Independent review of the role of metrics in research assessment: Call for evidence

Summary of areas for advice

1. This template provides a summary of the areas on which the steering group is seeking advice. You may use this template to respond with your advice if you would find it helpful.

2. Please send responses to **metrics@hefce.ac.uk** by **noon on 30 June 2014**. Independent review of the role of metrics in research assessment: Call for evidence

Independent review of the role of metrics in research assessment: Call for evidence	
Submission by UKCRC, UK Committee for Computing Research.	
Name: Ursula Martin	**Job title:** Professor of Computer Science
Organisation: University of Oxford	
E-mail: Ursula.Martin@cs.ox.ac.uk	**Phone:**

1. The UKCRC is an expert panel of the British Computer Society, the Conference of Professors and Heads of Computing, and the Institution of Engineering and Technology, for computing research in the UK. Its members comprise academic and industry research leaders.
2. Leads on this report for UKCRC are Professor Anthony Finkelstein FEng, University College London, and Professor Ursula Martin CBE, University of Oxford, who have had experience of REF/RAE both as panel members and in institutional leadership

Background

3. To frame our response we note that Computer Science, in the UK and internationally, is an exceptionally diverse and dynamic discipline, ranging from long lasting deep questions on the power and limits of computation, to rapid development and deployment of new technologies. A wide range of methodologies, experimental, mathematical, and qualitative, are breaking down disciplinary silos and traditional models of basic versus applied research. This diversity is reflected in the outputs, outcomes and economic and societal impacts of research: traditional journal papers; conference papers; software, services and devices; cultural artefacts; patents licenses and other forms of IP; and policy and regulatory material. Thus REF outputs include large suites of software, works of art and the like, alongside traditional journal and conference papers. Major universities enable vibrant research ecosystems – industry and government research labs, startups, incubators, venture funders and other intermediaries, with permeability between academia and industry vital for rapid exploitation, and fierce international competition for talented staff.
4. Computer science was among those disciplines arguing most strongly against the use of bibliometrics in the RAE and REF, drawing attention in particular to the variety of outputs of computing research, the very poor coverage of computer science

publications in standard commercial databases such as ISIS and SCOPUS, and the importance of interdisciplinary and applied work, often poorly cited, to the health of the discipline. In confirmation of this, only a third of the outputs submitted to the 2008 RAE in Computer Science had non-zero citation counts in SCOPUS, with the remainder either not listed at all, or listed with zero citations, as SCOPUS only counts citations within other articles listed in SCOPUS. HEIs report similar issues when preparing the 2014 REF submission. An international comparative study[6] by the Leiden group confirmed this, providing a detailed breakdown of the difficulties in bibliometric assessment of computer science, difficulties not overcome by techniques such as attempting to normalise between sub-disciplines with widely different patterns of citation behaviour, and compounded by the dynamic nature of a rapidly evolving discipline where new work may emerge that does not fit disciplinary silos.

5. In the light of the diversity of computer science research, and its importance for the economy and society, its assessment has been subject to much discussion among the international community. An authoritative and fully referenced study[7] by Informatics Europe, a federation of leading universities, concluded:

- Computer science is an original discipline combining characteristics of science and engineering. Researcher evaluation must be adapted to the specifics of the discipline.
- A distinctive feature of computer science publication is the importance of conferences, some of which are extremely selective, and books. Journal publication, while important for in-depth treatments of some topics, does not carry more prestige than top-quality conferences and books.
- An important part of computer science research produces artefacts other than publications, in particular software systems. In measuring impact, these artifacts can be as important as publications.
- In the computer science publication culture, the order in which a publication lists authors is generally not significant. In the absence of specific indications, it should not be used as a factor in evaluation.
- Publication counts, weighted or not, must not be used as indicators of research value. They measure a form of productivity, but neither impact nor research quality.
- Numerical impact measurements, such as citation counts, have their place but must never be used as the sole source of evaluation. Any use of these techniques must be subjected to the filter of human interpretation, in particular to avoid the many possible sources of errors. It must be complemented by peer review, and by attempts to measure impact of contributions other than publication.
- Any evaluation criterion, especially if it yields a quantitative result, must be based on clear and published criteria.
- Numerical indicators must not be used to compare research or researchers across different disciplines.
- In assessing publications and citations, the ISI Web of Science is inadequate for

