Meta-Psychology, 2025, vol 9, MP.2024.4601
https://doi.org/10.15626/MP.2024.4601
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: Daniël Lakens
Reviewed by: Invited Commentaries
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/SBPN5

Responsible Research Assessment I: Implementing DORA and CoARA for hiring and promotion in psychology

Felix D. Schönbrodt[1], Anne Gärtner[2,3], Maximilian Frank[1], Mario Gollwitzer[1], Malika Ihle[1], Dorothee Mischkowski[4,5], Le Vy Phan[6], Manfred Schmitt[7], Anne M. Scheel[8], Anna-Lena Schubert[9], Ulf Steinberg[10], and Daniel Leising[2]

[1]Ludwig-Maximilians-Universität München
[2]Technische Universität Dresden, Germany
[3]Freie Universität Berlin, Germany
[4]Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern, Germany
[5]Leiden University, The Netherlands
[6]Universität Bielefeld, Germany
[7]Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau
[8]Utrecht University, The Netherlands
[9]Universität Mainz, Germany
[10]Manres GmbH, Germany

The use of journal impact factors and other metric indicators of research productivity, such as the h-index, has been heavily criticized for being invalid for the assessment of individual researchers and for fueling a detrimental "publish or perish" culture. Multiple initiatives call for developing alternatives to existing metrics that better reflect quality (instead of quantity) in research assessment. This report, written by a task force established by the German Psychological Society, proposes how responsible research assessment could be done in the field of psychology. We present four principles of responsible research assessment in hiring and promotion and suggest a two-phase assessment procedure that combines the objectivity and efficiency of indicators with a qualitative, discursive assessment of shortlisted candidates. The main aspects of our proposal are (a) to broaden the range of relevant research contributions to include published data sets and research software, along with research papers, and (b) to place greater emphasis on quality and methodological rigor in research evaluation.

Keywords: DORA, CoARA, research assessment, research quality, impact

Several initiatives, such as the San Francisco Declaration on Research Assessment (DORA) or the Coalition for the Advancement of Research Assessment (CoARA) call on academic institutions to abandon the use of invalid quantitative metrics of research quality and productivity in hiring and promotion. Most prominently, this concerns the Journal Impact Factor (JIF). Although this metric never was intended to be used this way (e.g., Garfield, 2006), researchers and institutions often use the JIF as a proxy for scientific quality (Hrynaszkiewicz et al., 2024; McKiernan et al., 2019). However, there are convincing arguments that it should not be used for the assessment of individual achievements (e.g., Ramani et al., 2022). One reason is that it correlates *negatively* with multiple objective and subjective indicators of research quality, such as strength of evidence or replication success, but *positively* with reporting errors or the presence of QRPs and HARKing (Brembs et al., 2013; Dougherty & Horne, 2022; Kepes et al., 2022). That is, a higher JIF is – if anything – statistically associated with *poorer* research quality. Another reason is that quantitative indicators of "productivity" falsely imply that scientific quality is easy to quantify (e.g., Paulus et al., 2018). Furthermore, the use of relatively distal quantitative measures such as the JIF, the *h*-index, or simply the quantity of publications in hiring and promotion may have the unintended side-effect of fueling a "publish or perish" culture in which the use of questionable research practices is incentivized. This risk is significant, given the high incentive value of attaining a permanent position in academia (Leising et al., 2022a) and the fact that, at the same time, academia is largely lacking effective mechanisms of quality control and self-correction (Vazire & Holcombe, 2022).

The need for developing alternatives to existing metrics and indicators has been recognized by multiple ini-

tiatives that are currently working on research assessment schemes aiming to prioritize quality over quantity (European Commission 2021: Towards a reform of the research assessment system; Paris Call on Research Assessment 2022; Dutch public knowledge institutions and funders of research 2021: Recognition and Rewards: Room for everyone's talent; LERU 2022: A Pathway towards Multidimensional Academic Careers: A LERU Framework for the Assessment of Researchers; The Hong Kong Principles for assessing researchers, Moher et al., 2020; DFG: Package of Measures to Support a Shift in the Culture of Research Assessment; see also Moher et al. 2018). Most notably, the *Coalition for the Advancement of Research Assessment* (CoARA; https://coara.eu/) has more than 700 institutional members (by June 2024) who pledged to create an action plan with the goal to reform their research assessment procedures according to the commitments of the coalition. These commitments include to recognise the diversity of contributions to research, including practices that contribute to robustness, openness, transparency, and the inclusiveness of research, or to base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators (see https://coara.eu/).

The German Psychological Society (DGPs) signed DORA in 2021, became a signatory of CoARA in 2023, and tasked a group of experts among its members with preparing a proposal on how the key principles of these two initiatives may be practically implemented for the field of psychology: What should be the guiding principles of responsible research assessment? And how can we pragmatically replace the current, flawed metrics of research productivity with ones that more validly reflect reliable, incremental knowledge gain? The primary goal of such an assessment scheme would be to ensure that actual research quality is sustained (or even promoted) when evaluation metrics are being maximized – both actively, when researchers strategically decide how to behave in order to further their own careers (sometimes to the extent of gaming the system), and passively, when institutions select and reward individuals who scored highest in the rankings based on these parameters (Bakker et al., 2012; Franco et al., 2014; Müller & De Rijcke, 2017; Smaldino & McElreath, 2016; Tiokhin et al., 2021).

