Science and Public Policy, 34(8), October 2007, pages 538–542
DOI: 10.3152/030234207X256529; http://www.ingentaconnect.com/content/beech/spp

Introduction: Future pathways for science policy and research assessment: metrics vs peer review, quality vs impact

Claire Donovan

The idea for this special issue arose from observing contrary developments in the design of national research assessment schemes in the UK and Australia during 2006 and 2007. Alternative pathways were being forged, determined, on the one hand, by the perceived relative merits of 'metrics' (quantitative measures of research performance) and peer judgement and, on the other hand, by the value attached to scientific excellence ('quality') versus usefulness ('impact'). This special issue presents a broad range of provocative academic opinion on preferred future pathways for science policy and research assessment. It unpacks the apparent dichotomies of metrics vs peer review and quality vs impact, and considers the hazards of adopting research evaluation policies in isolation from wider developments in scientometrics (the science of research evaluation) and divorced from the practical experience of other nations (policy learning).

NATIONAL RESEARCH EVALUATION exercises are burgeoning and, until recently, one would have assumed this to be a time of stability, with established and emerging systems largely focused on quality assessment and modelled on: the UK Research Assessment Exercise (RAE); or a quantum informed by data on research funding, publications and higher-degree students; or a hybrid of these (von Tunzelmann and Mbula, 2003). Yet the UK and Australia have dramatically modified their approach to quality measurement. Curiously, both sets of research evaluation policies are premised on policy U-turns, albeit in opposing directions.

Post-2008, the UK is swapping its discipline and panel-based RAE for standard quantitative metrics. This will be the basis for allocating around £1.5 billion per year in block funding to the university sector and will be applied at the institutional level (although after fierce lobbying some form of peer review will be retained for the humanities, arts, social sciences, mathematics and statistics) (H M Treasury, 2006: 57).

In 2008, Australia is switching from an institutional-level metrics-only approach to a Research Quality Framework (RQF) aimed at research groups and based on the judgement of panels of expert peers and 'end users' (although a quantitative element will be retained as discipline-specific quality metrics will be provided to panels). The Australian process, which will allocate an estimated Aus$600 million a year, also includes a panel assessment of impact in the form of the social, economic, environmental and cultural returns of research beyond the academic peer community (DEST, 2007).

Australia's RQF courts novelty by embracing state-of-the-art trends in research evaluation towards more contextual assessments of quality (Butler and Visser, 2006; Moed, 2005) and socially embedded

Claire Donovan is in the Research Evaluation and Policy Project, Research School of Social Sciences, The Australian National University, Canberra, ACT 0200, Australia; Email: claire.donovan@anu.edu.au; Tel: +61 2 6125 2154; Fax: +61 2 6125 9767.

The authors of this special issue are grateful to Roy MacLeod, who encouraged the development of this project, Bill Page for his enthusiasm in taking on this special edition, and the referees who, in addition to their insightful comments on papers, considered issues raised by the collection as a whole and so helped to shape this introductory essay. Most thanks belong to the authors for their willingness to take aim at the constantly moving target of science policy and for entering into the spirit of engaging not only with colleagues in scientometrics and science governance but also with the broader academic community, research managers and evaluators and policy-makers.

Claire Donovan is a Research Fellow in the Research Evaluation and Policy Project, Research School of Social Sciences, The Australian National University. She previously held research posts at The Open University and Nuffield College, Oxford University. Her research focuses on social and political aspects of science, technology and innovation governance. She is a senior advisor to the Australian Government on evaluating the extra-academic returns (or public value) of university research, and is the author of a forthcoming book *The Governance of Social Science: New Foundations of a Science for Society* (Edward Elgar Publishing).

conceptions of research impact (FWF/ESF, 2007). Yet the UK's resolute pursuit of standard quality metrics is less 'messy' and more parsimonious. These major developments come from different ends of the Earth, both symbolically and literally. This perplexing inversion begs the core questions of this special issue: does the future of research evaluation rest with: metrics or peer review; the seemingly objective or the subjective; remote or embedded knowledge; serving disciplinary or societal ends?

