

# Is 'reproducibility' for all?

Jesper W. Schneider & Sven A. Ulpts

Danish Center for Studies in Research and Research Policy,  
Department of Political Science,  
Aarhus University

[jws@ps.au.dk](mailto:jws@ps.au.dk); [su@ps.au.dk](mailto:su@ps.au.dk)

# Danish Reproducibility Network: Launch Event

## Mission

- DKRN is a platform that connects Denmark-based researchers aiming to promote, facilitate and educate about **open, reproducible** and **robust** research
- Our work will ensure that Denmark remains an integral part of world-leading efforts contributing to the dissemination of **best research practices** and **positive culture change** in academia

Some important questions to consider

- What does open, reproducible and robust research mean ... and for whom?
- What does 'best practices' mean ... and for whom?
- What does a 'positive culture change in academia' mean ... and for whom?

**Is 'reproducibility' for all?**

# My main points

- **I am not going to spoil the party!**
- Just emphasize that regardless of the good intentions
  - such initiatives are not neutral, they have philosophical foundations
  - they tend to become normative and thus likely also ‘suppressive’
  - they can lead to epistemic injustice
- More concretely
  - Some background for where we are
  - Open science, meta-science as social movements
  - Conceptual confusions about ‘reproducibility’ and ‘replication’
  - Introduce a framework (work-in-progress) that aim to clarify the relevance and feasibility of ‘reproducibility’ for different modes of knowledge production

Background

# Superconductor LK-99, self-correction at work?

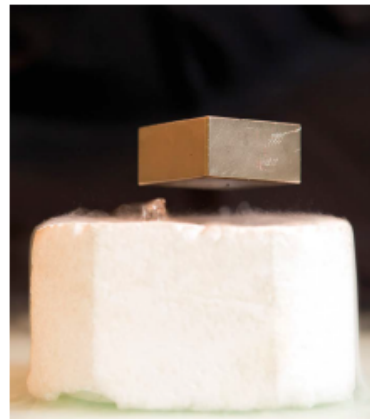
## REPLICATION EFFORTS FAIL FOR CLAIMED SUPERCONDUCTOR LK-99

Social media is abuzz with chatter about the material, but some scientists are pushing back on the hype.

By Dan Garisto

A South Korean team's claim to have discovered a superconductor that works at room temperature and ambient pressure has become a viral sensation — and prompted a slew of replication efforts by scientists and amateurs alike. But initial efforts to experimentally and theoretically reproduce the buzzworthy result have come up short, and researchers remain deeply sceptical.

The team, led by Sukbae Lee and Ji-Hoon Kim at the start-up firm Quantum Energy Research Centre in Seoul, reported in preprints published on the arXiv server on 25 July<sup>1,2</sup> that a compound of copper, lead, phosphorus and oxygen, dubbed LK-99, is a superconductor at ambient pressure and temperatures of up to at least 127 °C (400 kelvin). The team says that samples show two key hallmarks of superconductivity: zero electrical resistance and the Meissner effect, in which the material expels magnetic fields, leading samples to levitate above a magnet. Previous efforts have achieved superconductivity only in materials at very low temperatures or extremely high pressures. No



Levitation is a hallmark of superconductivity.

India in New Delhi<sup>3</sup> and Beihang University in Beijing<sup>4</sup> — reported synthesizing LK-99, but did not observe signs of superconductivity. A third experiment by researchers at Southeast University in Nanjing, China, found no Meissner effect, but measured near-zero resistance in LK-99 at  $-163\text{ °C}$  (110 K) — far below room

team's. "Our LK-99 is very similar to that as the reported superconducting LK-99," he says.

But Robert Palgrave, a chemist at University College London, says that both X-ray diffraction patterns obtained by these replication attempts are significantly different from the Korean team's patterns and from each other. (Members of the Beihang team did not respond to *Nature's* request for comment.)

The Southeastern University team obtained X-ray diffraction data that are more consistent with the Korean team's sample, according to Palgrave. But several researchers have questioned the claim that zero resistance was achieved at  $-163\text{ °C}$ . Evan Zalyss-Geller, a condensed-matter physicist at the Massachusetts Institute of Technology in Cambridge, says that the resistance measurement wasn't sensitive enough to distinguish between a superconductor and a low-resistance metal such as copper. (Members of the Southeastern University team did not respond to *Nature's* request for comment.)

Uncertainty about the structure of LK-99 limits the conclusions that researchers can draw from theoretical calculations, which assume a given structure.

On 31 July, a theoretical analysis posted on Twitter prompted excitement among online enthusiasts. Sinéad Griffin, who studies quantum materials at Lawrence Berkeley National Laboratory in Berkeley, California, shared her paper<sup>5</sup>, accompanied by a GIF of a 'mic drop'. The optimism was prompted by Griffin's use of DFT to find that LK-99 has 'flat bands', indicating that electrons in the material are strongly correlated with each other. "Flat-band systems tend to show interesting physics," Vishik says. "So when a material is predicted to have a flat



Pure LK-99 crystals made at a Max Planck Institute in Stuttgart, Germany.

## HOW SCIENCE SLEUTHS SHOWED LK-99 ISN'T A SUPERCONDUCTOR

Efforts to replicate the material explain why it displayed superconducting-like behaviours.

By Dan Garisto

Researchers seem to have solved the puzzle of LK-99. Scientific detective work has unearthed evidence that the material is not a superconductor, and clarified its actual properties.

superconductors function only at very low temperatures and extreme pressures.

The extraordinary claim quickly grabbed the attention of the science-interested public and researchers, some of whom tried to replicate LK-99. Initial attempts did not find signs of room-temperature superconductivity,

erasing doubts about the material's structure and confirming that it is not a superconductor, but an insulator.

The only further confirmation would come from the South Korean team sharing its samples, says Michael Fuhrer, a physicist at Monash University in Melbourne, Australia. "The burden's on them to convince everybody else," he says.

Perhaps the most striking evidence for LK-99's superconductivity was a video taken by the South Korean team that showed a coin-shaped sample of silvery material wobbling over a magnet. The researchers said that the sample was levitating because of the Meissner effect — a hallmark of superconductivity in which a material expels magnetic fields. Multiple unverified videos of LK-99 levitating subsequently circulated on social media, but none of the researchers who initially tried to replicate the findings observed any levitation.

### Half-baked levitation

Several red flags popped up to Derrick VanGennep, a former condensed-matter researcher at Harvard University in Cambridge, Massachusetts, who now works in finance but was intrigued by LK-99. In the video, one edge of the sample seemed to stick to the magnet, and it seemed to be delicately balanced. By contrast, superconductors that levitate over magnets can be spun and even held upside down. "None of those behaviours look like what we see in the LK-99 videos," VanGennep says.

He thought LK-99's properties were more likely to be the result of ferromagnetism. So he constructed a pellet of compressed graphite shavings with iron filings glued to it. A video made by VanGennep shows that his disc — made of non-superconducting, ferromagnetic materials — mimicked LK-99's behaviour.

On 7 August, the Peking University team reported<sup>3</sup> that this "half-levitation" appeared

