Research Evaluation, 2024, **33**, rvae055
https://doi.org/10.1093/reseval/rvae055
Article

OXFORD

Understanding the social and political dimensions of research(er) assessment: evaluative flexibility and hidden criteria in promotion processes at research institutes

Tony Ross-Hellauer [1,2,*], Noémie Aubert Bonn[3,4], Serge P.J.M. Horbach [5]

[1]Open and Reproducible Research Group, Know Center Research GmbH, Sandgasse 34, Graz, 8010, Austria
[2]Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36, Graz, 8010, Austria
[3]Department of Healthcare and Ethics, Faculty of Medicine and Life Science, Hasselt University, Martelarenlaan 42, Hasselt 3500, Belgium
[4]Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom
[5]Danish Centre for Studies in Research and Research Policy, Aarhus University, Bartholins Allee 7, Aarhus 8000, Denmark

*Corresponding author. Open and Reproducible Research Group, Know Center Research GmbH, Sandgasse 34, Graz, 8010, Austria.
E-mail: tross@know-center.at

Abstract

Debates about appropriate, fair and effective ways of assessing research and researchers have raged through the scientific community for decades, recently mostly concerned with discussing the merits and limitations of metric-based, quantitative assessments versus peer review-based, qualitative alternatives. Ample attention has been paid to formal assessment criteria, building to a consensus that less emphasis should be placed on quantification, while supporting open and diverse sets of criteria. Yet the theory and evidence upon which such policy reform depends is still surprisingly sparse. Based on qualitative free-text responses from 121 respondents gathered during an international survey of active researchers, this study examines researchers' perspectives on how criteria are applied in practice and how those being assessed perceive informal criteria to determine the outcomes of assessments. While confirming the general critique on over-emphasizing quantification, respondents particularly identify a mismatch between formal criteria and actual evaluation practices. Hidden criteria, including social, political, and demographic factors, are perceived important, especially in intransparent assessment procedures, opening up for assessors' evaluative flexibility. This adds to ongoing discussions on the performativity of assessment criteria and lays bare a tension between the rigidity and flexibility of criteria and the extent to which these can be transparently communicated.

Keywords: research assessment; academic careers; research metrics; assessment reform; transparency.

1. Introduction

Momentum for reform of research assessment processes is quickly gathering pace. Over the last decade, recognition of the need for more responsible use of metrics (DORA 2012; Hicks et al. 2015; Wilsdon et al. 2015), combined with observations that uptake of open and responsible research practices requires commensurate evaluation measures (Wilsdon et al. 2015; Munafò et al. 2017), has led policy actors to place research assessment reform at the top of their agendas (UNESCO 2021; CLACSO-FOLEC 2022; EUA 2022; Science Europe 2022). The recent formation of CoARA, the Coalition for Advancing Research Assessment, a global 'coalition of the willing' for reform underpinned by shared principles, commitments and timeframes (CoARA 2022), and the similar US Higher Education Leadership Initiative for Open Scholarship (HELIOS) initiative, arguably suggest a tipping-point.

Amongst the various ways in which researchers and their research are assessed, review, promotion and tenure (henceforth RPT) processes at research institutions have been identified as key for reform (CoARA 2022). RPT criteria are usually centred on three main types of activities: teaching, service and research, although the extent to which each of these domains are addressed in practice is disputed (Alperin et al. 2019). Considering criteria regarding research, although a quickly growing body of work demonstrates the need for greater support for open and responsible practices in RPT criteria (Moher et al. 2018, 2020; Schimanski and Alperin 2018; Rice et al. 2020; Alperin et al. 2022; Pontika et al. 2022a; Ross-Hellauer et al. 2023), or general support for the aim of reducing quantification, in terms of number of publications, research funding, or proxy measures for quality like the Journal Impact Factor (DORA 2012; Hicks et al. 2015; Wilsdon et al. 2015), less is known about how criteria are applied in practice.

This paper analyses qualitative free-text responses from 121 respondents gathered during an international survey of active researchers, to address two main research questions: (1) What are researchers' general perceptions towards research assessment criteria used in review, promotion and tenure processes at institutions? (2) What other factors (e.g. social, political, performance-based) which are not official criteria, do participants identify as nonetheless important in assessment processes?

2. Background and theory

Evaluation processes, no matter the veneers of objectivity applied, remain human processes at root. What is to be valued, and how, are continuously shaped by reflexively-evolving,

distinct, socially-constructed, practices, structures, infrastructures and situations (Krüger and Reinhart 2017). This partly explains the variation in formal assessment criteria across disciplinary contexts, career stages and history (Mantai and Marrone 2023). Moreover, decisions are also made by groups and individuals, and objectivity in decision-making is hence at risk from the social (Wennerås and Wold 1997; Bornmann, Mutz and Daniel 2007; Ginther et al. 2011; Teplitskiy et al. 2018), political (Wennerås and Wold 1997; Altbach, Yudkevich and Rumbley 2015; Torrance 2016) and cognitive (East 2016; Hatch and Schmidt 2019; Hom Jr and Van Nuland 2019; Juárez Ramos 2019) biases and preferences of evaluators. Furthermore, although reliance on quantitative indicators may have the veil of objectivity, they too are fundamentally affected by biases, either those baked into the metrics (Strathern 2000; Larivière and Gingras 2010; Larivière and Sugimoto 2019; Islam and Greenwood 2022), or those of the assessors using the metrics (Hammarfelt and Rushforth 2017). For example, some metrics measuring the adoption of open science practices can be biased against certain disciplines or even genders, because of social norms and practices (Strathern 2000; Zhu 2017). The case of gender inequality is another case in point. Despite formal policies to tackle this issue, the criteria for assessment, either qualitative or quantitative, by their design or use, may still result in disparity (Larivière et al. 2013; Macaluso et al. 2016; Jappelli, Nappi and Torrini 2017).

Rigidly defined metrics were initially promoted because of their ability to provide standardization and transparency, i.e. providing 'macro-level summary measure that can be evaluated independently and objectively allowing universities to make well-informed decisions' (Charlton and Andras 2007: 556). Such instrumentalist rationale is now heavily discredited as naive to the values encoded in qualities reduced to easily-comparable numeric quantities, and subsequent negative effects upon behaviours (Hicks et al. 2015; Wilsdon et al. 2015). Recent years have seen ample recognition of the ways in which poorly designed or poorly applied metrics can assume epistemic agency (Müller and de Rijcke 2017) to distort incentives (Smaldino and McElreath 2016), invite gaming (Siler and Larivière 2022), distort behaviours via goal displacement or task reduction (de Rijcke et al. 2016), threaten research integrity (Moher et al. 2020) and even transform academic values (Burrows 2012). In addition, the regime of transparency and quantification has had deep effects of restructuration of the experiential realities of research. Felt (2017: 55) has discussed how, in the context of the social sciences in Austria, 'the tempor(e)alities of academia' have been disrupted by 'the temporal and counting logic of funding agencies and universities'. Reliance on third-party funding has led to a 'projectification' of research, the 'packaging' of knowledge generation into 'three-year units', a rise in short-term contracts, and a focus on article publications (rather than books or other outputs) to more immediately demonstrate return-on-investment.

This, combined with recognition of the need for a broader range of criteria which incentivise open and responsible research (Wilsdon et al. 2015; Munafò et al. 2017), has meant recent years have seen renewed emphasis on the need for peer review and qualitative assessment, including new instruments like 'bio-sketches' (Curry 2018) or narrative CVs (Woolston 2022), for their potential to widen conceptions of achievement, broaden diversity and foster more holistic assessment and sustainable working environments (Robinson-Garcia et al. 2023). In turn, however, proponents of metrics argue that such moves 'will lead to randomness and a compromising of scientific quality' (Poot and Mulder 2021). Chawla (2021) sums recent discussions in the Netherland on this topic, and how critics suggest reforms to consider more holistic research criteria (e.g. teamwork, public engagement, leadership and Open Science practices) could lead to reduction of transparency and clarity in assessment, and the proliferation of political rather than scientific factors.

2.1 Transparent assessment criteria

Transparency has come to be prescribed as a crucial characteristic of either form of assessment. This has been fuelled by neoliberal philosophies and especially 'New Public Management' (NPM) organizational strategies, which centre efficiency, performance, and measurement as motivating principles in Higher Education. Although far from a unified theory, the basic characteristics of NPM are summed by Broucker and De Wit (2015) as use of private-sector organizational strategies, organizational distance between policy formulation and implementation, entrepreneurship, 'input and output controls' based on audit and evaluation, disaggregation of organizations, focus on costs and growth, and 'customer'-oriented framings for services. NPM hence frames transparency of decision-making, target-setting and evaluation as necessary conditions for accountability.