Thus, at the end of 2007, we encounter major changes in the national research assessment landscape, and the hiatus between design and implementation provides an opportunity for reflection. It has been argued elsewhere (Donovan, forthcoming) that the politics of constructing research assessment exercises resemble a "Pushmi-pullyu" — the two-headed llama of Dr Doolittle fame that tries to travel in opposite directions at once: the government 'push' towards external audit is offset by a 'pull' towards internal peer-based appraisal; the 'push' towards broader relevance is met with a 'pull' towards scientific autonomy; and this 'push' is sometimes forcefully directed towards the interests of industry and commerce, yet counterbalanced by an equally strong 'pull' towards broader public benefits.

This special issue of *Science and Public Policy* asks how these tensions become manifest, and why such different pathways are being trodden and retrodden by the same beast. It unpacks the apparent dichotomies of metrics versus peer review, and quality versus impact, and considers the hazards of adopting research evaluation policies in isolation from wider developments in scientometrics (the science of research evaluation) and divorced from the practical experience of other nations (policy learning).

The collection offers often radically different perspectives on how quantitative and qualitative approaches to research evaluation act as filters that connect the aims of science policy with the perceived value of research outcomes, be this tied to notions of scientific excellence or usefulness, and from the viewpoint of policy-makers and research-funding agencies, the scientometrics community, or the university sector and its assorted academic tribes. All contributors hold strong views about their preferred future pathway for research evaluation, so it was an easy task to persuade them to write in a more provocative manner than is the norm. So, while the papers are properly scholarly, they are written to stimulate discussion among specialist and non-specialist audiences within and beyond academia.

To tie the special issue together, authors were asked to address four questions concerning their favoured model of research evaluation, considerations central to designing national research assessment schemes:

- How does their preferred approach relate to the aims of government science policy?
- What are the predicted institutional and behavioural consequences?
- Does their model apply to a particular national context, or does it allow for a homogenised approach and international benchmarking?
- Do any research fields receive different treatment (for instance, the humanities, arts and social sciences)?

It is our intention that this collection of papers will not only contribute to the science policy and scientometrics literature, but will be of interest to the higher-education sector generally, and to research managers and policy-makers. While the UK and Australian cases take centre stage, this special issue is published at the cusp of fundamental reorientation in research evaluation practice, so will have international appeal.

Special issue overview

The papers in this volume run the gamut of academic opinion on future pathways for, and promising innovations in, national research evaluation systems. One referee commented on the "extreme tensions" that exist between the articles, so this introductory essay aims not only to give the flavour of the collection, but also to analyse the interplay of various 'pushes' and 'pulls'.

One major area of tension is described by Paul Nightingale and Alister Scott as the "relevance gap": the gulf between the research that society most requires and the research that is produced. They

It is our intention that this collection of papers will not only contribute to the science policy and scientometrics literature, but will be of interest to the higher-education sector generally, and to research managers and policy-makers

believe that a covenant exists between scientists and citizens that entails the purpose of publicly funded research being to solve social problems: yet they find research to be "a substitute for social action" that does not confront the political and complex problems society faces. They therefore offer ten suggestions to radically overhaul the peer-review system and thus close this relevance gap.

Nightingale and Scott identify a second point of tension: that research funders have failed to recognise the distinction between scholarly excellence and useful research — largely because quality is easy to audit and impact is not — and this focus on 'internal' discipline-led evaluation criteria has left science disconnected from decision-making. Yet the subsequent three papers (by Bruce Charlton and Peter Andras, Linda Butler, and Henk Moed) are explicitly and exclusively devoted to internal conversations about quality assessment, and how best to represent this.