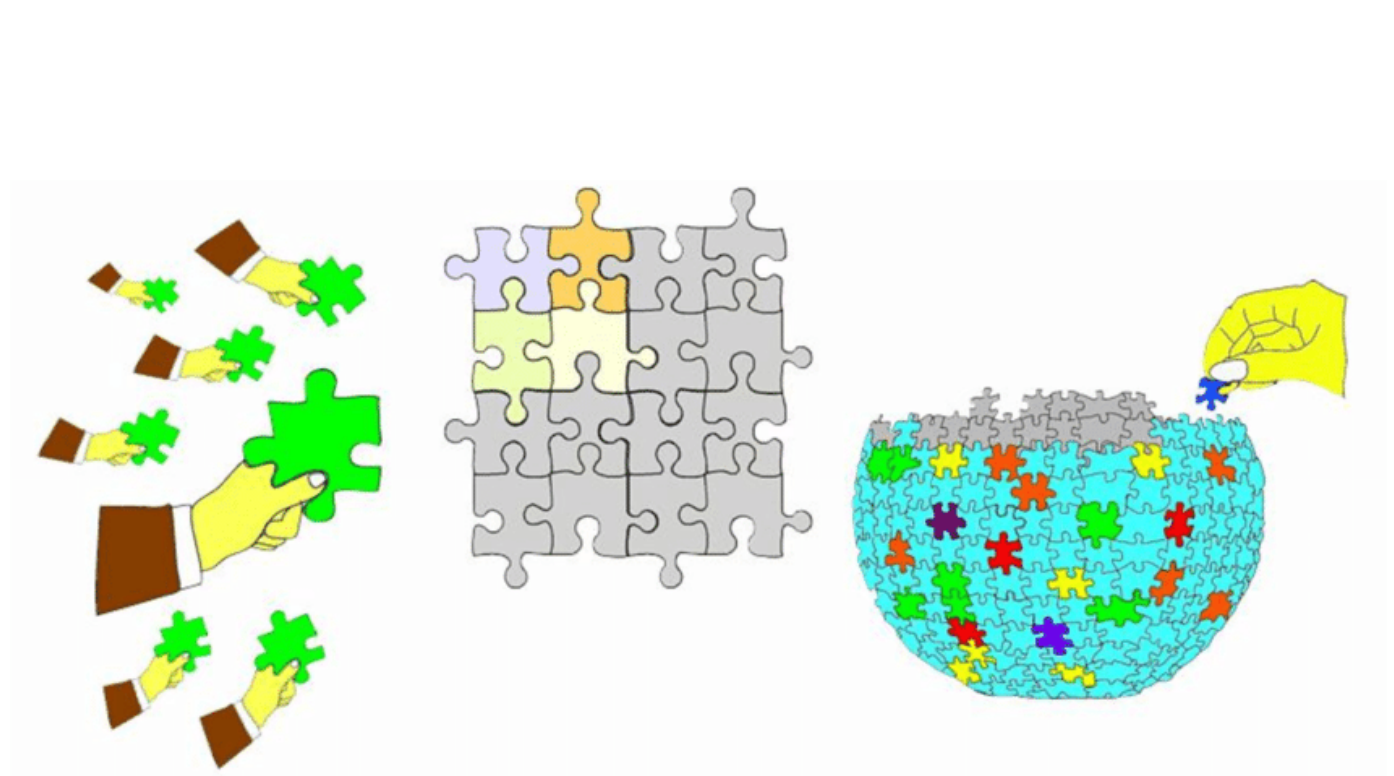
# How did they do it?

- Big claim = much interest
- More than 20 'replication' efforts in less than 2 months
- Hallmarks of superconductivity
  - Resistivity = 0
  - Meissner effect

} Theory (+ auxiliary assumptions) **failure to meet these criteria**
- Detective work (conceptually, experimentally)
  - The material is not a superconductor, its actual properties are clarified
  - Impurities in the material were responsible for sharp drops in its electrical resistivity and a display of partial levitation over a magnet, properties similar to those exhibited by superconductors

# That's how we think of 'replication' and self-correction', right?

- Part of ***the scientific method***
- Enabling scientific knowledge accumulation



But something is apparently 'wrong'?



# The 'reproducibility crisis' narrative

Open access, freely available online

Essay

## Why Most Published Research Findings Are False

John P. A. Ioannidis

### Summary

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

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

factors that influence this problem and some corollaries thereof.

### Modeling the Framework for False Positive Findings

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

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

should be interpreted based only on  $p$ -values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a  $2 \times 2$  table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let  $R$  be the ratio of the number of "true relationships" to "no relationships" among those tested in the field.  $R$

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is  $R/(R+1)$ . The probability of a study finding a true relationship reflects the power  $1 - \beta$  (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate,  $\alpha$ . Assuming that  $r$  relationships are being probed in the field, the expected values of the  $2 \times 2$  table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the  $2 \times 2$  table, one gets  $PPV = (1 - \beta)R/(R - \beta R + \alpha)$ . A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.

Copyright: © 2005 John P.A. Ioannidis. 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 work is properly cited.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: ioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

## RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

### Estimating the reproducibility of psychological science

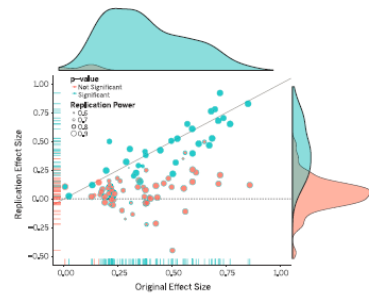
Open Science Collaboration\*

**INTRODUCTION:** Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

**RATIONALE:** There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a pre-

viously observed finding, and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

**RESULTS:** We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and  $P$  values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size ( $d$ ) of the replication efforts ( $M = 0.197$ ,  $SD = 0.207$ ) was half the magnitude of the mean effect size of the original efforts ( $M = 0.403$ ,  $SD = 0.186$ ), representing a



**Original study effect size versus replication effect size (correlation coefficients).** Original line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

SCIENCE | sciencemag.org

substantial decline. Ninety-seven percent of original studies had significant results ( $P < .05$ ). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 90% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated theoretical results in original results as 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.

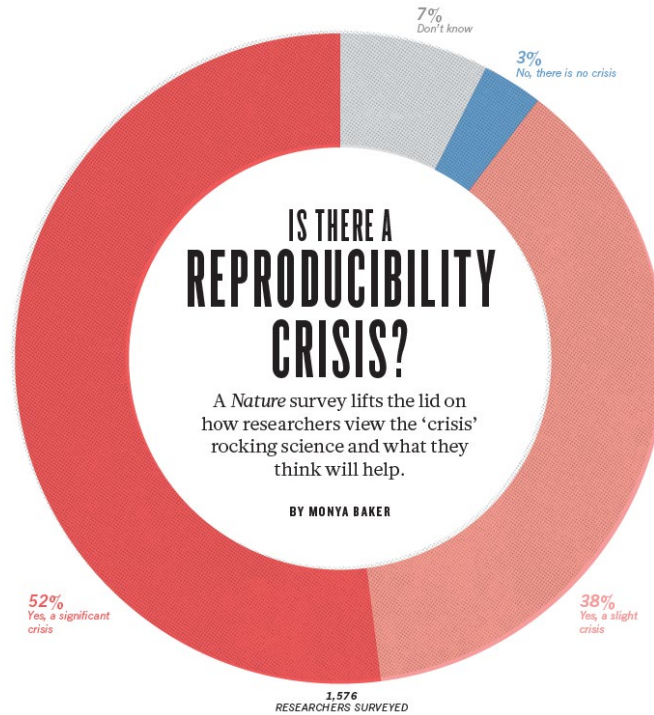
**CONCLUSION:** No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replication produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original  $P$  value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Innovation points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know. ■

Relevant author disclosures are available in the full article online.  
\*Corresponding author. E-mail: ioannid@cc.uoi.gr  
\*This article is Open Access Collaboration. Science 349, 464-476 (2015). DOI: 10.1126/science.1261191

30 AUGUST 2015 • VOL 349 | SCIENCE | 9-15

- 'Unexpectedly' many studies did not replicate in (social) psychology
- Questionable Research Practices were seemingly widespread
- Most published findings are false?



