Meta-Psychology, 2025, vol 9, MP.2024.4603
https://doi.org/10.15626/MP.2024.4603
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/Y5T7W

Responsible Research Assessment (Parts I and II): Responses to the Commentaries

Daniel Leising[1], Anne Gärtner[1,2], and Felix D. Schönbrodt[3]
[1]Technische Universität Dresden
[2]Freie Universität Berlin
[3]Ludwig-Maximilians-Universität München

We give a brief overview of our deliberations in responding to the commentaries on our two target papers presenting the RESQUE (Research Quality Evaluation) framework. While we were able to incorporate many suggestions for improvement directly, we acknowledge that other areas (e.g., quality in theorizing) will require further elaboration. In this paper, we specifically touch on the following issues: (a) eligible types of publication, (b) measurability, (c) quality criteria for software and datasets, (d) theoretical rigor, (e) quantity, (f) authorship, (g) potential bias (against certain methodologies, types of research contributions, or subdisciplines), (h) overall rigor score, (i) weighting of individual indicators, (j) types of data and samples, (k) impact, (l) interdisciplinary value, (m) teaching, (n) expertise, (o) gaming the new metrics, and (p) representativeness. The RESQUE framework has met with largely positive reception so far, but continues to evolve and will thrive best when community involvement stays strong.

Keywords: Metascience, Incentive, Evaluation

Preface

Research assessment is at the heart of our academic culture, as it defines what we value, what gets published, and which research projects get funded. As a consequence, it also plays a crucial role in determining who can stay and thrive in academia. Given that science will continue to be a highly collaborative endeavor, any reform of research assessment must rely on the wisdom of the crowd, consider diverse perspectives from many subfields, and seek a broad consensus in the community.

In light of a now widespread acknowledgment that research assessment needs to be reformed (Chapman et al., 2019) (https://sfdora.org/read,https://coara.eu/agreement/the-agreement-full-text,https://data.europa.eu/doi/10.2777/707440), a working group of the German Psychological Society (DGPs) proposed a framework for the evaluation of research quality in the context of hiring decisions in academic psychology: RESQUE, which stands for "RESearch QUality Evaluation". The working group now maintains a central landing page (https://www.resque.info/) to provide updates and versioning of the rating schemes, interactive web tools for data collection and for aggregating, enriching and visualizing the relevant information, as well as forthcoming expansion packs.

The first version of RESQUE was described in two target papers (Gärtner et al., 2022; Schönbrodt et al., 2022). Fifteen commentaries on these by various members of the community have been published in Meta-Psychology. Six additional commentaries in German were published in response to a project report (Gärtner et al., 2023) that we had written for Psychologische Rundschau (the main outlet of the DGPs). Here, we aim to provide some insight into the considerations that guided our responses to all of these commentaries. Sometimes we respectfully disagree with the commenters' viewpoints and explain our reasons for doing so. In many cases, however, we wholeheartedly embraced the suggestions and revised the target papers and the rating schemes accordingly.

Topics Considered in the Course of the Revision

(a) Eligible Types of Publication

In the first versions of our two target articles, we considered only published research articles, but multiple commenters convincingly argued that other types of text outputs should be considered as well (Brown, 2024; Fink-Lamotte et al., 2024; Karhulahti, 2024; Sparfeldt et al., 2024; Syed, 2024; Witte, 2024). We agree, and now suggest giving full consideration to preprints, conference proceedings, Stage 1 registered reports with or without "in principle acceptance", study protocols, monographs, book chapters, and test manuals – basically any text document that has a doi (or other persis-

tent identifier). All of these publication types have the same potential for scientific rigor and impact. Note that many or even most RESQUE indicators can be applied to meta-analyses as well (Fink-Lamotte et al., 2024). Given the intransparent nature and unknown quality of most pre-publication peer review (Bornmann, 2012; Cicchetti, 1991), we think it is reasonable to perform evaluations of methodological rigor regardless of whether a text has been accepted for publication or not. As a side effect, this practice also makes the assessment independent of (often long) publication delays. This will be particularly helpful for early career researchers on fixed-term contracts.

(b) Measurability

Some commenters (e.g., Niessen et al., 2023; Stengelin et al., 2024) offered lists of additional quality criteria for academic work that may also be considered in hiring and promotion procedures (e.g., building infrastructure; appropriate design and analysis choices). We agree that many of these things are important (Leising et al., 2022b), and we do not claim that the current set of indicators exhausts the full spectrum of what constitutes "good research". However, a system like RESQUE requires indicators that are not only valid but also objectively measurable with relative ease. This is particularly important in the first, more algorithmic phase of the assessment process. For practical reasons, we did not include additional criteria whose measurability was yet unclear to us. But given that RESQUE is conceived of as a living and versioned system, we remain open to including additional indicators later.

(c) Quality Criteria for Software and Datasets

It turned out that developing ready-to-use quality criteria for software and data will take significantly more time and effort than we had anticipated. In this regard, our first proposal was simply too optimistic. In their commentary, Brandmaier et al. (2024) suggested a whole set of highly appropriate indicators of quality in software development and maintenance. We have thus invited these authors to help us develop a working indicator set for this type of research output. This work is now well underway. A similar approach will be needed for assessing the quality of published datasets. We invite those of our colleagues who feel they have something to contribute in this regard to contact us and do just that.

(d) Theoretical Rigor

Several commenters (Dames et al., 2024; Niessen et al., 2023; Ulpts, 2024; Witte, 2024) criticized our proposal for giving issues related to "theory" too little weight. We wholeheartedly agree. Although good scientific work may be entirely descriptive and/or exploratory in nature, it goes without saying that much of what scientists do revolves around the goal of developing and refining abstract ideas about how the world works (i.e., building and testing theories). This type of work may also vary in quality, and those who are able and willing to produce the best theoretical work should be rewarded for doing so. Since we were unable to find an established set of criteria for evaluating the quality of people's theoretical work, we have now begun developing such a set ourselves, in cooperation with several colleagues. Basic features of the present version of that criterion set are briefly described in the revised version of the second target paper (Gärtner et al., 2025; Lange et al., 2025).

