Practicing responsible research assessment: qualitative study of faculty hiring, promotion, and tenure assessments in the United States

Rushforth, A.D.; Rijcke, S. de

Research Evaluation, 2024, **00**, 1–11
https://doi.org/10.1093/reseval/rvae007
Article

OXFORD

Practicing responsible research assessment: Qualitative study of faculty hiring, promotion, and tenure assessments in the United States

Alexander Rushforth [1,*] **and Sarah De Rijcke** [1]

[1]Centre for Science and Technology Studies (CWTS), Leiden University, Willem Einthoven Building, Kolffpad 1, Leiden 2300AX, the Netherlands
*Corresponding author. Email: a.d.rushforth@cwts.leidenuniv.nl

Abstract

Recent times have seen the growth in the number and scope of interacting professional reform movements in science, centered on themes such as open research, research integrity, responsible research assessment, and responsible metrics. The responsible metrics movement identifies the growing influence of quantitative performance indicators as a major problem and seeks to steer and improve practices around their use. It is a multi-actor, multi-disciplinary reform movement premised upon engendering a sense of responsibility among academic evaluators to approach metrics with caution and avoid certain poor practices. In this article we identify how academic evaluators engage with the responsible metrics agenda, via semi-structured interview and open-text survey responses on professorial hiring, tenure and promotion assessments among senior academics in the United States—a country that has so far been less visibly engaged with the responsible metrics reform agenda. We explore how notions of 'responsibility' are experienced and practiced among the very types of professionals international reform initiatives such as the San Francisco Declaration on Research Assessment (DORA) are hoping to mobilize into their cause. In doing so, we draw on concepts from science studies, including from literatures on Responsible Research and Innovation and 'folk theories' of citation. We argue that literature on citation folk theories should extend its scope beyond simply asking researchers how they view the role and validity of these tools as performance measures, by asking them also what they consider are their professional obligations to handle bibliometrics appropriately.

Keywords: responsible metrics; research assessment reform; hiring, promotion and tenure; responsible evaluation; journal Impact Factor.

1. Introduction

Evaluative bibliometrics emerged in the 1970s with the promise of providing rational, efficient, and effective means of judging the performance of research and researchers, that could complement—or even replace—peer review (Narin 1976; Wouters 1999; Sugimoto and Larivière 2018). However, counter-discourses to these promises are as old as evaluative bibliometrics itself, and have only grown in response to expanding academic performance regimes. Since the 2010s, reform movements for 'responsible metrics' and 'responsible research assessment' have emerged (Curry et al. 2020; Aubert Bonn and Bouter 2021), which emphasize not so much abandoning bibliometrics, but ensuring that they are used appropriately. Principally these movements have worked through campaigning to raise awareness of problems around metrics, through global standards and principles of good practice, through self-regulatory mechanisms like asking individuals and organizations to sign up publicly to pledges, and latterly through building resources to support community learning (Hatch and Curry 2020; Schmidt et al. 2021). Much emphasis is placed on *responsibilizing* professionals to become 'better citizens' and re-work what is meant by good evaluative practices (Leydesdorff et al. 2016; Hammarfelt and Rushforth 2017). In these regards, the responsible metrics movement has drawn influence from parallel responsibilization movements, especially Responsible Research and Innovation (Dorbeck-Jung and Shelley-Egan 2013; Davies and Horst 2015; Rip 2020).

Despite the growing visibility of assessment reform movements in some quarters, the fields of science studies and research on research have only just begun to venture into this emerging reform landscape (Pontika et al. 2022; Schönbrodt et al. 2022; Ross-Hellauer et al. 2023; Rushforth and Hammarfelt 2024). Prominent interventions in the responsible metrics movement include the San Francisco Declaration on Research Assessment (DORA 2013), the Leiden Manifesto (Hicks et al. 2015) and the Metric Tide report (Wilsdon 2016). Notable principles cited in these pro-reform documents include not relying upon journal-based publication indicators when making assessments for hiring, promotion, tenure or funding (DORA), and only using indicators to support rather than replace expert judgement (Leiden Manifesto and Metric Tide). More recently, calls for reforming assessment practices have extended to emphasize values promoted by parallel reform agendas including movements for open science, research integrity, and diversity, equity and inclusion (Curry et al. 2022). These statements and initiatives share concerns that excessive emphasis on quantitative research performance indicators overlooks and dis-incentivizes other important academic contributions, including teaching (Geschwind and Broström 2015), collegiality (Dawson et al. 2022), openness (UNESCO 2021) and integrity (Bouter 2020).

While global actors like UNESCO (UNESCO 2021) and Global Young Academy (GYA 2021) and regional actors like the Latin American Forum for Research Assessment (FOLEC-CLASCO 2021) have championed research assessment reform, arguably much momentum for such issues has come from European-based actors. Countries such as the Netherlands, Norway and Finland, have initiated national level policy initiatives to enact reforms of assessment

practices—with each drawing on the responsible metrics mantra that indicators should support, but not replace, expert peer judgement (VSNU et al. 2019; TJNK 2020; UiR 2021), while the UK devolved ministries and funders have commissioned the Future Research Assessment Programme (FRAP), "to understand what a healthy, thriving research system looks like and how an assessment model can best form its foundation" (UKRI 2022). The high profile European Commission-endorsed Agreement on Reform of Research Assessment (CoARA 2022), meanwhile, explicitly endorses the Leiden Manifesto, as does the League of European Research Universities (LERU 2022).

By contrast, the responsible metrics movement has had ostensibly less impact in the United States. If one takes, for instance, the indicator of DORA signatories, in March 2024, only three large United States research performing institutes had signed DORA (Syracuse University, Illinois Institute of Technology and Larkin University), with a small number of univeristy departments, research centers and academic libraries also having signed. This is not many, given at the time of writing, the Declaration is 10 years old. The United States is something of an anomaly among OECD nations, insofar as it has never had a government-led national assessment exercise (Cozzens 2007), which possibly accounts for the absence of a concerted 'national conversation' on the role of evaluative bibliometrics. Certainly, smaller research systems like the United Kingdom (which does have a national exercise), have had scores of academic institutes sign DORA, while our own country of work, the Netherlands (a relatively small research system, also with a national assessment exercise), has had as many universities sign DORA as the United States. In addition to these observations, a series of empirical, largely quantitative studies have described the continued presence of indicators like the Journal Impact Factor (JIF) in formal organizational documents like faculty handbooks in North America. In a cross-national comparison of hiring, promotion, and tenure documents, Rice et al. (2020) found 97% of the documents they sampled from North American academic institutes included 'traditional indicators' (peer reviewed publications, JIF, grant income) as explicit assessment criteria, compared to 50% of documents from European institutes (Rice et al. 2020). Similarly, McKiernan et al.'s (2019) study of North American research institutes' hiring, promotion and tenure documents, found the JIF or closely related terms featured in 40% of research-intensive institutes they sampled, with 87% of these coded as referring positively to its use, and 63% of all institutes that mentioned the JIF associating it with quality (McKiernan et al. 2019). While the influence of indicators like the JIF and H-index in hiring, promotion, and tenure assessments ultimately cannot be known (Ma and Ladisch 2019)[1], their continued presence in formal documentation, suggests attempts to discredit such indicators have not had the effects hoped for in the United States by reform campaigners. Removal of the JIF from faculty hiring, promotion, and tenure guidance and handbooks is, after all, one of main actions that organizations signing or aligning with DORA are expected to undertake (DORA 2013).

