

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

*Felix D. Schönbrodt<sup>1</sup>, Anne Gärtner<sup>2</sup>, Maximilian Frank<sup>1</sup>, Mario Gollwitzer<sup>1</sup>, Malika Ihle<sup>1</sup>, Dorothee Mischkowski<sup>3</sup>, Le Vy Phan<sup>4</sup>, Manfred Schmitt<sup>5</sup>, Anne M. Scheel<sup>6</sup>, Anna-Lena Schubert<sup>7</sup>, Ulf Steinberg<sup>8</sup>, Daniel Leising<sup>2</sup>*

*Contributorship Statement: There were two tiers of authorship for the white paper. Tier 1 authors wrote and revised the document (FS, AG, AMS, DL, USt). Tier 2 authors provided feedback on drafts and supported the call for responsible research assessment (MG, MI, MF, ALS, MS, DM, LVP).*

<sup>1</sup>Ludwig-Maximilians-Universität München

<sup>2</sup>Technische Universität Dresden

<sup>3</sup>Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern

<sup>4</sup>Universität Bielefeld

<sup>5</sup>Universität Koblenz-Landau

<sup>6</sup>Utrecht University

<sup>7</sup>Universität Mainz

<sup>8</sup>Manres AG

*Abstract: 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-step 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 rigor in research evaluation.*

The German Psychological Society (DGPs) signed the San Francisco Declaration on Research Assessment (DORA) in 2021. This declaration calls 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 it 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 (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, replication success, reporting errors, or the presence of QRPs and HARKing (Brembs et al., 2013; Dougherty & Horne, 2022; Kepes et al., 2022). That means, 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., 2022) 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, as identified in the original DORA statement, has been recognized by multiple initiatives 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; DFG: [Package of Measures to Support a Shift in the Culture of Research Assessment](#)).

In March 2022, the DGPs committees “Open Science and Data Management” and “Incentive structures in academia, abuse of power and scientific misconduct” were tasked with developing a whitepaper that specifies these considerations 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, Malhotra, & Simonovits, 2014; Müller & de Rijcke, 2017; Smaldino & McElreath, 2016; Tiokhin, Yan, & Morgan, 2021).

To this end, we propose *four principles of responsible research assessment* applicable to the hiring and promotion of individual researchers, and a proposal for a *two-step assessment procedure* for hiring committees that combines the objectivity and efficiency of indicators with a qualitative and narrative assessment of a shortlist of candidates. In a separate document, a specific, actionable way of implementing these principles is proposed (Gärtner et al., 2022).

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 position paper presents such an adjustment against the background 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 safeguard minimal standards of scientific rigor that are necessary for sustainable progress in the discipline (cf. Uygun Tunç & Pritchard, 2022). To correct course, the main aspects of our proposal are to broaden the range of academic contributions that count and to place greater emphasis on rigor in research evaluation.

## **Four principles of responsible research assessment in psychological science**

*Principle 1: Academic contributions are multifaceted. Regarding research contributions, do not only value (a) journal articles, but also (b) data sets and (c) research software development.*

Currently, the number of peer-reviewed publications (co-)authored by a candidate and the amount of grant money acquired by a candidate (“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<sup>1</sup> that researchers play in creating them. In line with the “Recognition and Rewards” initiative by Dutch research organizations and a recent LERU (2022) position paper, we call for de-emphasising the importance of publication numbers and third-party funding as hiring and promotion criteria. We argue that the following *five* areas of academic contributions need to be considered: Research, teaching, academic leadership, service to the academic institution/field, and societal impact (see Fig. 1).

However, in the remainder of the present paper, we do focus on the *Research* dimension, because this is the area in which an urgent need for alternative evaluation criteria has been

---

<sup>1</sup> Contributor roles can be made explicit using CRediT (Contributor Roles Taxonomy; <https://credit.niso.org>), 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.

most clearly articulated for many years (Abele-Brehm & Bühner, 2016; European Commission, Directorate-General for Research and Innovation, 2022; Leising et al., 2022; LERU, 2022)<sup>2</sup>. We suggest that three kinds of research contributions should be considered by hiring and promotion committees: (a) journal articles, (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.

