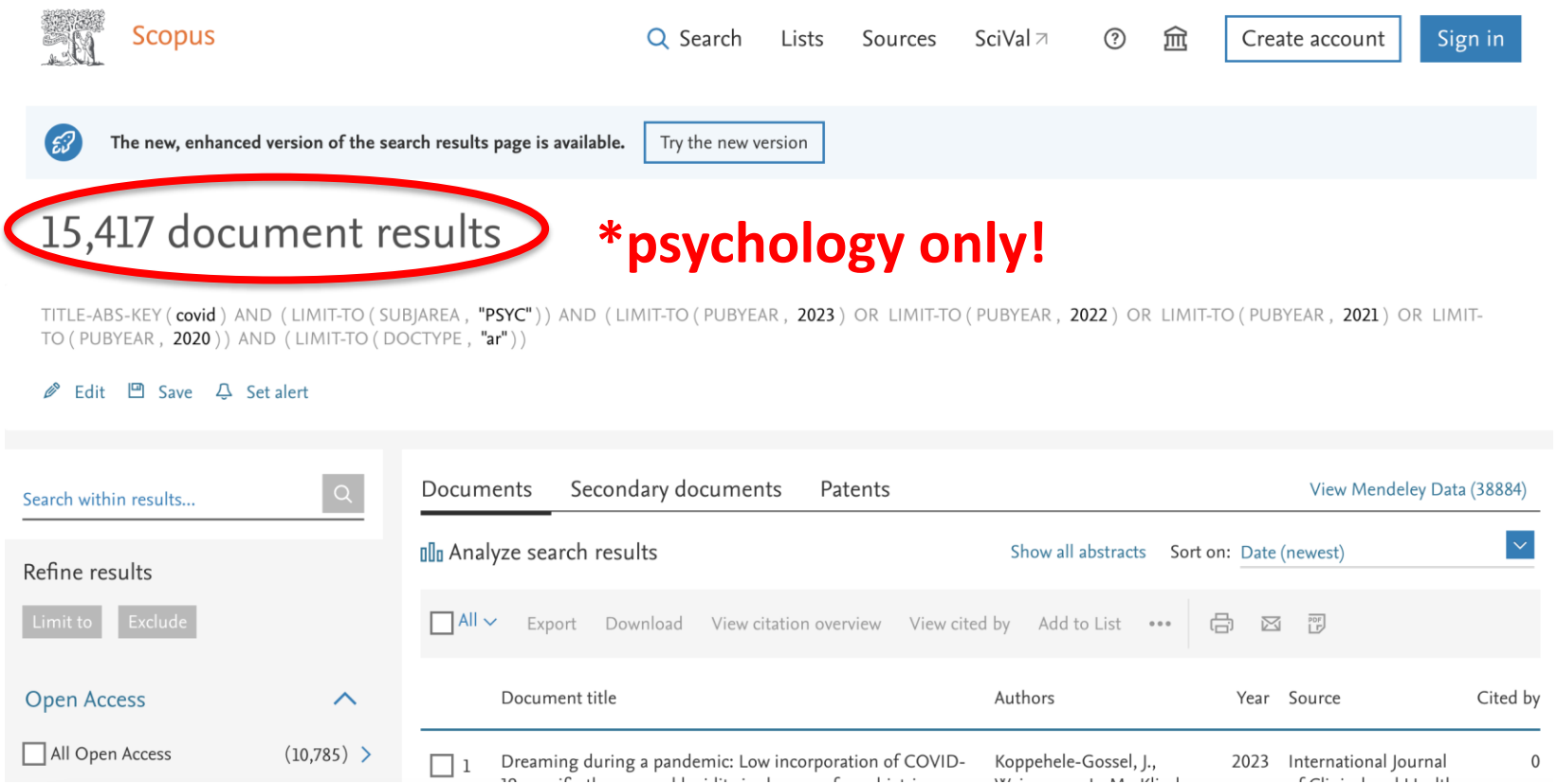
SAGE

R. Hans Phaf
University of Amsterdam

Abstract

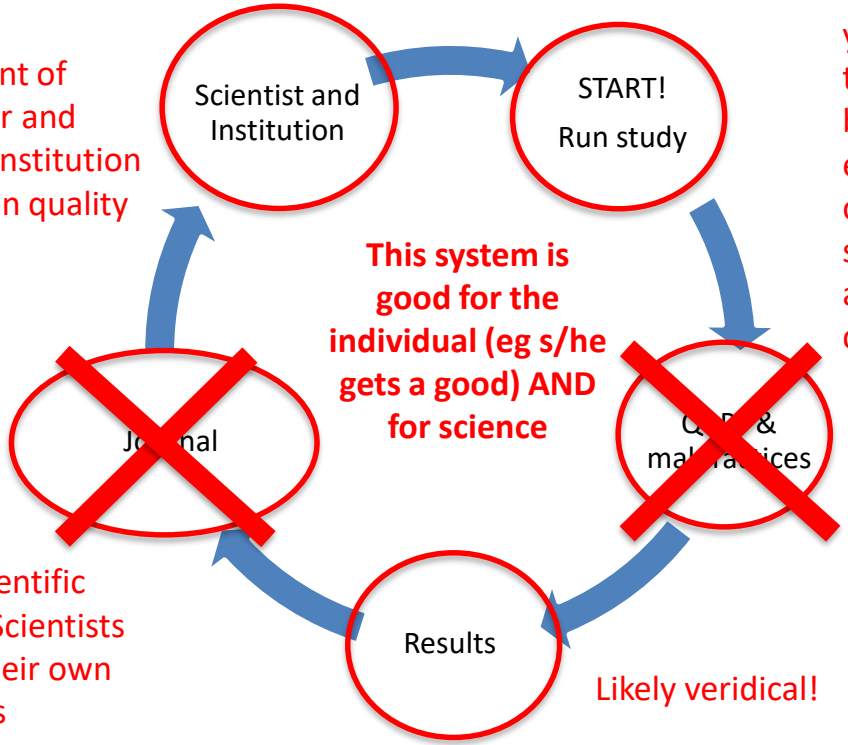
A publication deluge has impeded rather than advanced theory in experimental psychology. Many researchers rely more on null-hypothesis significance testing than literature studies to determine whether results are worthwhile. Four problematic publication practices are symptomatic for the theoretical deficit: (a) reinventing the wheel, (b) the Proteus phenomenon, (c) mechanical (non) replications, and (d) the survival of discredited hypotheses. Remedies include the development of AI tools recommending semantically related references, mandatory hypothesizing before and after results are known, and theoretical syntheses guided by meta-analyses and process models. The nonlinear theoretical development shows parallels to the optimization procedure of biological evolution. Theoretical hypotheses rather than experimental results are the elementary units of science. The fittest theories may survive alongside the least fit because they are not made to compete in research publications. Even if publication practices improve, winning hypotheses will often represent local optima and still cannot be taken with absolute certainty.

We are publishing too much: The COVID example



THE SCIENCE OF THE FUTURE

The possible “science of the future”



Assessment of researcher and research institution is based on quality

You begin the study only if you have the “GO” signal by the methodological review board. The idea of the study emerged from the joint effort of several groups of a given science community that will all participate in the data collection. It’s a multilab.

