Independent review of the role of metrics in research assessment

Summary of responses submitted to the call for evidence

Background

1. In April 2014, the Minister of State for Universities and Science launched a review of the role of metrics in research assessment. The review is chaired by Professor James Wilsdon (University of Sussex) and supported by an independent steering group, with HEFCE providing the secretariat and project management.

2. In 2008-09, HEFCE ran a pilot exercise to test the potential for using bibliometric indicators of research quality in the Research Excellence Framework (REF). At that time, it was concluded that citation information was insufficiently robust to be used formulaically or as a primary indicator of quality, but that there might be scope for it to inform and enhance processes of expert review.

3. The 2014-15 review will build on the previous pilot exercise to explore the current use of metrics for research assessment, consider the robustness of metrics across different disciplines, and assess their potential contribution to the development of research excellence and impact.

Context for the review

4. The quality and diverse impacts of scientific and scholarly research are commonly assessed using a combination of peer review and a variety of quantitative metrics. Peer review is the most established method of research assessment, and underpins the academic system in the UK and internationally. The use of metrics is a newer approach, but has developed rapidly as a potential method of measuring research quality and impact in some fields, though how best to do this is still the subject of considerable debate.

5. Metrics include the analysis of journal articles and their citations using a range of bibliometric methods, and have more recently expanded to include analysis of a more diverse range of research outputs. In addition, a growing array of social media and web-based alternative metrics have developed with potential to capture relevant dimensions of quality or impact. With the increasing capacity for real-time analysis based on large, linked datasets (‘big data’), some think that metrics could play an increasing role in the assessment of research.

6. Metrics may already be in use, or may be used in the future, either formally or informally, to inform judgements of research quality that are made inside higher education institutions (HEIs) for their own management purposes around human resources and research strategy development, and in decisions by funding agencies.

Call for evidence

7. The independent steering group issued a call for evidence to gather views and evidence from a wide range of sources on the use of metrics in research assessment. The call was open from 1 May to 30 June 2014. Respondents were asked to focus on four key issues:

- identifying useful metrics for research assessment
- how metrics should be used in research assessment
- 'gaming' and strategic use of metrics
- international perspectives.

8. A list of the questions posed in the call for evidence is provided at Annex 1.

Level and profile of responses

9. A total of 153 responses were received to the call for evidence; 67 (44 per cent) from HEIs; 42 (27 per cent) from individuals; 27 (18 per cent) from learned societies; 11 (7 per cent) from 'providers'[1]; three (2 per cent) from mission groups; and three (2 per cent) from other respondents.

10. The questions posed in the call for evidence were directly addressed by 139 of the respondents. The remainder identified relevant peer-reviewed literature, or online publications and blog-posts on the use of metrics for research assessment and management.

Analysis of responses

11. The responses were analysed to identify key themes, both in direct relation to the questions posed by the review and more generally. Additionally, significant individual comments and recommendations have been identified.

Headline themes

12. Where it was possible to classify views, 57 per cent of the responses expressed overall scepticism about the further introduction of metrics into research assessment (28 HEIs, 24 learned societies, 12 individuals, three mission groups, three providers and three other respondents). A number of arguments were made for this: the most frequently cited was that the use of metrics could unfairly disadvantage some disciplines, particularly in the arts, humanities and social sciences (35 responses: 23 HEIs, seven individuals, three learned societies, one mission group and one provider). Some responses felt that in a number of disciplines the use of metrics would not be possible at all (specific examples included law, English literature, nursing and criminology). A small

[1] The term 'providers' covers a range of publishers, consultants, and strategy companies.

number of responses (eight in total) also expressed concern that the use of discipline-led metrics could unfairly disadvantage interdisciplinary research.

13. Just under a fifth of responses supported the increased use of metrics in assessing research (11 HEIs, six individuals and six providers) and a quarter were ambivalent (20 HEIs, nine individuals, one learned society and one provider)[2]. Many supportive or ambivalent responses expressed the view that, in context, robust metrics could enhance the research assessment process, for instance suggesting that strong metrics could make it easier to establish a common understanding of 'what research is' across the subject disciplines.

