Meta-Psychology, 2025, vol 9, MP.2024.4312
https://doi.org/10.15626/MP.2024.4312
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: Carlisle Rainey, Leonie Dudda
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/FHB5N

The Replication Dilemma: Potential Challenges in Measuring Replication Value—A Commentary on Isager et al. (2025)

Adrien Fillon[1] and Subramanya Prasad Chandrashekar[2]
[1]ERA Chair in Science and Innovation Policy & Studies (SInnoPSis), University of Cyprus, Nicosia, Cyprus
[2]Department of Psychology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

The authors (Isager et al., 2025) start with the main assumption that researchers' efforts toward replications are constrained by resources, and they propose a simple, practically scalable framework of research replication value that guides the researchers and the scientific community at large intending to achieve bigger bang for the buck. Specifically, the authors propose a framework that combines citation impact and sample size of the original articles as a metric for assessing replication value. This implies that original studies with higher scores on this metric can be prioritized for replication efforts.
We thoroughly agree with the authors' assumption and indeed support the view of working towards an optimal framework that helps the community achieve maximum research impact from the replication efforts. In this commentary, we propose to discuss three important limitations that have to be considered before using such metrics.

Keywords: replication, value, open science, meta-science

Scientific citations are biased

Citations and the associated metric of journal impact factor are an overt reflection of the prestige of academic journals and significantly impact researchers' career advancements. Therefore, scientists, universities, and publishers all have strong interests in being associated with journals of high- impact factor. However, such a metric can be biased on several accounts. Firstly, data indicates that papers accepted in prestigious journals were first submitted and rejected by less prestigious journals, indicating randomness in the acceptance of a paper, and citations associated with it (Calcagno et al., 2012; Chorus and Waltman, 2016). Papers published in more prestigious journals tend to receive more citations, not only because of their inherent value but also as a result of chance. Secondly, the publication ecosystem increasingly includes journals that inflate their citation rate by increasing within-journal citations (Hanson et al., 2023). For example, Web of Science delisted the journal *Sustainability* in 2023 (Brainard, 2023), but Scopus decided to continue listing it (MDPI, 2024), despite concerns by the community that the journal does not provide enough rigor in the peer-review process. These profitable companies show that they want to gain from the economic benefits of science, while avoiding responsibility for the quality of scientific publications. We should tie the responsibility to the listing of publication to increase transparency and rigor in the scientific process.Fourthly, citations can be bought and sold by citation-boosting services. In a recent investigation, (Ibrahim et al., 2024), researchers found a service providing citations in 14 journals indexed in Scopus. Finally, there have been instances of covert practices by the journals to include "sneaked" references registered as metadata (e.g., Crossref) but these references do not appear in the actual publications (Besançon et al., 2023). Meanwhile, platforms such as Crossref maintain that they are not responsible for the quality of the metadata.

In summary, we acknowledge the simplicity and appeal of using citation counts as a measure of replication value in an ideal world. However, in the current academic ecosystem, citation counts are not an unbiased indicator of impact. To address this, citation counts should only be considered after verifying the paper's reliability by screening for low-quality journals, highly cited fraudulent scientists, problematic references, and papers from paper mills. A recent article by Ioannidis and Maniadis (2024) partially addresses this issue by exploring the use of metrics to counteract gamed metrics.

We have three additional remarks regarding citations. First, although citations are a measure of scientific impact, citations may or may not directly reflect the public interest, as such a connection is not straightforward. Second, the use of indexes such as citation could drive replication toward elitism. In the current ecosys-

tem, articles are not only cited based on their theoretical or practical possibilities but are also cited because of the status associated with the main author/s, or the media coverage. One of the main predictors of future citations is the degree to which the author is already an authority in the domain (Dong et al., 2016). Therefore, we should expect more replications for studies of the same set of authors that are authoritative in their domains. Diversity in researchers and in sample demographics is a core component of open science, and we think that the two metrics used in the proposal can go against this principle. Stated differently, we are concerned about the fact that, in order to streamline our response towards reproducibility crisis, this proposal can increase the generalizability crisis (Syed, 2023; Yarkoni, 2022). Finally, as suggested by Carlisle Rainey, one of our reviewers, while this has not been quantified to our knowledge, many papers are cited not necessarily for their substantive claims but for reasons such as similar measures or statistical tests. Such citations will need to be weighted in the equations for replication values.

