Science and Public Policy, 34(8), October 2007, pages 585–597
DOI: 10.3152/030234207X256538; http://www.ingentaconnect.com/content/beech/spp

The qualitative future of research evaluation

Claire Donovan

Science, technology and innovation (STI) policy aimed at technological advance, international competitiveness and wealth creation underpins the regulation of publicly funded research. Familiar quantitative evaluative 'metrics' fit snugly with these economic objectives. A re-imagined STI policy embraces wider intellectual, social, cultural, environmental *and* economic returns, using qualitative measures and processes to capture research outcomes.

IN COUNTRIES OF THE OECD (Organisation for Economic Co-operation and Development) we find, without exception, that the concerns of science, technology and innovation (STI) policy guide the frameworks and processes that regulate publicly funded research. The collective academic narrative of STI governance "resonates with the terms 'utility', 'commercialisation' and 'wealth creation'," and, according to this literature (Donovan, 2005: 599):

> Governments believe that scientific discovery creates social and economic progress, and so they desire to harness scientific research towards the twin causes of national technological advance and enhanced international competitiveness. In the pursuit of these goals, governments wish to derive maximum utility out of finite public funds while directing the research effort as efficiently as possible. This is the genesis of science governance.

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.

This paper is a condensed and reworked version of a report funded by the Council for Humanities, Arts and Social Sciences (CHASS) with the financial assistance of the Australian Government, through the Department of Education, Science and Training (DEST). The views expressed in this paper do not necessarily reflect those of CHASS or DEST. A version of this paper was presented at SPRU's 40th Anniversary Conference The Future of Science, Technology and Innovation Policy: Linking Research and Practice, 11–13 September 2006, University of Sussex. This paper has been improved by the thoughtful observations of *SPP* referees. The author is grateful to John Butcher for reproducing the graphics in this article.

Science governance and its underpinning STI principles are essentially concerned with research developments in science, technology, engineering and medicine (STEM). Yet this model is unreflexively applied to the governance of social science (Donovan, 2005: 604), and is extended even more incongruously to research in the humanities and creative arts.

Given this policy context, this paper addresses a particular aspect of science governance:[1] using quantitative indicators to evaluate the academic quality and extra-academic impact of publicly funded research. The paper demonstrates how standard quantitative indicators or 'metrics' fit with broad STI policy objectives, and produce a circularity that quite naturally favours the policy-makers' vision of excellence in STEM research.

However, this form of audit delivers an unnecessarily circumscribed view of the value of publicly funded research in STEM and beyond. The paper then outlines novel quantitative indicators that may be fairly applied both to STEM and to the humanities, arts and social sciences (HASS).[2]

Nevertheless, we find that novel quantitative HASS-friendly indicators encounter similar circularities faced by standard measures, and their promise is all too often diminished by reducing the worth of HASS research to either an instrumental end-product aimed at specific 'users' or to a crude economic rationalisation of its value. This overlooks the essence of HASS research and its major benefits to society at large; the returns of STEM research are similarly constrained, thus sustaining a false distinction between the public value of STEM and HASS research.

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).

The paper maintains that, although clearly desirable, the search for novel quantitative metrics is a palliative for the deficiencies of an outmoded STI policy framework. The focus of public policy is changing to accommodate social and environmental, as well as economic, considerations: this entails that the imperatives underpinning STI policy must similarly evolve and in tandem with how we might best account for STI returns.

The paper therefore outlines a re-imagined STI policy framework that embraces the 'triple bottom line' of social, economic and environmental returns, plus the intellectual and cultural payback from research. It concludes by advocating the use of qualitative impact modelling for research evaluation and STI foresight planning purposes. This approach captures the distinctive qualities of all disciplines and subfields, and allows a fair assessment of the diverse range of research outcomes derived from STEM and HASS alike.

We can thereby demonstrate that the value of publicly funded research may be conceived of in meaningful terms broader than "utility", "commercialisation" and "wealth creation" to which it is currently confined. In this respect, research evaluation should no longer aspire to the standardised use of blunt quantitative metrics: the future of STI policy and research evaluation is qualitative.

Pitfalls of popular quantitative indicators

The use of quantitative indicators or metrics to evaluate the quality of publicly funded scientific research is seen as desirable for several reasons, and the perceived benefits are often contrasted with supposed deficiencies of peer-based evaluation processes. For example, supporters might argue that metrics are more cost effective and less of a bureaucratic burden; that data may be used comparatively for international benchmarking exercises; and, crucially, that, because data is collected independently, results are transparent and verifiable, and thus unsullied by subjective and contingent peer judgement. Indeed, in the extreme, peer review is regarded as an unseemly process whereby academics self-regulate their own activities to serve their own esoteric pursuits rather than the public interest.