14. Although not explicitly sought in the call for evidence, a common theme that emerged was that peer review should be retained as the primary mechanism for evaluating research quality. Both sceptical and supportive responses argued that metrics must not be seen as a substitute for peer review (26 responses, 13 of which were from learned societies), which should continue to be the 'gold standard' for research assessment. Sceptical responses generally argued that metrics could never become a realistic substitute for peer review, while many supportive responses stated that robust metrics could support peer reviewers in making nuanced judgements about research excellence. Many responses argued that changes should only be made to the established methods for research assessment where they could be demonstrated to provide improvement. However, it was recognised that peer review is not without its own flaws or disadvantages, and suggestions for its improvement included increasing the transparency and representativeness of the process.

15. Many responses (22 in total: 15 HEIs, three individuals, three learned societies and one mission group) also argued that metrics are not robust enough at present to provide an adequate assessment of research quality (though a handful of responses noted that metrics have been improving in recent years). For example, 20 responses expressed concern about the reliability of journal impact metrics as a proxy for article quality, on the basis that journals perceived as 'high-quality' may still publish 'poor-quality' articles, and vice versa. Many responses also noted that using the number of citations as a research assessment metric could be problematic, because a piece of academic work may be frequently cited for negative reasons, such as its controversy, or through self-citation.

16. Several responses expressed concern that the principle of seeking to incorporate metrics into research assessment risked overlooking not only the quality but the purpose of research. Several responses felt that metrics could not capture the 'richness' of research quality, for instance arguing that 'the focus of HEFCE should be the less tangible concept of "knowledge"' (Tim Johnson, individual response). Several also expressed concern that the use of metrics would discourage risk-taking in research: for example, 'research inevitably has blind alleys and academics should not be punished for taking one of these' (N Hunt, individual response).

[2] A small number of responses did not express any opinion on the advisability of using metrics in research assessment.

Identifying useful metrics for research assessment

17. As many of the responses to the review were sceptical about the usefulness of metrics in research assessment in general, most did not advocate the utility of specific metrics. However, several metrics were discussed within responses, including the number of citations, the number of research outputs, staff and postgraduate research student numbers, and the amount of grant income or number of funding awards. Some responses expressed a view on the most effective metric criteria in assessing research quality. An example is the following:

> *'Data demonstrates that to accurately predict peer review, the required scientometric data include: journal article publications, book publications, conference-proceeding publications, citation data, honorific awards (esteem indicators bestowed), research grants (and pounds) won, and several key pieces of meta data. It is also important to measure "academic age" of scholars, levels of collaboration, success of research scholars, and so on.'* (Academic Analytics, provider response)

18. Many responses questioned the usefulness of citation numbers as a metric, when used as a proxy for quality (with more frequently cited articles viewed as being of higher quality). Several responses felt that such a proposal did not take account of differences between disciplines: for example, in some disciplines it is more common to produce long-form monographs than frequent articles, which would significantly reduce citations numbers in these fields.

19. Many responses cited Scopus (17 responses), Google Scholar (15 responses) and the Web of Science (18 responses), or a combination of these, as commonly used sources of bibliometric data, but there was no clear consensus as to which were most effective in providing robust metrics for research assessment, with some responses explicitly critical of the robustness of these sources' information.

20. Five respondents (four HEIs and one provider) , three of whom are part of the initiative, advocated for the usefulness of the Snowball Metrics Programme, which aims to develop metrics that complement peer review and expert opinion, and to normalise metrics by discipline to ensure the comparability of data and quality. A few respondents also recommended the use of Higher Education Statistics Agency data as a possible metric.

21. Several responses (19 in total, of which 15 were HEIs) proposed that altmetrics could be used as a research assessment tool; however, 12 responses (of which eight were HEIs) argued that altmetrics are not reliable enough to be used as a measure of research quality.