(e) Quantity

There seems to be a widely shared sentiment among scientists that the quality of one's work should be clearly prioritized over its mere quantity (e.g., Chapman et al., 2019). Unfortunately, the current incentive structure in academia largely rewards the opposite (Abele-Brehm & Bühner, 2016; Leising et al., 2022b). The RESQUE framework attempts to address this problem by asking authors to self-nominate a relatively small sample of what they consider to be their own best work. The average methodological quality of this small sample of research outputs is then determined using a detailed, algorithmic scoring procedure, and those applicants with the best work proceed to the second assessment phase.

The desired effects of this approach are: (a) a minimum level of productivity (in terms of sheer output quantity) is ensured, (b) applicants can not increase their chances by increasing the sheer number of their research outputs, but they will increase their chances directly by investing in the methodological quality of their work. Note, however, that we are not arguing against using quantity as a metric in general, but only against using quantity as a metric without establishing quality first (Ortner et al., 2013). For example, of two applicants with comparable quality scores in phase 1, a committee may still favor the one with the higher quantitative output in phase 2.

(f) Authorship

After careful deliberation, we decided to treat the issues of authorship and author order as follows: First, authorship represents a claim that one has contributed substantially to the process that resulted in a research article. However, we are fully aware that guest and honorary authorships constitute one of the most common

forms of scientific misconduct (Fong & Wilhite, 2017; Pruschak & Hopp, 2022). This may be explained by the fact that the mere length of an applicant's publication list continues to be a predictor of who will be hired (Abele-Brehm & Bühner, 2016; Chapman et al., 2019), and by the fact that there is currently no reliable mechanism for detecting this type of misconduct. Indeed, we are not aware of a single case in which a researcher has ever been sanctioned for this. Note that this is despite the fact that the practice itself has been clearly identified as a form of scientific misconduct for decades. Given this lack of effective oversight, we have no choice for now but to continue taking authorship claims at face value.

Second, we recommend using a version of the now-established CRediT system (Contributor Roles Taxonomy; https://credit.niso.org), which asks applicants to explicitly state the type and degree of their contribution to each submitted research output. The type is one of 14 standardized roles (such as "Conceptualization", "Data Curation", or "Writing – original draft"), and the degree is one of lead, equal, support or no role. As the CRediT standard does not yet provide precise definitions of these degrees, we developed the following working definitions for use with RESQUE:

- Lead: You were the single leader for this specific activity. If you are the only person in a specific role, choose 'lead'.
- Equal: You contributed equally to this activity, together with one or more other contributors. Choosing this option implies that no other co-author would choose the 'lead' option, and at least one other co-author would choose the 'equal' option.
- Supporting: You had a supporting role, which was lower than the contribution of a leading or equal role.
- No role: You did not contribute substantially to this activity.
- n/a: This role was not relevant for the research output at hand.

With this approach, we hope to reduce existing ambiguities surrounding the interpretation of authorships, and to make lying about contributions at least somewhat harder: The threshold for falsely claiming to have made a particular kind of contribution should be higher than the threshold for falsely claiming to have made "some" unspecified kind of contribution.

It should also be easier to disprove false claims of having made certain types of contributions. For example, a hiring committee could scrutinize a candidate who claims to have had the leading role in data analysis by asking in-depth questions about that analysis. However, false claims of having made a contribution will still be possible to some extent (Chapman et al., 2019), perhaps more easily for some CRediT categories (e.g., "supervision") than for others (e.g., data analysis). Despite this limitation, the CRediT system enables drawing a more fine-grained picture of an author's contributions than the order of authors does. Furthermore, if a department is specifically interested in hiring a scientist who is likely to make certain types of contributions (e.g., fundraising, coordinating large scale projects), this goal can be promoted by awarding bonus points to candidates who (claim to have) demonstrated the respective abilities, according to their previous CRediT roles.

(g) Potential Bias

With any assessment scheme, it is likely that some cases will not be adequately covered. The more important question, however, is whether an assessment scheme is systematically biased in favor of certain methodologies (e.g., experimental), certain types of research contributions (e.g., publications), or entire subfields of research (e.g., personality psychology). Several commenters raised the concern that this might be the case with the RESQUE framework.

Bias against certain methodologies

Some commenters criticized that our proposed rating scheme is not (fully) suitable for assessing research contributions based on qualitative methodology (Hostler, 2024; Karhulahti, 2024; Syed, 2024; Ulpts, 2024). As a consequence, researchers specializing in this type of research might face a systematic disadvantage. We respond to these concerns as follows:

First, as mentioned in some of the commentaries, some proposals on how to assess the quality of qualitative research do exist (e.g., Campbell et al., 2023; Johnson et al., 2020; Stenbacka, 2001). This suggests that assessing the quality of qualitative research is not per se impossible. We would thus like to encourage our colleagues working with qualitative methods to get in touch with us and help develop an expansion pack for the RESQUE framework that will do this type of research justice in evaluation contexts. Perhaps doing so would also help broaden the acceptance of quality standards for qualitative research more generally (a deficit highlighted by Hostler, 2024).

Second, several of the quality criteria present in the RESQUE Collector App are actually applicable to qualitative studies. For example, a researcher interviewing many participants about their experiences in a certain context may pre-register many aspects of her research design (e.g., sample size and recruitment strategy, research questions; Karhulahti et al., 2022) and make her research materials (e.g., interview forms) available. Although anonymization and provision of open qualitative data is often more difficult, it is not impossible in all cases.

Third, applicants who feel that they will be disadvantaged by the scoring algorithm have the option of declaring their entire body of research (or parts of it) to be exempt from scoring at all. Departments specifically interested in hiring a specialist in qualitative research may even ignore the RESQUE approach altogether until it becomes capable of adequately capturing the merits of good qualitative studies.

Bias against certain types of (research) contributions

Some commenters argued that certain types of research contributions (especially software development; see Auger and Claes, 2024) might be overvalued within the RESQUE framework, compared to other types (e.g., publications). We decided to postpone the development of indicator sets for published datasets and software because we realized that this is more complex than we had initially expected.