Using mathematical model to answer philosophical questions

The question of the value of replication requires assumptions concerning the place of replication in the scientific inquiry. The assumptions are that the goal of science is to discover regularities about nature, and that reproducible empirical findings indicate regularities. However, there are replicable but non- existent effects (for example, the field of homeopathy, Sigurdson et al., 2023), non-replicable but real effects (we can think, for example, about historical effects such as lockdown during Covid-19), and more importantly, there might be a linear relationship between the original and replication effect size (Devezer et al., 2021), with a potential reproducibility rate of true effect close to 0, and potential reproducibility rate of false effect close to 1. It leads to the question of using reproducibility criteria that are not aligned with epistemological reasons to conduct a replication, as "true" and "false" effects can only be disentangled with epistemological appraisal, and not solely on the replication rate. In Paul Meehl's words, increasing the verisimilitude of a theory is made by having "A good track record [..] of successful and almost-successful risky predictions, of "hits" and "near misses" for point or interval predictions of low tolerance, and predictions of function forms" (Meehl, 1990, p. 138-139). The non-consideration of the risk associated with testing the theory is an important shortcoming in the proposal for the value of replication.

If— as Devezer and Colleagues (2021) proposed— the replicability of true effects is low, it might not be surprising to have an observed probability of replication in the psychology of about 50% (See, for example, the replication database, Röseler et al., 2024). This low rate can be largely explained by the complexity of the experiment, the number of predictive factors manipulated in a particular experiment, and the difficulty of keeping entropy low (Fanelli, 2022). Successfully replicating an experiment increases the confidence we have in the model being tested, but not necessarily in the theory associated with the hypothesis, if the hypothetical effect is not properly captured by the model. In addition, an open question remains: what could be done with unsuccessful replications? Should we replicate them again based on the citation impact of the replication and the sample size (Boyce et al., 2024)? How much resources should we invest to further ascertain a model is replicated, instead of a theory? More crucially, how can we increase verisimilitude by doing only one replication of a study, if p-values should be understood based on their distribution?

We also note that most studies are designed to answer a theoretical question around the causal relationships (or co-relations) between variables, or practical questions about how useful an effect could be if the tested effect could be made available to the general population. Replication value should accommodate how a finding or an effect of interest maps onto the causality scheme of the theory, or has practical relevance, informing the usefulness and scalability of the findings. Guidelines to assess the potential pragmatism or exploration of study designs already exist (e.g., PRECIS-2, Loudon et al., 2015) but are not yet widely used in social sciences. Using these tools would help filter out studies that have poor theoretical or practical relevance.

Qualifiers for replication value metric

Although, in the earlier sections, we note concerns regarding the citation counts as a measure of impact, we still see the utility of the measure given certain qualifiers. One such qualifier could be to take account of the outliers in the citation rate. For instance, some citations could be "too good to be real" (see Figure 1a). Ibrahim and colleagues (2024) proposed to use of a citation concentration index to evaluate the outliers. Perhaps, including the citation concentration index as qualifying the citation count can help mitigate issues concerning citation measure.

We also believe that a framework such as PRECIS-2 (PRagmatic Explanatory Continuum Indicator Summary; Loudon et al., 2015) could serve as a valuable a priori filter. PRECIS-2 helps researchers design studies by considering them on a pragmatic (ecological) or explanatory (theoretical) continuum. By applying

Figure 1

Suspicious peer review proposed list of studies to include..

(a) *Note. This is a screenshot of a X post by Ian McCarthy published on the 12/05/2024. The red border highlighting the text of interest was added by Dr. McCarthy.*

this framework, poorly designed studies from original publications—those that are limited to inform theory or practice—can be filtered out. In cases where such evaluations don't meet the goals of informing theory or practice, replicators can then consider modifications to the original study to add value based on the requirement for informing theory or practice. Additionally, we propose creating replication selection policies through a consensus method. Such an approach would take into account the current practices, knowledge, and needs of a specific field. While achieving consensus initially requires effort, a subjective assessment can prioritize the replication of the most valuable studies while ensuring transparency and democratic decision-making. One example of a fast and worldwide consensus methodology was developed by Vickers and colleagues (2024) that gathers opinions of researchers regarding the fact that COVID-19 was caused by a virus. Using this method to assess the importance of replicating a particular paper or method within the field, by asking major contributors to this field with a survey could strengthen the argument for a need to replicate. Importantly, we do not advocate replacing a formal equation with purely subjective assessments of replication value. Instead, we recommend using a combination of a priori, systematized, and domain-dependent objective and subjective filtering systems before using an algorithmic approach. These a priori filters will improve the effectiveness of the proposed metric.

Author Contact

Correspondence concerning this article should be addressed to Adrien Fillon. Email: adrienfillon@hotmail.fr and ORCID: 0000-0001-8324-2715 Prasad Chandrashekar email: prasad.chandrashekar@ntnu.no and ORCID: 0000-0002-8599-9241

Conflict of Interest and Funding

Adrien Fillon is supported by the project SInnoPSis, funded by Horizon 2020 under grant agreement ID: 857636. The authors have no conflicts of interest to declare in this study.

