*Meta-Psychology*, 2025, vol 9, MP.2024.4604
https://doi.org/10.15626/MP.2024.4604
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/MU6ED

# Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology

Anne Gärtner[1,2], Daniel Leising[1], Nele Freyer[1], Philipp Musfeld[3], Jens Lange[4], and Felix D. Schönbrodt[5]

[1]Technische Universität Dresden, Germany
[2]Freie Universität Berlin, Germany
[3]University of Zurich, Switzerland
[4]HMU Health and Medical University Erfurt, Germany
[5]Ludwig-Maximilians-Universität München, Germany

Traditional metric indicators of scientific productivity (e.g., journal impact factor; *h*-index) have been heavily criticized for being invalid and fueling a culture that focuses on the quantity, rather than the quality, of a person's scientific output. There is now a wide-spread demand for viable alternatives to current academic evaluation practices. In a previous report, we laid out four basic principles of a more responsible research assessment in academic hiring and promotion processes (Schönbrodt et al., 2025). The present paper offers a specific proposal for how these principles may be implemented in practice: We argue in favor of broadening the range of relevant research contributions and propose a set of concrete quality criteria (including a ready-to-use online tool) for research articles. These criteria are supposed to be used primarily in the first phase of the assessment process. Their function is to help establish a minimum threshold of methodological (i.e., theoretical and empirical) rigor that candidates need to pass in order to be further considered for hiring or promotion. In contrast, the second phase of the assessment process focuses more on the actual content of candidates' research and necessarily uses more narrative means of assessment. The debate over ways of replacing current invalid evaluation criteria with ones that relate more closely to scientific quality continues. Its course and outcome will depend on the willingness of researchers to get involved and help shape it.

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

## Introduction

The San Francisco Declaration on Research Assessment (DORA) and the Coalition for Advancing Research Assessment (CoARA) call on academic institutions to abandon the use of invalid metrics of research quality and productivity in hiring and promotion. One metric that should no longer be used to assess individual research achievement is the Journal Impact Factor (JIF) as it correlates *negatively* with multiple indicators of research quality (Brembs et al., 2013; Kepes et al., 2022; Paulus et al., 2018). Other indicators such as a person's number of (co-)authorships, number of citations, and *h*-index have become the target of criticism as well, given their rather loose connection to scientific quality and their susceptibility to being gamed (Chapman et al., 2019; Leising et al., 2022a).

Despite their well-documented negative side-effects, these metrics continue to be among the most important and frequently used measures in hiring and promotion decisions (Abele-Brehm and Bühner, 2016) – a practice that has been massively, and repeatedly, denounced for a number of good reasons (e.g. Leising et al., 2022a; DORA: https://sfdora.org/read; The PLoS Medicine Editors, 2006). The critics of the current approach insist that, instead of focusing on predominantly quantitative measures of research productivity, more attention should be paid to markers of actual research *quality* (DFG: Package of Measures to Support a Shift in the Culture of Research Assessment; European Commission 2021: Towards a reform of the research assessment system; (Leising et al., 2022b, 2022a). However, it is still unclear *how* exactly a "better" approach would look like.

Schönbrodt et al. (2025) proposed four principles for a more quality-based research assessment in psychological science: *(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 reg-*

*istered reports with or without an "in principle acceptance", protocols, monographs, book chapters, psychometric test manuals), (b) data sets and (c) research software development. (2) Quantitative indicators do have practical advantages, but they have to be valid and need to be used responsibly. (3) Use (a) methodological rigor, (b) impact, and (c) quantity as independent evaluative dimensions in research assessment. (4) Value quality over impact and quantity.*

In our view, a researcher's commitment to responsible research practice should play a crucial role in making hiring and promotion decisions (Gärtner et al., 2023, 2024). The main reason is that hiring and promoting such researchers is most likely to contribute to the emergence of a credible scientific knowledge base. As a side-effect, such an approach would also document the institution's commitment to transparent and responsible research conduct, which may help enhance the institution's scientific reputation.

The goal of this paper is to put forward a concrete proposal on how to implement these rather abstract principles in practice, specifically in the field of psychology. In this, we aim to minimize the additional burden on committees (both in terms of time and expertise). We also recognise that the appropriateness of each individual indicator in our proposal may be vigorously debated. It will always be possible to find examples of articles, projects, or researchers to which the proposed evaluation system would not do perfect justice. Nevertheless, we believe that our proposal—if implemented—would represent a substantial improvement over the status quo.

## Phase 1: Establishing a minimum threshold for methodological rigor

Schönbrodt et al. (2025) proposed assessing the academic performance and potential of applicants in two successive phases: In Phase 1, the overall methodological rigor of an applicant's research is assessed by assigning quality scores to individual research outputs to which the applicant has contributed. In order to keep the workload of committees manageable, the assessment in this phase is mainly algorithmic and uses indicators that are relatively objective and easy to obtain. However, unlike most indicators currently in use, these alternative indicators focus more directly on research *quality*.