This is ironic indeed, as the harshest critics of peer review are advocates of metrics-only approaches to research evaluation, yet they overlook the fact that many 'quality' metrics are underpinned by peer-review processes (that is, refereed journal publications and competitive grant income). While degrees of trust or scepticism in the peer review of research quality vary, the political desire for comparative research quality metrics is clearly in the ascendancy internationally.

After 2008, the UK Research Assessment Exercise (RAE) will exchange peer review for standard quantitative indicators in STEM (from 2009–10), and will retain a light-touch peer-review exercise combined with as yet unspecified metrics for HASS subjects, mathematics and statistics (from 2013–14) (HM Treasury, 2006b: 57). The 2008 Australian Research Quality Framework (RQF), on the other hand, abandons a metrics-only approach in favour of a system-wide panel-based exercise where peer judgement is informed by standard quantitative indicators and the promise of novel, field-sensitive measures (DEST, 2006b: 20).

We are also witnessing a rise in the desire to evaluate the value of publicly funded research for 'end users' and industry, and the accompanying urge to construct quantitative measures to aid this assessment. However, these metrics are in their infancy (CHASS, 2005: 75) and their place in the 2008 RQF and post-2008 RAE remains unclear.

Stock criticisms of popular quantitative indicators

Popular quantitative indicators of research quality are subject to a set of stock criticisms.[3] The first, and by far the most damning, is that these metrics do not actually measure research quality. For example, research income is an input, rather than an output, measure: while competitive funds obtained is a popular proxy for research excellence, winning a grant or contract does not necessarily entail high quality outcomes (REPP, 2005: 30). Higher-degree student load or completions are related to teaching and supervision and have no bearing on research outcomes (REPP, 2005: 31).

Publication counts are productivity measures that do not gauge research excellence: while the number of peer-reviewed publications produced is often taken as an indicator of research quality, the bibliometrics literature finds 'quality' to be quantitatively inaccessible as the academic value of a publication is unknown until we can assess its influence on subsequent literature (REPP, 2005: 12). In this vein, bibliometricians take citation counts as indicators of research 'impact', or of the effect written work has on subsequent academic literature, but not of the inherent quality of publications: citations may be positive or negative, and work may be highly cited because its findings are contested (REPP, 2005: 2–4; 12–14).

Second, adopting standard measures may stimulate lower-'quality' research. For example, in Australia,

While the number of peer-reviewed publications produced is often taken as an indicator of research quality, the bibliometrics literature finds 'quality' to be quantitatively inaccessible as the academic value of a publication is unknown until we can assess its influence on subsequent literature

Linda Butler (2003; 2004) found a relationship between the introduction of publication counts in Australian performance-based block funding and a sharp rise in articles in lower-impact journals.

Third, there are various problems associated with data coverage, whereby standard metrics exclude much research output. For example, standard citation counts are compiled using data collected by Thomson Scientific (previously the Institute of Scientific Information); this metric is confined purely to citations between indexed journal papers and thus excludes all citations between books, chapters, non-indexed refereed journals, and 'grey literature' aimed at practitioners.

The database also has a relatively low representation of regional journals, small research fields and non-English papers (REPP, 2005: 17). The creative arts are virtually excluded, as is any field in which research outputs tend not to take the form of indexed journal publications,[4] making standard citation measures largely redundant for the majority of HASS subjects. While standard publication counts can include books, chapters and non-indexed refereed journals, there remains a tendency to disregard 'grey literature', creative works and performances.[5]

Popular quantitative measures of the extra-academic returns of research, or what policy circles (rather than bibliometricians) refer to as 'impact' indicators,[6] are less common and less routinely subject to critique: for example, funding from industry, and technometrics (number of patents, number of patent citations and number of intellectual property items that have been commercialised or for which protection is being sought). As supposed indicators of the public value of research, we may argue that standard quantitative impact measures do not actually measure research impact: they report activities relating to the early stages of commercialisation or technology transfer, and not actual research outcomes and benefits (CHASS, 2005: 75).

Response of the 2008 RQF and post-2008 RAE

There is a marked, and puzzling, difference between how the 2008 RQF and post-2008 RAE respond to the perennial criticisms of standard quality metrics. The 2008 RAE will present peer-review panels with standard data on research income and higher degree students, although these data will be treated as contextual information to inform peer deliberations, rather than metrics to feed into a funding formula.