How metrics should be used in research assessment

22. Many responses argued strongly that metrics should not replace peer review in the research assessment process. Many expressed concern that current metrics were not able to capture the nuance and richness of research quality and that metric-based assessment did not take account of differences between or even within disciplines, for example:

> *'[In mathematics] it is much easier to write journal articles in the sub-discipline numerics than in the sub-disciplines algebra or analysis, because algebra or analytic papers contain proofs, and proofs must be logically complete and correct (otherwise the proof simply does not exist). Therefore numericists have longer publication lists than analysts or algebraists with comparable academic age.'* (Dr Michael Dreher, individual response)

23. Several responses argued that if metrics were increasingly introduced into research assessment, it would be necessary to develop a 'basket' of metrics which could vary in their use by discipline and subject area. Some also suggested that if some form of metrics were introduced into research assessment, there would need to be a process of data normalisation, to ensure that metrics accurately reflected different disciplinary practice.

24. Some responses which supported the use of metrics in research assessment made specific recommendations about how metrics could complement current research assessment. For instance, 'there are many new (and old) sources of data that will be valuable in providing quantitative and qualitative evidence in supporting evaluative and resource allocation decisions associated with research assessment' (Public library of Science, provider response). A handful of responses suggested that citation numbers could be used to support peer reviewed research assessment, while one suggested that a judicious use of journal impact factors could be justified for the next REF or its equivalent (Anglia Ruskin University, HEI response).

25. Although a few expressed scepticism that this would actually be the case, some respondents acknowledged that greater use of metrics had the potential to reduce burden in research assessment. For instance, 'greater use of quantitative evidence (metrics) could be seen as a fairer and more objective method of assessment [because] metrics are arguably more transparent than peer review as the basis for the score/grading can be verified independently' (University of Southampton, HEI response).

'Gaming' and strategic use of metrics

26. Concern was expressed by 18 respondents that the increased use of metrics in research assessment would lead to increased 'gaming' of research assessment; specific concerns about metrics included making such strategic behaviours easier, and distorting individual behaviour. However, a few responses which were supportive of metric use argued that there is no evidence that metrics encourage gaming behaviours any more than any other form of research assessment.

27. Concerns about the risks of metrics being 'gamed' mostly focused on individual behaviours: for example, the risk of academics deliberately increasing their self-citation to boost their citation numbers. A handful of responses gave anecdotal reports of so-called 'citation clubs', with groups of academics agreeing to cite one another in a circular fashion to increase citation count. However, a few responses also questioned whether metrics use might encourage gaming among institutions, and also among publication journals.

28. A handful of responses argued that if metrics were introduced into research assessment, it would be necessary to 'police' gaming, in order to ensure that metrics

fairly represented research quality. Methods of discouraging gaming suggested included eliminating self-reported data elements from metrics, and ensuring maximum transparency in the metrics process.

29. A handful of responses objected strongly to the use of the word 'gaming' in the review literature, arguing that the use of metrics might well influence the behaviour of both institutions and individuals, who would naturally seek to improve their metrics, but that this was not the same as seeking to manipulate results to give an inaccurate impression of research quality.

International perspectives

30. Most respondents to the review focused on metrics in the context of the UK system of research and research assessment, but a few noted examples of metric use in international systems. Several respondents mentioned Excellence in Research for Australia's use of journal ranking as an indicator of research quality[3], and one mentioned the Norwegian BRI: one response noted that the approach in Australia had been to use peer review in some fields and metrics in others, to reflect disciplinary differences. It was also noted that international comparisons of metrics would be very difficult (excluding citation measures), because data are not collected on a comparable basis with the UK.

31. A few respondents expressed concern that the introduction of metrics would disadvantage the UK in comparison with other countries – for example, if UK academics were discouraged from engaging in international collaboration because it would fall outside the scope of metric assessment. One response also noted that the kinds of metrics currently under consideration were based on a citation system designed in the developed world, and would therefore be 'widely regarded as likely to disadvantage the developing world' (Million+, mission group response).

32. One respondent noted the particular disciplinary issues of international collaboration for linguistics:

> *'Linguistics is a highly international discipline, for obvious reasons, due to collaboration between UK researchers and native speaker linguists of many languages. Future REF evaluation mechanisms need to be sufficiently flexible to allow [the] outputs and activities of such international collaborations to be captured.'* (Linguistics Association of Great Britain, learned society response).