Regardless, the question of how to weigh different types of research contribution is ultimately to be decided by the respective evaluation committee. We trust that acceptable standards in this regard will emerge over time. It should be noted that, at present, neither the provision of datasets nor the development and maintenance of research software are in any way rewarded in typical hiring and promotion procedures. Their being valued in *some* way is almost certainly a step in the right direction.

Bias against certain subdisciplines

Several commenters expressed the concern that the proposed evaluation scheme might not be equally applicable to all subfields of psychology, and that certain criteria of particular relevance to some fields might be missing. For example, Brandt et al. (2024) pointed out that simulation studies play an important role in method development but that desirable features of simulation studies were not captured in the first version of our proposal. We agree that a one-fits-all version of the RESQUE Collector App is unlikely to be possible. There may be criteria that are more relevant to certain subfields than others. Therefore, we now suggest that the current, revised version should be seen as containing a "core set" of indicators that are likely to be applicable to most subfields, but should be complemented by more tailored "expansion packs" where necessary. Work on these expansion packs has already begun, with several task forces from DGPs divisions contributing to their development. Once finished and tested, they will be made available on the RESQUE website for free use.

Undue biases may also arise from the fact that meeting certain quality criteria may be more difficult for researchers working in a particular subfield. For example, several colleagues argued that they may not easily make their data available for a variety of valid reasons pertaining to data protection and privacy concerns (Fink-Lamotte et al., 2024; Niessen et al., 2023; Schwartz et al., 2023). Similar concerns came from researchers using qualitative methodology (e.g., Hostler, 2024), because the level of detail in such data is often high and may make individuals or organizations identifiable. Some of these concerns may be alleviated by making data available only as scientific use files (i.e., with access restrictions and contractual reuse agreements). This possibility is already incorporated in RESQUE.

The RESQUE Collector App offers the option to declare the entire set of open data indicators inapplicable to a given research output, as long as a plausible explanation is provided. In addition, members of a committee may decide in advance to treat certain indicators as generally inapplicable if they find them irrelevant or unsuitable to the decision-making process ahead of them.

(h) Overall Rigor Score

The question of whether the individual indicators should be used to form a single, overall score of methodological rigor has been discussed since the early stages of the RESQUE project. Many people seem to feel uncomfortable with the idea that such important decisions (e.g., whether to invite an applicant for an interview) may be made on the basis of a single metric. Would it not be better to compare the individual strengths and weaknesses of applicants in a more differentiated manner? This is indeed possible, as the RESQUE framework has now evolved to a point where comprehensive profiles for individual applicants can be easily derived. This, however, does not obviate the necessity to make decisions at some point. And in making these decisions, assessors will have to apply some kind of implicit or explicit cut-off to some kind of implicit or explicit continuum of suitability for the position in question.

This touches directly on the decades-long debate over whether diagnostic decisions should be made in a more intuitive or algorithmic fashion (a.k.a. "clinical vs. sta-

tistical prediction"; Meehl, 1954). Research has consistently shown that the latter approach is the more sound one, as it not only maximizes inter-individual comparability and transparency but also yields equal or better predictive validity (e.g., Grove et al., 2000). We thus propose embracing this view and argue that using a single broad score as a metric of the overall methodological rigor of an applicant's best work is acceptable, including the comparison of such a score with a cut-off for shortlisting. Of course, this claim rests on the assumption that the individual indicators feeding into the overall score have sufficient validity. We believe this is the case with the current version of the RESQUE criteria.

The RESQUE Collector App provides an overall methodological rigor score for each applicant, as well as a multi-dimensional rigor profile. This score is based on our admittedly subjective weighting of the individual criteria, but these weights may easily be adapted if necessary. Sparfeldt et al. (2024) argued that committees should openly discuss their evaluation schemes in advance and make them explicit. Lange et al. (2024) also suggested making these decisions in advance to maximize fairness and transparency. We agree with them. For example, it would be possible for a committee to define in advance that only applicants with hands-on experience in (a) preregistration and (b) replication attempts will be considered for the shortlist. The alternative to such an algorithmic approach would essentially be to make functionally equivalent decisions, just in a less traceable manner.

A committee that prefers not to make shortlisting decisions entirely dependent on a single methodological rigor score could instead adopt an approach in which this score is given a weight smaller than 1, and then make the rest of the judgment variance dependent on other sources. Even in these cases, however, we would recommend defining the selection rules in advance. Likewise, a committee may always choose to inspect the more differentiated rigor profiles for each candidate, and to consider this information in their decision making (e.g., in phase 2).

(i) Weighting of Individual Indicators

If one embraces the general idea of using an algorithmic approach in the first phase of the assessment process, one needs to employ some set of specific weights for each indicator. The exact point values we propose in the second target paper primarily reflect our own deliberations as to how much effort it takes to meet the respective quality standard, and of how much the value of a paper is increased by meeting it. Of course, there will always be a certain amount of subjectivity in such estimates. For now, the values are simply the result of discussions within our group. For the future, it would be desirable to determine these values in a more systematic and representative manner, by surveying a larger group of researchers. For example, several commenters (Auger & Claes, 2024; Brandt et al., 2024; Witte, 2024) were skeptical that the first version of our proposal gave too much weight to software development. A more satisfying solution in this regard might be derived by involving a larger number of expert judges (e.g., people who know how effortful it is to create and continuously update complex software packages).

(j) Types of Data and Samples

The original version of the empirical rigor criteria in RESQUE did not consider the type of data and the type of the samples that were used in an applicant's work. However, certain types of data and samples are usually more informative than others regarding a given research question (i.e., they have a higher value), and more informative data is often more effortful to collect. In recent years, the field of psychology has become increasingly - and rightfully - critical of studies that only use cross-sectional self-report data, especially when this data is gathered online. This type of data has played a very prominent role in the field, mainly because it can be collected very quickly and at low cost (e.g., from undergraduate students, in exchange for course credit). However, such data certainly is not ideal for answering many important research questions because of the inability to distinguish between the reality one is trying to assess and the participants' potentially biased views of that reality.

The revised version of RESQUE now assesses the type of data (e.g., self/other report, behavioral, physiological measures, content data, ...) and some relevant features of study samples (e.g., students, general population, rare condition; WEIRDness) in a more detailed manner. While these indicators do not by themselves contribute to the overall rigor score, they can be used to present a richer profile of an applicant's research. In addition, they can be easily used to define "red flags" (e.g., when all of an applicant's empirical papers only use cross-sectional self-report data).