[6] Developing Bibliometric Indicators of Research Performance in Computer Science: An Exploratory Study, 2007 http://www.cwts.nl/pdf/NWO_Inf_Final_Report_V_210207.Pdf

[7] Research Evaluation for Computer Science, 2008 http://www.informatics-europe.org/images/documents/research_evaluation.pdf

most areas of computer science and must not be used. Alternatives, imperfect but preferable, include Google Scholar, and (potentially) the ACM Digital Library.

- Evaluation criteria must themselves be subject to assessment and revision.

6. This is consistent with UK decisions on metrics and REF. Following a lengthy consultation and pilot exercises during the design of REF 2014, the notion of using metrics was rejected in favour of academic judgment of panel members, informed by bibliometrics where panels so chose. This was essentially due to concern that current metrics could not provide a transparent, accurate, accountable process at the proposed level of granularity: for example due to poor coverage by databases, edge effects in citation data, variations in behaviour between sub-disciplines, bias against interdisciplinary or industrial research and so on.
7. A particular concern was Equal Opportunities issues: HEFCE's own analysis[8] of the 2008 RAE results showed "Men were consistently more likely to be highly cited than women." (Para 10) and "the average number of papers per staff member for women was lower than that observed for men" (Para 45). Gender differences in publication and funding have been reiterated in a recent article in Nature[9].

Purpose of Research Assessment

8. In responding to the consultation we consider first the purpose and granularity of research assessment. Is it to establish the excellence of research, to show that institutions, disciplines or the nation as a whole reach some external standard when judged at local national or international scale? Or is it, in fact, intended as a practical means to determine the funding that UK universities will receive (QR), within the dual-funding scheme, to support research (though noting that it is not hypothecated)? Is research assessment intended to provide an authoritative indicator of research merit, a tool for delivering reliable league tables? Or is it, in fact, intended as a means to incentivise a particular set of behaviours by UK universities and to support a set of policy goals around research selectivity and concentration?
9. Excellence in research is fragile and evanescent - there can be no stable agreed judgment. Through the application of collective expert scrutiny some broad summative conclusions can be reached about a body of research but this is a necessarily complex, difficult and time consuming task. It is a matter that will always remain contested. There can be no consensus even on methodology. This is particularly true in a discipline like Computer Science, which thrives on radical innovation and breaking boundaries. Even the experts can get it wrong - for example Berners-Lee's paper proposing the web was rejected a number of times by referees.
10. By contrast it is generally recognized that it is necessary to make a rough and ready decision on government funding allocations, even by those who do not agree with the policies that funding allocation serves. Only when we separate, clearly in our minds, funding decisions from excellence, whatever that might mean, will we be able to have

[8] Analysis of data from the pilot exercise to develop bibliometric indicators for the REF The effect of using normalised citation scores for particular staff characteristics http://www.hefce.ac.uk/pubs/hefce/2011/11_03/11_03.pdf

[9] Bibliometrics: Global gender disparities in science Larivière,et al, *Nature* **504,** 211–213 (12 December 2013) http://www.nature.com/news/bibliometrics-global-gender-disparities-in-science-1.14321

a mature discussion of metrics.