From a theoretical perspective, research quality can be viewed as a multidimensional construct that encompasses both basic and more complex aspects. The basic aspects concern methodological rigor, while the more complex and elusive aspects concern concepts such as innovation, creativity, and ingenuity. Methodological rigor is a crucial component of research quality and can be measured relatively objectively, once a field has agreed upon common standards of good scientific practice. The concept refers to whether the research has been skillfully executed according to the standards of the field, and can be further subdivided into empirical and theoretical rigor. Rigor makes it more likely that research is reliable, reproducible, and of high scientific integrity.

In contrast, innovation, creativity, and ingenuity are more complex and difficult to measure. These aspects of research quality contribute to the actual advancement of knowledge, by offering new perspectives on important questions and intellectually sound ways of answering them. Importantly, these aspects of quality cannot be reduced to methodological rigor. In other words, methodological rigor is not a *sufficient* condition but only a *necessary* condition for high quality research. It is the *combination* of methodological rigor with innovation, creativity and ingenuity that drives high-quality research.

Phase 1 of the assessment process that we propose is supposed to help establish a minimum threshold of methodological rigor that candidates must pass in order to be further considered for hiring/promotion. In the case of hiring processes, this may mean being invited for an interview. Phase 2 of the assessment process then focuses more on the actual *content* of the candidates' research and necessarily uses more qualitative means of assessment.

As a rule, the assessments in Phase 1 should be provided by the candidates themselves and be independently verified for all candidates that proceed to Phase 2 (e.g., by committee members or by assistants). Again, this is to keep the workload for the committees within reasonable limits. For the same reason, we provide a ready-to-use online tool with assessment criteria for publication outputs, as well as an R package to automatically calculate multidimensional profiles for methodological rigor, author contributions, scientific impact and open science practices (see https://www.resque.info).

### Types of research output

According to Schönbrodt et al.'s (2025) first principle of responsible research assessment ("*academic contributions are multifaceted*"), there are other types of research contributions besides publications, namely published data sets and research software, that should be valued and taken into account in making hiring and promotion decisions. In what follows, we outline concrete evaluation criteria for how publications may be assessed in Phase 1. Criteria for research data sets and software

products are still under development at the time of writing.

Besides research, there are other types of academic contributions, such as teaching, leadership, service to the academic institution/field, and societal impact (cf. Schönbrodt et al., 2025), that should be considered in hiring and promotion decisions. A commission of the German Psychological Society is currently developing assessment criteria for the teaching domain, and we invite the scientific community to help develop measurable criteria for the other domains as well, to further enhance fairness, inclusiveness, transparency and standardization in the hiring and promotion of academics.

### *1) Publications / Research reports*

We suggest that candidates select what they consider to be the best (up to ten) of their own first-authored research reports that were published within the past five years. These should then be evaluated according to several quality criteria. The criteria are based on Leising et al. (2022b; 2022a) with some modifications (e.g., further specification of CRediT roles, FAIR format of data/scripts; engagement in theory-driven research; narrative statements on statistical power, sample size, scientific merit, and impact). The rationale behind this is to make visible the average methodological quality of the applicant's best, most recent work, while largely eliminating mere *quantity* as an assessment criterion.

The current set of assessment criteria for research reports can be found on the website of the RESQUE (Research Quality Evaluation) framework (https://www.resque.info). Note that they may still be subject to change in the course of ongoing discussions in the scientific community and based on the results of evaluation studies that are currently underway. The version that was current at the time this paper was written is provided in the appendix (https://osf.io/mu6ed/files/n2zdm). The criteria described in the following represent the *core set* of indicators that has emerged in the process so far. We expect them to be applicable to most areas of psychology. In addition, field-specific *expansion packs* that apply only to certain areas within psychology will be provided on the website.

First, aspects that should be specified for each nominated publication output are its DOI (title and year of publication are automatically retrieved), publication type (e.g., empirical vs. theoretical vs. computational), and publication format (e.g., published vs. preprint, paper vs. book vs. test manual). We recognize that the evaluation criteria are tailored to the assessment of original empirical research articles, which represent the vast majority of publications in most subfields of psychology. Nevertheless, other types of scientific publications (e.g., theoretical papers or simulation studies) do exist and are equally important. To prevent unfair disadvantages for candidates whose work mostly comprises these other types of publications, candidates may list these papers as well and specify their respective type of contribution. We suggest that candidates nominate no more than ten of their own papers, and rate these papers themselves, using the RESQUE Collector App. For papers that are not adequately covered by the given evaluation criteria, the required merit and impact statements (see below) may provide additional information as to why they are important and should be considered in the hiring/promotion process. For multi-study papers, it is important to specify which particular study is being rated, as studies may differ significantly with regard to research design and quality indicators.

Second, individual contributions to publications should be specified using a modified version of CRediT (Contributor Roles Taxonomy; https://credit.niso.org). Candidates are asked to make explicit the *type* and *degree* of their contribution to each submitted publication. Whereas the type refers to 14 standardized roles (such as “Conceptualisation”; “Formal analysis”, “Supervision” or “Writing – original draft”), the degree is specified as *lead*, *equal*, *support*, or *no role*. This information is not used in Phase 1 of the assessment procedure but provides a rich research profile to be discussed in Phase 2 (see below).