Only registered reports are published

Likely veridical!

No more scientific publishers. Scientists publish in their own autonomous repositories

ITALIAN REPRODUCIBILITY NETWORK

UPCOMING: SEMINARS ON OPEN SCIENCE

A remote educational course open to everyone, focused on Master and PhD Students covering open science topics and practices.

UPCOMING: ITRN KICKOFF EVENT

Online 14 October 2021

13:30-17:00

to be confirmed - work in progress

MISSION

The Italian Reproducibility Network (ITRN) is a peer-led consortium that aims to investigate the factors that contribute to robust research and disseminate them within the Italian scientific community. This is achieved by promoting initiatives and offering a hub for scientists to get in touch, exchange ideas and good practices, and promote collective learning. ITRN seeks collaboration with scientists in several disciplines, technical experts in relevant fields and stakeholders, so as to connect the widest possible spectrum of skills and knowledge.

Similarly to the other Reproducibility Networks in the world, ITRN investigates the factors that contribute to poor research reproducibility and replicability, and try to develop virtuous approaches to counter these factors and improve the quality of research. It also promotes collaboration among scientists and experts across a broad array of disciplines.

www.itrn.org

Join us!

(we are also among the signatories of the COARA agreement)

1222 • 2022
800
ANNI

UNIVERSITÀ
DEGLI STUDI
DI PADOVA