This paper is interested in the same empirical setting and broad problem area as these quantitative studies of assessment documentation, but takes a different route into the issue of quantitative indicators in United States academia: by taking a qualitative approach to how scientists and scholars construct accounts towards the responsible metrics agenda. Rather than describe quantitatively the relative presence or absence of bibliometric indicators, our aim is to surface points of friction between the emerging trans-national responsible metrics reform agenda and current understandings and practices around uses of quantitative research indicators in hiring, promotion and tenure. Such procedures are fundamental to the making of scientific careers and the reproduction of the institutions of science—and are procedures which campaigners consider to have been captured by the inappropriate influence of quantitative measures (DORA 2013; Moher et al. 2018; Schönbrodt et al. 2022).

Our empirical findings draw on interview and open-text survey responses from U.S.-based academic researchers. We explored how they engaged with questions of appropriate uses of research metrics in the context of hiring, promotion and tenure assessment activities, and whether these aligned with the emerging language of responsible metrics advanced by reform movements. To help theorize these dynamics we draw on insights from two previously separate lines of research within science studies—literature on citation 'folk theories' and literature theorizing Responsible Research and Innovation. Though our empirical materials concentrate on the United States, our findings may also be suggestive of points of friction trans-nationally-oriented assessment reform movements might encounter in other settings.

2. Citation folk theories

To address how scientists and scholars respond to calls to reform their practices, it is important to understand the roles and values scholars attach to metrics in evaluations and research. An existing line of research that speaks to such concerns focuses on *citation folk theories*. Folk theories of science and technology are "generalizations … based in some experience, but not necessarily systematically checked. Their robustness derives from their being generally accepted, and thus part of a repertoire current in a group or in our culture more generally" (Rip 2006: 349).

The folk theories concept was first applied to citations by Aksnes and Rip (2009), but this concept can be extended to include earlier studies of how scientists have engaged with citations and publication-based indicators as performance measures. Aksnes and Rip's survey of Norwegian authors with highly cited publications, focused on relationships scientists drew between quality of a paper and its citation history, what factors played a role in highly cited papers, and the fairness of the system. They concluded that citations are sought after because they are part of the reward system of science, but that there is also ambivalence in the views of scientists, who conveyed methodological shortcomings of indicators and noted factors such as 'over-citedness' (a publication is said to have acquired more citations than it 'deserves') and timeliness as means of deflating their validity as proxy measures for quality and impact. Hargens and Schuman's (1990) earlier study of biochemists and sociologists in the United States, likewise found 'dual use' of citations, with widespread scepticism towards such indicators in scientific communities being coupled with their continued reification by researchers in promoting themselves and in re-affirming or rationalizing the 'quality' of a journal or publication. Disciplinary orientations regarding the value of empirical data (e.g. quantitative sociologists were more favorable than qualitative sociologists towards citation counts), degrees of consensus in a scholar's

discipline, and the prestige of a scholar's department, also informed attitudes to citations as performance indicators (Hargens and Schuman 1990). In an earlier ethnographic study focusing on biomedicine, we examined scientists' folk theories of the Journal Impact Factor (Rushforth and de Rijcke 2015). Participants leant on citation folk theories to varying degrees when asked to account for how the JIF featured in their everyday research practices. Justifications and explanations that the JIF saves time for busy scientists conducting evaluations (Ma 2021) and helps them to mitigate limits of expert judgement in a context of hyperspecialization (Ma and Ladisch 2019) are further examples of folk theorizing on the role and value of the JIF. Such general awareness of the JIF's importance (which albeit carried some ambiguity), played into considerations around research collaborations, project planning and where and when to submit manuscripts for publication (Rushforth and de Rijcke 2015, see also Müller and de Rijcke 2017).

While folk theory accounts provide many important insights, we noticed upon revisiting this literature that much of it pre-dated or largely overlooked contemporary reform movements and initiatives like DORA (2013). Aksnes and Rip commented in 2009, for instance, that: "Today, the opposition against such [citation] measures seems to have weakened" (2009: 245), a statement that no longer holds. Furthermore, this literature has tended to ask scientists (often in the abstract) to reason if citations are valid and/or important tools for assessing quality and impact (Wouters 2014). They have *not* linked citations to questions of professional responsibility—as the responsible metrics reform movement prompts scientists to do. In addition to understanding how scientists understand the roles of citations, it is also important to ask: (when) is it appropriate to use bibliometrics? What are your professional obligations to use these tools responsibly? Do you recognize accounts of 'good' and 'bad' practice advanced within the responsible metrics reform movement?

To help unpack the significance of responsibility as a key concept in the contemporary research assessment reform landscape, we turn now to science studies accounts of other 'responsible' science reform movements, including Responsible Research and Innovation.

3. Responsibility and contemporary science reform movements

The Metric Tide report coined the term 'responsible metrics' in 2016, explicitly citing Responsible Research and Innovation as its inspiration (Wilsdon 2016). Popularized in European science policy in the 2010s (Owen et al. 2021), Responsible Research and Innovation (RRI) is a responsibilization movement aiming towards making research actors more aware and responsive towards the potential harms and uncertainties around their research and innovation activities (Stilgoe et al. 2013).[2] The responsible metrics movement's similarities to RRI go beyond simply having the word responsibility in common. These and other globally-oriented science reform movements, including research integrity (Davies and Lindvig 2021; Penders 2022), are 'normative projects' (Brundage and Guston 2019) that aim to re-make 'good' professional practices from a distance.

In our view, the concept of *responsibility languages*, developed in the context of RRI (Rip 2020), is productive for thinking about the travels of the responsible metrics movement's language into academic assessment. Responsibility languages sets out a 'grammar' for responsible action, packaged in the form of rules, standards, principles, mantras, narratives and so on. These languages seek to transform the world through pushing (maintaining, or proposing changes to) what Rip calls a 'division of moral labor'. Whereas the division of labor in industrial production refers to separation of tasks into specialized work divisions, here Rip extends the concept to incorporate the socio-moral order (e.g. roles and responsibilities actors have to one another). DORA and the Metric Tide for example, divide obligations out among research system actors (individuals, universities, publishers, funders etc) – each expected to play their part in a new division of moral labor whereby metrics are used appropriately in assessments. In doing so, these texts appeal to the *rights* of those being evaluated (and the rights of 'society') to have research assessed fairly and effectively (a right thought to be threatened by inappropriate uses of bibliometric indicators), and the *obligations* of research system actors to uphold their end of the science-society contract by handling bibliometrics appropriately. Rhetorically, responsibility languages also seek to persuade by constructing "evolving narratives of praise and blame" (e.g. the 'good' versus the 'cowboy' firm, or in this instance, the 'good' versus the 'bad' evaluator) (Rip 2020). The DORA signature is a good example of a device for cultivating praise: a means for individuals or organizations to signal to external stakeholders that, through publicly committing to self-regulate according to the statement's values, they are responsible evaluators (or good citizens). Texts like DORA, the Leiden Manifesto and Metric Tide also offer model languages which research funding organizations, research performing organizations, and other research organizations can copy or adapt within their internal documents, on job or funding applications, and on their websites to communicate they are responsible evaluators. General characterizations of 'bad' evaluators also circulate within the responsible metrics movement, for instance, individuals or organizations that persist in using the JIF, or replace entirely expert judgment with quantitative indicators (Rushforth and De Rijcke 2017). The Metric Tide Report's proposal to set-up a 'bad metric prize' (akin to awards for bad sex in movies) was another, tongue-in-cheek, means of cultivating shame, albeit one that did not capture the imagination of the UK research community (Curry et al. 2022).