The impact of NPM on Higher Education has been heavily criticized (Hammerschmid et al. 2013). A specific line of critique relevant for our purposes, is the extent to which NPM's 'will to transparency' is even possible or desirable. As Tsoukas (1997) discusses in regards to the 'information society', or Power (Power 1999) discusses related to the 'audit society', NPM-related philosophies require all that is measurable, to actually be measured, even if it requires proxies to stand in for the object or phenomenon to be audited. Transparency in the use of these metrics is considered to be key.

The degree to which transparency is desirable and attainable is not self-evident though. The application of criteria necessarily involves mediation and interpretation. Both the design and application of criteria involve hermeneutic activities, which are inherently individual and hence untransparent. In addition, those setting criteria, those applying criteria and those being assessed do not necessarily share a common reference framework, acts of translation are continuously required. Since criteria are selective and those interpreting them situated, there is no possibility that systems can be made fully transparent, 'no detached Olympian high ground from which it may be inspected' (Tsoukas 1997: 834).

As a case in point, research captured in the review by Schimanski and Alperin (2018) indicates that criteria are often considered unclear (Diamantes 2005; King et al. 2006; Smesny et al. 2007; Acker and Webber 2016), and that researchers perceive that criteria are flexibly applied by assessors (May 2005; Harley et al. 2010; Prottas et al. 2017). Kaltenbrunner and de Rijcke (2019) studied evaluative flexibility in the case of curricula vitae assessment in funding decision-making. They conceptualized such moments of assessment as a 'generative interplay' between a historically-contingent, formalized infrastructure (the CV), and 'a situated evaluative practice in which the representational function of that infrastructure is itself interpreted and established'

(Kaltenbrunner and de Rijcke 2019: 864). Here, standardization of the CV format comes in conflict with ever-evolving realities in and across diverse research contexts. Consequently, CV categories 'are too unspecific to make much sense on their own terms; they need to be contextualized on the fly' (Kaltenbrunner and de Rijcke 2019: 864).

As different epistemic communities and scholarly cultures are known to have diverse interpretations of quality and excellence (Moore et al. 2017; Hessels et al. 2019), this reflects fundamental tensions as to the extent that situated interpretation should be encouraged or contested. To complicate matters further, the representational and performative aspects of measuring tend to constantly interact. For instance, the distinction between those that measure and those that are measured is often blurry, as gatekeepers tend to be members of the community in which assessment is performed. In addition, boundaries between the communities they represent are increasingly foggy, as a result of fading disciplinary boundaries. Consequently, when proxies travel between contexts, their representational characteristics in one context might simultaneously act performatively in other academic settings, ultimately feeding back to their initial context, e.g. when metrics well-suited to one disciplinary community are used unreflectively in another.

2.2 Evaluative flexibility and hidden assessment criteria

Within the interplay between uniformity and flexibility, rigidity and transparency, it remains unclear to what extent evaluative flexibility of criteria enables social and political factors to manifest. Similarly, which hidden factors are at play remains an open question. We suggest this as an urgently underexplored dimension of the debate, especially as we might expect such factors to be highly consequential. For example, even within the tightly-controlled context of recruitment procedures in Swedish universities, which Hammarfelt and Rushforth (2017: 171) recognize are 'designed to be impartial and merit-based in that external reviewers assess the candidates', there are nonetheless 'many ways in which the recruiting department can influence the process'. The Swedish case is highly instructive here, with research on gender equity in hiring further exploring how 'interpretative flexibility' prevails even in seemingly tightly-controlled procedures (Helgesson and Sjögren 2019; Mählck, Kusterer and Montgomery 2020), with formalization realized via 'evolving and often vague metrics and standards, assessed through an opaque mix of procedural steps' (Helgesson and Sjögren 2019: 574). In the Netherlands, van den Brink, Benschop and Jansen (2010: 1477ff.) examined how attempts to bring transparency and accountability to academic appointment schemes were hindered by the fact that selection protocols 'remain toothless—paper tigresses that are fraught with implementation problems'. Their interviews 'contain many examples of political games and loose interpretations of the rules and regulations', with deviations justified as in the interest of meritocratic appointment and aversion to bureaucracy. Alternatively, studying researchers' perceptions of promotion and tenure criteria in the US and Canada, Morales et al. (2021) highlight the subjectivity inherent in commonly-used but vague terms such as 'quality', 'excellence' and 'prestigious'. These and other studies indicate that similar issues are at stake in various research contexts around the globe (e.g. Delgado, Tarango and Machin-Mastromatteo 2020; Shu, Liu and Larivière 2022).

Beyond these specific examples, there remains much to be known of how evaluative flexibility is experienced by those assessed. Hammarfelt and Rushforth (2017: 171) say that the 'broader politics and practices of academic recruitment is indeed a fascinating topic, which has so far only briefly been covered'. While interest in research assessment has surged in the past years, we contend that few empirical studies addressed the broader political and social factors involved in research assessment in review, promotion, and tenure. We here aim to broaden this understanding.

3. Methods

This study was an exploratory study of perspectives from researchers on research assessment criteria and how research assessments are operationalized in practice. Our research questions are:

1) What are the general perceptions towards research assessment criteria used in promotion processes at institutions?
2) What other factors (e.g. social, political, performance-based) not officially listed as criteria, do participants identify as nonetheless important in assessment processes?

The data for this paper derives from a survey targeting active researchers, distributed internationally from 29th June to 30th July 2021 (Pontika et al. 2022b; Ross-Hellauer et al. 2023). The survey aimed to examine participants' perceptions of review, promotion and tenure criteria and practices, with a focus on those related to Open Science and Responsible Research and Innovation.

3.1 Target population and survey instrument

The target population was active researchers from diverse academic disciplines globally. To compile our survey sample, we randomly selected email addresses from corresponding authors who had publications from 2014 to 2020 present in the CORE scholarly content aggregator service (Knoth and Zdrahal 2012). The full survey instrument included sections covering institutional context, respondents' views on their respective institutions' RPT policies, their views on the relative importance of various aspects of RPT criteria, their own Open Science and RRI practices, as well as demographic information. It covered a total of 31 questions, mostly quantitative in nature (the analysis of which is contained within a prior publication (Ross-Hellauer et al. 2023)). In addition, the survey contained open-ended questions regarding (1) participants' general perceptions towards research assessment criteria used in promotion processes at institutions, and (2) what other factors (e.g. social, political, performance-based) not officially listed as criteria, participants identified as nonetheless important in assessment processes. The responses collected in response to these latter two questions form the data analysed in this study.

3.2 Ethics and informed consent

No ethical approval was sought since the host institution (Graz University of Technology) of the lead researcher (Tony Ross-Hellauer) did not at that time require it, and the content of the survey was determined through internal consultation to be low-risk in terms of sensitive data or ethical issues. On

commencing the survey, all participants were presented with information required for informed consent, stating that participation was voluntary, and they could withdraw at any stage, that and how data would be anonymized and securely stored, and that (anonymized) results would be disseminated via research publications. Participants had to read and consent to these conditions before commencing the study. The survey was administered via LimeSurvey, with individualized invitations which used unique tokens to maintain anonymity.

3.3 Survey testing and distribution

The draft survey was tested in May/June 2021 with 11 participants (contacts of the study leads but not connected to the survey work), with feedback collected via email and Google form. In-depth cognitive interviews were conducted with two researchers. The instrument was revised and shortened in response to feedback. The survey ran from 29th June until 30th July 2021. Of 16,500 emails sent, the survey reached 11,463 participants (main reasons for emails not received were email bounce-backs). The survey received 323 responses (response rate 2.81%). Of these, a further 41 participants who indicated they were not active researchers, as well as 84 incomplete responses, were excluded for a total 198 full responses. Of these, 121 answered at least one of the two free-text questions analysed here (overall response rate for this study (1.05%)).

3.4 Data analysis and availability

Free-text responses were analysed in NVivo version R1. In a first step, the lead author of this study used an iterative bottom-up open coding approach based in Grounded Theory (Strauss and Corbin 1997), whereby initially very broad and freely-applied index codes were refined and grouped as themes emerged. This preliminary code-book was then discussed amongst all three authors, with codes revised and re-grouped collectively based on discussion of emergent themes. As the emphasis was on collaborative coding, no formal intercoder reliability checks were undertaken. Rather, consensus was reached through a series of iterative and interactive discussions and refinements. Any persistent disagreements were resolved by refining the definitions of codes and themes or, if necessary, by introducing new codes to better capture the nuances of the data. In the below presentation of results, obvious mistakes in spelling have been corrected, but the text is otherwise presented verbatim. We share the qualitative responses within Supplementary File, decoupled from the quantitative data, survey instrument and other materials included within the already published dataset (Pontika et al. 2022b), and with identifying information redacted to ensure the anonymity of respondents.