(k) Impact

It seems justifiable to demand that research should be impactful, especially when it is being paid for with public funds. But measuring impact properly is more difficult than it might seem at first sight (Greenhalgh et al., 2016). We were surprised to discover how murky the issue still is in conceptual terms and will thus begin with a brief conceptual analysis.

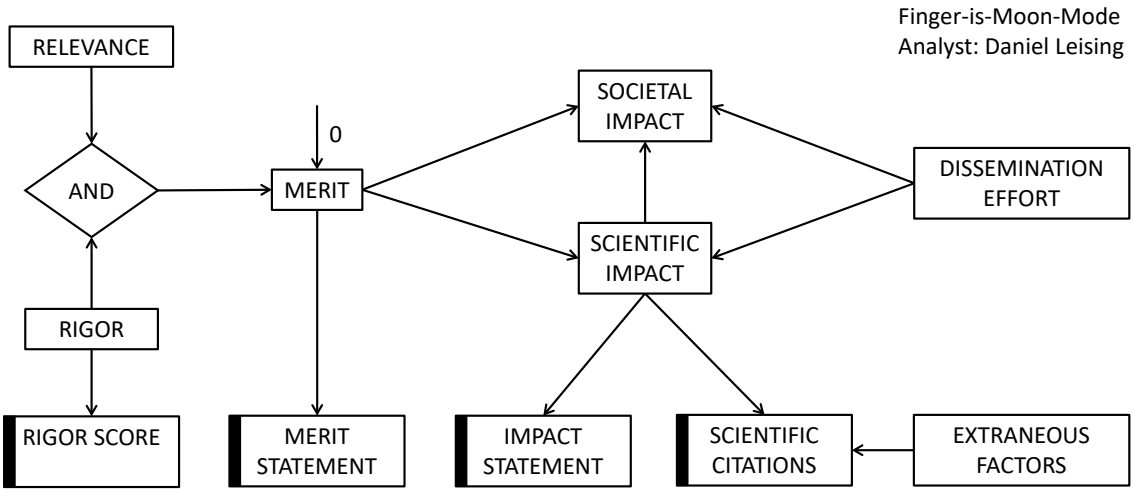

Figure 1

Conceptual model of how rigor, relevance, merit, impact and citations relate to one another. This analysis uses the VAST system (Leising et al., 2023). All paths are causal in nature. The zero coefficient implies that MERIT necessarily requires both RELEVANCE and RIGOR, and that no other relevant influences on MERIT exist. Boxes with thick borders denote measurements. "Finger-is-Moon Mode" implies that concepts are indexed directly by one of their names, instead of separating concepts from their names.

Figure 1 displays this analysis using the notation rules of the VAST system (Leising et al., 2023). Here, boxes symbolize concepts that may or may not apply to certain objects. The objects in this analysis are individual research papers, so the boxes in the display all symbolize potential properties of research papers. Boxes with thick black borders on the left-hand side symbolize something that can be measured. In the present analysis, this applies to RESQUE rigor scores, merit statements, impact statements, and the number of scientific citations to a paper (i.e., citations from other scientific articles).

The analysis captures the following ideas: First, scientific citations are a measure of scientific impact, which has to be distinguished from societal impact (Schwartz et al., 2023). For example, the invention of the steam engine had an enormous scientific and societal impact. In contrast, the discovery of the Higgs-Boson was a major scientific breakthrough but left most people's everyday lives unaffected so far. Within the RESQUE framework, both types of impact may be considered independently in phase 2 of an assessment process, after a basic level of methodological rigor has been established in phase 1. This is to avoid rewarding high citation counts for methodologically questionable papers. In the following, we will focus exclusively on scientific impact, because there is already a widely used - albeit quite controversial - metric for it: citation counts. In addition, we propose asking candidates for "impact statements".

Second, an article's scientific impact is a function of its scientific merit and of the dissemination efforts supporting it - by an article's authors (e.g., by attending conferences or promoting the paper on social media), or by others (e.g., publishers). These two influences on a paper's impact may compensate for each other: Highly meritorious papers may need little dissemination effort to become impactful, and papers with little merit may still be impactful (and thus get cited) if they are supported by strong dissemination efforts. In fact, we assume that even research with zero merit may become impactful. It should be noted, however, that

we do consider the independent influence of dissemination efforts on impact to be legitimate in principle: In our view, researchers do have a responsibility to inform their colleagues and the wider public about their research. Among two researchers with equally meritorious work, the one who has invested more in disseminating their work and therefore had more impact should be rewarded more.

It should also be noted that, in this model, scientific work may have great merit even if it has no impact (yet) at all. We assume that there may be applicants whose work is in fact highly meritorious but has not yet been recognized enough by their peers (Peter Higgs being a good example, at least in the first years after publishing his groundbreaking papers). Our framework is supposed to give these people a better chance than they would have under the current mainstream evaluation practices. For this purpose, we suggest using explicit merit statements that capture a paper's essential contribution from a candidate's perspective. These are supposed to be used as a starting point for in-depth discussions in phase 2 of the assessment process.

Third, we assume that the scientific merit of a research paper depends on a combination of (empirical and theoretical) rigor on the one hand, and the relevance of the given research question on the other hand. "Relevance" concerns the scope of the question or problem that is addressed (broader problems are more relevant than narrow ones), and the perceived urgency of finding an answer to the question or a solution to the problem. Other than merit and dissemination effort, which may compensate for each other entirely in creating impact, rigor and relevance affect merit in a multiplicative fashion. Accordingly, the VAST display (Figure 1) connects these two predictors using an AND diamond: Both rigor and relevance are necessary for achieving merit - as soon as either rigor or relevance is zero, there can be no merit.

