Meta-Psychology, 2024, vol 8, MP.2023.3735
https://doi.org/10.15626/MP.2023.3735
Article type: Commentary
Published under the CC-BY4.0 license

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

Edited by: Daniel Lakens
Reviewed by: Andreas Brandmaier, Roman Stengelin
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/27NF6

Responsible Research Assessment Should Prioritize Theory Development and Testing Over Ticking Open Science Boxes

Hannah Dames[1], Philipp Musfeld[1], Vencislav Popov[1], Klaus Oberauer[1], and Gidon T. Frischkorn[1]
[1]Department of Psychology, University of Zurich, Zurich, Switzerland

We appreciate the initiative to seek for ways to improve academic assessment by broadening the range of relevant research contributions and by considering a candidate's scientific rigor. Evaluating a candidate's ability to contribute to science is a complex process that cannot be captured through one metric alone. While the proposed changes have some advantages, such as an increased focus on quality over quantity, the proposal's focus on adherence to open science practices is not sufficient, as it undervalues theory building and formal modelling: A narrow focus on open science conventions is neither a sufficient nor valid indicator for a "good scientist" and may even encourage researchers to choose easy, pre-registerable studies rather than engage in time-intensive theory building. Further, when in a first step only a minimum standard for following easily achievable open science goals is set, most applicants will soon pass this threshold. At this point, one may ask if the additional benefit of such a low bar outweighs the potential costs of such an endeavour. We conclude that a reformed assessment system should put at least equal emphasis on theory building and adherence to open science principles and should not completely disregard traditional performance metrices.

Keywords: Research Assessment, Open Science, Theory Building

Academic assessment aims to identify people who will advance science by creating impactful knowledge and exhibiting strong leadership. Evaluating candidates' ability to achieve this in a rigorous and unbiased way is crucial in the hiring process. Schönbrodt et al. (2022) criticize current indicators used in evaluation procedure and propose a set of alternative metrics, implemented by Gärtner et al. (2022), focusing on scientific rigor (as measured mostly by adherence to open science practices) instead of research productivity. They also suggest hiring committees to additionally consider published data sets and research software when assessing researchers, and to abandon the use of the journal impact factors (JIF) and the h-index during assessment.

We welcome the initiative to discuss and re-evaluate the use of traditional indicators of scientific productivity (i.e., h-index) and to consider alternative metrics in the assessment of research quality. In our view, however, the complexity of evaluating a person's ability to significantly contribute to science dooms any toolbox approach to assessment. Substituting one flawed metric with another will not solve this problem. Although a fully optimal hiring process may never be reached, it is crucial to consider multiple metrics, including ones reflecting scientific rigor. We discuss the pros and cons of the suggested changes and highlight three main challenges of the current proposal.

Moving away from quantity towards quality

The proposal has several strengths: First, traditional quantitative performance metrics like citation count or h-index do not necessarily reflect the quality of a candidates' research, nor their personal qualifications. We thus support the inclusion of additional parameters to inform hiring decisions, particularly those reflecting scientific rigor, because researchers who promote methodological rigor and open science are rarely rewarded. Practicing open science may even be perceived as harmful, working against the goal of maximizing other research output (i.e., publications) that advance one's career. Second, Gärtner et al. (2022) proposed a structured assessment plan that includes explicit metrics for the first stage of evaluation. This approach not only improves transparency and reduces subjective biases, but also empowers applicants by enabling them to comprehend the hiring committee's decision-making process. Third, limiting the number of an applicant's papers submitted for evaluation aids efforts to move away from valuing quantity (i.e., number of publications) towards promoting quality (e.g., contribution to psychological research). This focus on quality over quantity conveys the message that success in academia is not determined by the number of publications in top journals but by meaningful and impactful research.

Nevertheless, we disagree that the initial selection process (intended to create a longlist of candidates) should mainly prioritize the proper use of open science methods (e.g., preregistration) while neglecting theory building and formal modelling. In our view, one of the greatest deficits in current psychological research is the lack of theory development (see also Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019) and efforts in counteracting this should be evaluated and rewarded. We therefore question whether the recommended measures, specifically the ones proposed in the evaluation sheets by Gärtner et al. (2022), improve the current imperfect status of assessment.