4 Results

4.1 Demographics

One hundred twenty-one respondents answered at least one of the two free-text questions. Respondents came largely from Europe and North America, although there was a long tail amongst the 34 countries represented: UK: 21 (17.36%); USA, 15 (12.40%); Canada, Italy 8 (6.61%); Netherlands, Switzerland: 7 (5.79%); Germany: 5 (4.13%); Australia, France, Indonesia, Spain, Sweden: 4 (3.31%); Austria: 3 (2.48%); Brazil, Hungary, Lithuania, Norway, Romania: 2 (1.65%); Burkina Faso, China, Ethiopia, Ghana, India, Ireland, Malaysia, Montenegro, New Zealand, Poland, Portugal, Russia, Slovakia, Slovenia, Taiwan, Tanzania, Uganda: 1 (0.83%).

The majority were male (83, 68.60%, vs. 38, 31.40% female, with 0% non-binary or other). 48 (39.67%) were professors, 44 (36.36%) senior lecturer or associate professor, 19 (15.70%) lecturer or assistant professor, 5 (4.13%) postdocs, 3 (2.48%) doctoral students, and 2 (1.65%) 'other'.

Regarding disciplinary coverage, the most prevalent disciplines tended to be from natural sciences, engineering or medicine. The full list was: Health: 15 (12.40%); Biology: 13 (10.74%); Engineering: 11 (9.09%); Computer Science: 9 (7.44%); Environmental Science: 8 (6.61%); Life Sciences, Medicine: 7 (5.79%); Education: 6 (4.96%); Physics: 4 (3.31%); Agriculture, Business, Economics, Linguistics, Psychology: 3 (2.48%); Astronomy, Chemistry, Climate, Materials Science, Statistics: 2 (1.65%); Agroforestry, Demography, Digital Humanities, Ethics, Film Studies, Geography, Health Science, History, Management Information Systems, Mathematics, Neuroscience, Philosophy, Public Health, Science Studies, Sociology, Sport, Transportation Systems: 1 (0.83%).

4.2 Overall attitudes to research assessment criteria

We first report on participants' responses to the question: 'How do you feel overall about the relevant research assessment indicators used in promotion processes at your institution?'. One hundred and eleven respondents provided an answer to this optional, free-text question (excluding non-answers like 'N/A').

4.2.1 Overall sentiments

Many respondents answered in such a way that their overall opinion on the subject could be coded for sentiment, either negative or positive, based on expressions of emotion (e.g. 'I feel frustrated' or 'quite happy') or evaluative statements (e.g. [processes are] 'totally biased' or 'fair enough'). Slightly more negative ($n = 35$) than positive ($n = 31$) responses were received (see Supplementary File S1).

Positive statements tended to be mild in their approval with statements like '[f]air', 'fair enough', 'OK' (including 'generally OK' and 'overall OK'), 'fine', '[s]atisfactory', 'generally suitable' and 'reasonable'. Some, however, were more positive with statements like '[g]ood', 'good and balanced' and 'quite happy' and 'I agree with all indicators'. Two of the responses indicating an overall positive sentiment mentioned a possible bias given their involvement in designing the assessment criteria: 'I feel it's quite ok. However, I may be biased, since I'm involved myself in deciding on the indicators we use'. and 'I think they are fair, but I am biased since I designed them'. Negative sentiments more frequently included strongly-worded statements. Among the 35 responses that we interpreted as clearly negative, we identified at least 15 as 'strongly negative' based on the language they used (see Supplementary File S1). These included terms such as 'totally biased', '[b]ad', 'mere hocus-pocus', '[t]wo-faced', 'not fair at all', 'a mess', and even 'almost bullshit'. In addition, respondents also reported emotions including feeling 'disappointed', 'very sad (I would say also demotivated, yet, I do research because I like it)', 'feeling overwhelmed', 'I feel frustrated', and '[v]ery uncomfortable'.

4.2.2 Over-reliance on quantitative metrics

By far the most commonly-mentioned factor for respondents' disappointment with research assessment was the use of quantitative indicators, mentioned by 27 respondents. In general, these respondents express well-documented critiques of assessment procedures' overreliance on metrics and tendency for being too 'number-focused'. One respondent even indicated that their institution's criteria were 'Utterly focused on metrics and quantity'. Another advised theirs were 'Mechanistic to the detriment of quality assessment. Largely a game of numbers'.

Such sentiments were related to a generally-stated feeling that metrics were ill-understood and misapplied: 'I feel that we are not highly devoted in learning about scientometrics and related fields and that we sometimes just apply some things that we do not understand well—I mean—I feel frustrated'. More optimistically on this point, however, some were aware of initiatives for responsible metrics in research assessment and specifically lauded developments towards changing metrics. One respondent from a university in eastern Europe saw a need to 'follow European trends in this area (e.g. asking for publications such as those in WoS, Scopus etc)' but also to acknowledge that the role of metrics differs across disciplines. This disciplinary distinction was mentioned by another respondent who stated that assessment in their institution was 'more balanced than in other institutions' since it acknowledged that 'not all fields of endeavour attract large funding opportunities' and that the absence of funded research was not a 'dealbreaker for promotion'. Respondents furthermore specifically mentioned that 'The Leiden Manifesto could serve as one of the crucial guidelines within this question'. Another respondent noted: 'It is better since DORA is being used. I don't think research assessment indicators can be considered in isolation from the other demands'. Similar awareness of, and support for, change was visible in the responses of others, including in responses reflecting a positive sentiment. For example, one respondent who mentioned that criteria were 'Largely standard and overall OK' also noted that 'criteria are currently shifting (e.g. in direction of OS, away from impact-based metrics), which I find positive', and one respondent who considered that 'For the most part they [assessment indicators] are fair as this is our profession. teaching, research and academic administration' still praised their university for having 'an open mind' and being 'proactive to advance and improve the evaluation by including slowly but surely all of those new considerations'.

Related to an overemphasis on metrics, many respondents complained about thoughtless counting of research funding obtained. In the words of one UK Lecturer, 'In my experience at my institution, the most important indicator is funding. Have you secured enough funding? What is your potential to secure funding? […] This tops any indicator above papers, where they are published or how many citations they have. Money is king'. Another respondent argued: 'It is more about getting funds than what you actually do with it'. Respondents who were generally positive regarding current assessment indicators also expressed this point, for example one respondent mentioned that indicators were 'Tending a bit toward metrics, but still emphasizing overall quality'. possibly indicating an agreement with the metrics used.

4.2.3 Evaluative flexibility

Another topic brought forward by a substantial share of respondents indicated that they perceived a disconnect at their institution between the officially-stated policies and the ways in which they were applied in practice. For example, one stated that some criteria 'matter officially but not in reality'. Another said 'I agree with the indicators that are SAID to be used; I'm not sure they are really used, though. Or only when it suits the evaluation committees'. In this context, some respondents raised an issue to which we return in our next subsection, that political positioning within the institution, and especially good relationships with heads of department or faculty, are often of great importance, despite the official criteria. A full professor from a Hungarian university said that '[u]sually, personal assessments are more important than numbers estimating the research quantitatively and qualitatively'. Two further examples from senior researchers from Romania and the USA, respectively, show the prevalence of this sentiment: 'There are several other aspects important to be promoted in our Academic institution such as being involved in politics, having relative or friend with influence in the management of the institution and so on' and 'My personal experience: I do not know that they [the indicators] are of importance unless promoted by your section chief'. Also, one senior researcher from the UK noted that 'I believe that the people who make the decisions on research promotions base them on the person they want to promote, rather than any actual metrics. They can justify any promotion based on any metric they want'.

Here the flexibility of research assessment criteria seems at issue, which comes in opposition to the longing for less quantitative criteria and more diverse indicators as described in the previous point. In fact, even if respondents valued assessments that are adaptive for discipline, fields, and team size, and even if they worried about quantitative assessment which 'objectivizes a comparison, but loses the human in the loop', the flexibility in how assessment approaches are implemented also raised substantial worry and frustration. One respondent advised that at their institution, criteria were 'used selectively for different people'. Another said: 'They can justify any promotion based on any metric they want'. In this sense, respondents appear to show that personal preferences, lack of transparency and personal biases seemed to occur despite the prominent place of metrics that was criticized above. Here, it is natural that good working-relationships with one's colleagues may be a factor in deciding who should progress. However, the fact that gatekeepers like heads of department or faculty hold such singular power might be taken to be of concern, especially since these people will often come from already privileged demographics and may (based on the principle of homophily) be more responsive to people like them in their personal relationships. One respondent, a female associate professor from a UK university, spoke directly to this in her assessment that processes at her institution are 'totally biased towards white men'.