IS THERE A  
**REPRODUCIBILITY  
CRISIS?**

A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

52%  
Yes, a significant crisis

38%  
Yes, a slight crisis

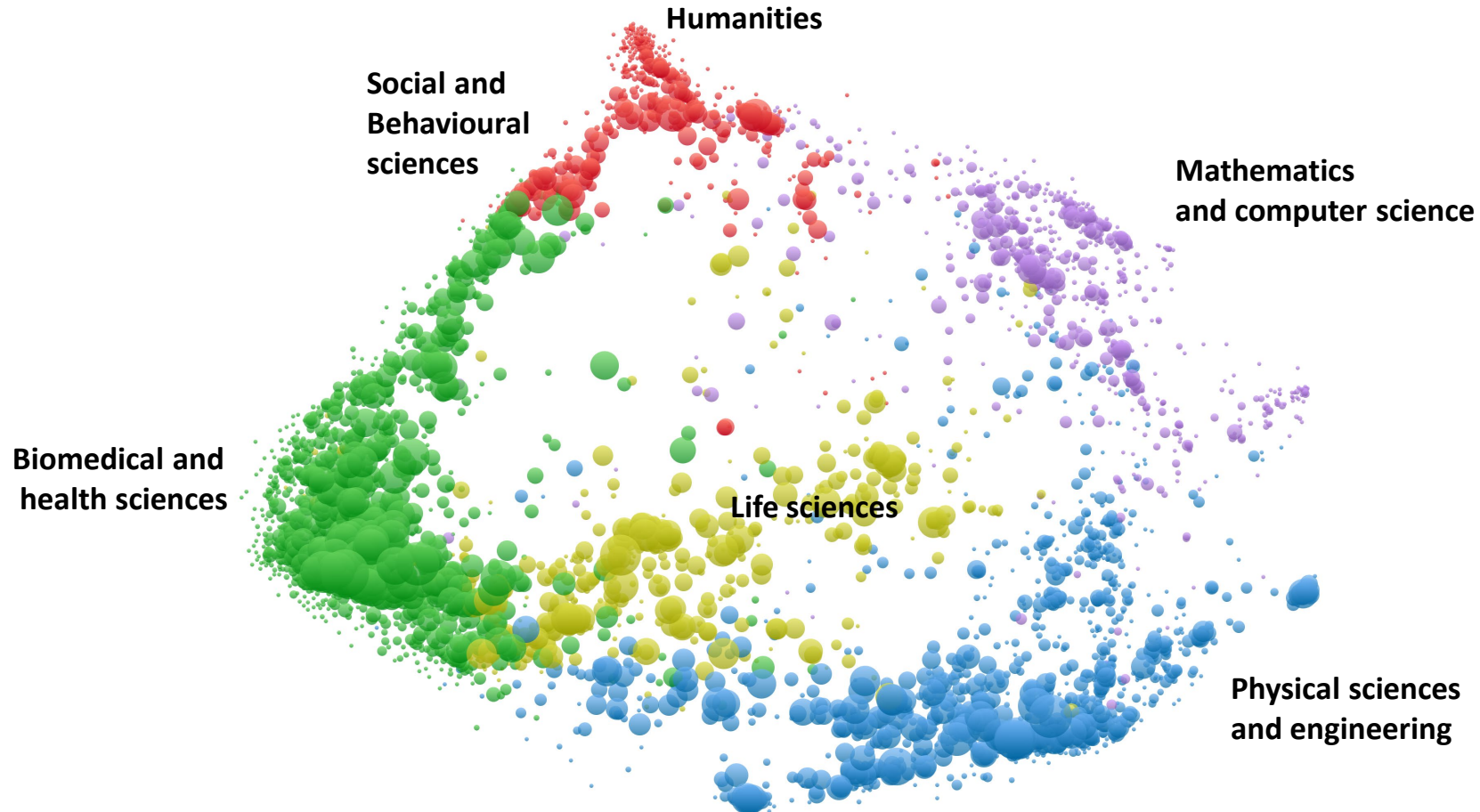
1,576  
RESEARCHERS SURVEYED

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 696-701. doi:10.1371/journal.pmed.0020124

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251)

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, 533(7604), 4 52-454.

So something is apparently 'wrong', but where?



# Not necessarily a problem related to lack of 'experimentation'

## Historical science, experimental science, and the scientific method

Carol E. Cleland

Department of Philosophy and Center for Astrobiology, University of Colorado, Boulder, Colorado 80309, USA

### ABSTRACT

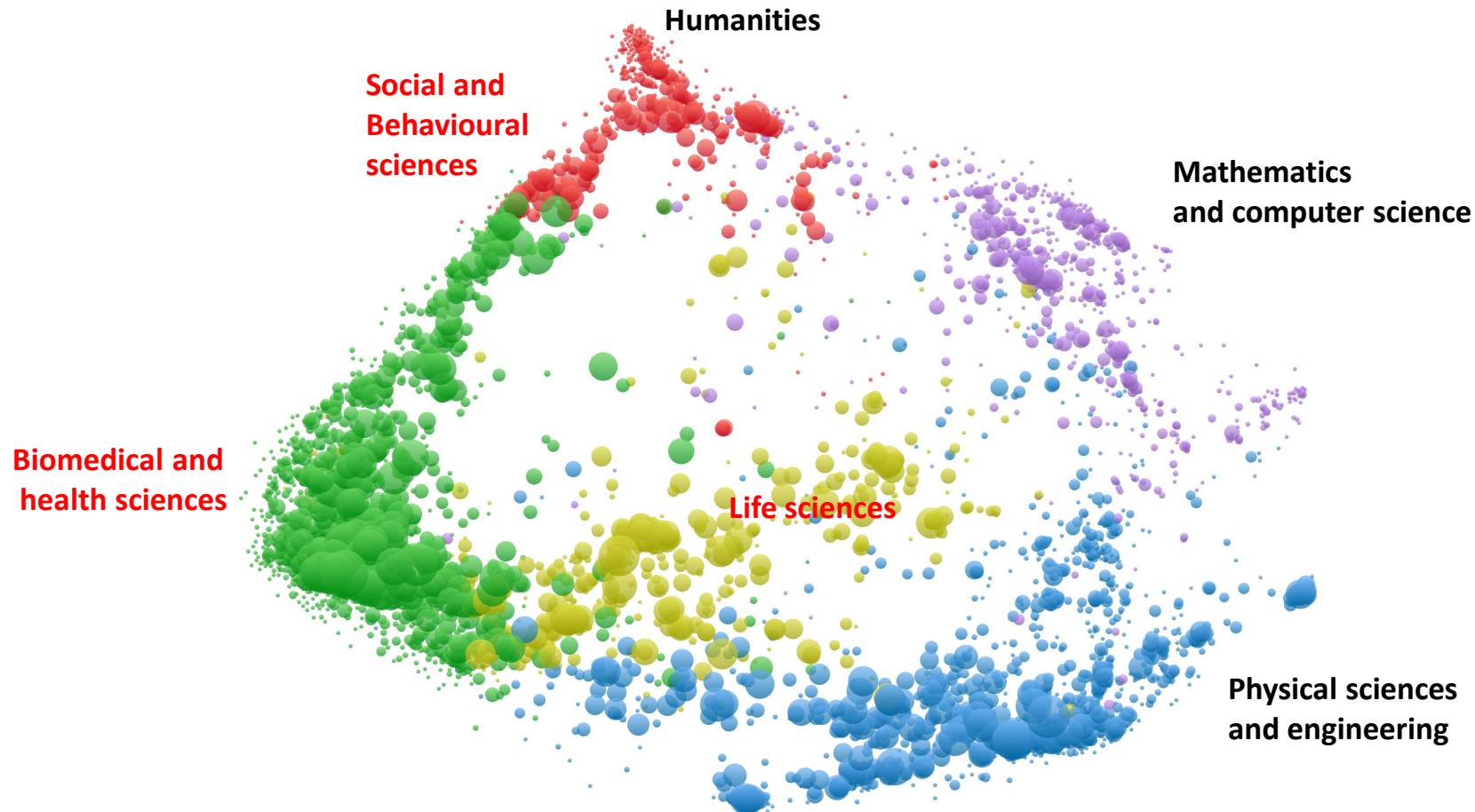
Many scientists believe that there is a uniform, interdisciplinary method for the practice of good science. The paradigmatic examples, however, are drawn from classical experimental science. Insofar as historical hypotheses cannot be tested in controlled laboratory settings, historical research is sometimes said to be inferior to experimental research. Using examples from diverse historical disciplines, this paper demonstrates that such claims are misguided. First, the reputed superiority of experimental research is based upon accounts of scientific methodology (Baconian inductivism or falsificationism) that are deeply flawed, both logically and as accounts of the actual practices of scientists. Second, although there are fundamental differences in methodology between experimental scientists and historical scientists, they are keyed to a pervasive feature of nature, a time asymmetry of causation. As a consequence, the claim that historical science is methodologically inferior to experimental science cannot be sustained.

Keywords: methodology, induction, history, experimental investigations.

“... causal overdetermination of past events by localized present events explains the practice of historical science, so the causal underdetermination of future events by localized present events explains the practice of experimental science”



# Something to do with the way we produce knowledge?



## What characterizes empirical quantitative knowledge production?

- Explanatory, exploratory or classificatory?
- Closed, open systems
- Interaction and causal density
- Can you run sophisticated experiments with reliable controls and run them many times over?
- Can you replicate own findings and adjust iteratively?
- Do you rely on “weak theories” with poor predictive power?
- What is the probability of hypothesis?
- Over-reliance on inferential statistics (mainly frequentists)
- Can you posit “true” null hypotheses?
- Can you posit plausible “alternative” hypotheses?
- Are “multiple testing”, “data dredging”, “optional stopping”, “garden of forking paths” challenges?
- Are 2 sigma (5 %) an epistemic threshold??
- ...