While the RESQUE framework does provide a method for assessing the rigor of research, we are not aware of an established method for assessing the perceived relevance of research questions. A consensus paper describing shared research goals may serve this function (Leising et al., 2024), but consensual accounts of important future research goals are relatively rare in the psychology literature (Leising et al., 2022a). As long as relevance may not be convincingly assessed in an algorithmic fashion, we recommend delegating discussions of relevance to phase 2 of the assessment process. The candidates' merit statements are supposed to be used as the basis for these discussions. As evident from Figure 1, merit statements may not only contain aspects of relevance, but also aspects of rigor. This is intended, because of two candidates who made it to the shortlist, the one with the more rigorous work shall still be rewarded more. In addition, merit statements may also contain information regarding additional aspects of rigor that are not captured by the current version of the RESQUE indicators. Note that we decided not to separately account for the novelty or creativity of a candidate's scientific work. The reason is that these are unexpectedly tricky conceptual issues. For example, it seems that the extent to which novelty and creativity are deemed praiseworthy is largely moderated by the respective project's success. Addressing these and other complexities in sufficient depth is beyond the scope of the present paper. For now, we recommend making novelty and creativity part of in-depth discussions in phase 2 of the assessment process.

Fourth, it is well known that a paper's citation count tends to be strongly influenced by a number of extraneous factors (e.g., Sandoval-Lentisco, 2024; Stroebe and Strack, 2023). Some of these influences are rather unavoidable. For example, papers in larger research fields tend to be cited more, older papers have more time to acquire citations, and meta-analyses tend to be cited instead of the original studies that they are based on. For some of these influences, effective corrections (i.e., age and field normalizations) are available (Stroebe & Strack, 2023).

In addition, there are extraneous factors of a more problematic nature. For example, some authors engage in excessive self-citation to boost their citation counts. This can be easily corrected by excluding self-citations. But then things quickly become even more difficult: For example, some authors are more or less forced (by reviewers and/or editors) to cite certain papers in the course of review processes, whereas other authors preemptively cite eminent figures in the field who might act as reviewers on their papers. Thus, citation counts may partly reflect the cited author's perceived power, even when controlling for the cited paper's merit. Citation networks may also reflect interpersonal and institutional networks (Blashfield & Reynolds, 2012): people may cite each other if they know and like each other. Moreover, articles appearing in journals with high impact factors may be cited more - not because of their better quality, but because of their greater visibility and/or because those who cite them use the JIF as a proxy for quality. Currently, these (and many other) extraneous influences on citation counts are not corrected for in any way, and often they go unnoticed.

Given the surprising lack of evidence for a positive association between research quality and citation counts (Aksnes et al., 2019; Dougherty & Horne, 2022), quality thus needs to be established first. Within the

RESQUE framework, explicit assessments of methodological rigor serve this purpose. If a committee decides that it wants to consider scientific impact as well, we would propose (a) using a citation metric that corrects for self-citation, field, and the age of a publication, and (b) collecting separate impact statements from applicants. The latter should explain a paper's scientific impact in narrative terms, without referencing citation numbers at all. For example, a candidate may explain how a methodological contribution they made was adopted as the go-to-approach by a larger number of colleagues who wrote a consensus paper on best practices in their field (accounting for just 1 citation). Note again that both of these assessments of scientific impact should only be used in phase 2: Impact may only be reasonably considered after rigor has been established.

(l) Interdisciplinary Value

Karhulahti (2024) suggested adding the interdisciplinary value of academic work as a "fifth principle". After considerable debate amongst us, we concluded that, while we share the general sentiment, this would not be feasible as part of the first (algorithmic) phase of the evaluation process. This is for several reasons: First, the boundaries between academic (sub-)fields are rather fuzzy and somewhat arbitrary. Consequently, using different categorization systems would lead to different assessments of interdisciplinarity. Second, in some cases, the methodological and theoretical differences within a field are greater than those between fields. For instance, researchers from quantitative vs. qualitative sociology may find it easier to collaborate with researchers from entirely different disciplines than with each other.

Third, it is not entirely clear what would qualify a work as interdisciplinary. Does an author's paper become interdisciplinary if it is published in a journal from "another field" and/or if it is co-authored with researchers from "another field", and/or if it gets cited by researchers from "another field"? Fourth, shall authors receive additional credit for every new field that they associate with? Fifth, given the current lack of oversight regarding deserved vs. undeserved authorship, naively rewarding any appearance of interdisciplinarity via co-authorship would create an additional incentive for adding guest authors from other fields.

Sixth, and most important, "true" interdisciplinarity is likely to be a necessity rather than a choice. For example, most psychologists lack the training to engage in complex modeling efforts and will thus have to rely on more competent colleagues when engaging in such a project. If successful, the outcome will be a paper of better quality (that is, with greater scientific merit), which in turn should increase the chances that the article will make an impact.

Summarizing these points, we conclude that, at present, the measurement of interdisciplinarity is too messy to be used in an algorithmic fashion.

We did incorporate an automatic assessment of interdisciplinarity into the RESQUE Collector App (using the article-based classification system in OpenAlex), but recommend using this information only in phase 2 of the assessment process, if deemed helpful by the respective committee. At this stage, it can be discussed how much a candidate's collaborations with colleagues in other fields actually contributed to the quality and/or scope and/or impact of their work.

(m) Teaching

In their commentary, Hansen et al. (2024) clearly stated that they consider current practices for assessing applicants' teaching abilities inappropriate. We fully agree with them. Hansen et al. also provided detailed suggestions for a better approach. As explained in the first target paper (Schönbrodt et al., 2025), we think that ultimately all five types of academic contribution (research, teaching, leadership, service to the academic field, societal impact) should be considered in academic hiring and promotion processes. In the second target paper, we focus only on evaluation criteria for research. Criteria for the other four types of contribution remain to be developed - if possible. A commission of the German Psychological Society has been tasked with developing criteria for the teaching domain. Here, we would like to suggest making this the focus of a more extensive debate in *Meta-Psychology*: What does "good teaching" (of psychology and other academic subjects) look like in the 21st century? What behaviors should "good" teachers exhibit and be rewarded for? And how should their performance be assessed?

(n) Expertise

Frischkorn (2024) argued that the second phase of our proposed assessment scheme is underdeveloped, and that an appropriate assessment of the quality of candidates' work during this phase would require the involvement of experts in the respective area. At present, this requirement is rarely met in typical hiring and promotion processes. Frischkorn argues that, as a consequence, committees often resort to using invalid metrics. We agree.