The post-2008 RAE acknowledges the problems of applying standard metrics to HASS (and mathematics and statistics) by pursuing a dual-assessment exercise: these subjects will receive a light-touch peer-review of research outputs "informed by a range of discipline-specific [that is, potentially novel] indicators".[7] Yet after 2008, STEM disciplines will no longer receive peer review, and research quality will be assessed solely by using standard metrics: research income; higher-degree student data; and a "bibliometric indicator of quality" (presumably standard citation counts). The outcomes of STEM and HASS assessments will be adjusted by a standard indicator — research volume — "to produce a funding allocation" (HM Treasury, 2006b: 57).[8]

Conversely, in Australia, as a direct response to the perceived shortcomings of the standard metrics upon which the country's research assessment was wholly based (DEST, 2004), the 2008 RQF introduces assessment panels and peer review of research outputs in all fields. Panels will be supplied with standard quality metrics in the form of competitive grant income and standard citation data; novel quality indicators and citation measures are promised for any discipline where standard data are inappropriate (DEST, 2006b: 20).

Thus, in encountering the pitfalls of popular quantitative indicators of research quality, the post-2008 RAE has chosen to apply standard measures to STEM, while engaging in a light-touch peer review for HASS, which may incorporate novel quantitative indicators. Yet, while the post-2008 RAE divides HASS and STEM, the 2008 RQF seeks to find novel quantitative metrics that can be fairly applied to both STEM and HASS.

In terms of research impact, the UK Treasury seeks to "maximise the economic impact of research", and so "the new system will provide greater rewards for user-focused research", and from 2007–08 the Higher Education Funding Council for England (HEFCE) will allocate £60 million for the "relative amount of research universities undertake with business" (HM Treasury, 2006b: 58).[9] Standard quantitative impact metrics are not discussed, but are well suited for this purpose.

Australia's 2008 RQF is unique in that it defines research impact as "the social, economic, environmental and/or cultural benefit of research to end users in the wider community regionally, nationally, and/or internationally" (DEST, 2006b: 21). Panels consisting of academic peers and end-users of research will consider impact statements and case studies, and "will be given generic indicators and will determine additional indicators of impact as appropriate for their discipline cluster" (DEST, 2006b: 20). No quantitative indicators have as yet been formally suggested,[10] although the limitations of standard and novel quantitative measures of research impact entail that metrics will play a secondary role to qualitative processes.

Meta-criticisms of popular quantitative indicators

By focusing on the relationship between quantitative metrics of research quality and impact, and the broad STI objectives these serve, this paper now offers an alternative critique of popular quantitative indicators. As noted above, there is agreement in the science governance literature that STI policy is premised upon the imperatives of enhancing international competitiveness, technological advance and wealth creation (Donovan, 2005: 599), a 20th century legacy of post-war reconstruction, expanded by Fabianism, consolidated by consensus politics and accelerated by neo-liberalism.

Given this historical and political context, it is little surprise that STI policy is framed in terms that match the bureaucrat's grand vision of STEM,[11] nor that markers of this policy's success map neatly onto a complementary 'economic rationalist' validation of scientific research excellence.[12] We find, therefore, that a more nuanced analysis of the wholesale application of popular quantitative metrics to all research enterprise reveals a circularity that naturally favours the bureaucratic vision of STEM at the expense of HASS in both the academic and public realm.

Indicators of research quality are clearly science-friendly. With respect to research income, STEM research often requires expensive equipment, and a greater proportion of national research budgets is devoted to STEM subjects, which renders this metric circular. Modelling of the likely effects of using research income as a core RAE metric, and particularly the inequities generated for HASS, was instrumental in the UK Treasury abandoning its plan to conduct the 2008 RAE purely on the basis of applying metrics at the institutional level, and thus to retain discipline-based peer review for HASS post-2008.[13]

Volume of higher-degree completions is an indicator that favours STEM subjects as laboratory-based and team-based research tends to produce faster doctoral submissions. Publication counts restricted to indexed journal papers will clearly benefit STEM over HASS, although we would normally expect books, chapters and non-indexed journals to be included in a productivity measure, and would hope to find a greater weighting allocated to books as opposed to journal papers. Yet it is likely that HASS publications aimed at the public or practitioners will be excluded,[14] and outputs from the creative and performing arts, and design, remain practically invisible.