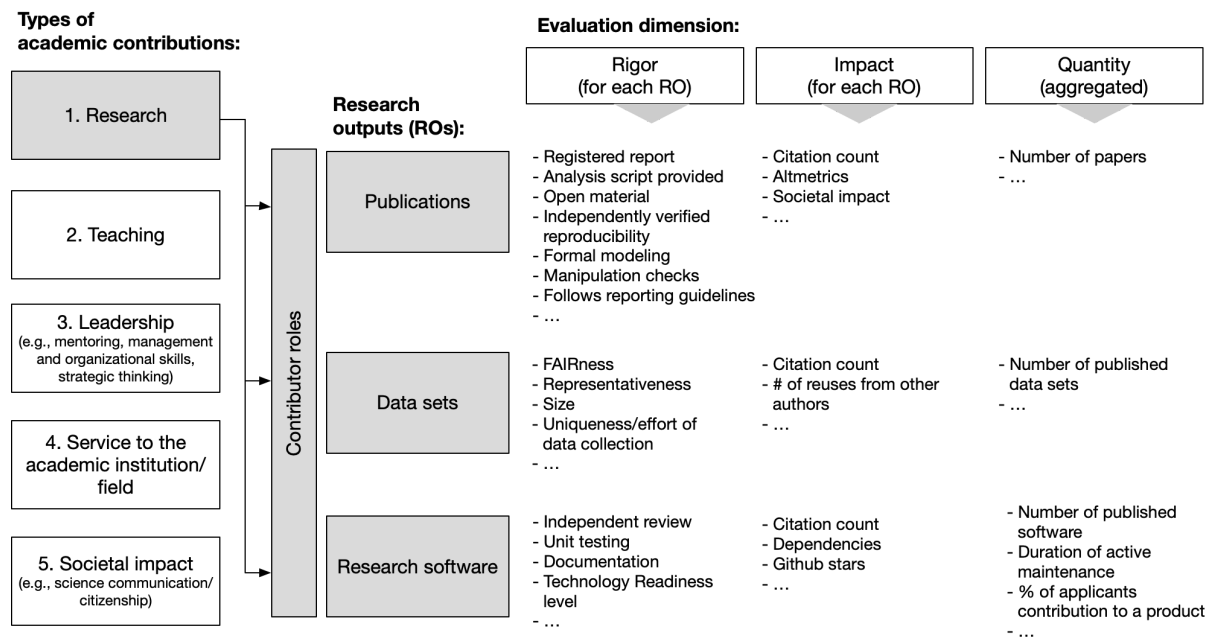


Fig. 1: Five types of academic contributions, three kinds of research outputs, three evaluation dimensions, and evaluation criteria for the latter.

*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 candidates 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). We predict that the pressure on hiring committees in Germany will not diminish in the foreseeable future, as many federal states are passing reforms to modernise the hiring processes for professorship positions resulting in a

<sup>2</sup> Several actionable suggestions for the other four assessment dimensions can be found in LERU (2022).

faster and more agile process (Bayerisches Staatsministerium für Wissenschaft und Kunst, 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<sup>3</sup>. 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 leaves, 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<sup>4</sup>. 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 is intransparent and thus not independently reproducible (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 special cases that do not fit proposed categories well. We therefore suggest that the use of objective indicators should be limited primarily to initial, first-step selections from a longlist of applicants in hiring processes, focusing on the basic skills and craftsmanship that every researcher needs to possess (“Two-step assessment”, see Fig. 2 and below). We further suggest that all applicants passing a certain threshold on these indicators should be considered in the next stage of the hiring process (instead of just selecting the “best” *n* candidates). This way, minor variations in scores will not unfairly disqualify applicants that are good enough. Even then, the (aggregated) indicators should not be used in a blind, strictly algorithmic way, but rather serve as a complement to human expert judgment. 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. We suggest asking candidates to submit short summary statements along with their applications, so that the hiring committee has a chance to read them before engaging in more algorithmic, indicator-based evaluation and selection.

---

<sup>3</sup> Inspired by “The Metric Tide” report, we use the label “indicator” instead of “metric” for the new indicators, reflecting that “data may lack specific relevance, even if they are useful overall.” (Wilsdon et al., 2015, p. 11).

<sup>4</sup> 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.

In contrast, evaluation of shortlist 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.