# But this is not new: Pretensions to be 'scientific' (NHST)

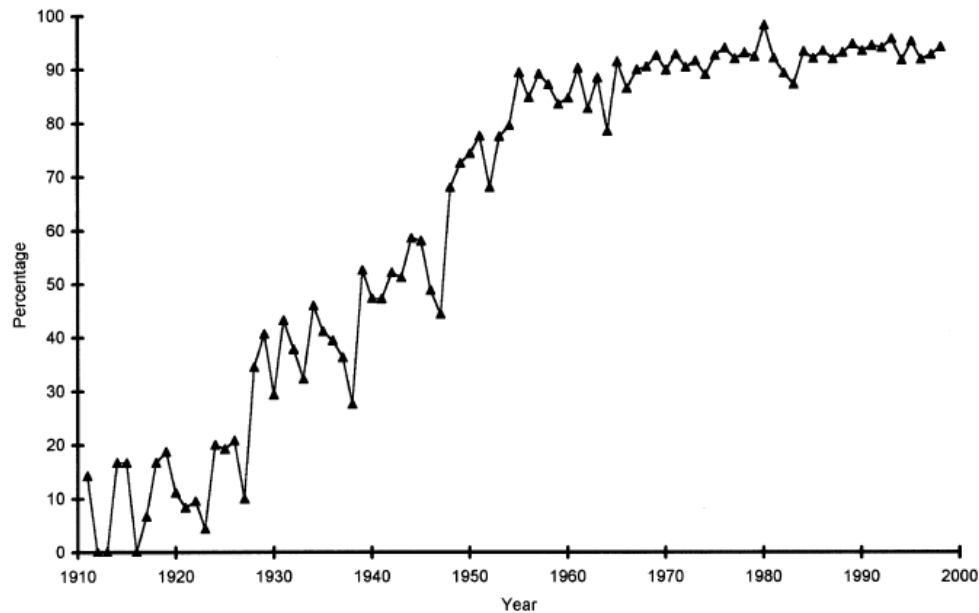
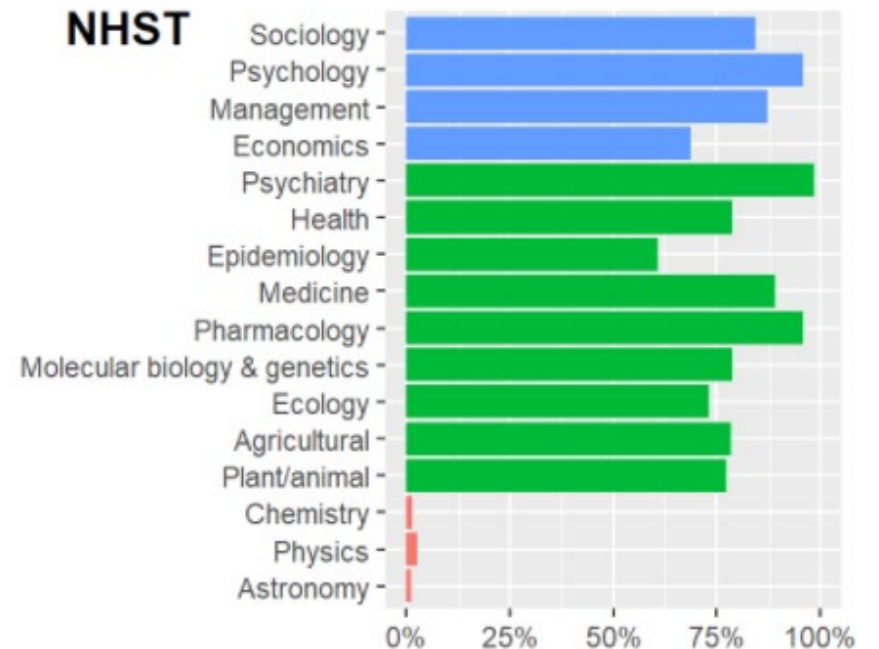


Figure 1. Percentage of total significance tests (*pes*, *crs*, and *p* values): All journals 1911 to 1998.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology - and its future prospects. *Educational and Psychological Measurement*, 60(5), 661-681.

## Statistical methods across fields



# But this is not new: Fallacious knowledge production

## TESTS OF SIGNIFICANCE CONSIDERED AS EVIDENCE\*

BY JOSEPH BERKSON, M.D.

*Division of Biometry and Medical Statistics, Mayo Clinic*

"After all, the higher statistics are only common sense reduced to numerical appreciation."—KARL PEARSON.

THERE WAS a time when we did not talk about tests of significance; we simply did them. We tested whether certain quantities were significant in the light of their standard errors, without inquiring as to just what was involved in the procedure, or attempting to generalize it. In recent years tests of significance have been more broadly conceived as tests of hypotheses, and they have been generalized as *t* tests, *F* tests and certain amplifications of these, such as analysis of variance or of covariance. It is hardly an exaggeration to say that statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance, though not always presented as such, some comparatively simple and forthright, others elaborate and abstruse. Behind this is a doctrine of analysis that consists of setting up what is called a "null hypothesis" and testing it. Indeed, in this conception not only does this procedure characterize the method of statistics, but it is considered to be the very essence of all experimental science. In his well known book, *The Design of Experiments*, R. A. Fisher wrote, "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."<sup>1</sup>

What is this null hypothesis procedure? I quote from a recent text.<sup>2</sup>

We have just set up the hypothesis that our sample of 900, which has a mean of 15,071 miles, is a random sample drawn from the population having a known mean of 15,200 miles. . . . Such a hypothesis is called a *null hypothesis* since our computations undertake to nullify it. The procedure may be summarized into three steps: (1) Set up the hypothesis that the true difference is zero. (2) Upon the basis of this hypothesis determine the probability that such a difference as the one observed might occur because of sampling variations. (3) Draw a conclusion concerning the hypothesis. If such observed difference could hardly have occurred by chance, we have cast much doubt upon the hypothesis. We therefore abandon the hypothesis and conclude that the observed difference is significant.

\* A paper presented at the 103rd Annual Meeting of the American Statistical Association, New York, December 29, 1941.

<sup>1</sup> R. A. Fisher, *The Design of Experiments*, Ed. 2, London, Oliver and Boyd, Ltd., 1937, p. 19.

<sup>2</sup> F. E. Croxson and D. J. Cowden, *Applied General Statistics*, New York, Prentice-Hall, Inc., 1940, p. 310.

*Psychological Bulletin*  
1960, Vol. 57, No. 5, 416-428

## THE FALLACY OF THE NULL-HYPOTHESIS SIGNIFICANCE TEST

WILLIAM W. ROZEBOOM  
*St. Olaf College*

The theory of probability and statistical inference is various things to various people. To the mathematician, it is an intricate formal calculus, to be explored and developed with little professional concern for any empirical significance that might attach to the terms and propositions involved. To the philosopher, it is an embarrassing mystery whose justification and conceptual clarification have remained stubbornly refractory to philosophical insight. (A famous philosophical epigram has it that induction [a special case of statistical inference] is the glory of science and the scandal of philosophy.) To the experimental scientist, however, statistical inference is a research instrument, a processing device by which unwieldy masses of raw data may be refined into a product more suitable for assimilation into the corpus of science, and in this lies both strength and weakness. It is strength in that, as an ultimate *consumer* of statistical methods, the experimentalist is in position to demand that the techniques made available to him conform to his actual needs. But it is also weakness in that, in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures learned more or less by rote from persons assumed to be more knowledgeable of statistics than he. There is, of course, nothing surprising

or reprehensible about this—one need not understand the principles of a complicated tool in order to make effective use of it, and the research scientist can no more be expected to have sophistication in the theory of statistical inference than he can be held responsible for the principles of the computers, signal generators, timers, and other complex modern instruments to which he may have recourse during an experiment. Nonetheless, this leaves him particularly vulnerable to misinterpretation of his aims by those who build his instruments, not to mention the ever present dangers of selecting an inappropriate or outmoded tool for the job at hand, misusing the proper tool, or improvising a tool of unknown adequacy to meet a problem not conforming to the simple theoretical situations in terms of which existent instruments have been analyzed. Further, since behaviors once exercised tend to crystallize into habits and eventually traditions, it should come as no surprise to find that the tribal rituals for data-processing passed along in graduate courses in experimental method should contain elements justified more by custom than by reason.

In this paper, I wish to examine a dogma of inferential procedure which, for psychologists at least, has attained the status of a religious conviction. The dogma to be scrutinized is the "null-hypothesis significance test" orthodoxy that passing statistical judgment on a scientific hypothesis by means of experimental observa-

## Philosophy of Science

June, 1967

### THEORY-TESTING IN PSYCHOLOGY AND PHYSICS: A METHODOLOGICAL PARADOX\*

PAUL E. MEEHL<sup>1</sup>

*Minnesota Center for Philosophy of Science*

Because physical theories typically predict numerical values, an improvement in experimental precision reduces the tolerance range and hence increases corroborability. In most psychological research, improved power of a statistical design leads to a prior probability approaching 1/2 of finding a significant difference in the theoretically predicted direction. Hence the corroboration yielded by "success" is very weak, and becomes weaker with increased precision. "Statistical significance" plays a logical role in psychology precisely the reverse of its role in physics. This problem is worsened by certain unhealthy tendencies prevalent among psychologists, such as a premium placed on experimental "cuteness" and a free reliance upon *ad hoc* explanations to avoid refutation.