The next criteria assess the *empirical rigor* of the publication, which includes considerations about open data, open scripts, open materials, and preregistrations. The open data criteria involve specifying the kind of data being used, whether it is original data, reused existing data, or simulated data. They also involve the sample characteristics, such as type and representativeness of sample, or its diversity. It also covers the type of data (such as questionnaire responses vs. behavioral observations or physiological measurements), study design (cross-sectional vs. longitudinal; experimental vs. observational), and specifying if the raw data is available in an institutional repository, providing the URL or DOI, along with details on access level and compliance with FAIR principles (Findable, Accessible, Interoperable, and Reusable).

The open scripts criteria involve a URL or DOI under which the scripts are provided, adherence to FAIR principles (including proper licensing, comprehensive documentation, and version control), compliance with reproducibility standards, and whether the correctness of computational results has been independently verified. The latter means that someone who is not a co-author has verified computational reproducibility. This applies

only if a verification report is published with a DOI or another persistent identifier. We are aware that this has rarely been implemented so far. However, some journals already conduct independent checks, and some universities have begun offering pre-submission certification of computational reproducibility.

The open materials criteria involve a URL or DOI under which study materials are provided. This means all materials that would be necessary to replicate a study (e.g., questionnaire items, experimental stimuli, texts for vignettes, or code for software that controls the experimental flow).

It should also be indicated whether the work was preregistered or contains a registered report, specifying *what* was preregistered (e.g., sample size planning, hypotheses, analysis plan) and whether it constitutes at least one preregistered replication attempt of an existing study.

Another group of criteria was added in response to several commentaries on the first version of this article. These address the *theoretical rigor* of an applicant's work. Theory development and testing play a crucial role in all of science and should thus be incentivised. The criteria apply not only to purely theoretical work but also to original empirical work that incorporates theoretical reasoning in generating its hypotheses. The criteria assess whether a study was theoretically motivated at all and, if yes, whether it adheres to certain standards for reporting theoretical work. This includes a description of the phenomena that are to be explained, one or more hypothetical mechanisms accounting for these phenomena, definitions of all elements in the theory, and, if applicable, a formalization of the theory (i.e., a formal model). We propose specifically rewarding formalization, as it comes with several desirable properties that narrative theories rarely provide (e.g., full objectivity of the predictions derived from the theory) (Lange et al., 2025; most recent version: https://osf.io/preprints/psyarxiv/e7tdc_v2). Additional criteria refer to how a theory was empirically tested. If a theory was tested, the work should provide 1) a description of all additional assumptions that were necessary to make the theory testable (e.g., estimation and/or fixation of parameters in a model, operationalizations), 2) an explanation of how the tested predictions were derived from the theory (e.g., by a narrative walk-through or formally), 3) an evaluation of the (relative) fit of the theory/model to the empirical data, and 4) a discussion of what the results mean for the theory. The version of the criteria current at the time of writing is provided in the appendix. Yet, these indicators are still undergoing validation and might change in response to empirical results. The most up-to-date version can always be found on the homepage (https://www.resque.info/).

The RESQUE Collector App further allows specifying sample size and statistical power considerations for each nominated publication. Applicants are asked to provide free text responses summarizing all their respective considerations (e.g., assumed effect size, a priori power calculations, significance threshold, corrections for multiple testing; planned sample size). To simplify, these responses may be directly copied and pasted from the manuscript. This information will not be used in Phase 1 of the assessment procedure but may be relevant for the more qualitative assessment of the candidate's work in Phase 2 of the assessment process (see below).

The Collector App also provides applicants with the opportunity to provide a “merit statement” and a “scientific impact statement”, as perceived by themselves. The merit statement briefly summarizes (max. 150 words) the contribution of an article to the field and why the article should be considered in the hiring/promotion process. A merit statement might highlight, for example, the interdisciplinary nature of the work (establishing connections with other fields), the development of a new theory that was developed or tested for the first time, innovative methodological choices, or the unusually great effort needed to collect the sample. The “scientific impact statement” briefly summarizes (max. 150 words) the scientific impact that the given work has made *beyond* bibliographic indicators (e.g., citations, which are automatically retrieved for each submitted work) (see below). For example, a contribution may be considered scientifically impactful if it has changed the way in which most researchers in the field analyze a certain type of data. Again, this information will not be used in Phase 1 of the assessment procedure but may be relevant for the more qualitative assessment of a candidate's work in Phase 2 of the assessment process (see below).

