CWTS
Meaningful metrics

Reproducibility concerns in scientometrics differ from other fields

Ludo Waltman
Centre for Science and Technology Studies, Leiden University

16th International Conference on Scientometrics & Informetrics
Wuhan, China
October 17, 2017

Universiteit Leiden

Reproducibility in psychology

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 previously 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.

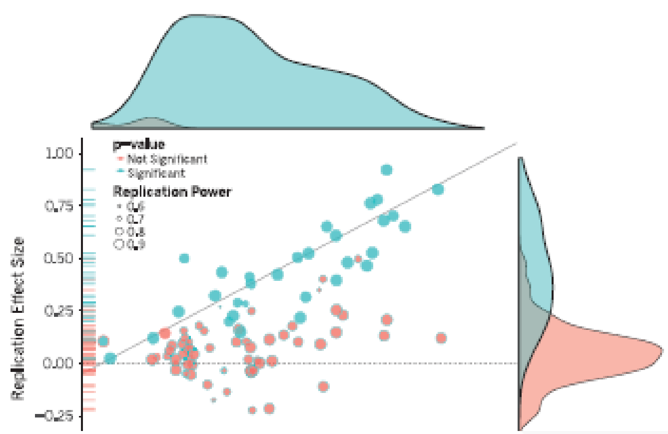
RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a 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 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.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716

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 replications 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 possible; replication 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



CWTS

Reproducibility in psychology

- Open Science Collaboration (2015):
  - “Scientific claims should ... gain credence ... by the replicability of their supporting evidence ... (Scientific) debates are meaningless ... if the evidence being debated is not reproducible”
  - “Direct replication ... is the means of establishing reproducibility of a finding with new data”
- Focus on exact reproducibility as opposed to conceptual reproducibility

CWTS

Why scientometrics is different from psychology

1. Large share of methodological (as opposed to empirical) research
2. Key data sources in principle available to all scientometricians
3. Many descriptive studies at population level, with no need for sampling
4. Journals put little pressure on publishing ‘surprising’ results

No reason to expect that scientometrics has major reproducibility problems similar to psychology

Biggest threat to reproducibility in scientometrics

- Mistakes in data analysis!
- Scientometricians perform relatively complex data analyses...
- ... but often lack a solid training in data analysis
- Which scientometric research teams do systematically try to replicate their own data analyses?
- Magnitude of the problem is unknown, but is likely to be significant

CWTS

Conceptual reproducibility: Debate on performance-based funding

Journal of Informetrics 11 (2017) 905–918

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Perverse effects of output-based research funding? Butler's Australian case revisited

Peter van den Besselaar[a,*], Ulf Heyman[b], Ulf Sandström[c]

[a] Network Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[b] Uppsala University, Uppsala, Sweden
[c] KTH Royal Institute of Technology, Stockholm, Sweden

A R T I C L E I N F O

Article history:
Available online 15 July 2017

A B S T R A C T

More than ten years ago, Linda Butler (2003a) published a well-cited article claiming that the Australian science policy in the early 1990s made a mistake by introducing output based funding. According to Butler, the policy stimulated researchers to publish more but at the same time less good papers, resulting in lower total impact of Australian research compared to other countries. We redo and extend the analysis using longer time series, and show that Butlers' main conclusions are not correct. We conclude in this paper (i) that the currently available data reject Butler's claim that "journal publication productivity has increased significantly… but its impact has declined", and (ii) that it is hard to find such evidence also with a reconstruction of her data. On the contrary, after implementing evaluation systems and performance based funding, Australia not only improved its share of research output but also increased research quality, implying that total impact was greatly increased. Our findings show that if output based research funding has an effect on research

Conceptual reproducibility: Debate on performance-based funding

- Van den Besselaar et al. (2017):
  - “Our aim is thus not to replicate Butler’s study”
  - “We conclude ... (i) that the currently available data reject Butler’s claim ... and (ii) that it is hard to find evidence for this also with a reconstruction of her data”
  - “Butler’s negative evaluation of the Australian funding system has been uncritically used in many other studies ... But science policy researchers have refrained from checking the data and from replication.”

CWTS

Conclusions

- Reproducibility concerns in for instance psychology do not translate directly to scientometrics
- Mistakes in data analysis probably represent biggest threat to reproducibility in scientometrics
- Replication of data analysis deserves more attention; may require sharing of data and computer code
- Key data sources in principle available to all scientometricians, but not in practice

CWTS