The purpose of the present paper is not so much to propound a doctrine or defend a thesis (especially as I should be surprised if either psychologists or statisticians were to disagree with whatever in the nature of a "thesis" it advances), but to call the attention of logicians and philosophers of science to a puzzling state of affairs in the currently accepted methodology of the behavior sciences which I, a psychologist, have been unable to resolve to my satisfaction. The puzzle, sufficiently striking (when clearly discerned) to be entitled to the designation "paradox," is the following: *In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data, is to increase the difficulty of the "observational hurdle" which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavior sciences, the usual effect of such improvement in experimental precision is to provide an easier hurdle for the theory to surmount.* Hence what we would normally think of as improvements in our experimental method tend (when predictions materialize) to yield

\* Received March, 1967.

<sup>1</sup> I wish to express my indebtedness to Dr. David T. Lykken, conversations with whom have played a major role in stimulating my thinking along these lines, and whose views and examples have no doubt influenced the form of the argument in this paper. For an application of these and allied considerations to a specific example of poor research in psychology, see [7].

# Is science really facing a ‘replication crisis’? Or is it more restricted?



COLLOQUIUM OPINION

## Is science really facing a reproducibility crisis, and do we need it to?

Daniele Fanelli<sup>a,1</sup>

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske  
November 3, 2017 (received for review June 30, 2017)

Efforts to improve the reproducibility and integrity of science are typically justified by a narrative of crisis, according to which most published results are unreliable due to growing problems with research and publication practices. This article provides an overview of recent evidence suggesting that this narrative is mistaken, and argues that a narrative of epochal changes and empowerment of scientists would be more accurate, inspiring, and compelling.

reproducible research | crisis | integrity | bias | misconduct

---

# Meta-research: Fixing suboptimal and wasteful applications of *the* ‘scientific method’



*Annual Review of Statistics and Its Application*

## Calibrating the Scientific Ecosystem Through Meta-Research

Tom E. Hardwicke,<sup>1,2</sup> Stylianos Serghiou,<sup>2,3</sup>  
Perrine Janiaud,<sup>2</sup> Valentin Danchev,<sup>2</sup> Sophia Crüwell,<sup>1,4</sup>  
Steven N. Goodman,<sup>2,3,5</sup> and John P.A. Ioannidis<sup>1,2,3,5,6</sup>

<sup>1</sup>Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité-Universitätsmedizin Berlin, 10178 Berlin, Germany; email: tom.hardwicke@charite.de

<sup>2</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California 94305, USA

<sup>3</sup>Department of Health Research and Policy, Stanford University, Stanford, California 94305, USA

<sup>4</sup>Department of Psychological Methods, University of Amsterdam, 1018 WS Amsterdam, Netherlands

<sup>5</sup>Department of Medicine, Stanford University, Stanford, California 94305, USA

<sup>6</sup>Departments of Biomedical Data Science, and of Statistics, Stanford University, California, USA



[www.annualreviews.org](http://www.annualreviews.org)

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Stat. Appl. 2020. 7:11–37

First published as a Review in Advance on November 1, 2019

The *Annual Review of Statistics and Its Application* is online at [statistics.annualreviews.org](http://statistics.annualreviews.org)

<https://doi.org/10.1146/annurev-statistics-031219-041104>

Copyright © 2020 by Annual Reviews.  
All rights reserved

### Keywords

meta-research, meta-science, methodology, bias, reproducibility, open science

### Abstract

While some scientists study insects, molecules, brains, or clouds, other scientists study science itself. Meta-research, or research-on-research, is a burgeoning discipline that investigates efficiency, quality, and bias in the scientific ecosystem, topics that have become especially relevant amid widespread concerns about the credibility of the scientific literature. Meta-research may help calibrate the scientific ecosystem toward higher standards by providing empirical evidence that informs the iterative generation and refinement of reform initiatives. We introduce a translational framework that involves (a) identifying problems, (b) investigating problems, (c) developing solutions,



# Is this a problem?

- No
  - There are many methodological practices linked to scientific ideals that are fallacious and should be improved where relevant
- Yes
  - There are many other useful methodological scientific practices that may be (more) 'suppressed' due to 'science reformers' strict focus on *a* scientific method
  - There are no 'best practices' *per se*
  - There are no scientific method *per se*
- Wider concerns about epistemic diversity

# Social movements

# Open Science, meta-science ... turned into activism

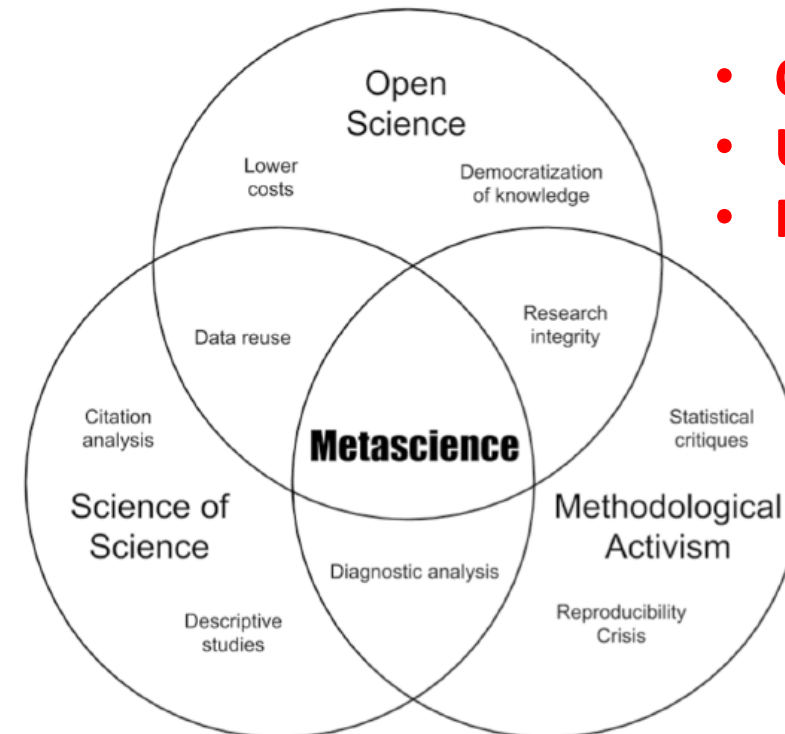


## Metascience as a Scientific Social Movement

David Peterson<sup>1</sup> · Aaron Panofsky<sup>2</sup>

Accepted: 7 March 2023 / Published online: 24 April 2023  
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

**Abstract** The “reproducibility crisis” has been one of the most significant stories in science in the past 15 years and has led to significant policy changes across the research landscape. Yet, scandals, irreproducible studies, and cries of crisis have occurred for decades in science. This article seeks to explain why the reproducibility crisis has taken root and become a force in science policy in ways previous crises have not. In short, we argue that it was through the scientific, institutional, and cultural efforts of a group of scientific activists we are calling metascientists. Metascience is a scientific social movement that seeks to use quantification and experimentation to diagnose problems in research practice and improve efficiency. It draws together data scientists, experimental and statistical methodologists, and open science activists into a project with both intellectual and policy dimensions. Metascientists have been remarkably successful at winning grants, motivating news coverage, and changing policies at science agencies, journals, and universities. The social movement lens is useful for understanding the popularization and impact of the reproducibility crisis narrative and suggests ways the institutions of science are adapting to meet a changing political and technological landscape.



- Change practice
- Use incentives
- Provide new norms

**Fig. 1** The strands of metascience

# Where does open science, metascience ... come from?

- Open science can be defined as:
  - ‘transparent and accessible knowledge that is shared and developed through collaborative networks’ (Vicente-Saez & Martinez-Fuentes, 2018, p. 434).
- It refers to a broad range of practices aimed at detecting scientific fraud and enhancing transparency and replicability of research
- Not neutral, such definitions have philosophical foundations and thus come with assumptions
- The motive for widespread propagation of open science practices is an honorable one: to improve the ‘quality’, ‘rigor’, and ‘credibility’ of science
- The concern is that, although these principles may benefit **some** ‘post-positivist’ research traditions, they may be detrimental to others

# Open Science stems from (post)positivism

Paradigm	Ontology	Epistemology	Axiology	Methodology
Positivism	There is truth!	We can know this truth!	Values should play no role in inquiry; researcher and the researched are separable.	Almost entirely quantitative, tightly controlled experiments
Post-Positivism	There is truth!	Alas, we are limited in our ability to know this truth, but we shall try our darndest!	Ok, so people clearly have values and biases, but we will completely remove their influence, and thus maintain separation.	Largely quantitative, including experimental and observational methods. But can also be qualitative and mixed methods.
Constructivism	There are multiple truths, varying by individuals and contexts.	Knowledge is co-constructed between researcher and participants, and thus cannot be independent of the researcher.	Values have strong influence on inquiry. These values should be discussed, described, and considered vis-à-vis the research.	Largely qualitative, especially via interviews, focus groups, and content analysis. But can also be quantitative and mixed methods.
Criticalism	There are multiple truths, and they are contoured by relative access to societal power.	Knowledge is co-constructed between researchers and participants, and it is the responsibility of researchers to empower participants.	All inquiry is embedded in a value system, and research should be used to improve the lives of those who are marginalized	Almost entirely qualitative, especially via participatory approaches. May also be non-empirical.

