ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Royal Statistical Society response to the Higher Education Funding Council for England's independent review of the role of metrics in research assessment

Response prepared by Professor William J. Browne on behalf of the Royal Statistical Society.

0. **Background**

0.1. HEFCE is currently reviewing the role of metrics in research assessment in England while the current Research Excellence Framework exercise is proceeding. We presume this is with an eye to changes for the next REF exercise.

0.2 Statisticians like other discipline scientists have a keen interest in how metrics can be used within our own discipline – here primarily the mathematics Unit of Assessment. We are also in a discipline with a large degree of interdisciplinary collaboration and so we have members who are aware of the issues of the use of metrics in other disciplines. Finally and probably most importantly the field of quantitative data analysis and the use of metrics generally should sit within statistics, where the importance of concepts like uncertainty can be best understood. Therefore we are well placed both in the specifics of the use of metrics for our discipline, and the appropriate use of metrics more generally.

0.3 For this response we have reviewed some of the most relevant literature as well as consulting among some of our membership.

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

1.1. Generally a research assessment will contain both qualitative and quantitative evaluations, and it is necessary for the REF to be based on both types of evidence. The weighting given to quantitative metrics will differ across disciplines and across situations. Quantitative (metric based) evaluations of research, research outputs and career decisions synthesise numbers of papers, grants, citations etc. into some measure of quality. Qualitative evaluations involve more assessment by peers or discipline 'experts'. Expert assessment is partly based on their opinion of quantitative measures, but also involves consideration of the raw data. For example experts will look at individual outputs and give their opinion based on what they read, rather than simply based on where the paper is published and how often it has been cited.

1.2. Factors that could influence the evaluation of individuals' research outputs include:

1.2.1 *Level of 'specific' expertise of the expert panel*. If the discipline to be judged is wide then 'experts' may feel unable to judge some work and they will have to rely on metrics for assistance.

1.2.2 *The purpose of the research evaluation and the time available to conduct it*. When the purpose is to form a research profile for an institution with 1000 or more research outputs, the precision of an individual grading is less of an issue as the gradings are

merged to form the profile. However when research is evaluated to decide on a staff member's promotion, the precision of individual grading matters more.

1.2.3 *The precision of grading at the aggregate level, which needs to be taken into account when ranking*. For fairness, the required size of submission for each staff member should be independent of their institution. However consideration should be given by the panel to sampling fewer outputs than are currently submitted for large institutions, in order to achieve acceptable accuracy before aggregation. This would also lead to better comparability with smaller institutions. It may alternatively be desirable to carry out more detailed, and therefore presumably more accurate, gradings of individual outputs (for example using more detailed readings) from smaller institutions.

1.3 As has also recently been highlighted by the Expert Advisory Group on Data Access, research metrics should also explicitly recognise other forms of output such as reports, books, software and the production of new high quality datasets as a valuable research output (EAGDA, 2014). The REF allows for this, and within social science panels the value of new datasets seems already to be recognised.

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

2.1. We lack hard evidence that the use of journal impact factors, researcher profiles like h-indices, and to a lesser extent individual article citation rates are useful. Metrics like numbers of PhD students in groups and research grant income are often used to assess university departments or schools, and their usefulness seems more widely agreed.

2.2 Adler et al. (2009) and subsequent discussants criticise the use of journal impact factors and h-indices, and in particular the use of comparisons across disciplines and sub-disciplines. There is no reason that publication in a high-impact journal necessary means the article is of higher research quality. The use of the mean as a summary measure is also problematic, as journal citations have a long 'tail' and are not well summarised by the mean.

2.3. The individual article citation rate has the advantage of at least being directly related to the output concerned. There are however differing rates of citation in different disciplines and in different types of publications (review articles are cited more frequently for example). More recent publications have also had less time to be cited compared to older publications. An often raised issue is that of 'poor quality' papers being extensively cited. This is sometimes dismissed on the grounds that 'on average' it has little effect. It is important not to under-estimate the importance of even a very small number of such extreme mis-gradings, since, if discovered and publicised, these are likely to reflect badly on the whole exercise.

2.4 It is also important to be aware of the flaws with current providers of bibliometric data. Several publishers lead the field in terms of visibility and adoption of their data, for example Thomson ISI and Scopus. However these sources have a specific set of journals in their core set which do not cover all of disciplines such as mathematics, and thus do not accurately reflect the citation rates of papers (Adler et al., 2009).