By conceptualizing responsible metrics as an emerging responsibility language, our aim is to enquire whether this language is penetrating and reconfiguring 'divisions of moral labor' around hiring, promotion and tenure of professorial faculty in United States academia—and whether this new responsibility language aligns with 'bottom-up' responsibilities articulated by respondents. Coined by Glerup et al. (2017), bottom-up responsibilities refer to scientists' propensities for articulating vocational senses of responsibility for doing what they consider 'good science' (or in this instance 'good evaluation'). Scientists provide such accounts, even if they are unfamiliar or uninterested in codified guidelines and standards for promoting responsible conduct (Glerup et al. 2017: 325). With these concepts in mind, we ask specifically:

- How 'fluent' were respondents with the 'responsibility language' around the DORA statement and wider responsible metrics movement?

- How did respondents engage with problems and solutions set out by the responsible metrics reform movement's responsibility language?

4. Data collection and analysis

Data for this study came out of Project TARA (Tools to Advance Research Assessment), a collaborative project with the San Francisco Declaration on Research Assessment (DORA) and Illinois Institute of Technology. We led on a workpackage aiming to understand the small numbers of United States-based research performing institutes signing DORA—a research puzzle co-conceived with our project partners—but led by us in terms of data collection, analysis and writing. The United States is the world's largest research system, and it is of interest for the responsible assessment reform movement that their cause has not resonated there yet. It is, we believe, a setting also of empirical interest for science studies researchers interested in the (non)spread of this evolving reform movement.

4.1 Research ethics

Ethical permissions for the survey and interviews were granted by the Ethics Review Committee for Social Sciences, Faculty of Social and Behavioral Sciences, Leiden University. Survey respondents were provided with an information sheet and consent 'button' at the beginning of the online questionnaire. Due to interviews being held online, informed consent was taken verbally at the beginning of each interview and recorded on a verbal consent form, that was shared with interviewees afterwards.

4.2 Survey

Partly inspired by Science Europe's (2019) and GRC's (2020) surveys, our survey focused on the current practices and (non)signing of DORA in American research performing institutions. As our emphasis was on exploration, the survey was open to a wide range of respondents. The survey was shared with members of DORA's mailing list, via American learned societies message boards, and through emails to senior faculty and administrators among American Association of Universities member institutes. This sampling strategy was guided by convenience and maximum variation considerations. The DORA mailing list comprises individuals from multiple sectors and positions, from professors and post-docs to publishers and librarians. Asking senior academic and administrative staff within DORA's mailing list to participate was deliberate, as we wanted to elicit at least some responses within our sample from U.S.-based individuals already knowledgeable or aware of the responsible metrics movement. Such responses would hopefully give useful insights into non-uptake of DORA in U.S. academia. Targeting learned societies and American Association of Universities members would in turn, we hoped, lead us to further the number of senior U.S. academics able to speak to faculty hiring, promotion, and tenure processes in their institute. Respondents from this pool would likely be less aware of DORA-related developments than those recruited via the DORA mailing list, thus broadening the range of responses to include more and less knowledgeable responses.

Within the information sheet and consent form participants were informed about the aim of the survey, namely to gain a broad understanding of research metrics policies and processes in the context of faculty hiring, promotion and tenure decisions. Respondents were asked to confirm they met our study inclusion criteria, namely: a) they were employed by a United States academic institute and b) considered themselves knowledgeable enough to answer questions about hiring, promotion, and tenure processes in their institutes (knowledge of DORA and the responsible metrics movement was not a prerequisite). We decided to use an anonymous survey link to ensure the survey could be passed on to additional appropriate persons within the organizations of recipients. Overall, 53 eligible responses were collected. The questionnaire and anonymized response data can be found here: https://doi.org/10.5281/zenodo.7882974

This paper draws upon qualitative data collected during the survey, from the following open text questions:

- Please reflect on whether you consider criteria and standards used in your institute to assess academic contributions in career advancement processes to be appropriate.
- Please add any further additional comments you have about the influence of quantitative publication and citation-based metrics in evaluating candidates in faculty hiring, promotion, and tenure processes within your institute.

These open text questions did not afford sufficient flexibility or insight to address our first research question: how 'fluent' were respondents with the 'responsibility language' around the DORA statement and wider responsible metrics movement? However, the open text questions elicited insightful responses for helping us explore and develop the second research question—how did respondents engage with problems and solutions set out by the responsible metrics reform movement's responsibility language? – which we address in the second section of our findings.

4.3 Interviews

Whilst responses to the survey pointed to much heterogeneity in the context of hiring, promotion, and tenure procedures across United States research institutes, the inflexible nature of the instrument and the general brevity of open-text responses prompted us to seek to explore emerging complexities in further depth. We therefore decided to approach individuals who had experience of hiring, promotion, and/or tenure procedures in United States research performing institutes for in-depth, semi-structured interviews. In total 18 individuals were interviewed.

Participants were recruited with a purposeful sampling approach, consisting of contacting individuals in the network of DORA staff and through snowballing. In doing so we looked for maximum variation in our sampling of individuals, covering for instance individuals comprising: both Carnegie-classified Research 1 (research intensive) and Research 2 (less research intensive) institutes, institutes that had and had not signed DORA, different disciplines, and at sufficient levels of seniority and experience to be able to inform us about their institute's processes. This would help to compose a multidimensional image of the various parties involved in hiring, promotion, and tenure procedures and compare different understandings and experiences across United States academia of the challenges for reforming assessment practices in line with emerging campaigns for responsible metrics.

4.4 Interview structure

Interviews lasted between forty-five minutes to one hour and covered issues from respondents' awareness of responsible metrics campaigns, to their uses of quantitative performance indicators in hiring, promotion, and tenure, to their views on the prospect of reforming hiring, promotion and tenure procedures in their institutes.

One methodological challenge we faced in approaching the question addressed in the second section of our findings—"how did respondents engage with problems and solutions set out by the responsible metrics reform movement's responsibility language?"—was defining the 'movement'. This after all is a coalition of various voices and statements that has evolved over time, whose message does not always add up to a fully consistent or coherent whole. Even a single document like DORA is a multifaceted text. To negotiate this complexity, respondents in interviews were asked questions that would steer them to consider four major themes of responsible metrics texts: whether JIF played a role in hiring, promotion and tenure decision-making processes in their institutes (following DORA's interest in the JIF); whether they agreed with the notion that quantitative indicators can inform but should not drive evaluation of candidates (in line with the Leiden Manifesto and Metric Tide); whether they agreed with proposition that appropriate uses of metrics are very important when it comes to defining what constitutes a 'good' evaluation and a 'good' evaluator (a consistent theme across the movement's discourse); and whether they agreed with the proposition that senior academics are (at least partially) accountable for enacting the responsible metrics agenda (another consistent theme in the movement). Responses were solicited through direct prompts, through sharing statements from DORA and the Metric Tide report on screen (using the MS Teams video screen sharing feature), or were provided by respondents in response to directed interview questions (e.g. "do indicators like the Journal Impact Factor or H-index play a role in hiring, promotion or tenure?"). These questions helped to elicit explanations, motivations, and justifications about 'good' evaluative practices.