- **Post-positivism:** A research paradigm holding that a knowable, tangible, and measurable reality exists (i.e., naïve or critical realism), knowledge claims about this reality can be developed objectively, and verification/falsification of a priori hypotheses is the most prevalent methodological choice.

- **Constructionism:** A scholarly movement that holds that reality is the result of communicative processes that create a sense of shared reality (i.e., it is locally co-constructed), emphasizes that objectivity is also co-constructed through communicative processes, and aims to examine taken-for-granted realities that might be oppressive or dysfunctional through future-forming, dialogic, hermeneutical, and dialectical research to generate new functional realities.

# Open Science stems from (post)positivism

- It's important to distinguish between different 'paradigms' and their associated research communities
- Undeclared assumptions that are, by definition, self-evident to a community can be seen as impositions by other communities that do not share them ... also within a paradigm
- A given community will defend as reasonable arguments that logically descend from those assumptions against competing communities
- For example, open science principles such as:
  - providing a verifiable distinction between hypothesis generation and hypothesis testing
  - reducing researcher bias
  - increasing reproducibility
- ... can be seen 'as reasonable arguments that logically descend from' a post-positivist epistemology anchored in notions of the scientific method where experimentation is paramount


# ... but

*Industrial and Organizational Psychology* (2022), 15, 525–528  
doi:10.1017/iop.2022.67



COMMENTARY

## Open science and epistemic pluralism: A tale of many perils and some opportunities

Andrea Bazzoli 

Washington State University Vancouver, Vancouver, WA, USA  
Email: [andrea.bazzoli@wsu.edu](mailto:andrea.bazzoli@wsu.edu)

Broadly, open science can be defined as “transparent and accessible knowledge that is shared and developed through collaborative networks” (Vicente-Saez & Martinez-Fuentes, 2018, p. 434). Hence, it refers to a broad range of practices aimed at detecting scientific fraud and enhancing transparency and replicability of research. In their focal article, Guzzo et al., (2022) highlighted several tensions between these values and applied research in organizations. In this commentary, we develop a slightly different argument: the open science movement, as a direct offspring of (post)positivist research paradigms<sup>1</sup>, has the potential to stifle epistemological and scientific pluralism and reproduce historical scientific hierarchies it purports to redress. In doing so, we distinguish between the spirit of open science (i.e., promoting participation, transparency, and access to science) and its implementations (e.g., OSF badges, TOP guidelines, and multi-laboratory research, but also sexist attacks on social media and podcasts by other scholars in the field [e.g., the Twitter pile-on in November 2021 regarding Roxanne Felig and her coauthors’ paper], and a general disregard of feminist epistemologies; Brabeck, 2021). In the first part of this commentary, we focus on open science’s ideals and examine a few unstated assumptions, advancing a set of equally valid assumptions based on constructionist thought, and then we discuss how unchecked implementations of open science practices can marginalize scholars that do not subscribe to its epistemic premises. We conclude with a few thoughts to improve the open science movement.

Bazzoli, A. (2022). Open science and epistemic pluralism: A tale of many perils and some opportunities. *Industrial and Organizational Psychology*, 15(4), 525-528. doi:10.1017/iop.2022.67

## Concern:

- Advocating open science also advocates certain post-positivist ideals, bolstering its dominance and, potentially, distorting and displacing different practices or minority research paradigms
- The open science movement has the potential to stifle epistemological and scientific pluralism and reproduce historical scientific hierarchies it purports to redress

# Nudging open science

- Link adherence to open science practices to tenure, promotion or publication can be problematic
- Increase the perception that open practices are not only normative, but also valued
- These ‘nudges’ may encourage the adoption of practices such as preregistration and data sharing where such practices are not sensible or feasible
- To be clear, there’s nothing wrong with ‘nudging’ post-positivist researchers to adopt open science research practices **within relevant research specialties**
- The concern is with ‘nudging’ researchers outside to adopt such specific practices




# ... example from within

*Industrial and Organizational Psychology* (2022), 15, 495–515  
doi:[10.1017/iop.2022.61](https://doi.org/10.1017/iop.2022.61)



FOCAL ARTICLE

## Open science, closed doors: The perils and potential of open science for research in practice

Richard A. Guzzo<sup>1\*</sup>, Benjamin Schneider<sup>2</sup> , and Haig R. Nalbantian<sup>1</sup>

<sup>1</sup>Workforce Sciences Institute, Mercer and <sup>2</sup>University of Maryland, Emeritus

\*Corresponding author. Email: [Rick.guzzo@mercer.com](mailto:Rick.guzzo@mercer.com)

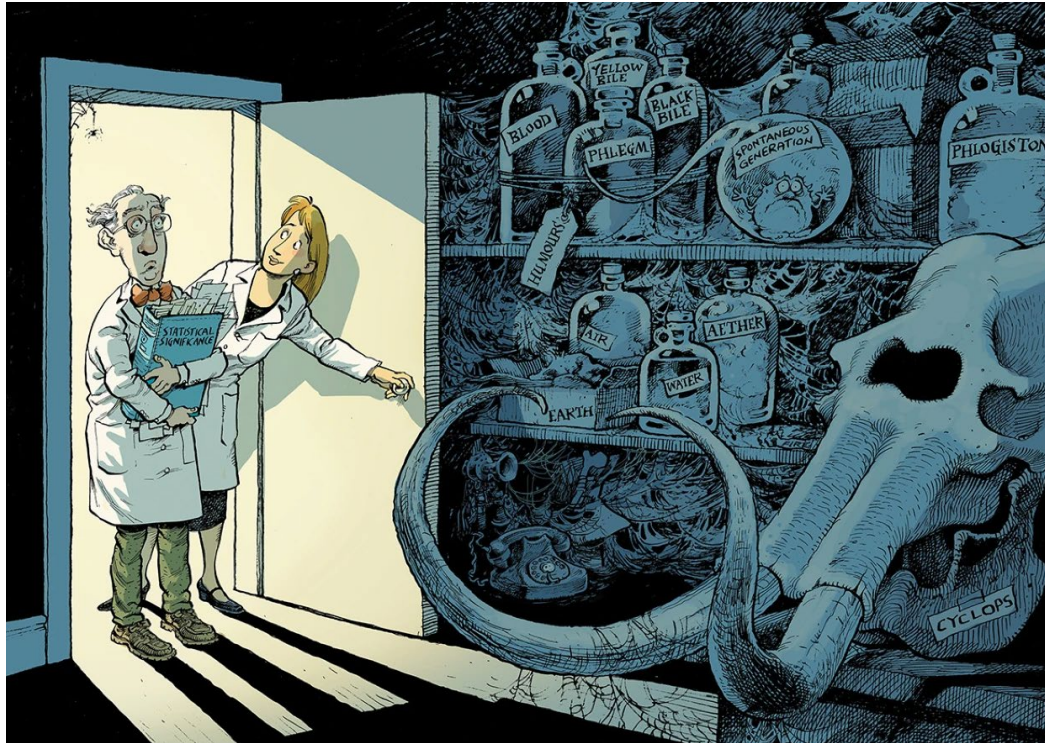
(Received 19 March 2020; revised 10 March 2021; accepted 31 May 2021)

### Abstract

This paper advocates for the value of open science in many areas of research. However, after briefly reviewing the fundamental principles underlying open science practices and their use and justification, the paper identifies four incompatibilities between those principles and scientific progress through applied research. The incompatibilities concern barriers to sharing and disclosure, limitations and deficiencies of overidentifying with hypothetico-deductive methods of inference, the paradox of replication efforts resulting in less robust findings, and changes to the professional research and publication culture such that it will narrow in favor of a specific style of research. Seven recommendations are presented to maximize the value of open science while minimizing its adverse effects on the advancement of science in practice.

- Incompatibility
  - #1: Disclosure and sharing
  - #2: Over-identification with the hypothetico-deductive model
  - #3: The paradox of replication
  - #4: Evolving cultural and professional norms

# Criticisms from within: No foundations or ‘best practices’?



- What is a successful replication?
- What's the role of NHST?
- What about inductive, abductive approaches?

