February 2017

Metrics in REF2021: advice from the UK Forum for Responsible Research Metrics

The Forum for Responsible Research Metrics has produced advice for the UK HE Funding Bodies on the use of quantitative indicators in the assessment of *outputs* in REF2021 (with further discussion planned later for assessment of *impact* and *environments*). This advice is given in response to some questions circulated by HEFCE as a basis for discussion at the Forum's first meeting, and the questions are given again here with our responses.

Q1: The extent to which different panels may benefit from quantitative data on research outputs may vary. How should Lord Stern's recommendation that all panels are provided quantitative data be implemented? Should panels be allowed to individually choose to receive metrics? Or should other criteria be used to decide which panels receive metrics?

Panels should be allowed to choose whether to adopt metrics/indicators. Metrics are of too low value to be useful in some panels so should not be universally adopted. In these panels the risk of misuse is greater than the small added value in some fields. Decision should be made at panel level because answers vary by discipline. It is also important to have panel-level standard practices – rather than for individual panel members – to give clear information to HEIs.

A suggested route for this might be as follows:

1) In first half of 2018, as part of its wider guidance on submissions and panel guidelines, HEFCE or its successor (Research England) should adopt a clear position on the use of metrics in REF2021, informed by work that the Forum for Responsible Research Metrics will undertake over the next 12 months, and by the views of panel and sub-panel chairs (who will then be in place).
2) Within this overall framework, main panels should then provide more targeted and strategic guidance about the value and limitations of quantitative data in their area, taking a view at the higher level of e.g. Panel C/social sciences, as well as individual discipline level.
3) Then the sub-panels should decide, in light of A & B, what exact approach they want to adopt in their discipline(s).

More generally, a key recommendation of Stern is to ensure that the REF has a strong institutional focus, so with this in mind, metrics might usefully play a more significant role than they played in

REF2014, which was based on individual outputs being assessed by panels. Note also that producing any extra measures will incur additional costs, which would be unwelcome for the sector.

Q2: In REF2014, a key principle was that all panels received the same citation data in the same format (an attempt to give Google Scholar data to Computer Science and Informatics fell through). Do you think it is important to maintain this principle or should alternative sources of data be considered for individual panels? If the latter, how might we ensure that different metrics are used consistently across the panels?

Yes, all panels should get the same data for simplicity of the process. Although computer science requested Google Scholar data for REF2014, this may be spammable and should not be used. In the arts and humanities: Book-based citation sources are more relevant but these are not comprehensive enough from Scopus and the Web of Science. The main alternative source of book citation data is not robust enough (Google Books) and field benchmarking is not developed for books so there is no real prospect of book-based citation indicators.

Alternative web indicators (altmetrics, web metrics, download indicators, etc.) are all spammable and should never be used for REF-style evaluations of outputs. For impact, there may be some scope for altmetrics to inform assessement, and the REF team should take a view on whether to actively prohibit its inclusion or allow it on a discretionary basis and then provide panel guidance on how it is used appropriately. For example, case study authors are often encouraged where possible to quantify elements of their impact narrative (exactly how many people attended that exhibition? How many schools did you visit and give talks at? etc). Altmetrics allow for much richer data on non-scholarly uptake, as some kind of proxy for impact - e.g. “there were 140,000 copies of my policy report downloaded in the first three months” – and in some ways this may be helpful. But there’s also a risk that if more of such data is available (which it clearly will be for REF2021, in ways it wasn’t before) those case studies that include it start to get viewed (even implicitly or unconsciously) by reviewers as “more robust” or “better-corroborated”. So some general and then main panel-specific guidance may be needed here, which we feel the Forum might be willing to help develop.