Frischkorn then went on to argue that the problem may be mitigated by establishing more permanent (e.g., lecturer) positions for people with the respective expertise at the local faculties. While there is nothing wrong with establishing more permanent positions for academics, we do disagree with this view. Giving "local

experts" the power to ultimately determine the scientific merit of a candidate's work bears a risk of inviting a strong influence of local micropolitics. To avoid this and the associated risk to scientific integrity, we would rather recommend increasing the proportion of *external* experts on hiring committees (i.e., experts who are not members of the hiring faculty but from a relevant subfield; Boessel-Debbert et al., 2025).

(o) Gaming the New Metrics

Several commenters (Brandt et al., 2024; Dames et al., 2024; Fink-Lamotte et al., 2024; Hostler, 2024; Ulpts, 2024) were concerned that the new metrics proposed here might be vulnerable to gaming efforts. This is a reasonable concern, given that the current metrics of scientific productivity have been shown to be so commonly gamed (Chapman et al., 2019; Fong & Wilhite, 2017; Pruschak & Hopp, 2022) that one may legitimately ask what variance they reflect *apart* from gaming efforts. We know of Goodhart's Law, and we are aware of the strong incentive to somehow "come out on top" in the competition for permanent jobs in academia. As this incentive is likely to continue to exist, the question is not whether the new metrics will be gamed, but whether they are harder to game than the current ones. We do think that this is the case, because the RESQUE indicators are more amenable to actually being checked (e.g., whether data were actually made available, whether analyses were actually pre-registered). We also see the possibility that "gaming" them actually leads to their fulfillment because they are closer to the actual practice they are supposed to measure [1]. Undeniably, there is room for fraudulent behavior concerning the RESQUE indicators, but we do think it is likely to be substantially smaller than with mere numbers of authorships and citations.

If certain indicator points principally cannot be achieved by certain types of research, this would incur an unfair disadvantage. Therefore, we introduced a "not applicable" option for most indicators. Using this option removes the points for that indicator from the maximum of attainable points. The overall rigor score is calculated as a "percentage of maximum possible" (POMP) across applicable indicators only. Thereby, applicants who are unable to achieve certain points for reasons beyond their control may still achieve the maximum of 100 percent.

However, this approach does allow for some gaming of the system, as well: Instead of selecting "not available" (e.g., open data), which would result in 0 points, applicants could simply select "not applicable", which would exclude the respective indicator from being counted. Taken to the extreme, applicants could do this with all the criteria they fail to meet. We therefore recommend that using the "not applicable" option should always require a justification. When using it, applicants should have to explain why the respective indicator is not applicable to this particular piece of work.

We propose checking all the self-reported characteristics of the papers submitted by candidates who are supposed to move from the longlist to the shortlist (and maybe for border cases). Preliminary research (Etzel et al., 2025) suggests that this would be feasible with acceptable effort. Justifications for declaring criteria to be "not applicable" may be included in these checks.

If the effort of checking the candidates' self-assessments is still considered to be too high, one may resort to only checking random subsamples of indicators for each shortlisted applicant instead. Even this approach would already imply greater diagnostic diligence as compared to typical contemporary committee work, and likely help improve the average level of methodological rigor among shortlisted candidates. To further reduce the effort associated with such checks, we are now exploring the possibility of establishing a central open access registry to which people may submit their scientific output. These would then be independently evaluated with regard to our quality criteria just once and then candidates would only have to send a committee the link to that evaluation.

(p) Representativeness

In their commentary, Stengelin et al. (2024) highlighted the current debate over a "generalizability crisis" in psychology, emphasized the importance of sample representativeness and suggested incorporating it as an indicator in the first assessment phase. We agree wholeheartedly and have amended the RESQUE Collector App accordingly. In fact, our initial proposal (Leising et al., 2022b) had already contained this as one of the ten most important quality indicators. In developing the RESQUE framework, however, we became a bit more skeptical regarding the feasibility of actually checking claims of representativeness for their validity. But given that representative data is still very rare in psychology, we now assume that any paper exhibiting this desirable quality will make that fact known in unambiguous terms (e.g., by detailing the strategy that was used for recruiting research participants).

It should be noted that sample representativeness is just one of two key dimensions in this regard. The other

[1] For example, the bonus points for an open license of source code (which is one of the FAIRness indicators) can be achieved within seconds. But once the code has an open license, its benefits of reusability are actually achieved.

- stimulus representativeness - may in fact be even more neglected, despite being equally important. Stimulus representativeness means that the stimuli that research participants are exposed to resemble those whose effects in the natural environment the researchers want to model. First and foremost, this concerns the ranges of the relevant stimulus characteristics, but it may also concern the central tendencies and the variance of the distribution(s) of those characteristics. A stimulus sample may be viewed as representative for the stimulus population if the sampling of the former from the latter took place at random. We assume that authors aiming for stimulus representativeness in their study are likely to highlight that desirable feature. Thus, we have now added this indicator to the RESQUE Collector App, too.

Outlook

We envision the development and improvement of RESQUE to be an ongoing collaborative process, shaped over time by shared community experiences and repeated evaluations. Hence, the state of the system described in this rejoinder and the revised versions of the two target papers can only be an intermediate step of a longer journey. Fortunately, at this point, community involvement is very strong, once more confirming that many colleagues also perceive an urgent need for reforming research assessment, and their great willingness to actively help with this reform.

We will close with an overview of current developments and a few suggestions for additional steps that may be taken to improve the RESQUE framework and make it more effective. First, as we stated above, the core set of indicators should be accompanied by expansion packs that account for the specifics of some of psychology's subdisciplines. The core set also has to incorporate reliable and valid indicators of what is good theorizing and indicators for other types of research contributions (i.e., data and software). Work on all of this is now well underway. Regarding quality indicators for contributions based on qualitative methodology, we would like to repeat our invitation to experts in this field to collaborate on this with us.