***2) Data sets and research software contributions***

The practice of making one's research data publicly available (“open data”) is increasingly supported by journals, publishers and funding agencies. Similar to research articles, data sets can now be assigned a digital object identifier (DOI), making them easy to find and cite. This in turn may help researchers get credit for the data they collect. However, in academic hiring and promotion processes, this type of research output is not (yet) valued as much as the traditional journal articles. This constitutes a disincentive and needs to change (DFG: Package of Measures to Support a Shift in the Culture of Research Assessment). Therefore, we decided to include “data sets” as a distinct type of re-

search output that should be considered in hiring and promotion processes. Note, however, that this category is intended to be used only for exceptional published data sets whose size, scope, or reuse potential clearly exceeds the scientific value of data sets that are primarily used for a specific publication (in the RESQUE Collector App, the latter type of data set should be listed under the "Open Data" item for the respective publication). Specific indicators for exceptional published data sets are still under development and will include the DataCRediT roles (e.g., Collection, Validation, or Curation; see https://www.wsl.ch/datacredit/) and FAIRness indicators.

Research software is another essential part of modern data-driven science, powering both data collection (e.g., *PsychoPy*, Peirce et al., 2019, or *lab.js*, Henninger et al., 2022) and data analysis (see, for example, *R* and the many contributed packages; R Core Team, 2024). In some cases, the functioning of entire scientific disciplines depends on the work of a few (often unpaid) creators and maintainers of critical software (Muna et al., 2016). Moreover, non-commercial open source software is a necessary building block for computational transparency, reproducibility, and a thriving and inclusive scientific community. Therefore, it is high time that research software development is properly acknowledged in academic hiring and promotion procedures.

Some research software is accompanied by a citable paper describing it (e.g., for the *lavaan* structural equation modeling package in *R*: Rosseel, 2012). However, these "one-shot" descriptions of software often do not adequately reflect the ongoing work and evolving teams required to maintain and develop research software over time. Furthermore, not all valuable software contributions have such an accompanying paper. Thus, we need to find other ways of acknowledging and appreciating these valuable contributions in research assessment. We suggest evaluating research software as a third type of scholarly output (besides research reports and datasets). Note that this category refers only to reusable research software (such as libraries or broadly reusable scripts) that represents a substantial scholarly effort on the part of the applicant. It is not intended to be used for specific analysis scripts that are written and used primarily for a specific research project (this latter type of software should be listed under the "Open reproducible scripts" item for the respective research paper). Specific criteria for evaluating software contributions are still under development, but they will contain contributor roles inspired by the INRIA Evaluation Committee Criteria for Software Self-Assessment [1], whether the codebase and the community is actively maintained, and whether code tests (e.g., unit tests, functional tests) and multiple types of documentation are available.

**Multidimensional profiles vs. making decisions**

There is a broad consensus in the scientometric literature that single indicators capture only limited aspects of a researcher's profile. Consequently, a multidimensional perspective has been widely recommended (Hicks et al., 2015; DORA: https://sfdora.org/read; CoARA: https://coara.eu/agreement/the-agreement-full-text/). However, when a multidimensional profile is employed to inform decisions—such as whether to invite a candidate for an interview—or to rank candidates, the multiple dimensions are inevitably reduced to a single evaluative metric. This dimensionality reduction can be performed either explicitly, by specifying the weights of each dimension or the algorithmic process used, or implicitly. In the latter case, evaluations lack transparency and may introduce inconsistency or bias, as different implicit weights might be applied to different candidates. To address these challenges, we propose (a) maintaining a multidimensional perspective for as long as possible to enable a holistic and nuanced evaluation of a candidate's research profile and, at the same time, (b) ensuring transparency by explicitly defining how the multiple dimensions are combined into a single evaluative metric when diagnostic decisions are made.

**Making decisions: The relative weight of individual indicators**

We offer some suggestions as to how many points should be assigned to research outputs that meet the proposed quality criteria. The appendix contains the version of the criteria and their individual weightings current at the time of writing (https://osf.io/mu6ed/files/n2zdm). For the most up-to-date version, please refer to the homepage (https://www.resque.info). Although these suggestions reflect the outcome of extensive discussions within our group and with many other colleagues, they still largely reflect our own opinions, based on our intuitions about (a) the relative *importance* of each criterion in promoting research quality, and (b) the *effort* typically involved in meeting a criterion. For the future, however, it will be desirable to determine these values in a more systematic and representative manner (e.g. by surveying a larger group of researchers).

Of course, each committee may easily adapt our suggestions to better meet their specific needs, depending

[1] https://www.inria.fr/sites/default/files/2019-10/Criteria_software_self_assessment.pdf

on the hiring or promotion process at hand. The numerical weights assigned to each indicator or category are a reflection of how much the committee values the individual criteria.

When it comes to making decisions in an assessment procedure, we advocate for a transparent and reproducible way of combining multiple dimensions into a decision criterion. To this end, we suggest the computation of an overall *relative rigor score* (RRS). We use a POMP scoring procedure (percent of maximum points), to ensure that the RRS for each research output can reach 100% even if some indicators are not applicable. Furthermore, we recommend weighting all submitted research outputs equally. For example, if two publications are available, each one contributes 50% to the overall score, even if they differ in their maximally attainable points. The reasoning behind this approach is that it should not be punished if some points cannot be obtained in principle.