The present whitepaper reports some results achieved by the task force, revised based on extensive peer-review, including 21 published commentaries in different outlets. We propose *four principles of responsible research assessment* applicable to the hiring and promotion of individual researchers, and a *two-phase assessment procedure* for hiring committees that combines the objectivity and efficiency of indicators with a qualitative and narrative assessment of the work of shortlisted candidates. In a separate document, the RESQUE framework (RESearch QUality Evaluation) is proposed as an actionable way of implementing these principles (Gärtner et al., 2025; see also the response to the commentaries of this special issue, Leising et al., 2024). Supplemental material and the current version of RESQUE (which is continuously developed) can be found at https://www.resque.info.

As a complex social system, science is constantly in flux. Because researchers react to institutional norms and incentives as well as to each other, any set of institutional rules will eventually require adjustments to remain relevant and effective. This whitepaper presents such an adjustment against the backdrop of the replication crisis (Nosek et al., 2022). While the assessment of scientific quality will always remain a challenge with imperfect solutions, we argue here that the past focus on quantitative measures of publication activity has failed to uphold minimal standards of scientific rigor that are necessary for sustainable progress in the discipline (cf. Uygun Tunc & Pritchard, 2022). To correct the course, the two main aspects of our proposal are to (1) broaden the range of academic contributions that count and to (2) place greater emphasis on methodological rigor in research evaluation.

Currently, the number of peer-reviewed publications (co-)authored by an applicant and the amount of grant money acquired by an applicant ("third-party funding") are among the most decisive criteria in making hiring decisions (Abele-Brehm & Bühner, 2016). However, the range of valuable academic contributions is much broader – both in terms of the "products" that are created and in terms of the contributor roles[1] that researchers play in creating them. We argue that the following *five* areas of academic contributions should be considered in assessments: Research, teaching, academic leadership (e.g., management and organizational skills, supervision, strategic thinking), service to the academic institution/field (e.g., contributing to committees and academic societies; editorial and reviewing activities), and societal impact (e.g., science communication, citizen science; see Figure 1). Acknowledging that a certain activity, such as science communication, is a valuable academic contribution does *not* imply that every researcher must engage in this activity. In contrast,

[1] Contributor roles can be made explicit using CRediT (Contributor Roles Taxonomy, see https://credit.niso.org; NISO CRediT Working Group, 2022), a high-level taxonomy with 14 roles (e.g., conceptualization, statistical analyses, writing the manuscript) that people may play in the production of scholarly output.

we assume that it is highly unlikely that a single person excels in all five areas of the multidimensional profile. Furthermore, not all academic jobs even require expertise in all five areas (e.g., purely research focused positions might ignore the teaching dimension). We advocate for employing realistic job profiles which do not expect such a "perfect" (but unrealistic) applicant.

In the remainder of the present paper, we exclusively focus on the *Research* dimension, because this is the area in which an urgent need for alternative evaluation criteria has been most clearly articulated (Abele-Brehm & Bühner, 2016; European Commission, Directorate-General for Research and Innovation, 2022; League of European Research Universities, 2022; Leising et al., 2022a). Several actionable suggestions for the other four assessment dimensions can be found in the whitepaper of LERU (2022).

Four principles of responsible research assessment in psychological science

Principle 1: Academic contributions are multifaceted. Regarding research contributions, do not only value (a) published journal articles, but also other research reports (including preprints, conference proceedings, Stage 1 registered reports with or without an "in principle acceptance", protocols, monographs, book chapters, psychometric test manuals), (b) data sets and (c) research software development.

We suggest that three kinds of research contributions should be considered by hiring and promotion committees for the Research dimension (see Figure 1): (a) narrative texts (including journal articles, preprints, conference proceedings, Stage 1 registered reports with or without an "in principle acceptance", protocols, monographs, book chapters, psychometric test manuals)[2], (b) published data sets, and (c) research software. Committees should encourage applicants to list all of their contributions in all three categories, preferably in separate sections of a structured CV.

Principle 2: Quantitative indicators do have practical advantages, but they have to be valid and need to be used responsibly.

We see two main reasons why metrics are so common in research assessment. First, metrics attempt to make research assessment more objective, to combat certain types of biases, and to facilitate a direct comparison between applicants in selection processes. Second, the use of metrics makes handling the sheer volume of applications manageable for hiring committees. For example, in Germany it is not uncommon for a hiring committee to receive more than 100 applications for a single tenured professorship position. This makes it likely that committees – contrary to the widespread ideal of focusing on the quality of the applicants' research – will ultimately resort to using the existing flawed quantity metrics, simply to be able to somehow complete their task (Schmitt, 2022).

However, problems arise when indicators are not valid, and research assessment focuses on "what can easily be counted" rather than "what really counts" (Abramo & D'Angelo, 2014, p. 2). Hence, being aware of the general risks of any metric (Goodhart's law), we call for a critical evaluation of existing indicators and the development and use of alternative and better indicators. The challenge is to preserve the undeniable advantages of quantitative indicators – *objectivity* and *efficiency* – while, at the same time, improving their *validity*.

Indicators should be transparent: it should be known how they are derived, and applicants should know which indicators will be used to evaluate them. The numeric values of each indicator should be reproducible and ideally based on an open and interoperable data infrastructure. Indicators also need to be fair (i.e., systematic bias should be avoided to the extent possible), for example by adjusting them for academic age, parental leave, or disadvantages (Wouters et al., 2019). Consequently, and in line with the DORA principles, we join the call for abandoning the use of the JIF and of the *h*-index (CWTS, 2021) in assessing individual papers or researchers[3]. Furthermore, proprietary black-box performance assessment tools (such as Elsevier SciVal, Interfolio ResearchFish, Clarivate InCites, or the now abandoned ResearchGate Score) should not be used in such assessments either, as their validity as measures of scientific merit/potential is at least as questionable, and their calculation cannot be independently reproduced (e.g., Copiello & Bonifaci, 2018).