The current framework – advantages and disadvantages

11. The current REF gives a transparent accountable process for computing two data points at the level of a UoA (discipline): the "research profile" (outputs, environment and impact), and the number of research active staff employed. The two points are linked in that a university submits four outputs within a time period for judgement for each "research active" member of staff, with appropriate protocols for reduced contribution due to early career status, personal circumstances and so on.
12. The first step must be an honest appraisal of the current system. On the positive side the current model seems to have served its purpose. It produces results that seem broadly accepted as a reasonable reflection of the research performance within the units of assessment. There are, of course, anomalies and injustices but they are not too egregious. It achieves a separation between academic judgment and policy makers while providing flexibility to political decision makers when the outcomes are translated into a funding formula. Arguably, the incentives have also led to an overall improvement in research performance.
13. **Costs of assessment** The negatives of the current model are however quite considerable. First and foremost it is difficult and expensive to run. The costs of actually performing the assessment are substantial, with the costs to HEFCE magnified by the costs to academics who are members of panels, and to institutions preparing submissions. Panel members achieve risible compensation for the considerable time devoted to the exercise that would otherwise be spent pursuing research. These costs pale into insignificance however, when the cost to universities in preparing submissions are factored in. Each institution making a submission must engage in a lengthy process adhering to complex rules and practices that vary subtly across subjects and require both academics and professional staff of universities a vast effort spread over at least two years leading up to submission. The welcome stress on Equal Opportunities and related matters has further increased the workload on institutions, requiring new HR mechanisms for "complex circumstances", and forsupporting those not submitted. The risk of error is high and the cost of errors, is potentially large. Institutions bear these costs to mitigate not just the financial risk of poor performance, but also the reputational risk as ratings feed into other league tables affecting matters such as student demand, thus multiplying the financial risk.
14. **Consequences within institutions** The importance of REF for institutions has, inevitably, affected management priorities. We hear increasing reports of REF driving priorities in hiring, probation, promotion, performance management and severance/redundancy policies, with the use of metrics being justified as an objective criterion to stand as proxy for likely panel decisions. Used at individual level this magnifies the weaknesses of metrics, for example gender or discipline bias. Thus, while HEFCE's rules for the REF encourage diversity of outputs, HEIs are explicitly driving behaviours towards the "safe " choices of highly rated journals and more academic work – thus diminishing exactly the kind of risky industrial or applied research that is likely to lead to innovation or commercial impact. The consequent decline in the more practical systems-based research is affecting computer science education, and international competitiveness.

Summary and recommendations

15. **Metrics are not appropriate for assessing computer science at a level of granularity suitable for judging Departments or individuals. This is confirmed by the thorough review by Informatics Europe cited above, augmented by evidence from RAE 2008 and institutions preparing for REF 2014**
16. **Metrics issues for further consideration**
    a. A particular issue for computer science, and other disciplines notably In the humanities, is the poor coverage of standard databases such as Scopus. Despite a commitment from HEFCE, it was not possible to reach contractual agreement with google over the use of google scholar for the 2014 REF. Resolving this would be a first step in providing more informative bibliometrics to augment human judgement.
    b. For the reasons indicated above it seems unlikely that it will ever be possible to remove the elements of human judgment required to take account of the full range of computer science research, especially as it impacts the economy and society. Nonetheless it might be useful to carry out a further investigation of ways to combine human judgment and metrics, allowing a group of experts in each subject to moderate and to weight elements of a basket of metrics, for example through using many sources of bibliometric data, altmetrics, grant income , PhD numbers and so on
    c. The three elements of the current REF are assessment of research; counting research active staff; and linking the two through assessing outputs associated to staff. It is the need for transparency, accountability and fairness in every detail of selecting staff and outputs that leads to much of the administrative overload for HEIs. If a basket of metrics approach could be developed, it might inform an alternative approach which used a metric based model of assessing research quality of a unit without such a close linkage to individuals or papers; and counted “research active” staff through data submitted to HESA by HEIs in an accountable and transparent manner.
17. **Responsible use of metrics** Notwithstanding the above, research metrics are increasingly being used in decision making. There is also increasing knowledge about the limitations of metrics, for example appropriateness of granularity, need for normalisation, and gender and other biases. Thus we propose that HEFCE, working with the Equality Challenge Unit, UUK, RCUK and other bodies,
    a. produce guidelines for the responsible use of metrics
    b. mandate that these are followed wherever metrics are used in decision making, and
    c. request bodies such as the Research Councils to publish a detailed analysis by gender and other groupings of their funding decisions.

    For example, HEFCE could require that In any rating activity based on metrics, evidence be provided, before the exercise commences, at the granularity at which they are to be used, of the impact of these metrics on women and under-represented groups.

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