A possible panel difference would be to exclude conference papers from benchmark citation data in fields that submit few conference papers? In REF2014 all panels benchmarked their data against journal articles and conference papers in Scopus. This showed, for example, whether the citation counts of articles were above or below average for the publishing field and year, and whether they were in various top percentiles. The inclusion of conference papers within the benchmark set is probably only relevant for disciplines where conferences are important, and the REF team could ask panels for their view on this. A logical solution would be to either exclude all conference papers from the benchmarking set (for consistency) or to only allow them for these disciplines (for their field-specific value).

Q3: Is there a role for the panels in determining what data they receive? How might the panels be involved in the development of the final approach to metrics in REF2021?

Individual panel-wide agreements of the meaning of metrics and acceptable uses will be needed. These agreements should be created in conjunction with citation analysis experts to warn against over-interpretation and dispel common myths about citation analysis. Panels may also help decide between citation data providers in terms of their field coverage. Panels should also help to decide whether to include conferences in the benchmarking data (see Q2).

Q4: What quantitative data should we provide? Should metrics be provided in the same way as for REF2014 (citation counts for each article and field/year-specific averages and percentiles) or should field-normalised indicators be provided instead, or in addition?

Background on field normalised indicators: Field-normalised citation counts typically use a formula to transform individual or sets of citation counts to so that the result is 1 when the article has the average number of citations for its field and year, or greater than 1 when it has above average citation impact. The most well-known is the Mean Normalised Citation Score (MNCS), which, for each article divides it by the average citation count of all articles from the same field and year:

MNCS: citation count transformation for each article: $citations/\overline{citations}$. The average of all transformed citation counts in the MNCS[1]

A better metric is the Mean Normalised Log-transformed Citation Score (MNLCS) because citation count data is highly skewed and so log-normalisation is needed to get reasonable results.

MLNCS: citation count transformation for each article: $\ln(1+citations)/\overline{\ln(1+citations)}$. The average of all transformed citation counts in the MLNCS. [2]

An alternative indicator proposed by a medical expert is the Relative Citation Rate (**RCR**).[3] This targets a problem with most existing field normalised indicators that they rely upon subject classifications from (usually) Scopus or the Web of Science, which are imperfect. This classifies each article's field through its co-citation network (papers connected to it by citation relationships). Evaluations are needed to assess the problems that are likely to arise with taking this approach to

[1] Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011a). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37-47.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011b). Towards a new crown indicator: An empirical analysis. Scientometrics, 87(3), 467-481.
[2] Thelwall, M. (in press). Three practical field normalised alternative indicator formulae for research evaluation. Journal of Informetrics. http://arxiv.org/abs/1612.01431
[3] B. Ian Hutchins, Xin Yuan, James M. Anderson, George M. Santangelo (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002541

define a field. The calculation is also not transparent as currently formulated, and so cannot be checked by institutions (a requirement of REF2014). The calculation also does not address the issue of skewed distributions for article citation counts and so it not useful for aggregate citation counts.

Any switch to a field-normalised indicator adds to complexity of the information and increase the risk that is it is misinterpreted. The current approach of the provision of benchmarking data seems adequate in the context of supporting judgements about individual outputs, as in REF2014. Despite this opinion, we should probably ask panel members if it is worth it for them. . On the one hand, there is the risk of misinterpretation, but on the other hand, panel members may not fully appreciate the importance of taking benchmark data into account. Essentially field-weighting incorporates the benchmarking data into a single number.

Much of the above would also depend on whether the panels will only use the bibliometric data on a paper-by-paper basis. If any data is used to help inform research group judgements (perhaps as part of the environment assessment), some sort of field normalised citation count could prove even more helpful.

Q5: REF2014 documents provided succinct guidance on the use and non-use of metrics (see Annex A). Is this level of guidance sufficient? If not, what additional guidance or contextual information might be required to ensure that they are used appropriately by the panels?