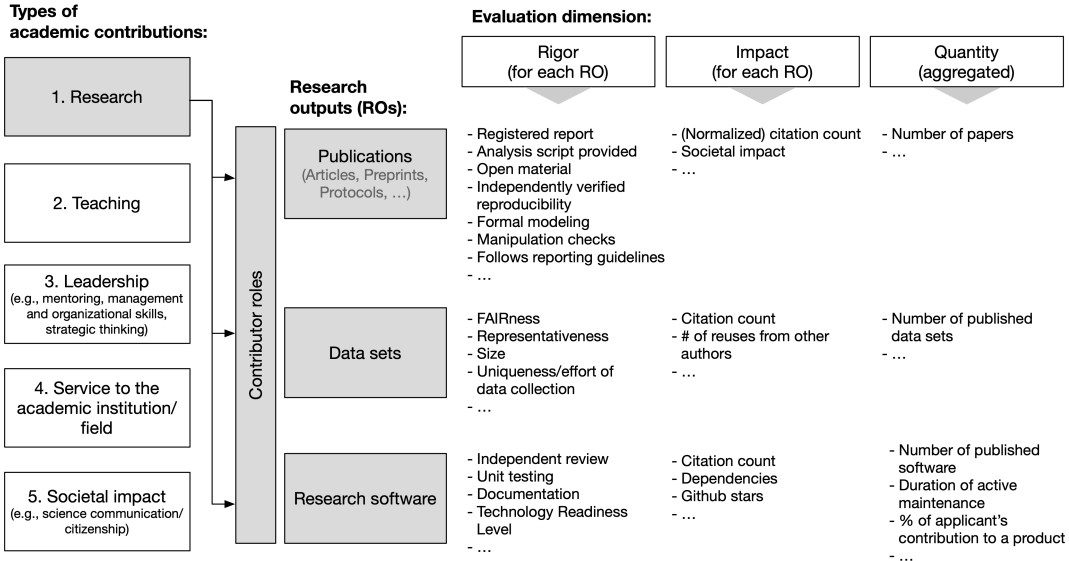
Using indicator-based evaluation systems usually implies a loss of nuance and a risk of not being able to capture certain cases that do not fit the proposed categories well. We therefore suggest that, in hiring processes, the use of indicators should be limited primarily to initial selections from a longlist of applicants, thus fo-

[2] We explicitly make no distinction between "peer-reviewed" and "non-peer-reviewed" contributions and suggest to give full consideration to the latter, because the quality-assuring function of peer-review is debatable in general (Bornmann, 2012; Cicchetti, 1991) and unknowable for any specific contribution due to its mostly closed nature.

[3] The original purpose of the JIF was to aid librarians to select journals for which they wanted to purchase institutional subscriptions. It might have some validity for this use case.

Figure 1

Five types of academic contributions, three kinds of research outputs, three evaluation dimensions, and exemplary evaluation criteria



cusing on the basic skills and craftsmanship that every researcher needs to possess ("Two-phase assessment", see Figure 2 and below). All applicants passing a certain threshold on these indicators should be considered in the next phase of the hiring process (instead of just selecting the "best" *n* applicants). This way, minor variations in scores will not unfairly disqualify applicants that demonstrate sufficient craftsmanship. Committees that do not want to employ a strictly algorithmic approach but prefer using a more holistic one even in the first selection stages, may simply use the same indicator scores as input to a more holistic human expert judgment. This at least ensures that methodological rigor will play *some* role in the process. In any case, applicants should be given the opportunity to explain in a few sentences if and why they think that something important is being overlooked when using these indicators.

In contrast, the evaluation of shortlisted candidates in hiring contexts, and candidates up for promotion should not rely on such an indicator-based algorithm and rather focus on a more qualitative, content-oriented assessment that explicitly considers all types of academic contributions.

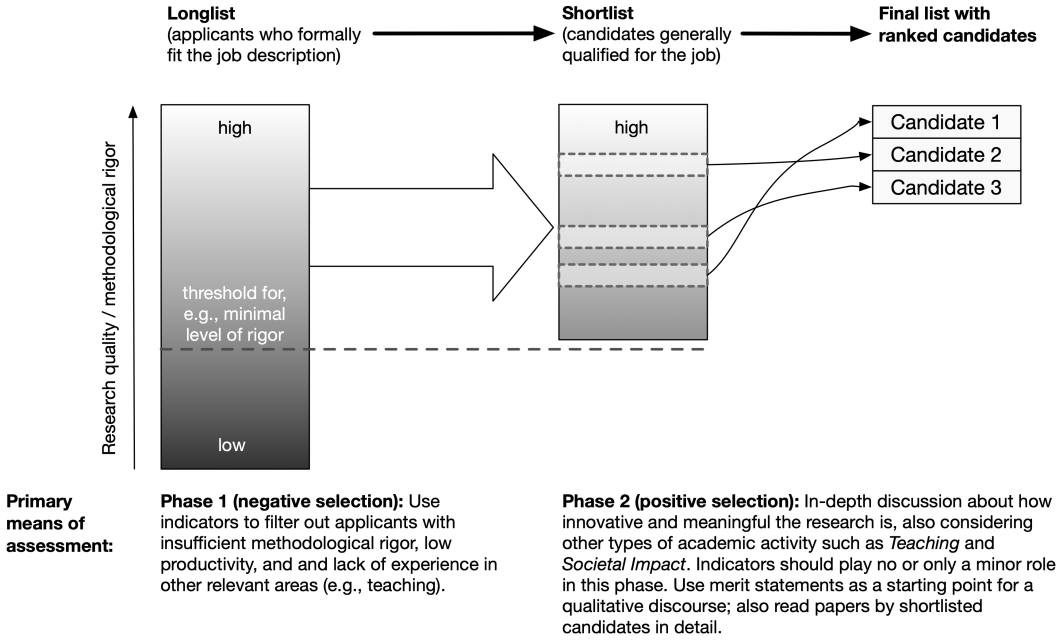
Principle 3: Use (a) methodological rigor, (b) impact, and (c) quantity as independent evaluative dimensions in research assessment.