While Nightingale and Scott believe in the redemptive potential of reforming peer review, Charlton and Andras choose to eschew it altogether in favour of a metrics-only approach: a simple count of indexed journal citations per university, perhaps augmented with a metric based on the distribution of science Nobel Prizes to identify "revolutionary science" institutions. Thus we find our third and fourth tensions: the view that peer review is subjective and contingent, whereas metrics are objective; and the search for simplicity in research assessment versus the pursuit of more 'messy' processes driven by complexity and diversity.

Charlton and Andras describe the UK RAE as a "highly complex, non-verifiable, un-checkable, evolving, bottom-up, discipline-based, peer-review process that lacks transparency". They are critical of its outcomes being the sum of internally generated and varying assessment criteria between disciplines, whereas scientometric evaluations share common criteria and can be executed "independently and objectively". Charlton and Andras also display a strong preference for top-down 'scientific' governance rather than bottom-up and contextual processes involving stakeholders. They view the use of metrics as "transparent, clear and cheap" and note that evaluations may be conducted by external experts without requiring the co-operation of those being measured: this will therefore minimise "distortion or corruption" and leave universities free to pursue their day-to-day business.

Butler and Moed write separately in favour of a middle ground where peer review is supported by a variety of quantitative quality metrics. Butler recognises our third tension, and maintains that lack of debate has polarised opinion in two camps: "red devils" who believe only peer review can assess research quality, and those wearing "rose-coloured glasses", for whom a metrics-only approach has no shortcomings. She notes the cyclical nature of recent proposed changes in research assessment and is concerned that there has been no attempt to learn from other contexts. She therefore calls for policy-makers to "pause and assess the vast wealth of experience that exists from research studies or evaluation exercises around the globe, and to take a more balanced approach to research assessment".

This is echoed by Moed, who too believes the future of research evaluation resides with an intelligent combination of advanced metrics and transparent peer review, so that the strengths of each approach may compensate for the limitations of the other. Both authors offer digests of essential information in applying quantitative measures to the assessment of research performance, Moed with especial reference to the UK RAE. Butler warns that stakeholders who lack this information are "amateur bibliometricians" placing absolute trust in metrics without knowledge of the methodological issues faced in the construction of such measures.

For Claire Donovan the future of research evaluation is qualitative: a new breed of science policy may extend beyond economic rationalism to embrace intellectual, social, cultural, environmental *and* economic returns, using qualitative measures and processes to capture research outcomes. Like Nightingale and Scott, her analysis focuses on the accountability of science to society; like Butler and Moed, she believes that relevant metrics are an aid within broader contextual approaches.

Donovan analyses quantitative measures as tools of science governance, and adds epistemological considerations to stock critiques of quality and impact indicators. She finds that standard metrics do not measure quality or impact, and neither do novel alternatives: in terms of our third tension, she states that metrics are as infused with human values as is peer review, and a nascent scientism aided and abetted by circular metrics has driven a false divide between science, technology, engineering and medicine (STEM) on the one hand and the humanities, arts and social sciences (HASS) on the other.

Donovan views the escalation of naïve quantification as a palliative for a dysfunctional science policy, and thus addresses our fourth tension: she describes an alternative holistic science policy that employs qualitative impact modelling to capture the public value of research, making better use of suitable metrics to inform the diffuse judgements of expert academic peers, end users and the beneficiaries of research.

Regarding the four questions posed to our authors, we find the tensions identified writ large, although this enables us to draw some conclusions about the causes; and hence detail the pathways most suited to the primary goals of national science policies and the logical modes of research evaluation these entail.

Matching research evaluation to science policy aims

The contributors to this issue were asked how their preferred mode of research assessment relates to the aims of government science policy. We find a push

towards broader societal relevance is met by a pull towards academic autonomy. This push is expressed most strongly by Nightingale and Scott's notion of the "relevance gap", and their belief that the needs of research users should override traditional disciplinary concerns. They therefore ask policy-makers to ignore the "three Sirens" of academic objectivity, academic autonomy and academic quality being invoked to avoid having to deal with relevance criteria.