4.2.4 Clarity of criteria

Related to the concern of criteria being used selectively or differently according to gatekeepers' personal preferences, several respondents reported that the assessment criteria or promotion guidelines at their institutions were unclear, intransparent or insufficiently communicated. Some respondents directly linked this lack of transparency to the allowance of the flexibility mentioned above, for example when commenting that criteria 'could be clearer', were 'not transparent', and even 'vague to the extent of obscurity' or 'intentionally vague'. Some respondents moreover highlighted the lack of any

criteria, such as one mid-career researcher from Romania, who advised 'The university does not have any internal policy for promotion. The promotions are at the free will of the dean'. Another respondent from Germany advised that since opportunities for internal promotion were scarce, promotion to Professorship entailed application to another institution, where 'the criteria depend on the preferences of the hiring committee'. Other respondents merely mentioned that criteria or guidelines were unclear, either in general: 'Okay but they could be clearer to faculty and better communicated', 'Guidelines are not entirely clear', 'They are not transparent', or (in the case of one) only in local languages and so inaccessible for non-native speakers.

In general, respondents mainly expressed a critical attitude towards unclear criteria: 'They are supposedly clear but in reality you need 'to do your time' (e.g an H index that needs time, not just number of publications and citations) and the reasons for non-promotion can be nebulous.' On the contrary, however, one respondent noted that the dominance of quantitative criteria at their institution meant that they were very clear, but that this itself was to the detriment of assessment of what matters: 'Mechanistic to the detriment of quality assessment. Largely a game of numbers.' This points to the fine balance between using rigid indicators that are very clear and transparent, or softer, more contextual or interpretative criteria that run the risk of creating leeway for gatekeepers to introduce favouritism or bias. Indeed, the respondent mentioning that criteria were 'vague to the extent of obscurity' also noted that '[t]hat is good in that quality can dominate numerics in decisions. But it's also intransparent for the candidate'.

4.2.4 Other factors

In addition, smaller numbers of respondents identified other factors as important. Firstly, linked to over-quantification (and misuse of the Journal Impact Factor especially), respondents thought that the venue of publication (e.g. journal title) took too much precedence over research quality *per se*. Others identified lack of emphasis on Open Science, e.g. 'I wish they emphasised open science more'. Finally, some found societal impact undervalued, although one UK respondent noted the positive influence of the Research Excellence Framework assessment exercise in this regard.

4.3 Perceptions of hidden factors in research (er) assessment

We next asked respondents 'Are there other factors (social, political, performance-based, etc) that are not officially used as promotion criteria, but that you nonetheless believe are important to getting promoted at your institution?'. Around a third of respondents (n = 68) answered this question, while 13 additional respondents answered 'N/A', 'no', 'don't know', or referred to previous answers (e.g. 'See comment above'), or some variant. These answers are hence excluded (see Supplementary File S1 for all responses).

4.3.1 Social and political connections

27 respondents highlighted political and social factors. The last section showed that respondents often saw a great degree of flexibility in the application of criteria, with social and political factors such as supportive networks and connections often influencing RPT decisions. This came across strongly in answers to this question, ranging from relatively benign acknowledgements that '[n]etworking within faculty' is important or that a 'candidate's existing and positive history in the institution can be an advantage', to stronger affirmations that '[p]ersonal relationships and endorsements play an important role', especially relationships with line-managers and heads of department or faculty 'who may lobby for you' (respondents from Australia, Norway and Italy respectively). One UK respondent framed this in terms of '[u]nderstanding the organisation politics, knowing the gatekeepers and the landscape'. Even when expressed in a more-or-less neutral way, there was acknowledgement from another UK researcher that such factors are often only unofficially applied: 'It is obvious that it is quite important to be well-connected with the important people involved in the decision making process. This is not officially used of course, but practically it is'.

While one New Zealand-based respondent framed this in terms of collegiality ('One has to get on with one's colleagues and be a team player to some extent'), most respondents were more negative. One respondent in France added that 'issues linked to conflict of interest are not addressed adequately enough in the evaluation committees'. This sentiment can also be observed in the responses of others. One respondent in Russia commented that 'loyalty to management' played a large role in promotions. Others, from Italy and France respectively, used pejorative words such as '[i]nbreeding' and 'cronyism' and linked this to perceived unjust outcomes. In the words of one mid-stage UK researcher: 'To be promoted, you have to be on the team of the person making the decision' adding that this led to 'lots of promotions to people who don't deserve them but are friends with the Director of Research'.

Many framed such sentiments in terms of internal politics and believed it had a corrosive influence. For instance, one respondent from the USA advised, 'Political factors are hugely important for promotion. That is why this institute is corrupted and many faculty members have been leaving'. Others, such as one respondent from Indonesia, framed it in terms of personal like and dislike or a need for belonging to the 'exclusive group'. This was seen by some as potentially detrimental to researchers working with less well-connected teams, with a UK-based researcher stating that 'you can only be promoted with a recommendation from your line-manager … but some line-managers hold more power than others'. Along the same lines, conducting 'research in the areas closest to the most important/powerful professors in the department' was also thought by a Netherlands-based respondent to help strengthen connections and advantage.

Adding to this perspective, two respondents, from Indonesia and Spain mentioned that local people are preferred to those from outside the region. In the words of one Spain-based Professor, it '[h]elps to be one of the locals' as there were '[s]till relatively few permanent staff from outside the region, neither nationally nor internationally'.

In addition, political affiliations were mentioned by two participants. The first, a researcher from a research-intensive university in the UK, said:

> Who are friends with who can have some influence and indeed some senior people may be affiliated to the same political party and can get footing to negotiate behind the scenes. This does appear to happen in some cases that is ultimately corrupt. (Senior Lecturer/Assoc. Prof, UK, Engineering, male)

The second, a researcher in Asia, advised that their country:

> is a very political environment, so things might be tough if you were from [REDACTED], and there is additionally a strong North-South divide in the country in terms of social and political attitudes. Oddly I suspect in some areas religion might be a factor in promotion within research groups.[1]

Opposing these statements, however, one UK postdoc advised that '[e]xternal applicants are preferred over promoting internal candidates'.

4.3.2 Research trends

Seven respondents cited the importance of research-focus and how this relates to current trends as a factor influencing promotions. One expressed that 'balanced research interests and diversity of topics' are an advantage. However, others advised that particular specializations were advisable. One advised: 'Definitely: there are two subfields outside of which it is almost impossible to get a promotion (and I'm not in these fields, of course …)', while another said that 'My impression is: Young scientists acting beyond certain mainstream fields have lower likelihood of getting promoted'. Interestingly, adherence to disciplinary-norms regarding research methods seemed important, as two researchers from opposite sides of the qualitative-quantitative divide in fields where those approaches are the norm had similar complaints:

> [R]esearch that does not validate held perspectives of trends in education are not valued. Qualitative research has preeminence in Education. (Australia, Professor, Health, male)

> There is a bias towards quantitative research in medical/health research rather than increasing our understanding of the nature of a problem via qualitative methods. "Women's research" (eg working with domestic violence), when framed as health inequalities is more likely to be understood and valued from my experience. (UK, Lecturer/Asst. Prof, Health, female)

4.3.3 Diversity and discrimination

Seven respondents also reflected on the role of gender and diversity criteria in promotion decisions. Within our sample, this was most heavily related to gender discrimination, although one pointed to 'discrimination against age and mental illness'. One advised thinking of gender equality as an important issue but was unaware if it was taken into account at their institution. Another advised '[t]here is rightly a push to get more under-represented groups into positions of leadership'. Two respondents reflected upon underlying structural reasons for inequity and argued that merely adjusting criteria will not compensate for this. The first (a male) pointed to how gender differences impact publication patterns, with women more likely to 'do things themselves' and therefore produce fewer papers with fewer co-authors, while the second (a female) extended the discussion to other characteristics by arguing that:

> It is not about the criteria, but opportunities to have produced research work. Disadvantaged groups are not given fair opportunities. The advantaged group, ie white are given opportunities for promotion such as fellowship, grant collaboration, and a permanent post. (UK, Senior Lecturer/Assoc. Prof., Health, female)

This latter contribution makes clear that promotion criteria are merely one barrier amongst many in fostering equity in research careers. Such structural factors, in addition to plain sexism and biases, perhaps underlie the pessimistic assessment of one female professor:

> In the end, my field and my institution are governed by (white) males. In spite of everything that is tried, or said, women are simply overlooked. I am afraid this especially applies to older women. The combi[nation] of ageism and sexism is maybe not additive, but even more harmful. (Netherlands, Professor, Life Sciences, female)

The agenda for change in this regard was not uniformly appreciated. As one male respondent advised: 'It seems that heterosexual males who are either unmarried or without children are actively discriminated against'. Another male respondent stated that '[t]here is rightly a push to get more under-represented groups into positions of leadership. However, this should not impact on those who are not under-represented. If two candidates are equally qualified and satisfy the criteria they should both be promoted'.