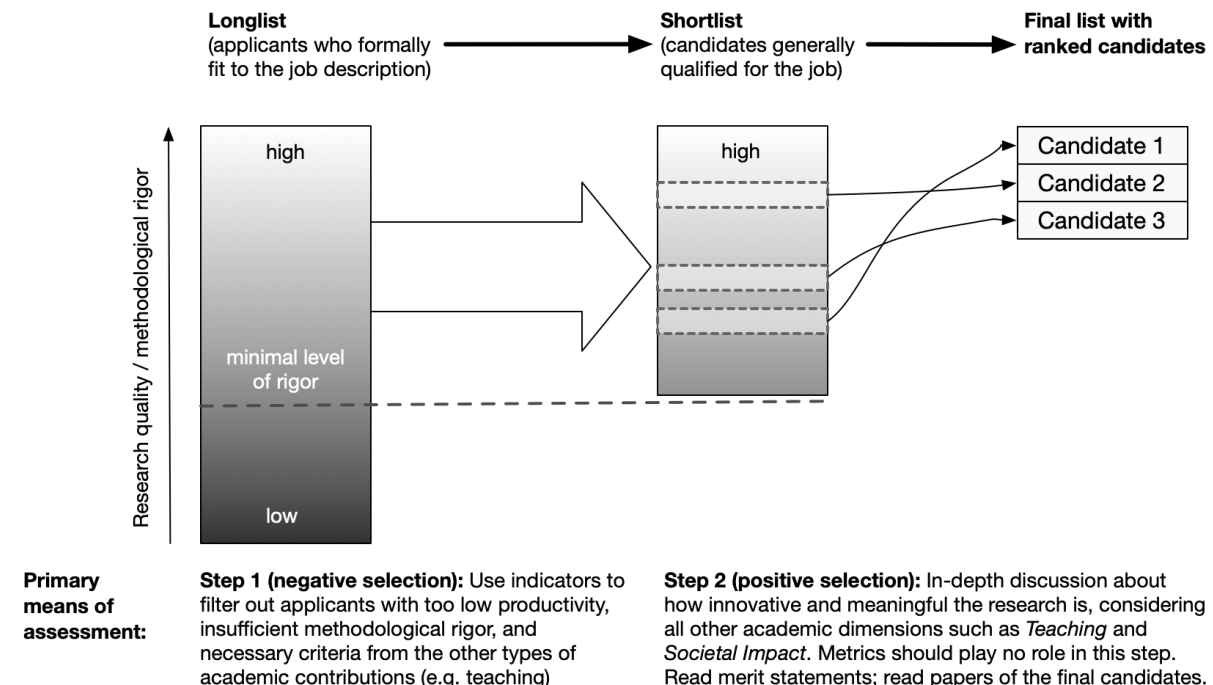


Figure 2: A two-step selection process.

*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 that the quality of research is often measured by its “impact”, which is operationalized in terms of the JIF or other citation metrics. 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 Fig. 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<sup>5</sup>, ranging from basic aspects such as “adhering to basic standards of good scientific practice” to more complex and sometimes

<sup>5</sup> [https://www.markhooper.io/taxonomy\\_draft.pdf](https://www.markhooper.io/taxonomy_draft.pdf)

elusive aspects such as “creativity, innovation, and ingenuity”. Even when researchers claim to “know good science when they see it”, it is hard to objectively operationalize such aspects. 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 has the potential 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.

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, 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; access to data is a logical precondition for independent reproducibility checks; and the presence of open code has been identified as the single largest predictor for successful reproductions of published results (Laurinavichyute et al., 2021). The more such features a research project espouses, the greater chances it will have to make any meaningful contribution to the scientific knowledge base. Unfortunately, these vital features of research have played a very minor role in research assessment so far. We argue that this needs to change, and that they have to be moved to the forefront instead (Leising et al., 2022).

Research outputs that do meet certain standards with respect to methodological rigor will then have to be further 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 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 espouse the necessary methodological rigor, its academic and/or societal impact may be determined. In our view, research that does make an impact is probably more valuable than research that makes none, all else being equal. On an indicator level, academic impact typically is measured via bibliographic metrics based on citations. Societal impact has, for example, been defined as “an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia” by the UK Research Excellence

Framework<sup>6</sup>. Impact on such a level is hard to operationalize by indicators and probably more validly assessed in narrative merit statements.

*(c) Quantity*

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.

Some general remarks: A *productivity* metric can be computed for individual candidates or institutions when the value of products is combined with their quantity (Abramo & D'Angelo, 2014). This operationalizes a quality-quantity-trade-off, as the same level of productivity can be achieved by either many low-quality outputs or few high-quality outputs. 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”, as these are mostly void of content or heterogeneously defined (usually as unclear mixture of quality, impact, and quantity).

*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 works lead 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, 2019, p. 6).

---

<sup>6</sup> <https://www.ukri.org/about-us/research-england/research-excellence/ref-impact/>



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, including open research practices, goes a long way in establishing these properties, and preliminary evidence suggests that assessing this merit is possible in a reliable fashion (for concrete suggestions for measurable quality indicators, see Fig. 1; Gärtner et al., 2022; Leising et al., 2022a, 2022b).

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 industriousness - 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., 2022). In fact, an often articulated impediment to implementing 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<sup>7</sup> This between-person variance is the main target of assessment procedures, and some level of quantitative productivity is certainly necessary for a researcher to be regarded as successful. 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 the field of psychology, and beyond (Nosek et al., 2022).

Especially for early career researchers (ECRs) it is thus 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) 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.

Quantity – as long as it does not come at the cost of quality – may have a limited role to play, however: For example, applicants for a permanent position in academia may have to demonstrate a quantitative minimum of outputs that surpasses the threshold of required

---

<sup>7</sup> 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 & Laham, 2010; see, however, Michalska-Smith & Allesina, 2017, for within-author comparisons. Forthmann et al., 2020, present a test of a theoretical model).

methodological rigor (e.g., a minimum number of published journal articles). Another procedure to turn the focus away from quantity is to restrict CVs to a maximum of, say, the 10 best research outputs selected by the applicants (as implemented, for example, by the German Research Foundation).