Many problems in research assessment have been identified as misuses of indicators for unintended goals, or uses of indicators that do not reflect the intended construct. One example is the frequent use of impact measurements like citation counts or Journal Impact Factor (JIF) as proxies for assessing quality. However, non-replicable publications are cited more than replicable ones (Serra-Garcia & Gneezy, 2021), and citation counts and impact factors have been found to be weak (and sometimes negative) predictors of research quality (Brembs et al., 2013; Dougherty & Horne, 2022). In order to avoid such misnomers, we call to clearly define and distinguish between three independent evaluation dimensions, namely: Methodological rigor, impact, and quantity (see Figure 1). Rigor and impact are separately assessed for each research output (e.g., a paper), whereas the quantity of research outputs is counted on the level of researchers or institutions. Evaluation criteria for each dimension will differ somewhat between types of research outputs.

(a) Methodological rigor (as one central aspect of quality). Research *quality* is a multidimensional concept (Hooper, 2022), ranging from fundamental aspects such as "adhering to basic standards of good scientific practice" to more complex and sometimes elusive aspects such as "creativity, innovation, and ingenuity". Even when researchers claim to "know good science when they see it", it is difficult to objectively operationalize it. Sometimes, whether a person's research activity has produced some valid and relevant contribution to knowledge can only be judged decades after publication. What we can do, however, is assess whether a given research output even only has the po-

Figure 2

A two-phase selection process



tential to make such a contribution. This may be assessed using indicators of methodological rigor, as one central and basic aspect of quality. "Rigor" refers to the research activities themselves (i.e., not their outcomes): Whether they have been skillfully executed according to standards of good scientific practice within the field. These standards nowadays include practices that contribute to robustness, openness, and transparency of research (see CoARA commitment 1, or the UK Research Excellence Framework[4]), but also, for example, aspects of theoretical rigor.

We explicitly do not suggest that quality may be *reduced* to rigor – it is easy to imagine research that has been performed rigorously and at the same time is completely irrelevant. But rigor can be seen as a necessary condition (or at least a probabilistic enabler) for the generation of impactful and valid knowledge.

There is a relatively high level of consensus regarding desirable features of empirical studies that will make robust knowledge gains more likely. Among these are the existence of replication attempts, a theoretical motivation for conducting research, independent verifications of computational correctness ("reproducibility checks"), good statistical power, and many more. For example, preregistration lowers the risk of bias in the analysis and interpretation of data, even more so when published as a registered report where additional quality control is performed by reviewers prior to data collection; access to and sufficient documentation of data is a logical precondition for independent reproducibility checks; the presence of open code has been identified as the single largest predictor for successful reproductions of published results (Laurinavichyute et al., 2022); and theoretically motivated research approaches are essential to drive accumulation of knowledge within a scientific field (Oberauer & Lewandowsky, 2019). High methodological rigor, which can include aspects of empirical and theoretical rigor, is a necessary (though not sufficient) condition for making a robust and substantial scientific contribution that also contributes to a cumulative science. Unfortunately, these vital features of research have played a very minor role in research assessment so far (Abele-Brehm & Bühner, 2016), although in recent years some professorship job ads contained these features as desirable or essential criteria (Nosek et al., 2022). We argue that aspects of intrinsic research quality have to be moved to the forefront of research assessment (Leising et al., 2022a, 2022b).

Research outputs that do meet certain standards with respect to methodological rigor will then have to be fur-

[4]See the report published in 2023 by a committee of the UK House of Commons, which asserts that the criteria in the Research Excellence Framework need "to assure that transparency is a prerequisite of top-scoring research" (p. 44; https://committees.parliament.uk/publications/39343/documents/194466/default/)

ther evaluated in terms of more complex quality criteria such as "innovation". This is the goal in the second stage of the two-stage process that we suggest here (see below). This second stage relies much more on holistic expert judgments and narrative discourse - both within a committee and between a committee and candidates.

(b) Impact. Once it has been established that a piece of research output does feature the necessary methodological rigor, its academic and/or societal impact may be determined.

In our view, high-quality research that has an impact is probably more valuable than high-quality research that has no impact, all else being equal. The paper by Leising et al. (2025) contains a relatively detailed discussion of what impact is (e.g., as opposed to merit) and how it may be measured. In a nutshell, we recommend considering impact *only* after a certain level of methodological rigor has been established. If impact is supposed to play a role in an assessment procedure, citation numbers (as one measure of scientific impact) should be adjusted for the age of the publication and the field in which an applicant works. Additional *impact statements* may also be gathered from applicants, explaining how (in terms of content, not citation numbers) some contribution has had, or could have, a relevant impact on the respective field. The same applies to societal impact.