3. **What new metrics, not available currently, might be useful in the future?**

3.1 There are many different types of metrics that could be devised, but all should be designed to reflect expert opinion in a given discipline. Adaptations are needed to 'scale' citation metrics according to what the norms are in a given discipline or sub-discipline. Varin et al.

(2014) for example look at constructing a measure that satisfies the ordering of specific journals based on expert opinion, while Bornmann et al. (2011) take account of differences in citation rates in different sub-disciplines in chemistry. Varin et al (2014: figure 2) use a hierarchical cluster analysis to identify clusters of journals. These to some degree represent common sub-disciplines, so the REF could consider normalising within groups of journals identified in this way, to put a citation rate in context. Time since publication and type of article would need to be factored in to these metrics. For example it should be easy for an expert to compare like for like, say to identify which articles are new research and which are review articles.

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

4.1 We would caution against cross-disciplinary comparisons in most cases, as it is very difficult to compare across disciplines that have very different publishing cultures. For example certain sub-disciplines of mathematics have lower publication rates which then result in lower citation rates and longer citation half-lives (Carey et al. 2007). It is therefore difficult to capture research quality in a metric approach in the timeframe that the REF exercise exists in. It might be feasible for disciplines with higher citation rates to make comparisons against other disciplines, if the REF sets out a method for norm referencing publications.

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

5.1 It is important for each 'discipline' to have some control of their use of metrics in research evaluation. This makes it necessary to group research according to discipline, but in some cases finer-grained distinctions seem necessary. For example individuals at the margins of a defined discipline, or those in marginal sub-disciplines, may not be represented on evaluation panels. Interdisciplinary researchers are also grouped under a discipline that is often but not exclusively governed by the department they belong to. For example medical statisticians might be returned under a medical unit of assessment and thus governed by the medical discipline's expected research culture. This may not always reflect well on their own research.

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

6.1 It is important to be aware of the flaws with current providers of bibliometric data. Several publishers lead the field in terms of visibility and adoption of their data, for example Thomson ISI and Scopus. However these sources have a specific set of journals in their core set which do not cover all of disciplines such as mathematics, and thus do not accurately reflect the citation rates of papers (Adler et al., 2009). Google Scholar may form a larger part of certain disciplines' REF returned outputs, for example in the social sciences. Google Scholar has a far wider scope and covers other forms of output such as reports, books, book chapters and software, however there seem grounds to question the reliability of its citation counts.

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

7.1 The use of metrics such as research income and numbers of PhD students often form part of the assessment of departments and seem to attract less criticism as indicators of research quality. There seems less hard evidence for the use of metrics with regard to journal articles and their impact factors.

8. **Is there evidence that the move to more open access to the research literature will enable new metrics to be used or enhance the usefulness of existing metrics?**

8.1 Although open access publishing will clearly influence university policy on publications to some degree it is not clear what impact it will have on metrics and their usefulness.

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

9.1 We would recommend the papers referenced on the final page of this document, which provide many useful examples.

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

10.1 Our summary view, based on papers cited here, are that metrics are useful as additional information in the assessment process and should complement, but not replace, expert opinion. The overall weighting assigned to such additional information should be a matter for individual disciplines to determine.

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

11.1 The use of metrics should be subject to some discretion from representatives of the specific disciplines, as they know their culture best. In mathematics (and statistics within the mathematics unit of assessment) there is a strong feeling that metrics are perhaps less useful than expert opinion. This is for reasons including that certain disciplines of mathematics have lower publication rates (Carey et al. 2007), and bibliometric sources do not reflect the citation rates of papers in the discipline (Adler et al., 2009).

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

12.1 Provided there is no systematic bias in the REF, aggregate metrics should broadly be favoured as they are generally more robust. For example Goldstein (2011) found that data could support institutional comparison (albeit showing large overlap in performance across many institutions) but not individual researcher comparison, in a study of research grant performance as attributed by reviewers of ESRC grants. However if the metrics are recorded simply at the higher level and are not formed from an aggregation of lower level outcomes, then there is more scope for ‘errors’ to have an effect i.e. an error in one lower level outcome will have little effect on an aggregate whilst an error on a outcome measured directly at the higher level could have a big effect.

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

13.1 It is perhaps hard to find documented evidence of gaming/strategic behaviour here but one has to assume that changes in research assessment do change the behaviour of

academics. In his discussion of Adler et al. (2009), Silverman mentions a 'boom-bust' mentality that existed prior and after the REF 2008.