Second, the applicability of a system like RESQUE (especially its more algorithmic phase 1) crucially hinges on the extent to which the given indicators can be objectively assessed at all. If the presence of certain desirable features of a piece of research were only in the eye of the beholder, then any attempt to assess them as part of hiring or promotion procedures would not be justifiable. At present, the number of empirical studies on this issue is still small (Etzel et al., 2025; Leising et al., 2022a), but they do show that inter-rater agreement in assessing some quality indicators of research publications is already quite acceptable, and high for an aggregated rigor score. However, the same research also shows that some quality criteria are much harder to judge than others, so psychology (like many other fields) would clearly benefit from establishing more consensual reporting standards, to make the job easier for raters of research quality.

Third, the procedure would become better legitimized if the weights of the individual indicators were derived in a more representative fashion. At present, they basically reflect the outcome of in-depth discussions among the members of the group developing the system. A more representative weighting scheme could be obtained by surveying a large sample of community members for their views on the matter. Lange et al. (2024) even argued that it would be desirable for committees hiring people for similar positions to use a joint set of indicators and weights for the given field. We agree, as this would establish greater comparability between the procedures applied at different institutions. As a side effect, such standardization would significantly reduce the effort associated with using the system.

Fourth, for the same reason (effort reduction) it would be desirable to establish a system in which each research product is scored once in a reliable fashion by an independent team of assessors, and the evaluation is then made publicly available. Parts of this assessment could (and should) be part of the traditional peer review process at journals. This database could be used by hiring and promotion committees as needed. On an opt-in basis, it could also be one data source for a public author's profile. We could imagine the Leibniz Institute for Psychology (ZPID) taking on this responsibility within the German academic landscape.

Fifth, we are pleased to be able to announce that the RESQUE project has been honored with the 2023 Einstein Award for Promoting Quality in Research (https://award.einsteinfoundation.de/award-winners-finalists/recipients-2023/early-career-award-2023). As a next step, we will test and evaluate the current set of quality indicators with regard to reliability, validity, usability, and efficiency in different appointment contexts. Several academic institutions have already agreed to participate and pilot the RESQUE framework, thus providing an invaluable real-world testing environment. All results and materials from the project will be made available with an open license in the RESQUE repository (https://github.com/RESQUE-Framework).

It seems that academia is in the process of seriously reconsidering several issues of vital importance to the functioning of the science system. We are delighted to see this happen, and proud that the RESQUE framework

continues to make a contribution to this ongoing discourse.

Author Contact

Corresponding author is Daniel Leising, Technische Universität Dresden, Germany (daniel.leising@tu-dresden.de).

ORCID:
Leising 0000-0001-8503-5840
Gärtner 0000-0003-4296-963X
Schönbrodt 0000-0002-8282-3910

Conflict of Interest and Funding

The authors state that they have no conflict of interest to declare.

This project and publication is supported by the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research - in cooperation with the BIH QUEST Center for Responsible Research. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, the Einstein Foundation or the award jury.

Author Contributions

Conceptualization: DL, AG, FS Writing - Original Draft: DL Writing - Review and Editing, DL, AG, FS

The order in which the authors are named reflects the perceived size of their respective contributions (first > last > middle).

Acknowledgments

We are most grateful to the Executive Committee of the German Psychological Society for their ongoing enthusiastic support, to the Editorial Board of Meta-Psychology for providing us with the forum for the present discussion, and to Daniël Lakens for overseeing the editorial processes for these three papers. We also would like to sincerely thank all members of the community who have helped us develop and improve the RESQUE framework to its current state. This includes not only the authors of the 21 commentaries, but also everyone who contributed to the project and with whom we have had fruitful discussions at conferences, after talks, or via email. It is a privilege to work in an environment where so many colleagues devote their valuable time and energy to improving research assessment with such dedication.

Open Science Practices

This article is a conceptual paper and as such is not eligible for Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Abele-Brehm, A., & Bühner, M. (2016). Wer soll die Professur bekommen? Eine Untersuchung zur Bewertung von Auswahlkriterien in Berufungsverfahren der Psychologie [Who should receive the professorship? A research on the evaluation of different hiring criteria for appointments in academic psychology]. *Psychologische Rundschau*, *67*(4), 250–261. https://doi.org/10.1026/0033-3042/a000335

Aksnes, D., Langfeldt, L., & Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*, *9*(1), 2158244019829575. https://doi.org/10.1177/2158244019829575

Auger, V., & Claes, N. (2024). Comment on "Responsible Research Assessment: Implementing DORA for hiring and promotion in psychology". *Meta-Psychology*, *8*, 2158244019829575. https://doi.org/10.15626/MP.2023.3779

Blashfield, R. K., & Reynolds, S. M. (2012). An Invisible College View of the DSM-5 Personality Disorder Classification. *Journal of Personality Disorders*, *26*(6), 821–829. https://doi.org/10.1521/pedi.2012.26.6.821

Boessel-Debbert, N., Kluge, A., Leising, D., Mischkowski, D., Phan, L. V., Richter, F., & Stahl, J. (2025). An analysis of functional relationships between systemic conditions and unethical behavior in german academia [Preprint]. https://doi.org/10.31234/osf.io/xj2m6_v1

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

Brandmaier, A. M., Ernst, M., & Peikert, A. (2024). Commentary: 'Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology'. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3715

Brandt, H., Henninger, M., Ulitzsch, E., Kleinke, K., & Schäfer, T. (2024). Responsible research assessment in the area of quantitative methods research: A comment on Gärtner et al. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3796

Brown, G. (2024). A broader view of research contributions: Necessary adjustments to DORA for hiring and promotion in psychology. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2022.3652

Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., & Goodman-Williams, R. (2023). Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data. *Advances in Methods and Practices in Psychological Science*, *6*(4), 25152459231205832. https://doi.org/10.1177/25152459231205832

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

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

Dames, H., Musfeld, P., Popov, V., Oberauer, K., & Frischkorn, G. T. (2024). Responsible research assessment should prioritize theory development and testing over ticking open science boxes. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3735

Dougherty, M., & 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). https://doi.org/10.1098/rsos.220334

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*. https://doi.org/10.31234/osf.io/4w7rb_v2

Fink-Lamotte, J., Hilbert, K., Bentz, D., Blackwell, S., Boehnke, J. R., Burghardt, J., Cludius, B., Ehrenthal, J. C., Elsaesser, M., Haberkamp, A., Hechler, T., Kräplin, A., Paret, C., Schulze, L., Wilker, S., & Niemeyer, H. (2024). Response to responsible research assessment I and II from the perspective of the DGPs working group on open science in clinical psychology. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3794