We encounter a pull expressed by Moed, who maintains that assessments of research quality stimulate research excellence, and a basic premise underpinning science policy is that "better quality science is more likely to contribute effectively to desired social outcomes than science that is somewhat less high quality". Yet we have no empirical proof of this, and Nightingale and Scott would reply that it is precisely these internally defined quality criteria that leave science detached from decision-making. Social relevance and scientific excellence therefore remain polarised within this special issue,[1] although there is agreement that qualitative approaches are suited to assessing research impact, and metrics (with or without peer review) to reporting research quality.

Stimulating (un)desirable behaviours

Authors were asked to outline the predicted behavioural and institutional consequences that would flow from their preferred model of research assessment. Both Butler and Moed mull over possible intended and unintended consequences of applying bibliometric indicators. For Moed, the issue is not that these measures may change researchers' behaviour, but whether any change enhances research performance and scholarly progress in general, albeit as defined by the same measures. For Butler, the key issue is that a balanced approach to research evaluation entails adopting a "basket of measures", which means many more signals to respond to, thus minimising game-playing.

For Charlton and Andras, a system based on total citation counts will lead to universities competing to attract the "most-cited research teams in the leading branches of the natural sciences and the quantitative social sciences". They see the benefits as "improving the pay, support and conditions of group members, and further increasing competition to succeed in highly-cited fields". Donovan, however, is critical of rewarding the "imagined hierarchy of science" and is concerned about the epistemological implications of diverting funds away from other disciplines: she believes that her preferred system would make visible the academic and public value of research in all fields.

Nightingale and Scott assert that their reform of peer review would stimulate more interaction among researchers, government, the private sector, non-governmental organisations (NGOs), think tanks and civil society organisations throughout all phases of the research process — not just the dissemination stage. They insist that this will "mean a close focus on society's real research needs".

International benchmarking

All authors believe their preferred approach to research evaluation may be applied comparatively to other countries. The underlying tension is the view that metrics-based systems are 'scientific', while peer-based appraisal is not replicable so cannot be used for benchmarking purposes: this encapsulates the search for simplicity in research assessment as opposed to more 'messy' approaches embracing complexity and diversity. For Charlton and Andras, a major flaw with the UK RAE is that its results cannot be used to track longitudinal changes or for international benchmarking, although this is the primary function of an evolving simple metrics scheme.

On the other hand, Moed insists that "[f]orming a quality judgement is not a mathematical problem, but the use of a system of weighted indicators can be a useful tool", and agrees with Butler that any research assessment exercise must use a range of metrics and retain peer review as a central element. Butler takes this one step further to add more complex discipline-specific indicators and demonstrates that we may still generate benchmarks against which performance can be judged. For Nightingale and Scott "there can be no quality control process quite so searching as that which involves people whose interests are being affected by the research", a theme taken up by Donovan's argument that research evaluation should incorporate the opinions of end users and beneficiaries. These assessment schemes entail sharing scientific power, and so are 'messy' yet comparative instruments.

Accounting for disciplinary differences

Contributors were asked to consider whether their preferred model of research evaluation should allow for separate treatment of different fields. The aim was to reveal preferences for 'one size fits all' or discipline-sensitive approaches, and hence continue to unpack the tension between simplicity and complexity in research assessment. Moed restricts his bibliometrics-based system to scientific fields with excellent bibliometric coverage (he excludes sociology, political science, anthropology, educational sciences and the humanities), whereas for Butler, "[t]he challenge facing policy-makers is to identify robust indicators, particularly for those disciplines not well-served by standard measures". She notes that because of the varying importance of different publication types in different research fields, the RQF and current RAE follow 'best practice' by making explicit allowance for field-specific characteristics, since panels may vary assessment methodologies within an overall framework.

Charlton and Andras explicitly adopt a 'one size fits all' metric as they are concerned with the overall performance of universities, and not how this performance is constituted. They believe that research evaluation should be restricted to "scientific research" in the form of "the mathematical and natural sciences, and the quantitative social sciences such as economics". They maintain "there are few compelling reasons for wishing to measure non-scientific research performance using metrics. To be blunt, non-scientific research is believed (by those outside it) to lack the critical national importance of science".