5. Discussion

Our paper describes results from two optional open text responses in a survey that examined participants' perceptions of review, promotion and tenure criteria and practices (cf, Ross-Hellauer et al. 2023). Out of the 198 respondents that completed the full survey, 121 responded to at least one of the two open text responses analysed in the present paper. Taken together, our findings indicate that, even though some of those who responded to the open text responses expressed an overall satisfaction with the current RPT criteria, several also expressed a general sense of frustration or discomfort with the criteria, or their application, at their institutions. This discomfort largely targets a perceived over-emphasis on quantification, with a focus on the number of publications and the amount of funding obtained. In particular, respondents mentioned that this focus on quantity came at the detriment of considerations of quality and broader research impact. This perspective aligns with the overarching concerns voiced by diverse stakeholders in discussions on research assessment that have been ongoing for over a decade (DORA 2012; Hicks et al. 2015; Wilsdon et al. 2015). However, we also found that for at least a few respondents, the perception was that the situation had improved somewhat in recent years, with some explicitly crediting DORA.

Righting the ship of over-quantification is a broad and ongoing project, which in recent years has become linked to a recognition that for structures of assessment to enable rather than obstruct embedding of open and responsible research, a much broader array of activities must be valued beyond brute numbers of publications or funding (Rushforth and Hammarfelt 2023). As we saw in our introduction, doing so requires deep examination of not only the 'epistemic agency' (Müller and de Rijcke 2017) that such metrics have acquired in shaping research behaviours and cultures, but also their effects of structuration upon current experiential realities of

research and what Felt (2017) has termed 'the tempor(e)alities of academia'.

To address these issues, current discussions on research assessment propose that we replace, or at least heavily complement, narrowly defined, largely quantitative indicators with broader indicators that, in the words of CoARA's first 'core commitment', 'recognise the diversity of contributions to, and careers in, research in accordance with the needs and nature of the research' (CoARA 2022: 4). In addition, research assessment 'should rely on qualitative judgement for which peer review is central, supported by responsibly used quantitative indicators where appropriate' (*ibid*., 3). Metrics are out of favour, and only mentioned as a complement to qualitative and narrative-driven indicators that enable evaluated individuals to explain how their contributions move their research, community, or research environments forward. While this change of approach opens new possibilities in recognizing the breadth and diversity of impact in researchers' careers, moving to this new approach cannot happen without caution. On the one hand, the weaknesses of peer review (Sivertsen and Rushforth 2023)—which are often missing from current discussions—need to be better understood, communicated, and addressed wherever possible. Indeed, it is well known that this process itself is prone to several biases and limitations. On the other hand, greater reliance on qualitative judgement will give greater room for 'evaluative flexibility' amongst assessors. We recognize that such flexibility can be valuable, even necessary, to identify legitimate and crucial elements that are missed by narrow metrics. However, we also recognise that greater flexibility may open the door to greater risk of bias and questionable decisions, just as an overreliance on quantitative indicators may lead to overly simplistic evaluations that fail to capture the nuanced and multifaceted nature of research and its impacts.

Our respondents expressed concerns about a mismatch between formal criteria and actual practices in current evaluation processes. In particular, respondents identified several 'hidden' criteria that are not formally listed as part of assessment processes, but are perceived to play an important role in the outcome of research assessments. Most prominently, these include social, political, and demographic factors, where people from certain backgrounds or with certain social connections are perceived to be favoured over others. Some respondents mentioned strong social connections and supportive networks as important elements in research assessment for career progression within their institution. Inasmuch as such factors are reflective of collegiality, defined by Cipriano and Buller (2012) as the relationships, respect, collaborative-spirit, common purpose and equitable distribution of responsibilities that support good-functioning of the research unit, we may see no problem. Yet, many of our respondents clearly perceived detrimental effects from this hidden hand of personal connection, extending far beyond collegiality towards 'cronyism', 'inbreeding' and unfair decisions based on who was friends with whom. While 'inbreeding', progression to faculty of graduates from the home institution, may be functional in fostering cohesion, stability and capacity-building in developing institutions (Horta and Yudkevich 2016; Balyer and Bakay 2022), broader issues of favouritism, cronyism and unfair promotions are listed by Osipian (2009) as examples of corruption in Higher Education. Inbreeding does seem to be associated with lower levels of diversity (Balyer and Bakay 2022) and productivity (Inanc and Tuncer 2011), and cronyism has been also been found to generally undermine social capital through mechanisms of ostracism within research institutions (Jawahar et al. 2021). Our respondents also noted experiences of discrimination based on demographics of gender, race and age. This finding aligns with the literature in this area which shows that discrimination persist even in contexts where strong anti-discrimination policies and legislation exist (Carr et al. 2000; Moss-Racusin et al. 2012).

Unclear or intransparent criteria were also flagged as a major issue by some in our sample, the implications of which merit further attention. Beyond unfair disadvantages and biases in the decision process, our respondents flagged potential psychological and social consequences (e.g. feelings of insecurity, the perceived need to network with powerful gatekeepers, or the perception that processes are 'biased against underrepresented groups') that could harm those assessed and are likely to affect some groups more than others. For example, it seems likely that if criteria are unclear and people perceive 'hidden' social and political factors to play a crucial role in promotion and tenure decisions, this might have profound implications for social dynamics within research groups, including risking the enforcement of traditional power hierarchies (Horbach et al. 2020; Ylijoki 2022). 'Hidden' assessment criteria may also leave researchers with perceived 'distributive or procedural injustice' which may in turn disrupt the collegiality and cultures in place and even impact research practices (Martinson et al. 2006, 2010).

We hence note a tension here between respondents' dissatisfaction with rigidly- and often narrowly-defined criteria that do not do justice to the full range of valuable academic activities and competences on the one hand, and the non-explicit, 'hidden', and loosely-defined criteria on the other. Yet criteria, however diverse or flexible, will remain constitutive of that which is to be measured, and entail mediation and interpretation (Tsoukas 1997); meaning that perspectives will to some extent remain partial, in both senses of that word. With no 'Olympian high ground' of full transparency from which to stand, assessors must still make quality judgments based on imperfect proxies contextualized 'on-the-fly' (Kaltenbrunner and de Rijcke 2019). In a world where criteria are 'used selectively for different people' and assessors 'can justify any promotion based on any metric they want' (to quote two of our respondents), increased diversity of criteria and use of qualitative assessment may increase the scope of 'evaluative flexibility' on behalf of the assessor, and potentially the scope for biases or the hidden-hand of personal connections to play out.

The tension identified here extends beyond criteria merely being transparent or opaque. In the context of gender discrimination, for example, some respondents reflected on the extent to which criteria themselves can be biased. As discussed above, transparency of criteria can itself actually be used to legitimize biases when evaluative practices misalign with prescriptive norms concretized as weak policy, so that transparent criteria may act as cover for biased evaluative practices to persist (Strathern 2000; van den Brink, Benschop and Jansen 2010). However, while it seems biases and personal preferences can persist even where officially legislated against, criteria that are 'vague to the point of obscurity' (quoting another respondent) clearly facilitate evaluative flexibility. Hence, the current discourse must not only avoid becoming an oversimplified dichotomy of quantitative vs.

qualitative, or rigid vs. flexible criteria, but rather seek to clearly account for the ways in which lack of clarity and flexibility can enable social and political biases, and identify ways to mitigate or avoid these effects.