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

14.1 It is generally recognised that any 'high stakes' system that encourages players to optimise their performance will change the nature of the system. Those promoting the REF should seek to monitor such effects and take them into account when making judgements.

15. **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?**

15.1 The risks are considerable and metrics should be used to support qualitative expert judgement rather than on their own. All metrics come with inherent uncertainty (Goldstein and Spiegelhalter, 1996, 2009). The uncertainty is often considerable, with the result that many of the differences found in comparisons are not statistically significant. Unavoidably, a major use of research assessment is the allocation of resources (money) to different research submissions. This requires multivariate assessments to be mapped to a single (univariate) scale so that resources can be allocated. What then happens next is that a similar mapping exercise is used to list submissions in a league table (generally without uncertainty estimates), which can skew the perception of whether differences are significant or not. Marginal sub-disciplines, disciplines with slower citation rate cultures, and those doing interdisciplinary work may be disadvantaged.

15.2 There also needs to be scope to recognise research excellence over different time frames (outside the specific time frame of the REF) to accommodate disciplines with longer publishing half-lives and also the ability to recognise 'academic impact'. REF2014 introduces a measure of research impact for the first time and here there is recognition that (non-academic) impact takes longer, so submissions can call on department work from a time frame before the current submission. Longer term citations can also often capture 'academic impact', i.e. the longer term influence of work both within the discipline and across disciplines. REF2014, as currently defined, places far less emphasis on 'academic impact'.

15.3 Considering the larger "half-life" of outputs in certain disciplines (i.e. time period until half the citations of an article have occurred), using solely outputs produced within the period of assessment may not be appropriate. Silverman has commented (to Adler et al. 2009) about the impact of the REF on researchers in different career stages. The REF has done much to try to be inclusive to younger career scientists (and career breaks) by allowing smaller numbers of outputs for certain groups. Silverman pointed out in the Stats panel of 2008 that a quarter of the faculty were new entrants.

16. **What can be done to minimise 'gaming' and ensure the use of metrics is as objective and fit-for-purpose as possible?**

16.1 Here there will inevitably be a conflict: metrics that make the assessment scheme more transparent will also make the scheme easier to 'game', and any 'high stakes' system that encourages players to optimise their performance will change the nature of the system.

Promoters of the REF must monitor these effects in order to take them into account when making judgements.

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

17. **Reference list**
In terms of evidence for this response we recommend in particular:
Adler, R., Ewing, J. and Taylor P. (2009) Citation Statistics (with discussion). *Statistical Science* 24, 1-14, and the various discussion contributions that follow from it.

We have also referred to:
Bornmann, L., Mutz, R., Marx, W., Schier, H., and Daniel, H-D (2011) A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? *J. Roy. Statist. Soc. Ser. A* 174: 857-879
Carey, A.L., Cowling, M.G. and Taylor P.G. (2007) Assessing research in the mathematical sciences. *Gazette of the Australian Maths Society*. 34: 84-89
Expert Advisory Group on Data Access (2014). *Establishing incentives and changing cultures to support data access* [Pdf], May 2014. London: Wellcome Trust. (Available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp056495.pdf)
Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. Roy. Statist. Soc. Ser. A* 159 385–443.
Goldstein, H. (2011). Estimating research performance by using research grant award gradings. *J. Roy. Statist. Soc. Ser. A* 174: 83-93
Hall, P. (2011). 'IMS Presidential Address by Peter Hall', 2 Sept 2011, *IMS Bulletin Online* (available at http://bulletin.imstat.org/2011/09/presidential-address-peter-hall/)
Varin, C., Cattelan, M. and Firth, D. (2014). *Statistical Modelling of Citation Exchange among Statistics Journals*. (available at http://arxiv.org/abs/1312.1794)
Reid, N. et al. (2012) *The Long Range Plan for Mathematical and Statistical Sciences Research in Canada 2013 – 2018* [website] (available at http://longrangeplan.ca/)

The RSS thanks the following Fellows for their contribution to this response:
- William J. Browne, Professor of Biostatistics, University of Bristol
- Harvey Goldstein, Professor of Social Statistics, University of Bristol
- Valerie Isham, Professor of Probability and Statistics, University College London
- Kevin McConway, Professor of Applied Statistics, The Open University
- Charles Taylor, Professor of Statistics, University of Leeds