(c) Quantity. Comparable to meta-analyses that use a risk-of-bias analysis to only include primary studies that surpass a certain quality threshold, we only look at the subset of papers by an applicant that fulfills certain quality standards. Hence, once a certain minimum level of methodological rigor has been established for a scholar's scientific contributions, we may actually start counting them. That is because we find it legitimate to consider scholar A more scientifically productive than scholar B when both have provided good quality contributions but A has produced more.

Importantly, both impact and quantity are highly confounded with academic age and other factors, such as the scientific field in which a person works. They should therefore be normalized against inputs relevant to the objective of the assessment, such as academic age (e.g., papers per year), third-party funding (e.g., papers per 100.000 € funding), or field (e.g., field-normalized citation rates). Finally, we refrain from using terms such as "research performance" or "excellence" because they are often vague or inconsistently defined (typically as an unclear mixture of quality, impact, and quantity, see also Moore et al., 2017).

Principle 4: Value quality over impact and quantity.

The goals of assessment and selection procedures in academia may differ from instance to instance: Do hiring institutions want to excel in university rankings? Accumulate as much third-party funding as possible? Maximize their publication volume, even if that may mean sacrificing quality? Shine in the realm of teaching or mentoring? The diagnostic tools that committees use should match the respective goals of assessment in each instance.

When the goal is scientific progress, defined as achieving valid and credible knowledge, it is important to differentiate *progress* and *quality*: "Quality is primarily an activity-oriented concept, concerning the skill and competence in the performance of some task. Progress is a result-oriented concept, concerning the success of a product relative to some goal. All acceptable work in science has to fulfill certain standards of quality. But it seems that there are no necessary connections between quality and progress in science. Sometimes very well-qualified research projects fail to produce important new results, while less competent but more lucky research leads to success. Nevertheless, the skillful use of the methods of science will make progress highly probable. Hence, the best practical strategy in promoting scientific progress is to support high-quality research." (Niiniluoto, 2024, p. 6).

Along these lines of argumentation, and assuming a low predictive validity when forecasting scientific progress, we argue that the first and most essential goal of evaluating individual researchers should be to select and promote researchers who skillfully demonstrate the ability to produce research that has a high intrinsic quality, according to standards of good scientific practice. Methodological rigor goes a long way in establishing these properties, and preliminary evidence suggests that assessing this rigor is possible in a reliable fashion (for an early version of the RESQUE indicators, trained student assistants achieved an ICC(1,1) of 0.91 for the overall empirical rigor score; Etzel et al. 2024). For concrete suggestions of measurable quality indicators, see Gärtner et al. (2025), Leising et al. (2022a); Leising et al. (2022b) and the RESQUE website.

But what about the pure quantity of a person's research activity? Producing a large number of publications, for example, may indeed reflect a scholar's scientific brilliance, efficiency, diligence, and hard work – but given the current lack of effective quality control in the academic system (Vazire & Holcombe, 2022), the same outcome may also be achieved by simply cutting corners in terms of methodological rigor or even honesty (Gopalakrishna et al., 2022; Leising et al., 2022a, 2022b). In fact, often articulated impediments to im-

plementing open science practices are the perceived extra effort and the associated opportunity costs (e.g., Houtkoop et al., 2018). This trade-off – a negative correlation between rigor and quantity on the *within*-person level – is independent from a potential *between*-person effect: Some researchers are arguably more capable of producing research outputs of high quality *and* higher quantity than others.[5] This between-person variance is the main target of assessment procedures.

Some level of quantitative productivity is certainly necessary for a researcher to be regarded as successful. Therefore, quantity – as long as it does not come at the expense of quality – may have a (limited) role to play. For example, applicants for a permanent position in academia may have to demonstrate a quantitative minimum of outputs that surpasses the threshold of required methodological rigor (e.g., a minimum number of published journal articles). However, the current practice of selecting competitors mainly via indicators of pure quantity, combined with a widespread lack of proper quality controls, sets an incentive for everybody to invest into the quantity, rather than the quality, of their own research. The bad scientific practices thus encouraged are one likely explanation for the low replicability rates that have now been well-established both within and beyond the field of psychology (Nosek et al., 2022).

Especially for early career researchers (ECRs), it is essential that the additional time and effort required by more rigorous research methods (e.g., pre-registering a study, sharing data in a FAIR way, writing reproducible code, engaging into formal modeling) is made visible and rewarded as part of an assessment process. An evaluation system that focuses almost exclusively on quantity sets the wrong incentives and lets researchers who are committed to sound scientific work fall short in relation to their colleagues who are more willing to maximize the quantity of their output at the cost of its quality.

A two-phase assessment for hiring professors: Methodological rigor and a multifaceted profile of academic contributions

We suggest assessing the academic merit and potential as professors in two consecutive phases. In Phase 1, primarily the overall methodological rigor of an applicant's research should be assessed. This may be accomplished in an algorithmic manner based on quality-based indicators (Leising et al., 2022a). The outcome should be compared against a threshold – a minimal level of rigor – to guide the selection of candidates to be considered for the shortlist. This negative selection (see Figure 2) builds on the empirical finding that interrater-agreement in peer review is higher at the low end of the quality scale (Cicchetti, 1991): We agree on what is *not* good research, but we have only low agreement on what good or excellent research is. This approach reflects (and makes explicit) the common assumption that *research* should be the most important criterion in hiring and promoting professors (Abele-Brehm & Bühner, 2016). Of course, indicators for other types of academic contributions, such as teaching, may be used as additional thresholds in Phase 1, depending on the priorities of the respective committee.