4.5 Data analysis

All interviews were recorded using Microsoft Teams, which provided an automatic audio transcript. The first author listened back to interviews to check for accuracy of the recorded texts and to begin to immerse themselves in the data. Open text survey responses were cleaned for spelling and typo errors and anonymized where necessary, before being uploaded together with interview transcripts onto AtlasTi Version 9 to support familiarization and coding of textual data.

We coded transcripts, initially using an open coding approach, followed by more refined mapping of emerging themes and categories. This process involved a constant comparative approach to move back and forth between data, discussions, and ongoing reading of the academic literature. Through this approach, we gradually were able to produce a composite narrative, that addresses first the questions of how 'fluent' interview respondents were with the responsible metrics responsibility language, and second to what extent survey and interview respondents' own accounts of metrics and bottom-up responsibilities aligned with the responsible metrics movements responsibility language. In this process, we identified various folk theories drawn on to explain and legitimate responses, which we subsequently sorted into three distinct clusters. Our narrative is supported throughout with illustrative quotations from interviews and the survey, which have been anonymized to protect the identities of respondents.

An important note on reflexivity: we are aware that as authors we are hardly neutral, passive observers of the reform movement we seek to analyze. Aside from collaborating with DORA on Project TARA, one of us (De Rijcke) was co-author of the Leiden Manifesto and both of us contributed to the independent report upon which the 2016 Metric Tide's recommendations were officially based. Beyond that, we serve on multiple assessment reform committees and projects, and are ourselves currently debating how our own institute can best practice responsible evaluation. The primary 'role' we aim to perform in this research is that of science studies researchers, but undoubtedly we bring other forms of knowledge and experience to the table—as academics, evaluators, administrators, campaigners, collaborators, and so on. We see such interests as enriching as much as 'biasing' our study findings.

5. Findings

5.1 Responsibility language 'fluency'

While there were varying levels of awareness of DORA and related statements and principles amongst our United States-based interview respondents, our overall impression was one of a lack of strong familiarity with the language of responsible metrics put forward by prominent campaign groups and good practice statements. Interviewees recruited via DORA's network were not as knowledgeable about DORA's statement and purpose as might be assumed. This is consistent with Davies's (2019) interview-based study regarding research integrity principles, in which researchers seldom knew the ins and outs of formal guidelines and principles, even if ostensibly sympathetic towards a general cause. Our findings supported emerging inferences regarding the continued prevalence of JIF and other indicators in hiring, tenure and promotion documents in the United States (e.g. McKiernan et al. 2019, Rice et al. 2020), which arguably is reinforced by the fact thatvery few U.S.-based academic institutes have signed DORA. The following quotation is illustrative of this emerging responsibility language not being 'on the radar':

> I'm trying to think if you know I've heard some people talk about, you know, the drawbacks of utilizing citations and impact factor and different things like that. You know there's some push to utilize other metrics. You know things like the H-index and things like that. But I haven't. I haven't heard it strongly. It's just a person here or there. It doesn't seem to have much momentum from what I've heard. (Interview DI1)

When asked about whether they recognized the phrase *responsible metrics*, most interviewees stated they did not. Among those with past encounters with DORA, respondents tended not to have read the statement frequently, less still have memorized it (per Davies, 2019). Researchers that had come into contact with statements like DORA tended to have formed broad impressions about their content, which were more-or-less accurate when compared to the original statements: one respondent, for example, mistook DORA as a

statement committing institutions to open access publishing. Others were candid that they had not heard of DORA or if they had, did not know what it was about. Where there was familiarity with the statement, it tended to be towards certain elements, particularly its critique of the JIF. For a number of respondents less familiar with the responsible metrics movement, hearing the term responsibility mentioned in relation to hiring, promotion and tenure, prompted them to reach for other modes of justification for how these procedures were organized responsibly. Particularly prominent were accounts of the impersonal nature of the procedures, with multiple checks and balances in place to ensure faculty are selected according to merit rather than say patronage. Another important reference point centered on their organization's efforts to ensure social justice and avoid discrimination on grounds of ethnicity, gender, religion, sexuality and so on. Overall, across most interview respondents, these responsibility languages appeared much more familiar and ready-to-hand than that of the responsible metrics movement.

All respondents—including those that had not crossed paths with the responsible metrics movement—were aware of at least some of the structural problems and issues around indicators problematized by the emerging global responsible metrics movements. Senior academics we approached were broadly familiar both with managerial discourses embracing the promises of metrics *and* with counter-discourses and critiques mobilized against them. As espoused in the folk theory literature, respondents could recite certain criticisms of citations and publications as indicators of research performance, as well as a discourse on perverse effects of quantitative indicators, including inhibiting interdisciplinarity, injustices owing to relative lack of coverage of certain disciplines' outputs in major bibliographic databases, and propensity for goal displacement (albeit unsurprisingly they did not use this social science lexicon). Examples of awareness of such 'bottom-up responsibilities' (Glerup et al. 2017) were evident in the account of one business and management professor. While not familiar with DORA or other responsible metrics statements prior to being approached for interview in this study, he exhibited broad awareness of the folk theory that citations drive self-interested behavior at the expense of the collective.

> You know, the problem is you know especially when you're in a really narrow silo, you and your buddies cite each other's papers and you basically build up each other's, you know, H-index and citation counts dramatically and you're kind of talking to each other. (Interview NF1)

Our respondents however did not bring up other diagnoses of problems around metrics that have been more visible around European policy discourses—including their associations with burnout, bullying, workplace environment, hypercompetitive career structures, and research misconduct. Likewise they did not raise technical criticisms of better known indicators like the JIF, such as lack of field-normalization or skewedness of citations within journals, suggesting either lack of awareness or lack of importance attached to them.

If the responses by our interviewees are indeed representative of a larger phenomenon, it seems that the responsibility language promoted by responsible metrics campaigns has not travelled as deep into United States-based institutions as champions of this cause would hope. We will now consider some justifications for how metrics were engaged with in hiring, promotion and tenure assessments, and how these (mis) aligned with efforts by reform actors to diagnose them as major, urgent problems threatening the fabric of academic life.

5.2 (Mis)alignments with the responsible metrics language

We now lay out three types of accounts respondents gave towards problems and solutions set out in responsible metrics reform discourses. Each is a 'cluster' (Rip 2006, 360) into which respondents' accounts and their supporting folk theories have been sorted. The clusters are a device for separating and comparing accounts according to common narrative patterns or themes (e.g. whether a respondent agrees or disagrees with prominent responsible metrics framings). One of these clusters—*strong endorsement*—largely agreed with the framing of problems set out by this movement, while two other kinds of responses—*moderate alignment* and *pragmatic rejection*—were more ambivalent towards the responsible metrics agenda. Across each cluster, multiple folk theories were mobilized to explain and legitimate claims and arguments. We will now detail each account, including how divisions of moral labor and characterizations of 'good' and 'bad' evaluators were drawn on to support each kind of explanation.