Donovan, on the other hand, believes that research in the humanities, arts and social sciences is undervalued or under-reported within standardised evaluation systems,[2] and the commonality of science fields is overplayed so that bibliometrics say little about engineering, computing and mathematics. Her solution is qualitative impact modelling, which captures the distinctive quality and impact of various research fields and allows for a fair comparative assessment. Nightingale and Scott, however, are more concerned that research quality is over-determined by disciplinary values and so metrics and peer review act against the sort of relevant multi-disciplinary research they believe is most needed by society.

This special issue of *Science and Public Policy* contains conflicting messages about the future of research assessment, although it is certain that the pathway to be chosen is contingent upon what the underpinning science policy is hoping to achieve. We might suppose that recent policy U-turns in the UK are symptomatic of a detachment from innovations in scientometrics and a lack of policy learning from research evaluation exercises around the world. We might even surmise that there is no clear logic connecting the broad aims of public policy to the *raison d'être* of publicly funded research to how this is best accounted for. Perhaps the tail has been wagging the dog and possible (cheap) technologies have been driving the policy.

If the aim of publicly funded research is purely to boost the nation's economic performance, then governments may be tempted to follow Charlton and Andras and use a simple metric, jettison peer review and disregard the value of any research that is 'non-science'. Yet, while this is a simple system, it remains uncertain how excellence criteria internal to science relate to wealth creation and international competitiveness, which remains a vital consideration if adopting Butler and Moed's balanced approach using a variety of metrics informing peer review for all disciplines.

If the aim of publicly funded research is to be relevant to end users and to solve 'wicked' social problems, then governments might follow Nightingale and Scott's suggestions for closing the "relevance gap". If relevance is taken to include broader social, environmental, cultural and intellectual (as well as economic) gains, then governments would also do well to adopt Donovan's recommendation to pursue qualitative impact modelling. We therefore find that the more broad, inclusive and democratic the vision of science policy, the more qualitative the appropriate evaluation process; and the more 'scientific' and quality-focused, the greater the need for quantitative methods.

Notes

1. Although Donovan (2006) maintains that a holistic approach to science policy and research evaluation reveals the quality/impact divide to be a false dichotomy.
2. See also The British Academy (2007: 31–37) for a similar critique of metrics applied to the humanities and social sciences.

References

FWF/ESF, Austrian Science Fund/European Science Foundation 2007. *Conference Report. Science Impact: Rethinking the Impact of Basic Research on Society and the Economy. Vienna: FWF/ESF*. Available at <http://www.science-impact.ac.at/index.html>, last accessed 2 October 2007.

British Academy 2007. *Peer Review: the Challenges for the Humanities and the Social Sciences*. London: The British Academy.

Butler, Linda and Martijn S Visser 2006. Extending citation analysis to non-source items. *Scientometrics*, **66**(2), 327–343.

DEST, Department of Education, Science and Training 2007. *Research Quality Framework: Assessing the Quality and Impact of Research in Australia. RQF Submission Specifications.* Canberra: Commonwealth of Australia.

Donovan, Claire 2006. Visible gains from research. *The Australian Higher Education Supplement*, 1 November, 33.

Donovan, Claire forthcoming. The Australian Research Quality Framework: a live experiment in capturing the social, economic, environmental and cultural returns of publicly funded research. In *New Directions for Evaluation: Reforming the Evaluation of Research*, eds. C L S Coryn and M Scriven. Los Angeles: Jossey-Bass.

H M Treasury 2006. *Investing in Britain's Potential: Building our Long-term Future*. Cmnd. 6984. London: The Stationery Office.

Moed, Henk F 2005. *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

von Tunzelmann, N and E Kraemer Mbula 2003. *Changes in Research Assessment Practices in Other Countries since 1999: Final Report To the Higher Education Funding Council for England*. Brighton: SPRU, University of Sussex.