Not all research that is methodologically rigorous will also contribute something innovative and important, but these latter aspects of research quality are much harder or even impossible to capture via simple indicators. Therefore, the primary means of assessment in the second phase, applied to the shortlist, should shift towards an in-depth discussion of the research's actual content. This would pertain to how innovative, creative, and meaningful the research is, how the work relates to previous and related work in the field, which problems it solves, and why we should care about that (Dougherty et al., 2019). Short narrative merit statements, provided by applicants themselves, should serve as input to this discussion.

As applicants may hardly be outstanding in all areas alike, assessments in Phase 2 should not result in one-dimensional rankings, but rather in multi-dimensional profiles of activity across the five types of academic contributions (Figure 1) and the multiple dimensions within each type. These profiles may then be compared in terms of quality and their respective fit to the given institution and position. Diversity considerations of the institution might guide candidate selection: A productive and inspiring department presumably has a good balance and diversity in competencies, and new colleagues might be selected in a way that they fill gaps in the department's profile.

As a consequence of this increase in complexity, comparisons between applicants will become messier and more difficult – which opens up room for potential bias due to groupthink, confirmation bias, motivated reasoning, and other processes.[6] Safeguards are needed

[5] Empirical studies trying to investigate the quality-quantity trade-off are often invalid (e.g., by operationalizing quality by JIF or citation counts), are inconclusive in their results (finding both evidence for a negative, no, a positive, or a nonlinear association), and mostly conflate within- and between-person effects (e.g., Abramo et al. (2010); De Rassenfosse (2013); Haslam and Laham (2010); see, however, Michalska-Smith and Allesina (2017), for within-author comparisons. Forthmann et al. (2020), present a test of a theoretical model).

[6] A progressive solution that combats multiple biases is to

to address such biases in Phase 2. For example, committees need to be explicit about their criteria of the different areas of academic contributions and how they operationalize and weigh them – decisions that should ideally be defined *a priori*, before they see the applications. For transparency and fairness, these competence profiles should be communicated to applicants in the job description. Another useful countermeasure can be the Delphi method, in which committee members first submit private evaluations of the applicants (ideally anonymously) and then discuss each other's evaluations in the group.

Previous funding as a criterion

It is common in research evaluations to give great weight to the amount of acquired grant money. This may reflect the hope that the decisions made by funders can be used as sufficiently valid proxies of research quality. However, if funding decisions are based on the same invalid indicators, such as the JIF of previous publications, they also inherit all of the problems outlined above. Ultimately, these problems might even get amplified in funders' review boards where many dozen proposals are processed in a single session (Schmitt, 2022): Although the direct comparison of many proposals might lead to a more stringent and consistent application of selection criteria, the sheer volume and the limited time may increase the need to rely on superficial and invalid indicators. If funding decisions themselves are influenced by previous funding success, a strong Matthew effect (Merton, 1968) may result as more and more funds are accumulated by fewer researchers, independent of quality or merit (Bol et al., 2018). Finally, grant sums differ hugely between fields. In conclusion, we recommend not using the (quantitative) sum of previous funding as an indicator that is directly compared between applicants. Nevertheless, funded projects can be assessed in a more qualitative way, depending on the career stage: Do applicants have experience in acquiring funding, which can document their experience writing successful grant proposals? Were applicants able to develop grant proposals from their research topics? Were funded projects completed in reasonable time?

Who should do all the work?

Doing research assessment the way that is proposed here requires more work than simply summing up impact factors or counting publications. But, judged from our experience, hiring committees - at least in psychology - have rarely resorted completely to such crude quantitative shortcuts. Instead, for example, they distributed papers of shortlisted candidates to committee members who then read, summarized, and graded them as input to the committee. Hence, the goal should be to channel such existing effort into more valid assessment procedures.

Still, in order to keep the burden on hiring committees within reasonable bounds and thus have a realistic chance for the new approach to actually be applied in practice, the procedure should be streamlined and technically supported to the extent possible. We do think, for example, that it will be legitimate to ask applicants to provide most of the information pertaining to Phase 1 indicators (pre-registrations, open data etc.) themselves. All of the data should be collected in online surveys so that it can easily be aggregated and presented to the committee. This self-reported information then should be verified on random samples from the longlist, and for everyone on the shortlist. This task (and more generally collecting the necessary information for Phase 1) may be performed by trained student assistants (see Etzel et al., 2024). In the long term, this data (at the level of individual publications) could be kept in a central database to avoid unnecessary duplication of work on the side of committees as well as applicants.

Alternatively, some of these tasks could be further outsourced. For example, the University of Bremen already commissions an external consulting firm to assess competencies of professorship applicants beyond research and teaching, such as leadership capabilities and organizational management skills. Software solutions are currently developed and tested that automatically extract some of the proposed indicators, such as the presence of open data and open code (see, for example, ScreenIT from Weissgerber et al. (2021), the Rigor and Transparency Index from Menke et al. (2022), or the DataSeer project). Such external input, also from commercial service providers, can be useful as long as the provided information on applicants is transparent and reproducible (i.e., no blackbox scoring algorithms), and both the selection of measurement instruments and the hiring decision itself is done solely within the faculty.

But even with all that technical and external help, members and in particular chairs of hiring and promotion committees (a) need to be selected based on their expertise, (b) need enough time to do their job, and (c) need systematic training to increase their diagnostic competence (e.g., they should know about concepts

perform a focal random selection or random ranking on the shortlist (Osterloh & Frey, 2019, 2020), an approach already implemented in some funding schemes (Luebber et al., 2023). If any person on the shortlist is equally qualified for a job, a random selection for the final list and a random order of the candidates might be as good (or even better) than long discussions based on invalid indicators and subjective biases.

such as reliability and validity, as well as common judgment biases both at the level of individual perceivers and of groups).