Author Contributions

Adrien Fillon: Conceptualization, Investigation, Methodology, Visualization, Writing - original draft, and

Writing - review & editing.

Subramanya Prasad Chandrashekar: Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, and Writing - review & editing.

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

Boyce, V., Prystawski, B., Abutto, A. B., Chen, E. M., Chen, Z., Chiu, H., Ergin, I., Gupta, A., Hu, C., Kemmann, B., Klevak, N., Lua, V. Y. Q., Mazzaferro, M., Mon, K., Ogunbamowo, D., Pereira, A., Troutman, J., Tung, S., Uricher, R., & Frank, M. C. (2024). Estimating the replicability of psychology experiments after an initial failure to replicate. https://doi.org/10.31234/osf.io/an3yb

Brainard, J. (2023). Fast-growing open-access journals stripped of coveted impact factors. https://doi.org/10.1126/science.adi0098

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & De Mazancourt, C. (2012). Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns. *Science*, *338*(6110), 1065–1069. https://doi.org/10.1126/science.1227833

Chorus, C., & Waltman, L. (2016). A Large-Scale Analysis of Impact Factor Biased Journal Self-Citations (W. Glanzel, Ed.). *PLOS ONE*, *11*(8), e0161021. https://doi.org/10.1371/journal.pone.0161021

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), rsos.200805, 200805. https://doi.org/10.1098/rsos.200805

Dong, Y., Johnson, R. A., & Chawla, N. V. (2016). Can Scientific Impact Be Predicted? *IEEE Transactions on Big Data*, *2*(1), 18–30. https://doi.org/10.1109/TBDATA.2016.2521657

Fanelli, D. (2022). The "Tau" of Science - How to Measure, Study, and Integrate Quantitative and Qualitative Knowledge. https://doi.org/10.31222/osf.io/67sak

Hanson, M. A., Barreiro, P. G., Crosetto, P., & Brockington, D. (2023). The strain on scientific publishing [Publisher: arXiv Version Number: 2]. https://doi.org/10.48550/ARXIV.2309.15884

Ibrahim, H., Liu, F., Zaki, Y., & Rahwan, T. (2024). Google Scholar is manipulatable [Version Number: 1]. https://doi.org/10.48550/ARXIV.2402.04607

Ioannidis, J. P. A., & Maniadis, Z. (2024). Quantitative research assessment: Using metrics against gamed metrics. *Internal and Emergency Medicine*, *19*(1), 39–47. https://doi.org/10.1007/s11739-023-03447-w

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

Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., & Zwarenstein, M. (2015). The PRECIS-2 tool: Designing trials that are fit for purpose. *BMJ*, *350*(may08 1), h2147–h2147. https://doi.org/10.1136/bmj.h2147

MDPI. (2024). Sustainability Passes Rigorous Scopus Reevaluation Process. Retrieved March 12, 2025, from https://www.mdpi.com/about/announcements/7352

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Röseler, L., Kaiser, L., Doetsch, C., Klett, N., Seida, C., Schütz, A., Aczel, B., Adelina, N., Agostini, V., Alarie, S., Albayrak-Aydemir, N., Aldoh, A., Al-Hoorie, A. H., Azevedo, F., Baker, B. J., Barth, C. L., Beitner, J., Brick, C., Brohmer, H., … Zhang, Y. (2024). The Replication Database: Documenting the Replicability of Psychological Science. *Journal of Open Psychology Data*, *12*(1), 8. https://doi.org/10.5334/jopd.101

Sigurdson, M. K., Sainani, K. L., & Ioannidis, J. P. (2023). Homeopathy can offer empirical insights on treatment effects in a null field. *Journal of Clinical Epidemiology*, *155*, 64–72. https://doi.org/10.1016/j.jclinepi.2023.01.010

Syed, M. (2023). The Slow Progress towards Diversification in Psychological Research. https://doi.org/10.31234/osf.io/bqzs5

Vickers, P., Adamo, L., Alfano, M., Clark, C., Cresto, E., Cui, H., Dang, H., Dellsén, F., Dupin, N., Gradowski, L., Graf, S., Guevara, A., Hallap, M., Hamilton, J., Hardey, M., Helm, P., Landrum, A., Levy, N., Machery, E., … Mitchell Finnigan, S. (2024). Development of a novel methodology for ascertaining scientific opinion and extent of agreement (N. Mubarak, Ed.).

PLOS ONE, *19*(12), e0313541. https://doi.org/10.1371/journal.pone.0313541

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685