5.2.1 Strong endorsement

Accounts that strongly endorsed the argument that metrics such as the JIF and H-index held too great a grip over academic assessments and—by extension—the research strategies of academics, were positive in their disposition towards DORA and the responsible metrics movement. These accounts were marked by positive dispositions towards the *potential* of reforms, were they to be implemented (cf Zuijderwijk et al. 2019).

Much accountability for what they considered the present state of affairs was laid at the door of 'bad' evaluators—resembling the figure of the 'bad expert' in science policy (Sweet and Giffort 2021). Bad evaluators were characterized, for instance, by pejorative use of the word 'traditional'. Closely coupled with traditional evaluators, are traditional indicators– publication and citation scores, particularly the JIF and other measurable indicators like grant money. This 'traditional' characterization was contrasted with 'modern' (and thus 'good') experts and indicators, that move with the times and embrace 'progress'.

> I think we need to advance further in identifying unique, non-traditional indicators of research impact. We remain too stuck in traditional metrics of grant money and academic publications. Many of our department faculty are not interested in considering social media or other "new" ways in which faculty can make an impact with their work. (Survey Response, East University 1)

Traditional evaluators are characterized by being stuck in the past and recalcitrant to progress—sometimes served by structures of self-interest (e.g. they benefitted from such approaches themselves), preferring an easier way out (at worst, depicted as laziness), or being in an institutional echo-chamber ('it's all they know', 'they're stuck in their ways').

> They, the administrators, like having numbers to base their determination on, so it's easier to base it on quantitative analysis than qualitative analysis. I mean, yeah. So I think they in … and it's just been historically used in some departments, so they haven't gone away from that or they haven't thought of different ways of looking at the data that could maybe be a little more subjective than the numbers can. They don't also realize that the numbers don't tell the whole story, especially in early career researchers and things like that. (Interview BK1)

In strong endorsement accounts, it is partly individuals that are held accountable for the persistance of traditional indicators, and partly it is the 'wider culture' which individuals are said to reproduce. The word culture was evoked to describe recurrent patterns of behaviour which were systemic and multi-level:

> I was reviewing the Faculty handbook in advance of this conversation. Like it came up in my mind a couple of times, that one of the most difficult things with American universities' structure is it becomes so hard to get faculty to do anything that they really don't want to do. It is, you know, the system of incentives, especially at research universities, it gets to the point where it literally becomes all about research, only about, you know, grant size and publication. And that sort of outweighs, you know, any sort of considerations about teaching or service or the other things. (Interview ET1)

This quote points to how indicators are interwoven with behavior but also with the structure of university bureaucracies. It is not either individuals or culture that are accountable in *strong endorsement* accounts—but rather both, with accounts oscillating back and forth between emphasizing one or the other. 'Traditional' individuals, for instance, were depicted as the carriers of a backward culture which frustrated change and progress. A recurring example was the figure of supervisors or advisors, who were held responsible for reproducing cultures of metrics by introducing young researchers to these measures.

> Interviewer: Where and when do scientists learn about the importance of quantitative research metrics?
>
> Respondent: I think you know when they're graduate students, right it's conveyed to them by their advisors (Interview EPE1)

Despite citing cultural and systems level accountability for the problems, *strong endorsement* accounts ultimately cited individuals in senior positions as accountable agents for affecting better practices towards indicators. In the following interview excerpt, the respondent signals they are a responsible agent, who recognizes and fulfils obligations as a senior academic to challenge any uses of inappropriate indicators when encountered:

> I would shut it down if someone on their CV wrote it [JIF score]. I, needless to say, have to write a lot of letters of for promotions and I will see that on people's CV. I've never seen anyone … none of my faculty have ever put that on their CV. And I would have it removed. This is the kind of thing where I'd go back to them before we sent the stuff out and say "Take that off" (Interview TC1)

The propensity to associate bad evaluation practices with individual failings was referred to by some natural scientists through the individualized language of 'bias' as an epistemic vice that needs to be overcome. One interview respondent, a medical researcher, recognized himself as negotiating the tightrope between good and bad evaluator, by maintaining attention towards his own 'biases':

> We have a bias. I think we have, we all have this bias and some of us realize that and we say, OK, well, we can't do that. That's not the best thing. I mean, we all see a paper published in *Nature*. We get excited. That's just habitual. That isn't necessarily the best thing for us to do, but it's something we have to overcome. (Interview BD1)

Such accounts of 'good evaluators' are, we would suggest, largely compatible with ideals of 'good citizenship' that calls for assessment reform seek to promote. Citizens are imagined as autonomous, reflexive, but duty-bound social agents. These are the kinds of divisions of moral labor that DORA and other reform actors are seeking to cultivate and promote further—and if one paid attention only to strong endorsement accounts, this message would appear to have reached a receptive audience. How though does this emerging responsibility language come unstuck among senior United States academics who are more cautious or sceptical towards the responsible metrics arguments?

5.2.2 Partial alignment

Our data also elicited clusters of accounts and folk theories that did not subscribe fully, or even partially, to the problematizations of research metrics in hiring, promotion, and tenure assessments and new divisions of moral labor the responsible metrics movement has put forward. Several respondents casted doubts of how widespread the problem actually is:

> I frequently hear of arbitrary publication assessments being used, but these are not codified anywhere and thus difficult to specifically address within my institution. (Survey Response South East University 2)

Much of the faculty hiring, promotion, and tenure practices in the U.S. takes a 'portfolio' approach, asking applicants to account for research, teaching, and service activities in the application materials. Partial alignment accounts tended to reject claims that assessments were skewed towards quantitative indicators—arguing that they appear in certain parts of the decision-making processes, alongside other considerations, but do not unduly dominate these assessment process overall, or the assessment of an applicants' research achievements:

> We take a balanced approach to reviewing candidates research productivity and impact. Citation counts, impact factor, etc are part of the review but they are only part of the assessment. (Survey Response MidWest University 4)
>
> [After reading out a section DORA statement calling for abandoning JIFs, the interviewee responds] Yeah, exactly.

> I would say that that that's in line with what I was saying, although you know we like. The difference is that if it [impact factor] is, if it is there and noteworthy, we would say something about it. But we don't base the assessment on that. (Interview UQ1)

While the proposition that indicators should support but not lead assessments was largely agreed with, there was reluctance to accept full-scale denunciations of certain indicators like H-index and JIF. Demonstrating an understanding of the limitations of well known indicators is used as a means of defending its presence and signaling one's own status as a responsible evaluator:

> We occasionally discuss H-index but we also recognize that different disciplines and sub-disciplines are cited differently. We pay much more attention to narrative evaluations of publications. (Survey Response East University 1).