Concluding remarks

With the signature of the DORA declaration and joining the CoARA coalition, many scientific organizations express their goal to change research assessment towards greater validity and embracing quality over quantity. But we also have to walk the talk. Based on the proposed principles, multiple implementations can be envisioned, in particular for the new rigor indicators. Along with this position paper, one concrete suggestion for an implementation in hiring committees is provided in Gärtner et al. (2025). However, we invite the community to develop additional or alternative implementations and to evaluate them in practice. Research assessment is at the heart of our academic culture, as it defines what we value, what gets published, what research projects get funded, and which researchers are able to stay in academia. Therefore, any research assessment tool has two complementary functions: It should be a valid diagnostic tool for selection or evaluation – but it also shapes the incentive structure that fosters certain types of research practices.

Author Contact

Felix D. Schönbrodt
Correspondence concerning this article should be addressed to Felix D. Schönbrodt, Email: felix.schoenbrodt@psy.lmu.de

Conflict of Interest and Funding

All authors declare that they have no conflicts of interest. This project and publication is supported by the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research - in cooperation with the BIH QUEST Center for Responsible Research. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, the Einstein Foundation or the award jury.

Author Contributions

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: Felix D. Schönbrodt: conceptualization, visualization, writing; Anne Gärtner: conceptualization, visualization, writing; Maximilian Frank: writing; Mario Gollwitzer: writing; Malika Ihle: writing; Dorothee Mischkowski: writing; Le Vy Phan: writing; Manfred Schmitt: writing; Anne M. Scheel: writing; Anna-Lena Schubert: writing; Ulf Steinberg: writing; Daniel Leising: conceptualization, visualization, writing

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

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? *Psychologische Rundschau*, *67*(4), 250–261. https://doi.org/10.1026/0033-3042/a000335

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics*, *101*(2), 1129–1144. https://doi.org/10.1007/s11192-014-1269-8

Abramo, G., D'Angelo, C. A., & Costa, F. D. (2010). Testing the trade-off between productivity and quality in research activities. *Journal of the American Society for Information Science and Technology*, *61*(1), 132–140. https://doi.org/10.1002/asi.21254

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, *115*(19), 4887–4890. https://doi.org/10.1073/pnas.1719557115

Bornmann, L. (2012). The Hawthorne effect in journal peer review. *Scientometrics*, *91*(3), 857–862. https://doi.org/10.1007/s11192-011-0547-y

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

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*(1), 119–135. https://doi.org/10.1017/S0140525X00065675

Copiello, S., & Bonifaci, P. (2018). A few remarks on ResearchGate score and academic reputation. *Scientometrics*, *114*(1), 301–306. https://doi.org/10.1007/s11192-017-2582-9

CWTS. (2021). *Halt the H-index*. https://doi.org/10.5281/ZENODO.4635649

De Rassenfosse, G. (2013). Do firms face a trade-off between the quantity and the quality of their inventions? *Research Policy*, *42*(5), 1072–1079. https://doi.org/10.1016/j.respol.2013.02.005

Dougherty, M. R., & Horne, Z. (2022). Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. *Royal Society Open Science*, *9*(8), 220334. https://doi.org/10.1098/rsos.220334

Dougherty, M. R., Slevc, L. R., & Grand, J. A. (2019). Making Research Evaluation More Transparent: Aligning Research Philosophy, Institutional Values, and Reporting. *Perspectives on Psychological Science*, *14*(3), 361–375. https://doi.org/10.1177/1745691618810693

Etzel, F. T., Seyffert-Müller, A., Schönbrodt, F. D., Kreuzer, L., Gärtner, A., Knischewski, P., & Leising, D. (2024). *Inter-Rater Reliability in Assessing the Methodological Quality of Research Papers in Psychology*. https://doi.org/10.31234/osf.io/4w7rb

European Commission, Directorate-General for Research and Innovation. (2022). *Agreement on Reforming Research Assessment*. https://eua.eu/downloads/news/2022_07_19_rra_agreement_final.pdf

Forthmann, B., Leveling, M., Dong, Y., & Dumas, D. (2020). Investigating the quantity–quality relationship in scientific creativity: An empirical examination of expected residual variance and the tilted funnel hypothesis. *Scientometrics*, *124*(3), 2497–2518. https://doi.org/10.1007/s11192-020-03571-w

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90. https://doi.org/10.1001/jama.295.1.90

Gärtner, A., Leising, D., Freyer, N., Musfeld, P., Lange, J., & Schönbrodt, F. D. (2025). Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology. *Meta-Psychology*, *9*. https://doi.org/10.15626/MP.2024.4604

Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE*, *17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Haslam, N., & Laham, S. M. (2010). Quality, quantity, and impact in academic publication. *European Journal of Social Psychology*, *40*(2), 216–220. https://doi.org/10.1002/ejsp.727

Hooper, M. (2022). *A Taxonomy of Research Quality v1.7.pdf* (pp. 563170 Bytes). figshare. https://doi.org/10.6084/M9.FIGSHARE.20113565.V2

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Hrynaszkiewicz, I., Novich, B., Harney, J., & Kiermer, V. (2024). *A survey of how biology researchers assess credibility when serving on grant and hiring committees*. https://doi.org/10.31222/osf.io/ht836