ROYAL SOCIETY  
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research



Cite this article: Buzbas EO, Devezer B, Baumgaertner B. 2023 The logical structure of experiments lays the foundation for a theory of reproducibility. *R. Soc. Open Sci.* 10: 221042. <https://doi.org/10.1098/rsos.221042>

Received: 11 August 2022  
Accepted: 2 February 2023

Subject Category:  
Mathematics

Subject Areas:  
statistics

Keywords:  
reproducibility, replication, open science,  
metascience, experiment, statistical theory

## The logical structure of experiments lays the foundation for a theory of reproducibility

Erkan O. Buzbas<sup>1</sup>, Berna Devezer<sup>1,2</sup> and Bert Baumgaertner<sup>3</sup>

<sup>1</sup>Department of Mathematics and Statistical Science, <sup>2</sup>Department of Business, and <sup>3</sup>Department of Politics and Philosophy, University of Idaho, Moscow, ID 83844, USA

EOB, 0000-0003-1446-3447; BD, 0000-0002-5979-2781

The scientific reform movement has proposed openness as a potential remedy to the putative reproducibility or replication crisis. However, the conceptual relationship among openness, replication experiments and results reproducibility has been obscure. We analyse the logical structure of experiments, define the mathematical notion of idealized experiment and use this notion to advance a theory of reproducibility. Idealized experiments clearly delineate the concepts of replication and results reproducibility, and capture key differences with precision, allowing us to study the relationship among them. We show how results reproducibility varies as a function of the elements of an idealized experiment, the true data-generating mechanism, and the closeness of the replication experiment to an original experiment. We clarify how openness of experiments is related to designing informative replication experiments and to obtaining reproducible results. With formal backing and evidence, we argue that the current ‘crisis’ reflects inadequate attention to a theoretical understanding of results reproducibility.

Open practices in science have been intuitively proposed as a key to solving the issues surrounding reproducibility of scientific results. However, a formal framework to validate this intuition has been missing and is needed for a clear discussion of reproducibility.

To whom could 'reproducibility' be relevant?

# First we need to ask what ‘reproducibility’ actually means?

- Barba (2018) three categories of usage for ‘reproducibility’ and ‘replicability’:

- **A:** The terms are used with no distinction between them
- **B1:** ‘Reproducibility’ refers to instances in which the original researcher’s data and computer codes are used to regenerate the results, while ‘replicability’ refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study
- **B2:** ‘Reproducibility’ refers to independent researchers arriving at the same results using their own data and methods, while ‘replicability’ refers to a different team arriving at the same results using the original author's artifacts

<i>A</i>	<i>B1</i>	<i>B2</i>
political science economics	signal processing scientific computing econometry epidemiology clinical studies internal medicine physiology (neuro) computational biology biomedical research statistics	microbiology, immunology (FASEB) computer science (ACM)

# First we need to ask what ‘reproducibility’ actually means?

- ‘reproducibility’, ‘replicability’, ‘repeatability’, ... ‘trustworthiness’, ‘robustness’, ‘generalizability’, ... ‘transparency’

## PERSPECTIVE

### SCIENTIFIC INTEGRITY

## What does research reproducibility mean?

Steven N. Goodman,\* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of “research reproducibility” are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for “truth.”

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps312-341ps312.

- Methods
  - Results
  - Inferential
- } Reproducibility

## Replication and trustworthiness

Rik Peels<sup>a</sup> and Lex Bouter<sup>b</sup>

<sup>a</sup>Philosophy Department and Faculty of Religion and Theology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; <sup>b</sup>Department Of Epidemiology And Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

### ABSTRACT

This paper explores various relations that exist between replication and trustworthiness. After defining “trust”, “trustworthiness”, “replicability”, “replication study”, and “successful replication”, we consider, respectively, how trustworthiness relates to each of the three main kinds of replication: reproductions, direct replications, and conceptual replications. Subsequently, we explore how trustworthiness relates to the intentionality of a replication. After that, we discuss whether the trustworthiness of research findings depends merely on evidential considerations or also on what is at stake. We conclude by adding replication to the other issues that should be considered in assessing the trustworthiness of research findings: (1) the likelihood of the findings before the primary study was done (that is, the prior probability of the findings), (2) the study size and the methodological quality of the primary study, (3) the number of replications that were performed and the quality and consistency of their aggregated findings, and (4) what is at stake.

### KEYWORDS

Replication; trustworthiness; trust; replicability; reproducibility

- Three kinds of replication:
  - a reproduction
  - a direct replication
  - a conceptual replication

Peels, R., & Bouter, L. (2021). Replication and trustworthiness. *Accountability in Research*, 1-11. doi:10.1080/08989621.2021.1963708

# First we need to ask what ‘reproducibility’ actually means?

- Reproducible in which sense – eight categories
- Where kinds of ‘reproducibility’ is linked to types of research?

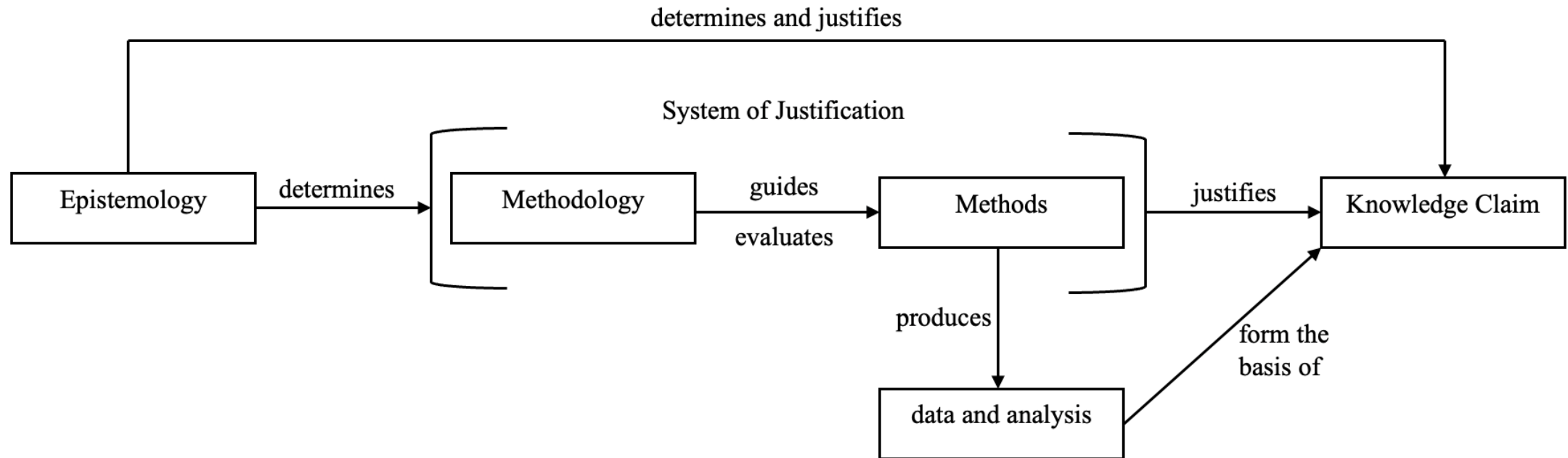
*Table 1.* Synoptic View of Types of Research Design/Methods and Related Understanding of Reproducibility Discussed in “Beyond the Ideal of Direct Reproducibility” Section.

Type of Research	Example	Degree of Control on Environment	Reliance on Statistics as Inferential Tool	Reproducible in Which Sense?
Software development	Computer engineering, informatics	<i>Total</i>	High	<i>Computational R:</i> Obtain same results from the same data.
Standardized experiments	Clinical trials, environmental safety controls	Very high	High	<i>Direct R:</i> Obtain same results from different runs of the same experiment.
Semistandardized experiments	Behavioral economics, experimental psychology, research on model organisms	Limited	Variable	<i>Scoping R:</i> Use differences in results to identify relevant variation. <i>Indirect R:</i> Obtain same results from different experiments. <i>Hypothetical R:</i> corroborate results implied by previous findings.
Nonstandard experiments and research based on rare, unique, perishable, inaccessible materials	Research on experimental organisms, archeology, paleontology, history	Low	Low	<i>Reproducible Expertise:</i> Any skilled experimenter working with same methods and materials would produce similar results
Nonexperimental case description	Case reports in medicine, (types of) multisited ethnography	None	Low	<i>Reproducible Observation:</i> Any skilled observer would pick out similar patterns
Participant observation	Ethology, participant observation in anthropology	None	None	<i>Irreproducible Observation:</i> different observers are assumed to have different viewpoints and produce different data and interpretations

# First we need to ask what 'reproducibility' actually means?

- **No consensus in sight**
- **More claims at taxonomies will follow**
- **Confusion will remain**
  
- Overall its has to with actually 'redoing' something or 'enabling the redoing of something'
- The importance of 'redoing something' is clearly linked to the pivotal element of experiment in notions of the scientific method
- Moving beyond 'experiment' the idea of 'redoing' becomes much more challenging
- The idea of 'redoing' is linked to objectivism, realism and (post)positivism

# To whom could 'reproducibility' be relevant? Pivotal role of epistemology



Inspired by: Carter, S. M., & Little, M. (2007). Justifying Knowledge, Justifying Method, Taking Action: Epistemologies, Methodologies, and Methods in Qualitative Research. *Qualitative Health Research*, 17(10), 1316-1328.