Partial alignment accounts relate to elements of the responsible metrics language in ambivalent ways: while there was wide agreement with mantras like "metrics should not drive decision making processes rather than drive them", there was nonetheless persistence with uses of certain indicators which many in the reform movement would consider too flawed to play any kind of legitimate role. While *partial alignment* accounts acknowledge a generalized risk that metrics can become too influential and thereby lead to poor decision-making in the wrong hands, they do not believe this characterizes their own practices. Likewise—and more forcefully pushing back against responsible metrics discourse and "reaffirming established norms" (Zuijderwijk et al.., 2019) - they do not agree that well-known indicators like the JIF or H-Index should be discredited as evaluative tools: continuing to use these tools does not, in partial alignment accounts, equal being a 'bad evaluator':

> Certainly H-index, counts for something you know, we certainly don't want to see, especially if we're hiring someone at the assistant or associate professor level. I think it's important to look to see if there are significant gaps in their publication record and you know, sometimes that could signal some things we, you know, sometimes it might not for a junior assistant professor, you know, you want someone to be productive, you want someone to come out of a post doc or two postdocs demonstrating some productivity, you know, in developing a research program that you know, there's cohesion … I mean funding is helpful, but we want to make sure that this person is publishing, will be recognized in their field for their contributions, and it doesn't necessarily mean that they need to publish 10 or 15 papers. We don't measure excellence in that way. Excellence is measured by quality and we look for quality more so than quantity. (Interview BD1)

The quote starts by portraying the H-index as an effective indicator for weighing-up scholarly productivity and impact within a candidate's track record. This endorsement of the indicator is, however, also accompanied by reassurances that the individual and their colleagues are aware that such indicators are not the only criteria that should be taken into consideration. Responsibility, for this respondent, is ensured by the use of such an indicator within a 'basket of indicators': they are responsible users of the indicator because they do not allow it to *drive* the decision process. We are reassured they know better than to use the H-index alone to determine impact or a candidate's worth. This is a different (bottom-up) account of responsibility than that espoused in responsible metrics language—the latter problematizing the technical properties of such indicators[3], and in so doing undermining their legitimacy almost entirely. Neither technical arguments against the indicator's reliability or robustness, nor attempts to advance a division of moral labor ordered around 'good' versus 'bad' evaluation, seem to infiltrate, let al.one upend, the sorts of justifications articulated within *partial alignment* accounts.

5.2.3 Pragmatic rejection

Another perspective that was ambivalent towards responsible metrics discourse was pragmatic rejection—named so because these accounts provided 'pragmatic' justifications to not sign-up to responsible metrics solutions. Justifications centered on indicators being de facto 'rules of the game' and part of the background infrastructure for academic assessment. As tools that were taken-for-granted, it was seen as undesirable to bring problems they 'resolved' (temporal and epistemological constraints) to the foreground and create more work for colleagues doing thankless service work in time-poor academic settings. To remove metrics through reforms is to invite uncertainties (Zuijderwijk et al. 2019). *Pragmatic rejection* tended not to justify the enduring presence of quantitative indicators in epistemological terms (ie by arguing they are credible proxies of quality or impact) – on the contrary sometimes *pragmatic rejection* accounts even acknowledged they have flaws. The core emphasis of these accounts was on the need to 'live with' their imperfections.

> I think for the most part, people kind of begrudgingly are OK with it, in the sense that you know, it may not be the best, but it's probably the best alternative that we have. And so if there if there were other alternatives or things you know. Maybe some of the types of metrics that we've talked about that gain more traction [open science indicators were discussed earlier in the conversation] … and maybe it could. They could gain some momentum, but otherwise I would say that people are kind of … this is kind of the way that we've done it, and it's worked out OK. So we're just going to keep going down that path. (Interview DI1)

The continuing legitimacy of the use of indicators the responsible metrics movement seeks to discredit, is premised here on a collective agreement that they are the rules of the game.

The pragmatic rejection accounts also divert past the good-bad evaluator dichotomy explored in the previous two accounts, and offer instead the figure of the 'pragmatic evaluator', doing what they can in the circumstances and accepting compromises. Indicators, in these accounts, played a pragmatic role in negotiating temporal constraints of assessments which must process large volumes of applicants in scarce amounts of time. Metrics were cited as a screening tool and deadlock breaker between otherwise highly skilled and credentialed candidates (consistent with arguments in Reymert 2021, Ma 2021). Likewise, the JIF was also appealed to as a pragmatic solution to other structural problems, namely

epistemological challenges in making expert judgments across hyper-specialized disciplinary borders. Indicators such as the JIF are claimed to help negotiate this problem by allowing comparison between otherwise heterogeneous entities:

> Interviewer: And so I highlighted one bit of text [from the DORA statement shared on the Zoom screen] which is their general recommendation that says "do not use journal-based metrics like journal impact factors, surrogate measures of the quality of individual research articles and to assess an individual's scientist contributions in hiring promotional funding decisions". So just, you know, if you could share any thoughts or impressions about this?
>
> Respondent: So this is a lovely idea, but very difficult to implement. Umm, what I face … So again I am the most senior person doing innovation, entrepreneurship, commercialization, innovation, ownership, actually and even finance within the school of management, right? So across those domains, I am the most senior person and I'm an associate professor. So the people who are judging me, there's not a single person in finance or innovation or entrepreneurship who is judging my work. So therefore, they have a very hard time. They are human resource management. They are accounting. They are, you know strategy. They are leadership. You know, there are all sorts of other domains of business, but nobody in my area, you know, supply chain. You know, there's people in lots of other areas, but nobody that's really actually reading any of my journals or any of my work or my colleagues. So. So it's very hard for them in all fairness to them, to like look and say this is a really great article. (Interview NF1)

Like the strong endorsement accounts, pragmatic rejection accounts set-out the multi-level, systematic nature of problems around quantitative indicators and research reward systems more generally. Pragmatic rejection, however, constructs a much more passive form of agency and accountability: in the *strong endorsement* account (see above), individuals are imputed with moral obligations to challenge poor practice and enact cultural changes. In *pragmatic rejection* accounts, the systemic nature of the problems justifies the issues being too big for individual academics, departments or even academic institutes to take on. In the meantime, indicators like the JIF are considered a serviceable, ready-to-hand solution that constitute a stable convention experts from disparate research communities can settle upon. Problems and solutions put forward by the responsible metrics movement do not appear able to pierce through the armor of *pragmatic rejection* accounts.

6. Discussion and conclusion

Responsible metrics is an ongoing reform movement with a concern to make academic 'citizens' more responsive to concerns about (mis)uses of bibliometrics. Tentatively, and without trying to claim generalizability, our findings suggest there is not yet a deep level of familiarity with international reform movements for responsible metrics and assessment in the United States. The lack of familiarity with the responsible metrics movements' 'responsibility language' was manifest in: the lack of referencing specific points in responsible metrics statements; lack of awareness of the wider range of actors involved in enacting performative powers of metrics (e.g. nobody mentioned publishers); the propensity to present their own 'bottom up' responsibilities which were different from the reform movements' language, or were similar only by coincidence because all actors inhabit the same professional world. We also observed that the responsible metrics agenda did not command the visibility or sense of shared urgency for hiring, promotion and tenure, processes, as concerns over merit-based advancement (meritocracy), impersonal authority, or social justice carried among our United States respondents.

We utilized social science concepts to help theorize and open-up responses to the responsible metrics movement to further inform reflection and debate. Our study *not only draws on* the folk theories of citations literature, but has *extended this literature* to the present juncture. The citation folk theories literature suggests that scientists and scholars are more-or-less knowledgeable about citations as performance indicators (e.g for quality and impact of published works), and that they support their accounts through theories or generalizations picked up as members of the professional world of academic research. In principle, this literature ought to provide useful theoretical insights to assessment reform champions concerned with the persistent presence of bibliometric indicators in academic evaluation and research. Mostly it has posed questions about how scientists understand and enact the value of such indicators as performance measures, but has not hitherto posed questions of professional obligation and responsibility to handle such indicators with care.