Science Progresses by Theory

The authors argue that bad scientific practices are "one likely explanation for the low replicability rates" in psychology (Schönbrodt et al., 2022, page 9). A stronger emphasis on methodological rigor during research assessment is meant to counteract this. Methodological rigor is mostly evaluated by the implementation of open science conventions such as preregistration and FAIR data (Gärtner et al., 2022). Although we generally support the idea of incentivizing scientific rigor, we worry that it is insufficient for measuring good psychological research. One of our primary concerns is that the strict compliance to open science principles is neither a sufficient nor a valid indicator for a "good scientist."

Theorizing is undervalued

Science advances by good theories and their iterative testing and correction (Deutsch, 2011; Popper, 1959). The lack of proper theorizing in psychological research is a major cause of its low replicability (see Oberauer & Lewandowsky, 2019; Szollosi et al., 2020; Van Rooij, 2019). If the aim is to identify researchers that will significantly contribute to science, efforts in building and testing strong theories need to be valued and rewarded in the assessment process. Yet, the current proposed solution greatly undervalues the role of theorizing and formal modelling (e.g., more items reward open science than theorizing). A reformed assessment system should put at least as much emphasis on theory building (e.g., formalizing theories as computational models) as on adherence to open science principles.

Open science and good science are not necessarily the same

A researcher can follow all open science principles, while still doing irrelevant research (as noted by Schönbrodt et al., 2022, themselves). To give an extreme example: Without any theorizing, a candidate makes a random prediction (Szollosi et al., 2020) and they preregister an experimental design and analysis plan to test it. Subsequently, they upload the collected data in a FAIR format. The work gets published in a peer-reviewed conference proceeding. This candidate would receive a high score on the proposed evaluation sheets (Gärtner et al., 2022), without having any impact on scientific progress. Thus, a narrow focus on open science conventions is not a useful metric to discriminate researchers who do "good" vs "bad" science. It may even encourage researchers to opt for easily preregisterable studies and to refrain from time-intensive theory building. Yet, the interests of science might be better served by (non-preregistered) work testing and developing strong theories.

A "Toolbox Approach" of Psychological Science? The Risk of Gamification

While we appreciate methodological rigor, a rather narrow focus on a metric around it could lead to harmful consequences. The very act of defining a metric to quantify research quality or scientific rigor encourages researchers to "game" their way to better evaluations (Macdonald, 2022). Traditional metrics incentivize early career researchers to publish in high-impact journals to advance their academic careers. We worry that the proposal will simply shift candidates' focus to maximize the new, arguably easier-to-achieve, metrics (e.g., publish more data sets), rather than encouraging theoretical development in psychology. The current imbalance may give the impression that adherence to open science principles is more important than theory development and may even reinforce a lack of theorizing in psychology.

The metrics Gärtner et al. (2022) proposed aim to establish a minimum level of methodological rigor in the first assessment stage (Figure 2 in Schönbrodt et al., 2022, page 6), which is extended by a more sophisticated evaluation of scientific impact in a second stage. However, given the possibility to "game the system" in this stage, most applications will soon surpass this threshold. As the first selection stage then does not filter out many applications, it has become ineffective for research assessment

How to move forward?

We want to make clear that we do not argue against the role of open science in improving the replicability of psychological science. We routinely use open science practices ourselves and encourage collaborators and trainees to do the same. Our goal is to point out the risks and shortcomings in narrowly focusing on the proposed criteria in the academic hiring process. We understand that we offer more criticism than we provide solutions. Based on the challenges described here, we see three points of improvements:

The scientific hiring process is complex; simply replacing traditional metrics with new ones is unlikely to adequately capture the quality of a candidate's research. While acknowledging the limitations of concentrating on a limited number of performance indicators (e.g., relying on total citation counts, the h-index, or the JIF, Barnes, 2017; Brembs et al., 2013; Serra-Garcia & Gneezy, 2021), we caution against completely dismissing currently used ones. For example, JIF (e.g., Bornmann & Williams, 2017; Waltman & Traag, 2020), early publication success (Laurance et al., 2013; Lee, 2019), the h-index (Hirsch, 2007), and a linear combination of an author's past productivity and the past citation rate of their average paper (Hönekopp & Khan, 2012) show some predictive power for future research success (e.g., future publication output or number of citations). In light of the sparse and mixed findings concerning the predictive validity of existing performance indicators, and a lack of empirical studies demonstrating an added value of alternative measures, it appears premature to dismiss all traditional performance indicators when evaluating research quality and potential future research success. Nonetheless, we support the notion that these performance metrics should not be used as the sole criterion for assessing research quality, given the valid criticism for individual performance metrics. Instead, a composite measure including and appropriately weighing scientific rigor (e.g., adherence to open science principles), research quality (e.g., theory building and computational modelling), and impact as well as productivity (e.g., citation rate, h-index or alternative measures, Bihari et al., 2023) would cover many of the aspects considered by Schönbrodt et al. (2022).