Should such assessments of methodological rigor serve to establish a lower threshold, or should they rather be used to rank candidates? We assume that (a) at present, there is still considerable variation among applicants in how seriously they take methodological rigor in their own scientific work, but also that (b) almost 15 years after the beginning of the so-called replicability crisis, there should be a sufficient number of applicants who have actually improved the rigor of their own research. Based on these assumptions, we recommend that committees use the RESQUE framework to establish a lower threshold of methodological rigor that is required to enter the second phase of the assessment procedure. In an evaluation study, where researchers from personality and social psychology submitted their three best papers ($k = 52$ scorable papers; Etzel et al., 2025), an average RRS of 31% was found. However, other subfields of psychology have not yet adopted rigor enhancing practices in the same way and have probably lower RRSs on average.

When comparing (e.g., tenure-track) promotion with hiring processes, an important difference is that in the former cases there is no selection of a few short-listed candidates from many more who applied. Instead, the research output of a given researcher over a given period of time (e.g., the last six years) must be evaluated. Again, the quality criteria proposed in this paper may be used to support committees in making such decisions. In the future, they may even be used to define certain field-specific quality standards for a candidates' research in advance (e.g., at least one key hypothesis test must be pre-registered, all data must be made openly available, at least one replication attempt). This would be an effective way to raise methodological standards while at the same time making the promotion process much more transparent and predictable for candidates.

The correctness of applicants' self-reports can be verified by the respective committee at any time. To encourage truthful reporting, we recommend informing candidates that such checks will take place and that providing incorrect information will lead to exclusion from further consideration for the position or promotion. To keep this feasible, we recommend checking only small samples of the information provided by candidates in the first phase (longlist) of the process. However, for candidates who proceed to the second phase (shortlist), *all* of their self-assessments should be checked. The (limited) experience we have with such processes does suggest that this is feasible. However, institutions might also consider *paying* someone who is sufficiently qualified to carry out such checks in a professional and independent manner.

### Phase 2: In-depth analysis and discussion of research content

Phase 2 of the evaluation process focuses more on the actual content and merit of a candidate's research. Here, criteria such as scientific impact, relevance, innovation and creativity come into play. This requires in-depth discussions of an applicant's work with them, among committee members, and with external reviewers.

The ratings on different indicators collected in Phase 1 of the assessment process may be used to create a multidimensional profile for each candidate. A sample profile can be seen in Figure 1. It reflects a person's performance in various domains of methodological rigor (open data, open materials, preregistration, reproducible code and verification, theorizing and formal modeling) across all of their nominated papers. Note that the profile is calculated only for those criteria that could be met in principle (without the "not applicable" option). For example, "open data" is only scored if sharing was possible and not restricted by data security concerns.

Several additional indicators are collected in Phase 1 of the evaluation process but are only supposed to be considered in more detail in Phase 2. These include: merit statement, scientific impact statement, CRediT roles, sample size and power considerations, type of data, type of sample, interdisciplinarity, and internationality (Leising et al., 2025).

As laid out in Schönbrodt et al. (2025), the further assessment of shortlist candidates should focus on the actual *content* of their research. This is likely to include a discourse over how innovative, creative, and meaningful the research is, whether and how it contributes

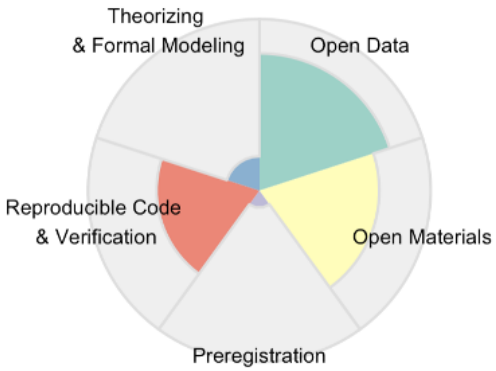


**Figure 1**

*Methodological rigor profile. Exemplary output from the RESQUE tool, showing a summary of the methodological (i.e., empirical and theoretical) quality of an applicant's work in various domains.*

something valuable to the knowledge base in the respective field, and how it fits with the hiring institution's strategic vision and resources. To enable this discourse, it will be necessary to actually read research papers that the candidates nominated in Phase 1. For this task, applicants can nominate their three best papers (among the, say, 10 best papers they submitted in the RESQUE tool). The content of the merit statements that the candidates provide for their nominated publications may become a basis for in-person discussions between candidates and the committee. Committees may send candidates a few key questions regarding their research in advance to enable them to properly prepare for these discussions. Adopting this practice might not only help mitigate unwanted influences of (e.g.) candidates' personal charisma or ability to "dodge" certain questions on the committee's judgment, it would also demonstrate that at least some committee members *have actually read* the respective papers and thus foster the integrity of the process overall.

As another way of emphazising the actual *content* of candidates' research, we explicitly advise that Journal Impact Factors and applicants' *h*-indices must not be collected or utilized *at any stage* of the assessment process.

