Meta-Psychology, 2025, vol 9, MP.2024.4194
https://doi.org/10.15626/MP.2024.4194
Article type: Commentary

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Felix D. Schönbrodt
Reviewed by: Gilad Feldman, Steven Verheyen
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/UFQHM

Replication Value Increases With Transparency, Test Severity, and Societal Impact

René Bekkers[1]
[1]VU Amsterdam

This comment argues that replications should not be prioritized as a function of citation count and sample size, but instead as a function of societal impact, test severity, and the information value of a replication.

Keywords: replication; transparency; test severity; impact; citation count; sample size

Which claims in science should receive more thorough scrutiny in replication attempts? Isager et al. (2025) approach replication value from a utilitarian perspective, and propose a formula to quantify it. In the formula, the value of a replication for a claim increases with the scientific impact of the initial study reporting the claim measured by the average number of citations per year since the study appeared, and reduces with the certainty of the finding, measured by the number of observations supporting the initial claim. In this comment, I offer two criticisms on the proposal for the replication value formula, and I propose an alternative: that the empirical content of a claim and the opportunity to improve the severity of the test of a transparently reported, reasonably large effect size that is societally relevant makes claims more important targets for replications.

Criticism 1: Societal impact deserves more weight in the value of replication attempts

Claims are more valuable as targets for replication attempts if they have larger consequences for science and society. The more we can learn from a replication, viewed from both a scientific knowledge and societal application and practical utility perspective, the more valuable is its replication. The utility of replicating a claim depends on its information value, how strongly it is believed to be true before the replication, and the outcome of the replication.

The potential consequences of a replication are more serious if people in practice make more important decisions on claims that turn out to be false positives or false negatives. The societal consequences of false claims are more grave if the health and wealth of a larger proportion of humanity is affected, for instance if more lives are at stake, if the economic value of phenomena that the claim is about are larger, and especially if current policies do not achieve desired outcomes.

Criticism 2: Citation counts and sample size are bad indicators of replication value

The replication value formula proposed by Isager et al. (2025) multiplies citation counts with sample size, implying that more highly cited studies are more valuable targets for replication attempts, and even more so when they rely on small samples. However, both citation counts and sample size are invalid indicators of the scientific utility of replication attempts.

A large body of scholarship has documented a plethora of problems with the number of citations as an indicator of scholarly impact (Bornmann & Daniel, 2008). Citation counts are not only biased by time and field effects, but also by journals, editors, authors, readers, article factors, and technical problems. Citation counts are susceptible to gaming. They are inflated by guest authorships (Meursinge Reynders et al., 2024) and by coercive citation practices (Wilhite & Fong, 2012).

Citation counts also convey information about the likelihood of replication success: studies that do not replicate are cited more frequently than studies that do replicate (Serra-Garcia & Gneezy, 2021).

Using citation counts as an ingredient for replication value would make studies that are less likely to replicate in the first place more likely to become targets of replication attempts. In that case we can be sure of replication failures for years to come. Sample size is also invalid as an indicator of replication value. On one hand, significant findings in small samples can reflect stronger effect sizes. Once a researcher has an inkling that the effect is large, the sample size can justifiably be kept small. On the other hand, at a given effect size, findings in small samples are more likely to be flukes, and result in less precise estimates. In the presence of incentives for publishing significant findings, false positive small

n studies are more likely to be selected for publication when they achieve statistical significance, resulting in publication bias. We do not learn much from failures to replicate candidate gene findings in underpowered studies. Billions of dollars wasted in the hunt for effects of specific genes in small samples could have been prevented (Harden, 2021).

An alternative

I propose that studies are more useful to replicate when they report a practically meaningful effect size with transparently reported and available data and code, and when a replication can provide a more stringent test of a regularity with higher information value. Practically meaningful. The value of a replication should not only be determined by scientific criteria. The societal implications of actionable knowledge should also be taken into account when targets for replication are identified. Before policies are put into clinical or medical practice, the claims on which they are based should be replicated successfully. Especially when the falsehood of a claim could do more harm in practice if it is used to inform a wide-ranging policy, and especially when the evidence from the initial study is not particularly strong, it should be replicated. Transparently reported. Candidates for replication should be of sufficient quality such that the reported findings actually have a chance of replication. Claims based on poorly documented interventions, measurements, data collections and models of analysis are less worthy of replication attempts. Claims based on materials that are documented transparently enough to enable replication can be evaluated with respect to the strength of the evidence. A replication is more informative if the methods of the original study leave more room for doubt. Given the ubiquity of questionable research practices, the default attitude to a claim of general regularity based on a single study should be a skeptical one until it is replicated independently (Feynman, 1974).

Test severity

Observation of guidelines for high quality research including preregistration, collecting large samples for sufficient statistical power, and transparent and complete documentation of research materials and analysis procedures, are likely to enhance the likelihood that claims can be reproduced successfully. Claims based on studies observing these guidelines are less uncertain and hence less important to replicate. Claims in registered reports are particularly less likely to be false positives based on questionable research practices, and should therefore be viewed as presenting stronger evidence (Scheel et al., 2021).

A lack of statistical power is one of key characteristics of uncertainty. As claims based on larger original effect sizes are more likely to replicate (Open Science Collaboration, 2015), studies reporting strong effects at moderate statistical power are more valuable as targets for replication attempts than similarly powered studies reporting weak effects. Larger effect sizes are more likely to have meaningful implications in practice.

Information value

In addition to opportunities for methodological improvements, the scientific value of replication attempts should also be based on the Popperian criterion of the amount of empirical information conveyed by a theory (Popper, 1972). Scientific consequences of claims are more grave if the claim is stated in more general terms, and if the claim is a more fundamental part of theories. If an empirical discovery necessitates the reformulation of established theories, it should check out in independent replication attempts. The scientific and practical value of replications should be evaluated jointly. Reducing the uncertainty about a claim with more stringent tests is more important when the societal implications of false beliefs about the claim are more grave. Instead of citation counts and sample size, I suggest that the degree of uncertainty resulting from a lack of severity of the test should guide decisions to attempt replication of claims that have larger practical and theoretical implications.

Author Contact

René Bekkers, r.bekkers@vu.nl

Conflict of Interest and Funding

The author declares no conflicts of interest. The author received funding from the Netherlands Organization for Scientific Research (NWO), grant 406.23.SW.042.

Author Contributions

Conceptualization and writing: RB

Open Science Practices

This article is purely conceptual and as such is not eligible for Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Feynman, R. P. (1974). Cargo cult science: Some remarks on science, pseudoscience, and learning how not to fool yourself.

Harden, K. P. (2021). "reports of my death were greatly exaggerated": Behavior genetics in the postgenomic era. *Annual Review of Psychology*, *72*, 37–60. https://doi.org/10.1146/annurev-psych-052220-103822

Isager, P., van 't Veer, A., & Lakens, D. (2025). Replication value as a function of citation impact and sample size. *Meta-Psychology*, *9*. https://doi.org/10.15626/MP.2022.3300

Meursinge Reynders, R. A., Cavagnetto, D., ter Riet, G., Di Girolamo, N., & Malicki, M. (2024). Automatically listing senior members of departments as co-authors is highly prevalent in health sciences: Meta-analysis of survey research. *Scientific Reports*, *14*, 5883. https://doi.org/10.1038/s41598-024-55966-x

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford University Press.

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Wilhite, A. W., & Fong, E. A. (2012). Coercive citation in academic publishing. *Science*, *335*(6068), 542–543. https://doi.org/10.1126/science.1212540