Our study's approach and findings help to bridge the gap between the citation folk theory literature and concerns animating the responsible metrics reform movement. In particular, concepts borrowed from science studies accounts of Responsible Research and Innovation like *division of moral labor* and *responsibility language* suggest that this reform movement seeks to enroll new recruits to its cause by persuading them of the shortcomings of professional practices of evaluation, brought on by certain kinds of uses of quantitative indicators and asserting new roles and responsibilities (a new division of moral labor) towards such tools. Whether new divisions of moral labor imagined and advanced by the reform movement aligns with prevailing evaluative practices, is an important empirical question, which our proposed use of these concepts can help to guide.

The adoption of these concepts is particularly useful for understanding notable ambivalences towards the responsible metrics agenda. Our data has unpacked, for example, how folk theories of citations inform defenses of the JIF or H-index, with respondents arguing they are useful proxies for likely citation impact of publications and productivity of candidates—a position that responsible metrics campaigns like DORA deem intellectually incoherent and damaging to science. Furthermore, respondents stressed the responsible uses of these indicators, for example, on the grounds they were used in 'moderation' and with reflexive awareness of their limitations (partial alignment accounts). In such accounts, core principles of the responsible metrics agenda, like ensuring that multiple rather than single indicators are drawn on to inform decisions, or that peer review deliberations occur that place indicator scores into context, were already being practiced. Some of these justifications appear thus to be

already 'responsible' in the terms of the responsible metrics movement, while other elements aligned much less.

Professionals also reassured us of their reflexive awareness towards limitations of indicators, when defending their uses on 'pragmatic' grounds (mobilizing folk theories that these indicators help overcome epistemological and temporal constraints of hiring, promotion and tenure processes). A related folk theory under this pragmatic rejection cluster, stated that even if the indicators are not epistemically robust, their legitimacy is ground in community and organizational consensus that they are de facto rules of the game. This is a different repertoire to the technical arguments against the JIF mobilized by reform advocates, and before them, scientometricians. Accounts defending persistence of the JIF and H-index provided an implicit rejection of the 'bad expert' construction being articulated within the language of the responsible assessment reform movements: reasoning that given the legitimacy of these tools' presence is established and widely accepted by those around them, can such measures really be that important a problem? And is potential disruption caused by poking at this issue, the most urgent problem that should be fixed right now?

These findings suggest that the problem-solution package set out by the reform movement and the shifts in division of moral labor academics are expected to enact, did not fully resonate among these United States academics. It is not simply that there is information deficit around the responsible metrics' cause (though no doubt this is a problem for the movement). An original and important finding of our study is that even when presented information on the movement, scientists and scholars may still construct the division of moral labor around indicators in different ways to the movement—meaning the responsible metrics problematization does not disrupt or displace already embedded 'bottom-up responsibilities'. Raising awareness and providing further information on the technical limitations of certain indicators, is thus a necessary, but on its own, insufficient step for mainstreaming this agenda.

No doubt our empirical findings are specific to this geographic area, but we believe that our insights and approach could have wider relevance science studies scholars, in studying empirically changes these campaigns are (un)able to initiate in other settings. Given that reform movements have an international (or at least a large regional) focus, our study provides a useful blueprint for posing questions of responsibility to scientists and scholars elsewhere. We hope our analytic approach provides useful insights into the arguments scientists and scholars mobilize for and against assessment reform causes, and that, finally, the findings may catalyze further discussions in emerging assessment reform movements as to why their messages are not automatically cutting through.

Acknowledgements

The authors wish to thank the 'Tools to Advance Research Assessment' (TARA) project team members Stephen Curry, Haley Hazlett, Marta Sienkiewicz, Zen Faulkes and Ruth Schmidt for their inspiring collegiality, and for their comments to an earlier version of this article. We also thank Ludo Waltman and members of the CWTS Evaluation and Culture Focal Area for comments on an earlier draft, and DORA's former program manager Anna Hatch for support during the project. Finally, we are grateful to Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin, for their generous funding of Project TARA.

Conflict of interest statement. None declared.

Notes

1. Given the confidential and ubiquitous nature of hiring, promotion and tenure procedures around the world, this is a largely insurmountable problem.
2. Responsible Research and Innovation does not seek to 'control' uncertainty through rationalized logics of accountability, liability, and evidence, but encourages research actors to attend to "future-oriented dimensions of responsibility—care and responsiveness" in science and technology (Stilgoe et al. 2013: 1359).
3. For example, how citation and publication statistics are aggregated into a single H-index number has been shown to produce inconsistent rankings, making comparisons of researcher's scores problematic. The number is also vulnerable to differences in densities of citations and collaboration patterns across fields, and towards gaming (e.g. including one's name on as many publications as possible) (CWTS 2021).

References

Aksnes, D. W., and Rip, A. (2009) 'Researchers' Perceptions of Citations', *Research Policy*, 38: 895–905.

Aubert Bonn, N., and Bouter, L. (2021) 'Research Assessments should Recognize Responsible Research Practices—Narrative Review of a Lively Debate and Promising Developments', in: *Handbook of Bioethical Decisions—Vol. II Scientific Integrity and Institutional Ethics*. Dordrecht: Springer.

Bouter, L. (2020) 'What Research Institutions Can Do to Foster Research Integrity', *Science and Engineering Ethics*, 26: 2363–9.

Brundage, M., and Guston, D. H. (2019) 'Understanding the Movement (s) for Responsible Innovation', in: *International Handbook on Responsible Innovation*. Edward Elgar Publishing.

COARA (2022) 'Agreement on Reforming Research Assessment'.

Cozzens, S. E. (2007) 'Death by Peer Review? The Impact of Results-Oriented Management in US Research', in: R. Whitely, and J. Glaser (eds), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, 225–242. Dordrecht: Springer.

Curry, S., de Rijcke, S., Hatch, A., Pillay, D., Van der Weijden, I., and Wilsdon, J. (2020) 'The Changing Role of Funders in Responsible Research Assessment: progress, Obstacles and the Way Ahead', RoRI Working Paper 3. <https://rori.figshare.com/articles/report/The_changing_role_of_funders_in_responsible_research_assessment_progress_obstacles_and_the_way_ahead/13227914> accessed 5 February 2024.

Curry, S., Gadd, E., and Wilsdon, J. (2022) 'Harnessing the Metric Tide: indicators, Infrastructures & Priorities for UK Responsible Research Assessment', *Research on Research Institute Report*.

CWTS. (2021) 'Halt the H-index', Leiden Madtrics. <https://www.leidenmadtrics.nl/articles/halt-the-h-index> accessed 12 May 2023.

Davies, S. R. (2019) 'An Ethics of the System: Talking to Scientists about Research Integrity', *Science and Engineering Ethics*, 25: 1235–53.

Davies, S. R., and Horst, M. (2015) 'Responsible Innovation in the US, UK and Denmark: Governance Landscapes', in: *Responsible Innovation 2: Concepts, Approaches, and Applications*. Springer.

Davies, S. R., and Lindvig, K. (2021) 'Assembling Research Integrity: negotiating a Policy Object in Scientific Governance', *Critical Policy Studies*, 15: 444–61.

Dawson, D., Morales, E., Mckiernan, E. C., Schimanski, L. A., Niles, M. T., and Alperin, J. P. (2022) 'The Role of Collegiality in Academic Review, Promotion, and Tenure', *Plos One*, 17: e0265506.