# To whom could 'reproducibility' be relevant?

- Ontology

- E.g., **there is truth**  $\Leftrightarrow$  there are multiple truths varying by individuals, contexts or contoured by relative access to societal power

- Epistemology

- E.g., **we can know this truth, or we are limited in our ability to know this truth, but we shall try our darnedest**  $\Leftrightarrow$  knowledge is co-constructed between researcher and participants and cannot be independent?

- Systems of Justification

- What are the established epistemic criteria for the type of research in a study that indicate trustworthiness or quality?

- Research Goals

- What is the motivation or goal behind the study?

# Example: Epistemic characteristics about the quality or trustworthiness (in qualitative settings)

Position (and key source)	Characteristic	Defining questions	Illustrative practices
Naturalistic inquiry (Lincoln and Guba, 1985)	Credibility	To what degree has the investigator given voice to the different constructions of reality found in one's data? Credibility is assessed by those one has studied.	"Prolonged engagement" (p. 301); "persistent observation" (p. 304); triangulation (e.g., different data sources, methods, investigators, etc.); "peer debriefing" (p. 308); "negative case analysis" (p. 309); "referential adequacy" (p. 313); "member checks" (p. 314)
	Transferability	Is there contextual similarity between the context one is studying and other contexts? The burden of proof for such a comparison lies with those who want to compare findings to other contexts more than with the original investigator.	Providing a lot of details (e.g., thick description) to "show" not "tell" the reader the findings
	Dependability	Has the investigator taken into account "both factors of instability and factors of phenomenal or design induced change"? (p. 299)	All the practices of credibility plus "stepwise replication" within the dataset (p. 317) and "inquiry audit" (p. 317)
	Confirmability	Was there a process for verifying the data? Confirmability is a characteristic of the data, not the investigator.	Inquiry audit; triangulation; "reflexive journal" (p. 319); "audit trail" (p. 319); "audit process" (p. 320)
Case studies / positivism (Yin, 2003)	Construct validity	Are your measures operationalizing your concepts correctly?	"Use multiple sources of evidence; establish a chain of evidence; have key informants review draft" (p. 34)
	Internal validity	Is there a causal relationship between variables or constructs?	"Do pattern-matching; do explanation-building; address rival explanations; use logic models" (p. 34)
	External validity	Can findings be generalized and to what domain?	"Use theory in single-case studies; use replication logic in multiple case studies" (p. 34)
	Reliability	Can it be replicated across cases in the study?	"Use case study protocol; develop case study database" (p. 34)
Ethnography (Locke and Golden-Biddle, 1997)*	Authenticity	Communicating that the author was in the field and did not do violence to the experience of the informants	"Particularizing everyday life" (p. 601); "delineating the relationship in the field" (p. 603); "depicting the disciplined pursuit and analysis of data" (p. 604); "qualifying personal biases" (p. 605)
	Plausibility	Does the academic audience "buy" it in that it (a) makes sense and (b) makes a contribution? (p. 600)	"Normalizing unorthodox methodologies" (p. 605); "drafting the reader" (p. 606); "legitimizing the atypical" (p. 606); "smoothing the contestable" (p. 608); "differentiating findings—a singular contribution" (p. 609); "building dramatic anticipation" (p. 610)
	Criticality	Does the study make the author rethink assumptions about the field or their own work?	"Carving out room to reflect" (p. 610); "provoking the recognition and examination of differences" (p. 610); "imagining new possibilities" (p. 611)
Process research (Langley, 1999; Gehman et al., 2018)†	Longitudinal data	Has the author studied things over time?	Showing that the data fit with the time span of the examined process; interviewing people about facts or events if asking them to be retrospective or, if interviewing them in real time, trying to understand how their interpretation of events evolves; using one or a combination of different analytical strategies: narrative, quantification; attending to risk of retrospective reconstruction

\* It is important to note that these authors are arguing why ethnographic work is convincing, not trustworthy. We include their arguments here as they are about what makes for good qualitative research.

† Langley does not use the term "trustworthiness" in her descriptions but does lay out the fundamentals of process research.

Figure 3 - Note. Reprinted from "Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research", by Pratt, M. G., Kaplan, S., & Whittington, R., 2020, *Administrative Science Quarterly*, 65(1), 1–19. <https://doi.org/10.1177/0001839219887663>

# If it is relevant to some degree, how feasible is it?

- Should we link kind of ‘reproducibility’ to types of research?
- What about ‘epistemology’?
- E.g. qualitative and mixed methods can be based on post-positivist paradigms
- E.g. are clinical trials and experiments in HE-particle physics similar types of research?

“In clinical trials aimed to test ... drugs ... [t]he degree of controls and standardization being implemented is among the most impressive and tightly scrutinized within the biomedical realm. Experiments conducted in particle accelerators in physics are similarly controlled, the focus on one centralized experimental apparatus being particularly helpful in establishing a fixed framework within which experiments can be successfully repeated ... typically evaluated through recourse to statistical inference, thus privileging statistics as a key validating tool for reasoning from evidence

*Table 1.* Synoptic View of Types of Research Design/Methods and Related Understanding of Reproducibility Discussed in “Beyond the Ideal of Direct Reproducibility” Section.

Type of Research	Example	Degree of Control on Environment	Reliance on Statistics as Inferential Tool	Reproducible in Which Sense?
Software development	Computer engineering, informatics	Total	High	<i>Computational R:</i> Obtain same results from the same data.
Standardized experiments	Clinical trials, environmental safety controls	Very high	High	<i>Direct R:</i> Obtain same results from different runs of the same experiment.

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Vol. 36B, pp. 129-146): Emerald Publishing Limited.

# If it is relevant to some degree, how feasible is it?

<b>Relevance</b>	<b>Epistemology</b>	<b>System of Justification</b>	<b>Research Goal</b>	<b>Feasibility Subject of Investigation</b>	<b>Research Setup/ Resource Dependence</b>	<b>Methodological Uncertainty</b>	<b>Theoretical Uncertainty</b>
<b>Drug Trials</b>	Post-Positivism	Validity; reliability; control	Applied; profit; drug effectiveness; matters of fact	Interactive; high plasticity; high historicity; living kinds	Lab-experiment; AB-test; small-N; less control, low-medium resource dependence	High	High; low theoretical guidance
<b>Particle Physics</b>	Post-Positivism	Validity; reliability; control	Matters of fact; Nomothetic; theory-testing	Indifferent; non-living kinds; low-medium plasticity; low historicity	Lab-experiment; predictive test; large-N; strong control; high resource dependence	Low	Low; high theoretical guidance

# If it is relevant to some degree, how feasible is it?

Relevance	Epistemology	System of Justification	Research Goal	Feasibility Subject of Investigation	Research Setup/ Resource Dependence	Methodological Uncertainty	Theoretical Uncertainty
<b>Drug Trials</b>	Post-	Validity;	Applied;	Interactive;	Lab-experiment;	High	High;
<p>Redoing less feasible, more noise, more variation, more uncertainty → epistemic evidence restricted</p>							
<b>Particle Physics</b>	Post-Positivism	Validity; reliability; control	Matters of fact; Nomothetic; theory-testing	Indifferent; non-living kinds; low-medium plasticity; low historicity	Lab-experiment; predictive test; large-N; strong control; high resource dependence	Low	Low; high theoretical guidance

# Aim to develop a framework

- Clarify what is wanted irrespective of conceptual preferences
  - What should vary?
- Determine degree of relevance in relation to 'redoing' and 'enabling' based on epistemology, systems of practice and goals for a knowledge production mode
  - Not relevant, relevant to some degree, highly relevant
- Determine the feasibility in relation to 'redoing' and 'enabling' based on simple coding of subject, research setup and theoretical and methodological uncertainties
  - High, medium or low feasibility
  - Degree of feasibility → epistemic expectancy

Is 'reproducibility' for all?

# Thanks for your attention

[jws@ps.au.dk](mailto:jws@ps.au.dk); [su@ps.au.dk](mailto:su@ps.au.dk)

Visit: <https://tier2-project.eu/>



Funded by the European Union.

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission.

Neither the EU nor the EC can be held responsible for them.