### **A two-step 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 steps. In Step 1, 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, 2022). The outcome should be used as a threshold – a minimal level of rigor – to guide the selection of candidates to be considered for the shortlist (negative selection, see Fig. 2). 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 in Step 1 as well, 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 step, 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 Step 2 should not result in one-dimensional rankings, but rather in multi-dimensional profiles of activity across the five types of academic contributions. These profiles may then be compared in terms of quality and their respective fit to the given institution and position.

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.<sup>8</sup> Safeguards are needed to address such biases in Step 2. For example, committees need to be explicit about their criteria of the different areas of academic contributions and how they operationalize and

---

<sup>8</sup> A progressive solution that combats multiple biases is to perform a focal random selection or random ranking on the shortlist (Osterloh and Frey, 2019, 2020), an approach already implemented in some funding schemes. If any person on the shortlist is equally qualified for a job, a random selection for the final list and the order of the candidates might be as good (or even better) than long discussions based on invalid indicators.

weigh them – decisions that should ideally be defined *a priori*. Another useful countermeasure can be the Delphi method, in which committee members first submit private evaluations of the candidates (ideally anonymously) and then discuss each other's evaluations in the group.

### **Previous funding as an indicator**

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 increases the need to rely on superficial and invalid indicators. If funding decisions themselves recur to previous funding success, a strong Matthew effect (Merton, 1968) is installed as more and more funds are accumulated by fewer researchers, independent of quality or merit (Bol, de Vaan, & van de Rijt, 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. Notwithstanding, funded projects can be assessed in a qualitative way, depending on the career stage: Do applicants have experience in acquiring funding, documenting 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 means more work than simply summing impact factors or counting publications. But, to our experience, hiring committees - at least in psychology - have rarely resorted completely to such crude quantitative shortcuts. Instead, for example, they distributed papers of shortlist candidates to committee members who then read, summarized, and graded them as input to the committee. Hence, the goal should be to channel such 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 Step 1 indicators (pre-registrations, open data etc.) themselves. All of the data should be collected in online surveys so that it will be easy to then aggregate and present the data 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 step 1) may be performed by trained student assistants. 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, or the Rigor and Transparency Index from Menke et al., 2022). 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 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, the DGPs continues its effort to foster good scientific practices in the scientific community. But we also have to walk the talk. We therefore call on all DGPs members to discuss the principles of responsible research assessment we propose here, with the ultimate goal of hopefully reaching a broad consensus within our academic society. Based on these 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 (Gärtner et al, 2022). However, we invite the community to develop additional or alternative implementations and to evaluate them in practice.

### **References**

- Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen?: Eine Untersuchung zur Bewertung von Auswahlkriterien in Berufungsverfahren der Psychologie. *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>
- Bayerisches Staatsministerium für Wissenschaft und Kunst (2022). *Factsheet zum Bayerischen Hochschulinnovationsgesetz*. [https://www.stmwk.bayern.de/download/21827\\_Factsheet\\_BayHIG.pdf](https://www.stmwk.bayern.de/download/21827_Factsheet_BayHIG.pdf)
- 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>
- 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>
- 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. *Zenodo*. <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>
- 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](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., & Schönbrodt, F. D. (2022). *Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology*. URL to be added.

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>

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>

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*, job.2623. <https://doi.org/10.1002/job.2623>

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2021). 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 [Preprint]. *PsyArXiv*.  
<https://doi.org/10.31234/osf.io/hf297>

League of European Research Universities [LERU]. (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., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. D. (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. D. (2022b). Ten steps toward a better personality science – a rejoinder to the comments. *Personality Science*, 3, e7961. <https://doi.org/10.5964/ps.7961>

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Meta-Research: 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, 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>

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. (2019). Scientific progress. In E. N. Zalta (Hrsg.), *The Stanford encyclopedia of philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2019/entries/scientific-progress/>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., 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), annurev-psych-020821-114157. <https://doi.org/10.1146/annurev-psych-020821-114157>

Osterloh, M., & Frey, B. S. (2019). Dealing with randomness. *Management Review*, 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*, *amle.2021.0177*. <https://doi.org/10.5465/amle.2021.0177>

Schmitt, M. (2022). Improving research quality: The roles of the timing and scope of changes in the incentive structure and the quality of committee work. *Personality Science*, *3*, e9227, 16-18. <https://doi.org/10.5964/ps.9227>

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*, 212-223. doi: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>

Wilsdon, J., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. DOI: 10.13140/RG.2.1.4929.1363

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*, 621–623. <https://doi.org/10.1038/d41586-019-01643-3>