As a general rule, we recommend considering the scientific *impact of an applicant's work* only after a certain level of methodological rigor has been established, i.e. only for candidates that have passed the respective threshold in Phase 1 of the assessment procedure. Scientific impact is typically measured in terms of citation numbers, but these should always be adjusted for the age of the publication and the field in which a candidate works, to prevent unfair biases. A detailed discussion of what scientific impact is and how it can be measured can be found in the paper by Leising et al. (2025).

Within the RESQUE framework, we suggest measuring scientific impact in a two-fold manner: (1) The RESQUE tool provides age- and field-corrected BIP! scores for each nominated paper. BIP Scholar (https://bip.imsi.athenarc.gr/site/home) is a non-commercial open-source service intended to facilitate fair researcher assessment; it provides impact scores (and five impact classes) for each publication. (2) Candidates are also asked to provide "scientific impact statements", explaining how (in terms of content, not citation numbers) their research has had, or could have, a relevant impact on the respective field.

CRediT roles provide a clear and detailed account of a candidate's specific contributions to a research output. By examining credit roles, the committee may better understand the candidate's skill set and determine if it matches the requirements of the position. For example, roles such as "Funding Acquisition" or "Project Administration" indicate a candidate's ability to initiate and manage projects, while roles such as "Writing: Original draft" and "Visualisation" indicate the writing skills of a candidate and their ability to interpret data and highlight key findings through visual representation.

The RESQUE tool also allows users to indicate the *degree* of a candidate's contribution: lead, equal, support, no role. A sample illustration of a candidate's contributions across ten nominated publications is shown in Figure 2.

In fields where multi-author papers are common, credit roles may help distinguish the individual contributions of several authors to the same paper more clearly (e.g., one candidate may have developed the software and have performed most of the analyses, whereas another candidate may have conceptualized the research project and acquired the funding). These specifics would remain hidden when only the *order* in which authors are listed were inspected.

While Phase 2 emphasizes a detailed assessment of research content and the criteria discussed here, it is important to note that additional aspects should also be evaluated at this stage of the selection process. These include aspects such as the candidate's fit for the position, contributions to teaching, leadership skills, service to the field/committee work, and societal impact (cf. Schönbrodt et al., 2025). Exploring these additional dimensions in detail goes beyond the scope of this paper. Nevertheless, we encourage our colleagues to actively discuss and develop such criteria, striving for consensus and further standardization in the selection process to ensure these elements are systematically integrated.