We also suggest moving away from the disproportionally strong focus on preregistration and open science. The count of a candidate's preregistered experiments is not a good metric to evaluate their ability to advance science (see Szollosi et al., 2020, for various arguments). Scientific rigor deserves attention in the evaluation process but should not be valued more than theory building. Furthermore, scientific rigor does not boil down to following open science practices.

Finally, we expect that with time most applicants will pass the first evaluation stage. If so, assessing a candidate's ability to contribute to science in the second evaluation stage becomes most critical. Yet, this stage remains underdeveloped in the current proposal. Excluding most quantitative information from this stage could open the door to subjective preferences and biases of the committee members. In a second commentary, we outline how the second evaluation stage could be improved through changes to the composition of hiring committees (see Frischkorn et al., 2023).

Author Contact

Corresponding author: Hannah Dames; Mail: hannah.dames@outlook.com; Address: Binzmühlestrasse 14/22, CH-8050 Zürich, Switzerland

Conflict of Interest and Funding

The authors declare no conflict of interest.

Author Contributions

HD prepared the initial draft of this comment. Apart from that, all authors made equal contributions to its content as well as the reviewing and editing of it.

Open Science Practices

This article is theoretical and as such was not eligible for any Open science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Barnes, C. (2017). The h-index debate: An introduction for librarians. *The Journal of Academic Librarianship*, *43*(6), 487–494. https://doi.org/https://doi.org/10.1016/j.acalib.2017.08.013

Bihari, A., Tripathi, S., & Deepak, A. (2023). A review on h-index and its alternative indices. *Journal of Information Science*, *49*(3), 624–665. https://doi.org/10.1177/01655515211014478

Bornmann, L., & Williams, R. (2017). Can the journal impact factor be used as a criterion for the selection of junior researchers? a large-scale empirical study based on researcherid data. *Journal of Informetrics*, *11*(3), 788–799. https://doi.org/https://doi.org/10.1016/j.joi.2017.06.001

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00291

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. penguin uK.

Frischkorn, G. T., Dames, H., Musfeld, P., Popov, V., & Oberauer, K. (2023). Responsible research assessment requires structural more than procedural reforms.

Gärtner, A., Leising, D., & Schönbrodt, F. (2022). Responsible research assessment ii: A specific proposal for hiring and promotion in psychology.

Hirsch, J. E. (2007). Does the <i>h</i> index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198. https://doi.org/10.1073/pnas.0707962104

Hönekopp, J., & Khan, J. (2012). Future publication success in science is better predicted by traditional measures than by the h index. *Scientometrics*, *90*(3), 843–853. https://doi.org/10.1007/s11192-011-0551-2

Laurance, W. F., Useche, D. C., Laurance, S. G., & Bradshaw, C. J. A. (2013). Predicting Publication Success for Biologists. *BioScience*, *63*(10), 817–823. https://doi.org/10.1525/bio.2013.63.10.9

Lee, D. H. (2019). Predicting the research performance of early career scientists. *Scientometrics*, *121*(3), 1481–1504. https://doi.org/10.1007/s11192-019-03232-7

Macdonald, S. (2022). The gaming of citation and authorship in academic journals: A warning from medicine. *Social Science Information*, *61*(4), 457–480. https://doi.org/10.1177/05390184221142218

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.

Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., Schubert, A.-L., et al. (2022). Responsible research assessment i: Implementing dora for hiring and promotion in psychology.

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

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in cognitive sciences*, *24*(2), 94–95.

Van Rooij, I. (2019). Psychological science needs theory development before preregistration. *Psychonomic Society Featured Content*.

Waltman, L., & Traag, V. A. (2020). Use of the journal impact factor for assessing individual articles: Statistically flawed or not? *F1000Research*, *9*.