Equality and diversity

33. Many responses expressed concern that the increased use of metrics would specifically disadvantage under-represented groups in research, in particular early-career researchers (24 responses) and women (24 responses). It was noted that a system of metrics assessment which was based on citation numbers was likely to favour more established researchers, as they would have had more time to produce articles and other outputs, and would therefore have a negative impact on younger and newer researchers, or those who had taken career breaks. Several responses noted that there is some research evidence to suggest that women researchers are less likely to be cited than

[3] Australia's use of journal rankings was halted before the second exercise in 2012 – see www.theaustralian.com.au/higher-education/story-e6frgcjx-1226065864847

their male peers, and also less likely to use self-citation, which means they are disadvantaged by citation metrics. One response in particular argued that HEFCE should provide clear evidence of how metrics would impact upon women and under-represented groups before implementing any metrics-led research assessment (University of Bristol, HEI response).

34. A handful of responses (eight in total) pointed out that the predominance of English-language publications in most academic fields meant that non-English speaking academics were likely to be disadvantaged in terms of citation numbers, regardless of the quality of their research.

Recommendations (including for the next REF, or its equivalent)

35. Many discipline-specific responses, especially for the arts, humanities and social sciences, explicitly stated that they did not feel that metrics could be adequately used as a research assessment tool within their field.

36. Several responses, both sceptical and supportive, stated that they would welcome a thorough review of the role of metrics in the REF, both past and present. Comments on the nature of a review of research assessment included the following:

> *'An analysis of REF 2014 impact case studies might present a picture of whether there are in fact a set of metrics which could be applied to impact; nevertheless whether these could be collected in a way which is any less time consuming than preparing a case study is clearly too early to say.'* (University of Hertfordshire, HEI response)

> *'A full and strategic review would include all funders and understand the inter-relationship and dynamics of a system that extends far beyond the UK. Such a review might look at how the REF will direct funds to university researchers and feed into aggregate measures of institutional worth.'* (Academy of Social Sciences, learned society response)

37. Several responses called for a lighter-touch REF in general, and highlighted the increased burden that producing impact case studies for REF 2014 had placed on institutions. In addition to scepticism about the general use of metrics to accurately capture research quality, several responses argued that the wording of the review conflated impact and quality, and disputed that these were automatically the same thing in terms of research assessment.

38. More than half of respondents (57 per cent) stated that they would be willing to engage with a HEFCE-led workshop on the future role of metrics in research assessment.

Research Policy Team

HEFCE

October 2014

Annex 1: List of questions

Identifying useful metrics for research assessment

What empirical evidence (qualitative or quantitative) is needed for the evaluation of research, research outputs and career decisions?

What metric indicators are currently useful for the assessment of research outputs, research impacts and research environments?

What new metrics, not readily available currently, might be useful in the future?

Are there aspects of metrics that could be applied to research from different disciplines?

What are the implications of the disciplinary differences in practices and norms of research culture for the use of metrics?

What are the best sources for bibliometric data? What evidence supports the reliability of these sources?

What evidence supports the use of metrics as good indicators of research quality?

Is there evidence for the move to more open access to the research literature to enable new metrics to be used or enhance the usefulness of existing metrics?

How should metrics be used in research assessment?

What examples are there of the use of metrics in research assessment?

To what extent is it possible to use metrics to capture the quality and significance of research?

Are there disciplines in which metrics could usefully play a greater or lesser role? What evidence is there to support or refute this?

How does the level at which metrics are calculated (nation, institution, research unit, journal, individual) impact on their usefulness and robustness?

‘Gaming’ and strategic use of metrics

What evidence exists around the strategic behaviour of researchers, research managers and publishers responding to specific metrics?

Has strategic behaviour invalidated the use of metrics and/or led to unacceptable effects?

What are the risks that some groups within the academic community might be disproportionately disadvantaged by the use of metrics for research assessment and management?

What can be done to minimise ‘gaming’ and ensure the use of metrics is as objective and fit-for-purpose as possible?

International perspective

In addressing the issues and questions above, please include relevant evidence and examples from outside of the UK, where appropriate.

Additional question

Would you be interested in participating in a workshop/event to discuss the use of metrics in research assessment and management?