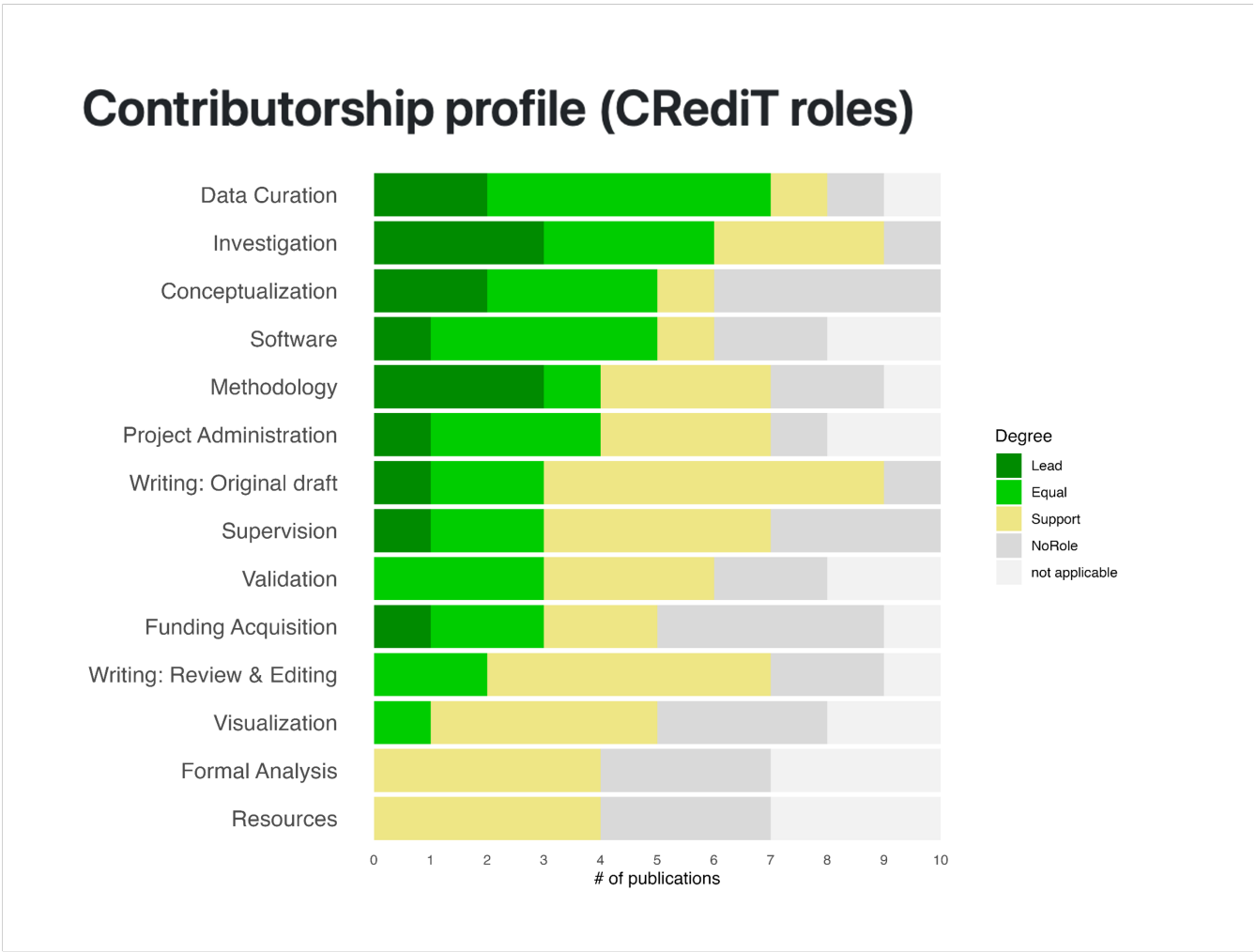


**Figure 2**

*Sample bar chart summarizing a candidate's contributions across ten nominated publications. The chart accounts for both the type of contribution and the respective degree of involvement/strength of contribution.*

### "Red flags" in evaluating research performance

To avoid rewarding candidates for engaging in "bad scientific practice", additional indicators ("red flags") for fraud, QRPs, and sloppiness may be checked for all research papers nominated by applicants on the shortlist. This is not meant to imply that most researchers engage in QRPs willingly and with full awareness. *P*-hacking, for example, is not necessarily a deliberate attempt at gaming the system. It may just as well be a product of human fallibility (e.g., conveniently forgetting about one's previous attempts to test the same hypothesis by alternative means). Nevertheless, *p*-hacking often does have important (and detrimental) consequences for the quality of research (see Stefan and Schönbrodt, 2022) and should therefore be screened for during the hiring process. Ultimately, it is the responsibility of each individual scientist to guard against the potential influence of *p*-hacking on his or her own research, and scientists who take this responsibility seriously should be rewarded for doing so.

The following tools might be considered in this regard: Statcheck (http://statcheck.io), the GRIM test (Brown and Heathers, 2017), and the SPRITE test (Heathers et al., 2018) are tools that can be used to check research papers for errors in statistical reporting. We suggest using them to check the papers nominated in Phase 1 of the process for consistency between reported and recomputed *p*-values. Other aspects that may be considered are a) whether a highly cited study has ever been subjected to a direct, independent replication attempt, b) sweeping claims, counterintuitive and

shocking results, and c) results that seem to be too good to be true, "excess significance" (i.e., the observed number of statistically significant results is too large compared to their expected number).

Needless to say, none of these techniques is perfect and thus none should result in the "automatic" exclusion of candidates. However, when several red flags converge, caution is warranted. If concerns arise, the committee may send follow-up questions and ask applicants to respond in writing.

### Implementation in practice

The use of explicit and measurable criteria for responsible research assessment in hiring and promotion decisions is likely to have a number of desirable effects: 1) It emphasizes the critical importance of high quality (as opposed to mere quantity) of scientific research. 2) It makes the evaluation standards used by a committee (at least in Phase 1 of the process) transparent to everyone involved, thus enhancing accountability and fairness. 3) It helps establish a basis for relatively objective assessments of research quality, which should improve inter-rater reliability, which in turn is a prerequisite for validity. Indeed, in a pilot study, we found very good inter-rater agreement (ICC(1,1) = .81) for assessments of empirical rigor using an earlier version of this rating scheme (Leising et al., 2022b). In a recent study, the estimate was even higher (ICC(1,1) = .91) for the empirical rigor criteria (Etzel et al., 2025). Notably, both assessments were provided by research assistants with very limited research experience of their own. This suggests that independent verification of candidates' self-assessments should be relatively straightforward, especially when the number of research papers nominated by each candidate is relatively small (e.g., 10). Note, however, that indicators for theoretical rigor have not yet been validated and that results might look differently. Still, we believe that it is possible to implement an assessment scheme such as the one proposed here with little to no additional effort for committees. The threshold for implementing a process like the one described here should be further lowered by the fact that we provide a ready-to-use online tool for automated creation of multidimensional profiles.

We would like to reiterate, however, that the details of this proposal are not set in stone. Rather, we see them as a starting point of a discussion over what criteria, if not impact factor and *h*-indices, should be used in hiring and promoting academic scholars. Each committee may develop and use its own version of such an evaluation scheme, including modified or even entirely new criteria and different rules for weighing criteria. We welcome any feedback on such attempts to implement responsible research assessment.

Regarding the chance of such alternative assessment schemes ever becoming mainstream, it should be noted that evaluation criteria for academic work are already changing significantly. Good scientific practice, sustainable data management, and overall transparent and reproducible science are increasingly imposed as standards by national and international funding agencies (e.g., DFG, European Horizon, NIH, SNF, to name a few). Therefore, we see our present proposal as a contribution to a movement that is already well underway.