Kepes, S., Keener, S. K., McDaniel, M. A., & Hartman, N. S. (2022). Questionable research practices among researchers in the most research-productive management programs. *Journal of Organizational Behavior*, *43*(7), 1190–1208. https://doi.org/10.1002/job.2623

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. https://doi.org/10.1016/j.jml.2022.104332

League of European Research Universities. (2022). A Pathway towards Multidimensional Academic Careers - A LERU Framework for the Assessment of Researchers. https://www.leru.org/publications/a-pathway-towards-multidimensional-academic-careers-a-leru-framework-for-the-assessment-of-researchers

Leising, D., Gärtner, A., & Schönbrodt, F. D. (2025). Responsible Research Assessment (Parts I and II): Responses to the Commentaries. *Meta-Psychology*, *9*. https://doi.org/10.15626/MP.2024.4603

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022a). Ten steps toward a better personality science – how quality may be rewarded more in research evaluation. *Personality Science*, *3*, e6029. https://doi.org/10.5964/ps.6029

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022b). Ten steps toward a better personality science – a rejoinder to the comments. *Personality Science*, *3*, e7961. https://doi.org/10.5964/ps.7961

Luebber, F., Krach, S., Martinez Mateo, M., Paulus, F. M., Rademacher, L., Rahal, R.-M., & Specht, J. (2023). Rethink funding by putting the lottery first. *Nature Human Behaviour*, *7*(7), 1031–1033. https://doi.org/10.1038/s41562-023-01649-y

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife*, *8*, e47338. https://doi.org/10.7554/eLife.47338

Menke, J., Eckmann, P., Ozyurt, I. B., Roelandse, M., Anderson, N., Grethe, J., Gamst, A., & Bandrowski, A. (2022). Establishing Institutional Scores With the Rigor and Transparency Index: Large-scale Analysis of Scientific Reporting Quality. *Journal of Medical Internet Research*, *24*(6), e37324. https://doi.org/10.2196/37324

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Michalska-Smith, M. J., & Allesina, S. (2017). And, not or: Quality, quantity in scientific publishing. *PLOS ONE*, *12*(6), e0178074. https://doi.org/10.1371/journal.pone.0178074

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, *18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, *16*(3), e2004089. https://doi.org/10.1371/journal.pbio.2004089

Moore, S., Neylon, C., Paul Eve, M., Paul O'Donnell, D., & Pattinson, D. (2017). "Excellence R Us": University research and the fetishisation of excellence. *Palgrave Communications*, *3*(1), 16105. https://doi.org/10.1057/palcomms.2016.105

Müller, R., & De Rijcke, S. (2017). Thinking with indicators. Exploring the epistemic impacts of academic performance indicators in the life sciences. *Research Evaluation*, *26*(3), 157–168. https://doi.org/10.1093/reseval/rvx023

Niiniluoto, I. (2024). Scientific Progress. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2024). Metaphysics Research Lab, Stanford University.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

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

Osterloh, M., & Frey, B. S. (2019). Dealing With Randomness. *Management Revue*, *30*(4), 331–345. https://doi.org/10.5771/0935-9915-2019-4-331

Osterloh, M., & Frey, B. S. (2020). How to avoid borrowed plumes in academia. *Research Policy*, *49*(1), 103831. https://doi.org/10.1016/j.respol.2019.103831

Paulus, F. M., Cruz, N., & Krach, S. (2018). The Impact Factor Fallacy. *Frontiers in Psychology*, *9*, 1487. https://doi.org/10.3389/fpsyg.2018.01487

Ramani, R. S., Aguinis, H., & Coyle-Shapiro, J. A.-M. (2022). Defining, Measuring, and Rewarding Scholarly Impact: Mind the Level of Analysis. *Academy of Management Learning & Education*, *21*(3), 470–486. https://doi.org/10.5465/amle.2021.0177

Schmitt. (2022). Open peer commentaries to Leising et al., Ten steps toward a better personality science: How quality may be rewarded more in research evaluation. *Personality Science*, *3*, e9227. https://doi.org/10.5964/ps.9227

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

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Tiokhin, L., Yan, M., & Morgan, T. J. H. (2021). Competition for priority harms the reliability of science, but reforms can help. *Nature Human Behaviour*, *5*(7), 857–867. https://doi.org/10.1038/s41562-020-01040-1

Uygun Tunc, D., & Pritchard, D. (2022). *Collective epistemic vice in science: Lessons from the credibility crisis* [Preprint]. http://philsci-archive.pitt.edu/21120/

Vazire, S., & Holcombe, A. O. (2022). Where Are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, *26*(2), 212–223. https://doi.org/10.1177/10892680211033912

Weissgerber, T., Riedel, N., Kilicoglu, H., Labbé, C., Eckmann, P., Ter Riet, G., Byrne, J., Cabanac, G., Capes-Davis, A., Favier, B., Saladi, S., Grabitz, P., Bannach-Brown, A., Schulz, R., McCann, S., Bernard, R., & Bandrowski, A. (2021). Automated screening of COVID-19 preprints: Can we help authors to improve transparency and reproducibility? *Nature Medicine*, *27*(1), 6–7. https://doi.org/10.1038/s41591-020-01203-7

Wouters, P., Sugimoto, C. R., Larivière, V., McVeigh, M. E., Pulverer, B., De Rijcke, S., & Waltman, L. (2019). Rethinking impact factors: Better ways to judge a journal. *Nature*, *569*(7758), 621–623. https://doi.org/10.1038/d41586-019-01643-3