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS One*, *12*(12), e0187394. https://doi.org/10.1371/journal.pone.0187394

Frischkorn, G. T. (2024). Responsible Research Assessment requires structural more than procedural reforms. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3734

Gärtner, A., Leising, D., Freyer, N., Musfeld, P., Lange, J., & Schönbrodt, F. D. (2022). Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology [Preprint]. https://doi.org/10.31234/osf.io/5yexm_v1

Gärtner, A., Leising, D., & Schönbrodt, F. (2023). Empfehlungen zur Bewertung wissenschaftlicher Leistungen bei Berufungsverfahren in der Psychologie [Recommendations for the evaluation of academic performance for hiring and promotion in psychology]. *Psychologische Rundschau*, *74*(3), 166–174. https://doi.org/10.1026/0033-3042/a000630

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

Greenhalgh, T., Raftery, J., Hanney, S., & Glover, M. (2016). Research impact: A narrative review. *BMC Medicine*, *14*, 78. https://doi.org/10.1186/s12916-016-0620-8

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hansen, M., Beitner, J., Horz, H., & Schultze, M. (2024). Indicators for teaching assessment. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3763

Hostler, T. (2024). Research assessment using a narrow definition of "research quality" is an act of gatekeeping: A comment on Gärtner et al. (2022). *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3764

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education*, *84*(1), 7120. https://doi.org/10.5688/ajpe7120

Karhulahti, V.-M. (2024). Interdisciplinary value. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3679

Karhulahti, V.-M., Branney, P., Siutila, M., & Syed, M. (2022). A primer for choosing, designing and evaluating registered reports for qualitative methods. *MetaArXiv*. https://doi.org/10.31222/osf.io/2azkf

Lange, J., Degner, J., Gleibs, I. H., & Jonas, E. (2024). Faires und valides shortlisting in Phase 1 [Fair and valid shortlisting in phase 1]. *Psychologische Rundschau*, *74*(3), 187–189. https://doi.org/10.1026/0033-3042/a000641

Lange, J., Freyer, N., Musfeld, P., Schönbrodt, F., & Leising, D. (2025). A checklist for incentivizing and facilitating good theory building. *Zeitschrift für Psychologie*, *233*. https://doi.org/10.1027/2151-2604/a000604

Leising, D., Grenke, O., & Cramer, M. (2023). Visual Argument Structure Tool (VAST) Version 1.0. *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2021.2911

Leising, D., Liesefeld, H., Buecker, S., Glöckner, A., & Lortsch, S. (2024). A tentative roadmap for consensus building processes. *Personality Science*, *5*. https://doi.org/10.1177/27000710241298610

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

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

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Niessen, C., Melchers, K. G., Ohly, S., Fay, D., Handke, L., & Kern, M. (2023). Ein Plädoyer für breit gewählte und anforderungsbezogene Leistungsindikatoren [a call for using broad and demand-focused achievement indicators]. *Psychologische Rundschau*, *74*(3), 180–182. https://doi.org/10.1026/0033-3042/a000637

Ortner, T. M., Kretzschmar, A., Rauthmann, J. F., & Tibubos, A. N. (2013). Berufungsverfahren unter einer diagnostischen Perspektive durchführen. *Psychologische Rundschau*, *74*(3), 187–189. https://doi.org/10.1026/0033-3042/a000638

Pruschak, G., & Hopp, C. (2022). And the credit goes to — Ghost and honorary authorship among social scientists. *PLOS One*, *17*(5), e0267312. https://doi.org/10.1371/journal.pone.0267312

Sandoval-Lentisco, A. (2024). Commentary: "Responsible Research Assessment: Implementing DORA for hiring and promotion in psychology". *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2022.3655

Schönbrodt, F. D., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., & Leising, D. (2022). Responsible Research Assessment I: Implementing DORA and CoARA for hiring and promotion in psychology [Preprint]. https://doi.org/10.31234/osf.io/rgh5b_v1

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. *Meta-Psychology*, *9*. https://doi.org/10.15626/MP.2024.4601

Schwartz, B., Szota, K., Schmitz, J., Lueken, U., & Lincoln, T. (2023). Mehr Differenzierung nach Fachgebieten [More differentiation by subdisciplines]. *Psychologische Rundschau*, *74*(3), 184–185. https://doi.org/10.1026/0033-3042/a000639

Sparfeldt, J. R., Spörer, N., Greiff, S., & Schneider, R. (2024). Ein Plädoyer für valide(re) Bewertungen der wissenschaftlichen Leistungen in Berufungsverfahren [A call for (more) valid evaluations of scientific achievement in hiring processes]. *Psychologische Rundschau*, *74*(3), 185–187. https://doi.org/10.1026/0033-3042/a000640

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision*, *39*(7), 551–556. https://doi.org/10.1108/EUM0000000005801

Stengelin, R., Bohn, M., Sánchez-Amaro, A., Haun, D., Thiele, M., Daum, M., Felsche, E., Fong, F., Gampe, A., Giner Torréns, M., Grueneisen, S., Hardecker, D., Horn, L., Neldner, K., Pope-Caldwell, S., & Schuhmacher, N. (2024). Responsible Research is also concerned with generalizability: Recognizing efforts to reflect upon and increase generalizability in hiring and promotion decisions in psychology. *Meta-*

Psychology, *8*. https://doi.org/10.15626/MP.2023.3695

Stroebe, W., & Strack, F. (2023). Zweierlei Maß? Warum manche Psychologen den Gebrauch von quantitativen Indikatoren der Forschungsqualität ablehnen[A double standard? Why some psychologists reject the use of quantitative indicators of research quality]. *Psychologische Rundschau*, *74*(3), 175–189. https://doi.org/10.1026/0033-3042/a000631

Syed, M. (2024). Valuing preprints must be part of responsible research assessment. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3758

Ulpts, S. (2024). Responsible assessment of what research? Beware of epistemic diversity! *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3797

Witte, E. (2024). Comment on: responsible research assessment I and responsible research assessment II. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3685