Again, it should be noted that the present proposal focuses on research, which is only one of many valuable contributions that (aspiring) members of the academic community can make (Schönbrodt et al., 2025). We have chosen to focus on research because (a) it is clearly of central importance and (b) many quality criteria for this area of academic activity are relatively well defined and can be assessed with good objectivity (Etzel et al., 2025). However, much work remains to be done to develop similarly objective and valid quality criteria for the other domains: teaching, academic leadership, service to the field/institution, and societal impact. We thus call on our colleagues to begin a public discourse (possibly in this journal) on how to properly assess the value of these other types of contributions.

### Concluding remarks

In a previous report, we outlined four basic principles of responsible research (Schönbrodt et al., 2025). In the current report, we present a proposal for putting these principles into practice, focusing on research as a key area of academic activity. Specifically, we offer concrete evaluation criteria, including a ready-to-use online tool, for research publications (soon to be followed by criteria for data sets and research software). We hope that by adopting these (or similar) evaluation schemes, committees will be able to give greater weight to quality (and lesser weight to quantity) in their hiring or promotion decisions. This is crucial because the credibility of academic institutions, and of science more broadly, depends on the quality of their scientific output more than on anything else. However, the present proposal should still be seen as a starting point rather than the conclusion of an important debate. We welcome criticism, alternative proposals, and suggestions for improvement from our colleagues in the field. We would be particularly grateful for any feedback from colleagues who have actually tried to implement (some of) our suggestions in their own committee work.

### Conflict of Interest

All authors declare that they have no conflicts of interest.

### Authors Note

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.

### 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.

### Author Contributions

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: Anne Gärtner: conceptualization, visualization, writing; Daniel Leising: conceptualization, visualization, writing; Nele Freyer: writing; Philipp Musfeld: writing; Jens Lange: writing; Felix D. Schönbrodt: conceptualization, visualization, writing.

Correspondence concerning this article should be addressed to Anne Gärtner, Email: anne_gaertner@tu-dresden.de

Anne Gärtner

### 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

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*, 291. https://doi.org/10.3389/fnhum.2013.00291

Brown, N. J. L., & Heathers, J. A. J. (2017). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Chapman, C. A., Bicca-Marques, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., Guo, S., Hemingway, C. A., Leendertz, F., Li, B., Matsuda, I., Hou, R., Serio-Silva, J. C., & Stenseth, N. C. (2019). Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1916), 20192047. https://doi.org/10.1098/rspb.2019.2047

Etzel, F. T., Seyffert-Müller, A., Schönbrodt, F. D., Kreuzer, L., Gärtner, A., Knischewski, P., & Leising, D. (2025). *Inter-rater reliability in assessing the methodological quality of research papers in psychology* [PsyArXiv Preprint]. https://doi.org/10.31234/osf.io/4w7rb_v2

Gärtner, A., Leising, D., & Schönbrodt, F. D. (2023). Empfehlungen zur berücksichtigung von wissenschaftlicher leistung bei berufungsverfahren in der psychologie. *Psychologische Rundschau*, *74*(3), 166–174. https://doi.org/10.1026/0033-3042/a000630

Gärtner, A., Leising, D., & Schönbrodt, F. D. (2024). Towards responsible research assessment: How to reward research quality. *PLoS Biology*, *22*(2), e3002553. https://doi.org/10.1371/journal.pbio.3002553

Heathers, J. A., Anaya, J., Van Der Zee, T., & Brown, N. J. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (sprite). https://doi.org/10.7287/peerj.preprints.26968v1

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The leiden manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

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

Lange, J., Freyer, N., Musfeld, P., Schönbrodt, F., & Leising, D. (2025). A checklist for incentivizing and facilitating good theory building. *Zeitschrift für*

*Psychologie*, *233*(4), 279–283. https://doi.org/10.1027/2151-2604/a000604

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

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

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

Muna, D., Alexander, M., Allen, A., Ashley, R., Asmus, D., Azzollini, R., Bannister, M., Beaton, R., Benson, A., Berriman, G. B., Bilicki, M., Boyce, P., Bridge, J., Cami, J., Cangi, E., Chen, X., Christiny, N., Clark, C., Collins, M., & Zonca, A. (2016). The astropy problem. https://doi.org/10.48550/arXiv.1610.03159

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

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

R Core Team. (2024). *R: A language and environment for statistical computing* [[Computer software]]. R Foundation for Statistical Computing. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Schönbrodt, F. D., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., Schubert, A.-L., Steinberg, U., & Leising, D. (2025). Responsible research assessment I: Implementing DORA and CoARA for hiring and promotion in psychology. *MetaPsychology*, *9*. https://doi.org/10.15626/MP.2024.4601

Stefan, A. M., & Schönbrodt, F. D. (2022). Big little lies: A compendium and simulation of p-hacking strategies. https://doi.org/10.31234/osf.io/xy2dk

The PLoS Medicine Editors. (2006). The impact factor game. *PLoS Medicine*, *3*(6), e291. https://doi.org/10.1371/journal.pmed.0030291