Ultimately, any model of research assessment remains a human process, and hence at risk of bias. What respondents seem to be implicitly longing for is a sense of objectivity in research assessments. This has commonly been understood as a sense of what Daston (1992) refers to as 'mechanical objectivity', seemingly commensurable with the protocolized use of rigid criteria and metrics. However, it is probably more fruitful to seek a different kind of objectivity in this inherently human and hence subjective process. As per Longino (1990) and Daston and Galison (2007), objectivity in review and assessment processes could be conceptualized as being in agreement with community standards. Acknowledging the diversity of norms and values across diverse academic contexts (Lamont 2009), fair assessments then arise, in the words of Mallard, Lamont and Guetzkow (2009, 576), 'when they use standards that are most appropriate to the object of evaluation. Rather than applying a single universal criterion indiscriminately, they specify which criteria, or lenses, are most appropriate to assess the strengths and weaknesses of the object under evaluation'. This perspective directs attention from asking what assessment criteria or processes are to be used, to asking why the assessment is conducted, how such tacit community-driven criteria and processes are negotiated, who gets to decide on them, and what social and political hierarchies matter in providing closure to such processes.

6. Conclusion

Our results testify to core tensions in the ongoing quest to find the most appropriate ways of assessing research and researchers. While discussions seem perennially at risk of getting bogged down in unproductive dichotomization of qualitative vs. quantitative modes of assessment, our study further problematizes characteristics of assessment procedures, hitherto less frequently discussed. These include the evaluative flexibility on the side of the assessors and the extent to which assessment criteria can and should be transparent.

Whatever system of assessment, based on whatever set of assessment criteria, our respondents signal a need to 'walk the talk'. A misalignment between formal criteria and actual practices was flagged as a major source of frustration and insecurity, leaving respondents with feelings of unfairness. This also underlines that, whatever set of criteria is deemed preferable, only changing criteria without changing assessment practices accordingly is meaningless. In fact, through creating more room for evaluative flexibility, this might exacerbate perceptions of unfairness and inequality.

Our study provides empirical evidence on changes in research assessments and reveals a tension between stated principles of research assessment and how these principles are employed in practice. Our findings therefore contribute to the ongoing debates on how research and researchers should be assessed, shedding light on emerging assessment practices and providing various actions to foster responsible and fair research(er) assessment procedures. While we believe that the recommendations from our findings may be applicable to a diversity of settings, we recognize that constraints in our sample and data collection methodology may have limited the breadth of perspectives captured. The context from which these recommendations were generated and possible gaps and limitations should therefore be considered.

First, while it provides a broad diversity of insights from researchers from a wide variety of academic contexts and disciplines, our study is subject to several limitations, including those related to the kind of data we collected. Our findings come from free-text responses collected as part of a broader survey. Since responses on the two questions included in this paper were not mandatory, it is possible that respondents who disagreed with current research assessment were more likely to respond than respondents who agreed with current research assessment. For this reason, quantitative details of responses should be considered carefully. Free-text responses also allowed for no follow-up to request elaboration or clarification, and subtle clarifications may have been missed in the short answers captured. Similarly, we considered only the views of those being assessed, not those responsible for assessment (although in many cases, especially for senior staff, experience of both roles is possible). This last point is especially important since narrative and qualitative approaches to research assessment increase expectations of responsibility and fairness from assessors. Their perspective should be addressed in future research.

In addition to these, we should mention limitations related to the size and composition of our sample that limit the generalisability of our findings. First, our sample was relatively small, with a low response rate and uneven geographical and demographic coverage. In particular, our sample contained an overrepresentation of scholars from Europe and the USA, as well as of male and senior STEM researchers. The limited number of responses does not allow for detailed analysis of differences across the very different contexts and backgrounds of respondents. The rapidly changing assessment contexts in different settings may have had an important impact on responses we cannot account for. Second, our sample is likely to be subject to self-selection bias, potentially overattracting scholars that are particularly satisfied or dissatisfied with the evaluation processes they are subjected to. This applies to both the overall sample of respondents to our survey and to the subset of respondents providing input to the optional open-ended questions analysed in this article. While we hence do not claim generalisability of our findings, we still believe our results add valuable context to the current debates about research assessment.

Looking forward, our results suggest various actions to foster responsible and fair research(er) assessment procedures. As a first step, it is crucial to question and detail the dimensions, concepts, and purposes that form the core of research assessment. Creating a shared understanding and a common reference framework—to the extent possible—between assessors and those assessed should be a priority, including replacing 'slogan terms' which lack fixed meaning (Hatch 2019) such as 'excellence', 'innovation', and 'impact' by concrete definitions (cf, Moore et al. 2017). This shared understanding could also be promoted through training of assessors covering, for example, ways of preventing bias, fostering metric-literacy, and transparently communicating criteria and motivations for decisions. It is also essential to acknowledge that reflective and responsible research assessment takes time and resources. For qualitative assessment to be sustainable, research communities will need to be given time and recognition for high quality assessment and may need to reduce the frequency at which researchers are

assessed, including through the introduction of longer-term grants and more secure research contracts. More research is also needed to identify the elements at play in research assessment, including from the perspective of research assessors. This should particularly expand our understanding about the inherent biases at play in research assessment processes, ways of identifying these biases and subsequently ways of mitigating their consequences if deemed undesirable. This requires moving beyond the dichotomy of qualitative and quantitative ways of assessment, and acknowledging that transparency alone is insufficient to address biases.

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Funding

None declared.

Note

1. Further demographic information withheld to avoid risk of reidentification given political sensitivity of topic.

References

Acker, S., and Webber, M. (2016) 'Discipline and Publish: The Tenure Review Process in Ontario Universities', in L. Shultz and M. Viczko (eds) *Assembling and Governing the Higher Education Institution: Democracy, Social Justice and Leadership in Global Higher Education*, pp. 233–55, Palgrave Studies in Global Citizenship Education and Democracy. London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-52261-0_13

Alperin, J. P. et al. (2019) 'How Significant Are the Public Dimensions of Faculty Work in Review, Promotion and Tenure Documents?', *eLife*, 8: e42254. https://doi.org/10.7554/eLife.42254

Alperin, J. P. et al. (2022) 'The Value of Data and Other Non-Traditional Scholarly Outputs in Academic Review, Promotion, and Tenure in Canada and the United States'. in A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister (eds) *The Open Handbook of Linguistic Data Management*, pp. 171–82. The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0017

Altbach, P. G., Yudkevich, M., and Rumbley, L. E. (2015) 'Academic Inbreeding: Local Challenge, Global Problem', *Asia Pacific Education Review*, 16: 317–30. https://doi.org/10.1007/s12564-015-9391-8

Balyer, A., and Bakay, M. (2022) 'Academic Inbreeding: A Risk or Benefit for Universities?', *Journal of Education and Learning*, 11: 147. https://doi.org/10.5539/jel.v11n1p147

Bornmann, L., Mutz, R., and Daniel, H.-D. (2007) 'Gender Differences in Grant Peer Review: A Meta-Analysis', *Journal of Informetrics, The Hirsch Index*, 1: 226–38. https://doi.org/10.1016/j.joi.2007.03.001

Broucker, B., and De Wit, K. (2015) 'New Public Management in Higher Education'. In *The Palgrave International Handbook of Higher Education Policy and Governance*, edited by Jeroen Huisman, Harry de Boer, David D. Dill, and Manuel Souto-Otero, 57–75. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-137-45617-5_4

Burrows, R. (2012) 'Living with the H-Index? Metric Assemblages in the Contemporary Academy', *The Sociological Review*, 60: 355–72. https://doi.org/10.1111/j.1467-954X.2012.02077.x

Carr, P. L. et al. (2000) 'Faculty Perceptions of Gender Discrimination and Sexual Harassment in Academic Medicine', *Annals of Internal Medicine*, 132: 889–96. https://doi.org/10.7326/0003-4819-132-11-200006060-00007

Charlton, B. G., and Andras, P. (2007) 'Evaluating Universities Using Simple Scientometric Research-Output Metrics: Total Citation Counts per University for a Retrospective Seven-Year Rolling Sample', *Science and Public Policy*, 34: 555–63. https://doi.org/10.3152/030234207X254413

Chawla, D. S. (2021) 'Scientists at Odds on Utrecht University Reforms to Hiring and Promotion Criteria'. *Nature Index*. https://www.nature.com/nature-index/news-blog/scientists-argue-over-use-of-impact-factors-for-evaluating-research, accessed 10 June 2024.

Cipriano, R. E., and Buller, J. L. (2012) 'Rating Faculty Collegiality', *Change: The Magazine of Higher Learning*, 44: 45–8. https://doi.org/10.1080/00091383.2012.655219

CLACSO-FOLEC. (2022) *A New Research Assessment Towards A Socially Relevant Science In Latin America And The Caribbean*. Mexico City, Mexico: Latin American Council of Social Sciences (CLACSO). https://biblioteca-repositorio.clacso.edu.ar/bitstream/CLACSO/169747/1/Declaration-of-Principes.pdf, accessed 10 June 2024.