*[**Example guidance from MPA**] 20. Citation data will inform the assessment as follows:*

*a. Where available and appropriate, citation data will be considered as a **positive indicator of the academic significance of the research output**. This will only be one element to **inform peer-review** judgements about the quality of the output, and will not be used as a primary tool in the assessment.*

b. The sub-panels recognise that the citation count is sometimes, but not always, a reliable indicator. They are also aware that such data may not always be available, and the level of citations can vary across disciplines and across UOAs. Sub-panels will be mindful that citation data may be an unreliable indicator for some forms of output (for example, relating to applied research) and for recent outputs. Sub-panels will take due regard of the potential equalities implications of using citation data.

*c. Sub-panels will use citation data only where provided by the REF team, and will **not refer to any additional sources of bibliometric analysis**, including journal impact factors.*

The REF2014 was good but a couple of refinements might help.

1) The Forum was split on whether the assertion in the guidance that "sub-panels will take due regard of the potential equalities implications of using citation data" was useful in practice. On the one hand, this may not be practical for individual outputs and so it could be deleted in favour of explicitly ruling out the use of career-based metrics, such as total citation counts or h-index, as these have equal opportunities implications – e.g., part time working, disability and fulltime childcare reduce h-indexes and career citation counts. On the other hand, evidence for REF2014 suggested that there was a paper-by-paper effect[4], with studies finding both the presence[5] and absence[6] of gender bias in disciplines.
2) At least one panel member claimed to have used Journal Impact Factors in a banned way so the existing cautions against using them could perhaps be strengthened.

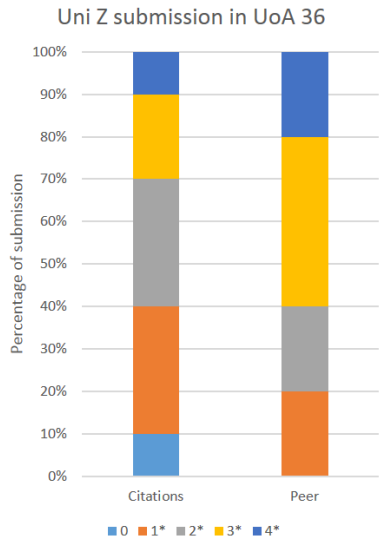
A more general observation that the information provided could to be expanded to include explicit reference to the wider debates that have taken place across the sector over the past 5 years, which make awareness and sensitivity to these issues far more important. The framework developed by the Wilsdon Review for responsible metrics may be helpful here, as would the principles in the Leiden Manifesto, if the implications in a REF2021 context were spelled out.

Q6: Lord Stern discusses the potential use of metrics to cross-check overall output profiles. What is the forum's response to this? Should it be optional or compulsory for panels? What kind of contextual information or expert bibliometric advice might be provided to help interpretations?

Cross-checking profiles might take the form of comparing the range of output scores for a submission against the scores predicted by (field normalised) bibliometrics (e.g., as in the graphs below) in order to check for anomalies. This seems like a useful exercise – especially to double check that areas of high quality research have not been overlooked. If this creates issues with identifying grade boundaries in the bibliometric analysis, an alternative suggestion would be to use bibliometric indicators to rank submissions and compare that to peer review rankings (i.e. one would need to do %4*, %4*+3*, GPA) and look for major discrepancies.

It may also be useful to check scoring for interdisciplinary outputs, with any cross-referred outputs to be recorded in order that an analysis can be done of whether interdisciplinary outputs are more/less likely to be cited, and whether this matches up with panel member scores.

[4] http://www.hefce.ac.uk/media/hefce1/pubs/hefce/2011/1103/11_03.pdf
[5] https://www.cambridge.org/core/journals/international-organization/article/div-classtitlethe-gender-citation-gap-in-international-relationsdiv/3A769C5CFA7E24C32641CDB2FD03126A 10.1017/S0020818313000209
[6] https://www.nceas.ucsb.edu/meta/Koricheva/Leimu_Koricheva_2005_TREE.pdf



OR

Useful to double-check whether a successful *area* of research has been ignored