DORA. (2013). *The Declaration*. <https://sfdora.org/read/> accessed 29 Aug 2022.

Dorbeck-Jung, B., and Shelley-Egan, C. (2013) 'Meta-Regulation and Nanotechnologies: The Challenge of Responsibilisation within the European Commission's Code of Conduct for Responsible Nanosciences and Nanotechnologies Research', *Nanoethics*, 7: 55–68.

Science Europe. (2019). 'Research Assessment in the Transition to Open Science'.

FOLEC-CLASCO. (2021) *The Latin American Forum on Research Assessment*. <https://www.clacso.org/en/folec/> accessed 17 Apr 2023.

Geschwind, L., and Broström, A. (2015) 'Managing the Teaching–Research Nexus: Ideals and Practice in Research-Oriented Universities', *Higher Education Research & Development*, 34: 60–73.

Glerup, C., Davies, S. R., and Horst, M. (2017) '"Nothing Really Responsible Goes on Here": scientists' Experience and Practice of Responsibility', *Journal of Responsible Innovation*, 4: 319–36.

GYA (2021) *Research Assessments that Promote Scholarly Progress and Reinforce the Contract with Society*. Global Young Academy.

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.

Hargens, L. L., and Schuman, H. (1990) 'Citation Counts and Social Comparisons: Scientists' Use and Evaluation of Citation Index Data', *Social Science Research*, 19: 205–21.

Hatch, A., and Curry, S. (2020) 'Changing How We Evaluate Research is Difficult, but Not Impossible', *Elife*, 9: e58654.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015) 'Bibliometrics: The Leiden Manifesto for Research Metrics', *Nature*, 520: 429–31.

LERU (2022) *A Pathway towards Multidimensional Academic Careers: A LERU Framework for the Assessment of Researchers*. League of European Research Universities.

Leydesdorff, L., Wouters, P., and Bornmann, L. (2016) 'Professional and Citizen Bibliometrics: complementarities and Ambivalences in the Development and Use of Indicators—a State-of-the-Art Report', *Scientometrics*, 109: 2129–50.

Ma, L. (2021) 'Metrics as Time-Saving Devices', in: F. Vostal (ed.) *Inquiring into Academic Timescapes*. Emerald Publishing Limited.

Ma, L., and Ladisch, M. (2019) 'Evaluation Complacency or Evaluation Inertia? A Study of Evaluative Metrics and Research Practices in Irish Universities', *Research Evaluation*, 28: 209–17.

Mckiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., and Alperin, J. P. (2019) 'Use of the Journal Impact Factor in Academic Review, Promotion, and Tenure Evaluations', *eLife*, 8: e47338.

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P., and Goodman, S. N. (2018) 'Assessing Scientists for Hiring, Promotion, and Tenure', *PLoS Biology*, 16: e2004089.

Müller, R., and De Rijcke, S. (2017) 'Exploring the Epistemic Impacts of Academic Performance Indicators in the Life Sciences',. *Research Evaluation*, 26: 157–68.

Narin, F. (1976) *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Citeseer.

Owen, R., Pansera, M., Macnaghten, P., and Randles, S. (2021) 'Organisational Institutionalisation of Responsible Innovation', *Research Policy*, 50: 104132.

Penders, B. (2022) 'Process and Bureaucracy: Scientific Reform as Civilisation', *Bulletin of Science, Technology & Society*, 42: 107–16.

Pontika, N., Klebel, T., Correia, A., Metzler, H., Knoth, P., and Ross-Hellauer, T. (2022) 'Indicators of Research Quality, Quantity, Openness, and Responsibility in Institutional Review, Promotion, and Tenure Policies across Seven Countries', *Quantitative Science Studies*, 3: 888–911.

Reymert, I. (2021) 'Bibliometrics in Academic Recruitment: A Screening Tool Rather than a Game Changer', *Minerva*, 59: 53–78.

Rice, D. B., Raffoul, H., Ioannidis, J. P., and Moher, D. (2020) 'Academic Criteria for Promotion and Tenure in Biomedical Sciences Faculties: cross Sectional Analysis of International Sample of Universities', *BMJ*, 369: m2081.

Rip, A. (2006) 'Folk Theories of Nanotechnologists', *Science as Culture*, 15: 349–65.

Rip, A. (2020) 'Technology and Evolving and Contested Division of Moral Labour', in: B. Beck, and M. Kühler (eds), *Technology, Anthropology, and Dimensions of Responsibility*, 23–32. Dordrecht: Springer.

Ross-Hellauer, T., Klebel, T., Knoth, P., and Pontika, N. (2023) 'Value Dissonance in Research(Er) Assessment: individual and Perceived Institutional Priorities in Review, Promotion, and Tenure', *Science and Public Policy*, 1–15.

Rushforth, A., and de Rijcke, S. (2015) 'Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in The Netherlands', *Minerva*, 53: 117–39.

Rushforth, A., and de Rijcke, S. (2017) 'Quality Monitoring in Transition: The Challenge of Evaluating Translational Research Programs in Academic Biomedicine', *Science and Public Policy*, 44: scw078.

Rushforth, A., and Hammarfelt, B. (2024) 'The Rise of Responsible Metrics as a Professional Reform Movement: A Collective Action Frames Account', *Quantitative Science Studies*, 1–19.

Schmidt, R., Curry, S., and Hatch, A. (2021) 'Creating SPACE to Evolve Academic Assessment', *Elife*, 10: e70929.

Schönbrodt, F. D., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., and Leising, D. (2022) 'Responsible Research Assessment I: Implementing DORA for Hiring and Promotion in Psychology', *PsyArXiv*. doi:10.31234/osf.io/rgh5b.

Stilgoe, J., Owen, R., and Macnaghten, P. (2013) 'Developing a Framework for Responsible Innovation', *Research Policy*, 42: 1568–80.

Sugimoto, C. R., and Larivière, V. (2018) *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

Sweet, P. L., and Giffort, D. (2021) 'The Bad Expert', *Social Studies of Science*, 51: 313–38.

TJNK, T. (2020) *Good Practice in Researcher Evaluation. Recommendation for the Responsible Evaluation of a Researcher in Finland*. Helsinki: The Committee for Public Information (TJNK) and Federation of Finnish Learned Societies (TSV).

UIR (2021) *NOR-CAM: A Toolbox for Recognition and Rewards in Academic Careers*. Oslo: Universities Norway.

UKRI (2022) *Future Research Assessment Programme*. <https://www.ukri.org/about-us/research-england/research-excellence/future-research-assessment-programme-frap/> accessed 21 Mar 2023.

UNESCO (2021) *Recommendation on Open Science*. Paris: UNESCO and Canadian Commission for UNESCO.

VSNU, NFU, KNAW, NWO, and ZONMW (2019) *Position Paper 'Room for Everyone's Talent'*. The Hague: NWO.

Wilsdon, J. (2016) *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. London: HEFCE.

Wouters, P. (2014) 'The Citation: From Culture to Infrastructure', in: B. Cronin, and C. R. Sugimoto (eds), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, 47–66. Cambridge, MA: MIT Press.

Wouters, P. F. (1999) *The Citation Culture*. Amsterdam: Universiteit van Amsterdam.

Zuijderwijk, J., Dix, G., and Benedictus, R. (2019) The Evaluative Breach. *WCRI*, Hong Kong.