CoARA. (2022) 'Agreement on Reforming Research Assessment'. https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf, accessed 10 June 2024.

Curry, S. (2018) 'Let's Move beyond the Rhetoric: It's Time to Change How We Judge Research', *Nature*, 554: 147. https://doi.org/10.1038/d41586-018-01642-w

Daston, L. (1992) 'Objectivity and the Escape from Perspective', *Social Studies of Science*, 22: 597–618.

Daston, L., and Galison, P. (2007) *Objectivity*. Princeton University Press.

Delgado, R.-M., Tarango, J., and Machin-Mastromatteo, J. D. (2020) 'Scientific Evaluation Models in Latin America and the Criteria for Assessing Researchers', *Information Development*, 36: 457–67. https://doi.org/10.1177/0266666920943966

Diamantes, T. (2005) 'Online Survey Research of Faculty Attitudes Toward Promotion and Tenure', *Essays in Education*, 12: 1–12. https://openriver.winona.edu/eie/vol12/iss1/3

DORA. (2012) *San Francisco Declaration on Research Assessment*. DORA. https://sfdora.org/

East, R. (2016) 'Bias in the Evaluation of Research Methods', *Marketing Theory*, 16: 219–31. (https://doi.org/10.1177/1470593115609797.

EUA. (2022) *The EUA Open Science Agenda 2025*. Brussels: European University Association. https://eua.eu/downloads/publications/eua%20os%20agenda.pdf

Felt, U. (2017) 'Under the Shadow of Time: Where Indicators and Academic Values Meet', *Engaging Science, Technology, and Society*, 3: 53–63. https://doi.org/10.17351/ests2017.109

Ginther, D. K. et al. (2011) 'Race, Ethnicity, and NIH Research Awards', *Science*, 333: 1015–9. https://doi.org/10.1126/science.1196783.

Hammarfelt, B., and Rushforth, A. D. (2017) 'Indicators as Judgment Devices: An Empirical Study of Citizen Bibliometrics in Research Evaluation', *Research Evaluation*, 26: 169–80. https://doi.org/10.1093/reseval/rvx018

Hammerschmid, G. et al. (2013) 'Trends and Impact of Public Administration Reforms in Europe: Views and Experiences from Senior Public Sector Executives'. COCOPS Policy Brief, no. 4. https://lirias.kuleuven.be/1867383

Harley, D. et al. (2010) *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines*. Center for Studies in Higher Education.

Hatch, A. (2019) 'To Fix Research Assessment, Swap Slogans for Definitions', *Nature*, 576: 9. https://doi.org/10.1038/d41586-019-03696-w

Hatch, A., and Schmidt, R. (2019) *Rethinking Research Assessment: Unintended Cognitive and System Biases*. DORA. https://sfdora.org/wp-content/uploads/2020/09/DORA_

UnintendendedCognitiveSystem Biases.pdf, accessed 10 June 2024.
Helgesson, K. S., and Sjögren, E. (2019) 'No Finish Line: How Formalization of Academic Assessment Can Undermine Clarity and Increase Secrecy', *Gender, Work & Organization*, 26: 558–81. https://doi.org/10.1111/gwao.12355
Hessels, L. K. et al. (2019) 'Variation in Valuation: How Research Groups Accumulate Credibility in Four Epistemic Cultures', *Minerva*, 57: 127–49. https://doi.org/10.1007/s11024-018-09366-x
Hicks, D. et al. (2015) 'Bibliometrics: The Leiden Manifesto for Research Metrics', *Nature*, 520: 429–31. https://doi.org/10.1038/520429a
Hom, H. L., Jr., and Van Nuland, A. L. (2019) 'Evaluating Scientific Research: Belief, Hindsight Bias, Ethics, and Research Evaluation', *Applied Cognitive Psychology*, 33: 675–81. https://doi.org/10.1002/acp.3519
Horbach, S. P. J. M. et al. (2020) 'On the Willingness to Report and the Consequences of Reporting Research Misconduct: The Role of Power Relations', *Science and Engineering Ethics*, 26: 1595–623. https://doi.org/10.1007/s11948-020-00202-8
Horta, H., and Yudkevich, M. (2016) 'The Role of Academic Inbreeding in Developing Higher Education Systems: Challenges and Possible Solutions', *Technological Forecasting and Social Change*, 113: 363–72. https://doi.org/10.1016/j.techfore.2015.06.039
Inanc, O., and Tuncer, O. (2011) 'The Effect of Academic Inbreeding on Scientific Effectiveness', *Scientometrics*, 88: 885–98. https://doi.org/10.1007/s11192-011-0415-9
Islam, G., and Greenwood, M. (2022) 'The Metrics of Ethics and the Ethics of Metrics', *Journal of Business Ethics*, 175: 1–5. https://doi.org/10.1007/s10551-021-05004-x
Jappelli, T., Nappi, C. A., and Torrini, R. (2017) 'Gender Effects in Research Evaluation', *Research Policy*, 46: 911–24. https://doi.org/10.1016/j.respol.2017.03.002
Jawahar, I. M. et al. (2021) 'Does Organizational Cronyism Undermine Social Capital? Testing the Mediating Role of Workplace Ostracism and the Moderating Role of Workplace Incivility', *Career Development International*, 26: 657–77. https://doi.org/10.1108/CDI-09-2020-0228
Juárez Ramos, V. (2019) *Analyzing the Role of Cognitive Biases in the Decision-Making Process*. IGI Global. https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/978-1-5225-2978-1. https://www.igi-global.com/gateway/book/179223
Kaltenbrunner, W., and de Rijcke, S. (2019) 'Filling in the Gaps: The Interpretation of Curricula Vitae in Peer Review', *Social Studies of Science*, 49: 863–83. https://doi.org/10.1177/0306312719864164
King, C. J. et al. (2006) 'Scholarly Communication: Academic Values and Sustainable Models', 126, Berkeley, CA: University of California, Berkeley.
Knoth, P., and Zdrahal, Z. (2012) 'CORE: Three Access Levels to Underpin Open Access', *D-Lib Magazine*, 18 Nov. 2012. https://doi.org/10.1045/november2012-knoth
Krüger, A. K., and Reinhart, M. (2017) 'Theorien Der Valuierung—Bausteine Zur Konzeptualisierung Von Valuierung Zwischen Praxis Und StrukturTheories of Valuation—Building Blocks for Conceptualizing Valuation between Practice and Structure', *Historical Social Research/Historische Sozialforschung*, 42: 1. https://doi.org/10.12759/HSR.42.2017.1.263-285
Lamont, M. (2009) *How Professors Think. Inside the Curious World of Academic Judgement*. Cambridge, MA: Harvard University Press.
Larivière, V. et al. (2013) 'Bibliometrics: Global Gender Disparities in Science', *Nature*, 504: 211–3. https://doi.org/10.1038/504211a
Larivière, V., and Gingras, Y. (2010) 'The Impact Factor's Matthew Effect: A Natural Experiment in Bibliometrics', *Journal of the American Society for Information Science and Technology*, 61: 424–7. https://doi.org/10.1002/asi.21232
Larivière, V., and Sugimoto, C. R. (2019) 'The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects', in W. Glänzel, H. F. Moed, U. Schmoch, and M. Thelwall (eds) *Springer Handbook of Science and Technology Indicators*, pp. 3–24. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_1
Longino, H. E. (1990) *Science as Social Knowledge*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/paperback/9780691020518/science-as-social-knowledge
Macaluso, B. et al. (2016) 'Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship', *Academic Medicine*, 91: 1136–42. https://doi.org/10.1097/ACM.0000000000001261
Mählck, P., Kusterer, H. L., and Montgomery, H. (2020) 'What Professors Do in Peer Review: Interrogating Assessment Practices in the Recruitment of Professors in Sweden', *Gender, Work & Organization*, 27: 1361–77. https://doi.org/10.1111/gwao.12500
Mallard, G., Lamont, M., and Guetzkow, J. (2009) 'Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review', *Science, Technology, & Human Values*, 34: 573–606. https://doi.org/10.1177/0162243908329381
Mantai, L., and Marrone, M. (2023) 'Academic Career Progression from Early Career Researcher to Professor: What Can We Learn from Job Ads', *Studies in Higher Education*, 48: 797–812. https://doi.org/10.1080/03075079.2023.2167974
Martinson, B. C. et al. (2006) 'Scientists' Perceptions of Organizational Justice and Self-Reported Misbehaviors', *Journal of Empirical Research on Human Research Ethics*, 1: 51–66. https://doi.org/10.1525/jer.2006.1.1.51
Martinson, B. C. et al. (2010) 'The Importance of Organizational Justice in Ensuring Research Integrity', *Journal of Empirical Research on Human Research Ethics*, 5: 67–83. https://doi.org/10.1525/jer.2010.5.3.67
May, D. C. (2005) 'The Nature of School of Education Faculty Work and Materials for Promotion and Tenure at a Major Research University', Doctoral Thesis, University of Pittsburgh. http://d-scholarship.pitt.edu/7274/1/DansETD2.pdf
Moher, D. et al. (2020) 'The Hong Kong Principles for Assessing Researchers: Fostering Research Integrity', *PLOS Biology*, 18: e3000737. https://doi.org/10.1371/journal.pbio.3000737
Moher, D. et al. (2018) 'Assessing Scientists for Hiring, Promotion, and Tenure', *PLOS Biology*, 16: e2004089. https://doi.org/10.1371/journal.pbio.2004089
Moore, S. et al. (2017) 'Excellence R Us": University Research and the Fetishisation of Excellence', *Palgrave Communications*, 3: 1–13. https://doi.org/10.1057/palcomms.2016.105
Morales, E. et al. (2021) 'How Faculty Define Quality, Prestige, and Impact of Academic Journals', *Plos One*, 16: e0257340. https://doi.org/10.1371/journal.pone.0257340
Moss-Racusin, C. A. et al. (2012) 'Science Faculty's Subtle Gender Biases Favor Male Students', *Proceedings of the National Academy of Sciences*, 109: 16474–9. https://doi.org/10.1073/pnas.1211286109
Müller, R., and de Rijcke, S. (2017) 'Thinking with Indicators. Exploring the Epistemic Impacts of Academic Performance Indicators in the Life Sciences', *Research Evaluation*, 26: 157–68. https://doi.org/10.1093/reseval/rvx023
Munafò, M. R. et al. (2017) 'A Manifesto for Reproducible Science', *Nature Human Behaviour*, 1: 0021. https://doi.org/10.1038/s41562-016-0021
Osipian, A. L. (2009) 'Feed from the Service": Corruption and Coercion in State-University Relations in Central Eurasia', *Research in Comparative and International Education*, 4: 182–203. https://doi.org/10.2304/rcie.2009.4.2.182
Pontika, N. et al. (2022a) 'Indicators of Research Quality, Quantity, Openness and Responsibility in Institutional Review, Promotion and Tenure Policies across Seven Countries', *Quantitative Science Studies*, 3: 888–911. https://doi.org/10.1162/qss_a_00224
Pontika, N. et al. (2022b) 'Data and Code for "Value Dissonance in Research(Er) Assessment: Individual and Institutional Priorities in Review, Promotion and Tenure Criteria Related to Research

Quality, Quantity, Openness and Responsibility"'. Zenodo. 10.5281/zenodo.7472276.
Poot, R., and Mulder, W. (2021) 'Banning Journal Impact Factors Is Bad for Dutch Science', *Times Higher Education (THE)*, 3 Aug. 2021. https://www.timeshighereducation.com/opinion/banning-journal-impact-factors-bad-dutch-science
Power, M. (1999) *The Audit Society: Rituals of Verification*. Oxford, New York: Oxford University Press.
Prottas, D. J. et al. (2017) 'Relationships among Faculty Perceptions of Their Tenure Process and Their Commitment and Engagement', *Journal of Applied Research in Higher Education*, 9: 242–54. https://doi.org/10.1108/JARHE-08-2016-0054
Rice, D. B. et al. (2020) 'Academic Criteria for Promotion and Tenure in Biomedical Sciences Faculties: Cross Sectional Analysis of International Sample of Universities', *BMJ*, 369: m2081. https://doi.org/10.1136/bmj.m2081
de Rijcke, S. et al. (2016) 'Evaluation Practices and Effects of Indicator Use—A Literature Review', *Research Evaluation*, 25: 161–9. https://doi.org/10.1093/reseval/rvv038
Robinson-Garcia, N. et al. (2023) 'Valuation Regimes in Academia: Researchers' Attitudes towards Their Diversity of Activities and Academic Performance', *Research Evaluation*, 32: 496–514. https://doi.org/10.1093/reseval/rvac049
Ross-Hellauer, T. et al. (2023) 'Value Dissonance in Research(Er) Assessment: Individual and Institutional Priorities in Review, Promotion and Tenure Criteria', *Science and Public Policy*, 51: 337–51. https://doi.org/10.1093/scipol/scad073
Rushforth, A., and Hammarfelt, B. (2023) 'The Rise of "Responsible Metrics" as a Professional Reform Movement: A Collective Action Frames Perspective', *Quantitative Science Studies*, 4: 879–97. https://doi.org/10.1162/qss_a_00280
Schimanski, L. A., and Alperin, J. P. (2018) 'The Evaluation of Scholarship in Academic Promotion and Tenure Processes: Past, Present, and Future', *F1000Research*, 7: 1605. https://doi.org/10.12688/f1000research.16493.1
Shu, F., Liu, S., and Larivière, V. (2022) 'China's Research Evaluation Reform: What Are the Consequences for Global Science?', *Minerva*, 60: 329–47. https://doi.org/10.1007/s11024-022-09468-7
Science Europe. (2022) *Research Assessment*. Science Europe. https://www.scienceeurope.org/our-priorities/research-assessment/
Siler, K., and Larivière, V. (2022) 'Who Games Metrics and Rankings? Institutional Niches and Journal Impact Factor Inflation', *Research Policy*, 51: 104608. https://doi.org/10.1016/j.respol.2022.104608
Sivertsen, G., and Rushforth, A. (2023) 'The New European Reform of Research Assessment | R-QUEST Policy Brief No. 7', R-QUEST Policy Brief no. 7. Oslo, Norway: Center for Research Quality and Research Impact Studies. https://www.r-quest.no/news/the-new-european-reform-of-research-assessment/
Smaldino, P. E., and McElreath, R. (2016) 'The Natural Selection of Bad Science', *Royal Society Open Science*, 3: 160384. https://doi.org/10.1098/rsos.160384
Smesny, A. L. et al. (2007) 'Barriers to Scholarship in Dentistry, Medicine, Nursing, and Pharmacy Practice Faculty', *American Journal of Pharmaceutical Education*, 71: 91. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2064889/
Strathern, M. (2000) 'The Tyranny of Transparency', *British Educational Research Journal*, 26: 309–21. https://www.jstor.org/stable/1501878
Strauss, A., and Corbin, J. M. (1997) *Grounded Theory in Practice*. SAGE.
Teplitskiy, M. et al. (2018) 'The Sociology of Scientific Validity: How Professional Networks Shape Judgement in Peer Review', *Research Policy*, 47: 1825–41. https://doi.org/10.1016/j.respol.2018.06.014
Torrance, H. (2016) 'Political Aspects of Assessment', in M. A. Peters (ed.) *Encyclopedia of Educational Philosophy and Theory*, pp. 1–5. Singapore: Springer. https://doi.org/10.1007/978-981-287-532-7_392-1
Tsoukas, H. (1997) 'The Tyranny of Light: The Temptations and the Paradoxes of the Information Society', *Futures*, 29: 827–43. https://doi.org/10.1016/S0016-3287(97)00035-9
UNESCO. (2021) *UNESCO Recommendation on Open Science*. Paris, France: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
van den Brink, M., Benschop, Y., and Jansen, W. (2010) 'Transparency in Academic Recruitment: A Problematic Tool for Gender Equality?', *Organization Studies*, 31: 1459–83. https://doi.org/10.1177/0170840610380812
Wennerås, C., and Wold, A. (1997) 'Nepotism and Sexism in Peer-Review', *Nature*, 387: 341–3. https://doi.org/10.1038/387341a0
Wilsdon, J. et al. (2015) 'The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management'. https://doi.org/10.13140/RG.2.1.4929.1363
Woolston, C. (2022) 'Time to Rethink the Scientific CV', *Nature*, 604: 203–5. https://doi.org/10.1038/d41586-022-00928-4
Ylijoki, O.-H. (2022) 'Invisible Hierarchies in Academic Work and Career-Building in an Interdisciplinary Landscape', *European Journal of Higher Education*, 12: 356–72. https://doi.org/10.1080/21568235.2022.2049335
Zhu, Y. (2017) 'Who Supports Open Access Publishing? Gender, Discipline, Seniority and Other Factors Associated with Academics' OA Practice', *Scientometrics*, 111: 557–79. https://doi.org/10.1007/s11192-017-2316-z

Research Evaluation, 2024, 33, 1–12
https://doi.org/10.1093/reseval/rvae055
Article