It has long been recognised that citation counts give STEM research an advantage as citation metrics were developed using indexed journal articles as the basis for scholarly communication. Research methods and orientations in HASS are distinct from those of STEM and communication practices or literatures are differently structured, which produce unfavourable bibliometric consequences (Glänzel and Shoepflin, 1999: 31; Hicks, 2004: 473; Luwel *et al*, 1999: 13; Moed, 2005; Nederhof *et al*, 1989: 427; Nederhof, 2006; van Leeuwen, 2006).

The impetus to create quantitative indicators that capture the extra-academic impact of research within the public realm is a more recent development, and we have seen that efforts tend to focus on funding from industry, and technometrics tied to commercialisation activities. These metrics accord with a clear economic rationalisation of the value of public-funded research. However, figures on industry funding, number of patents and number of spin-off companies created are of a rather low-order: these relate to prospective impact rather than actual public benefit from the outcomes of research, and privilege channelling public funds to create private value rather than public value. To date, such metrics have largely excluded HASS except when researchers can account for their activities in commercial terms; in this respect, impact evaluation tends to be confined to a narrow view of the benefits that research may bring.

It has long been recognised that the driving force behind applying popular quantitative indicators of STEM performance to HASS has been the desire of funding agencies and policy-makers to develop methodologies to evaluate the research efforts of the whole university sector (Katz, 1999: 1; Luwel *et al*, 1999: 13). The tendency has been to take measures designed to evaluate STEM research and apply these to HASS, rather than to generate HASS-specific metrics. This unreflexive practice has simply made HASS less visible when it is judged in STEM terms.[15]

In this respect, we find that metrics have become part of the armoury of STI governance, yet science indicators are imbued with human values masquerading as neutral markers of what science should aspire to be and do, and hence what constitutes scientific excellence in the public arena. Thus, the distinctive features and contributions of HASS research are undervalued or unreported within standardised evaluation systems, while the distinctive features of the bureaucrats' vision of STEM are overplayed.

Stock critiques of popular quantitative indicators of research quality and impact overlook the fact that metrics constructed to fit with broad STI policy objectives place severe limits on the perceived value of publicly funded research. On the quality side, deep epistemological consequences flow from adopting metrics that quite naturally favour STEM research. For example, within HASS, citation indicators designed to measure the academic quality of research in the natural and experimental sciences will detect research most like that in the natural and experimental sciences. This is illustrated by the greater coverage of highly quantitative social science research, particularly in economics and psychology, which share a journal-based and international orientation that mirrors STEM publishing practice (van Raan, 1998: 3; Katz, 1999: i).

There is an inherent danger that undetected qualitative social science literature may be viewed as 'soft' or less mature, and hence a lower order of knowledge. An imagined hierarchy of science is thereby created, and becomes the tautological basis upon which funds will be distributed (Donovan, 2007). However, we must note that it is increasingly recognised that citation measures based on indexed journal papers do not necessarily complement all fields of STEM research. For example, engineering has a greater focus on producing technical reports that are excluded in standard citation counts (Bourke *et al*, 1996), and computing, environmental science, mathematics, nursing and statistics are also poorly served (REPP, 2005: 16–17).

On the impact side, we find that the public value of HASS research is reduced to the extent to which it can report against quantitative indicators designed to detect investment from industry, commercialisation and technology transfer. However, while these supposedly 'science-friendly' metrics fit snugly with STI policy objectives, might we not also suggest that this form of audit delivers an unnecessarily circumscribed view of the public value of STEM research?

Promise of novel quantitative indicators[16]

International practice and trends in measuring the quality and impact of publicly funded STEM and HASS research are reviewed by CHASS (2005: 56–81), which identifies a range of potential novel quantitative measures that should, in theory, give more equitable treatment to STEM and HASS. Indeed, the future of research evaluation is thought to rest upon the development of field-sensitive quantitative indicators more favourable to HASS, as STEM is held to be well catered for.

However, while such novel measures are increasingly promoted within research policy networks, there has been little systematic testing of these new metrics to gauge their usefulness, or indeed, validity. This paper now reviews the latest developments in producing these novel quality and impact metrics, and considers whether more representative quantitative indicators provide the solution to the pitfalls exposed by stock and meta-criticisms of traditional measures.

We give particular attention to the most advanced efforts to include novel quantitative data in national evaluative schemes, but these novel data are not used as stand-alone metrics, but rather as information that feeds into a peer-review process

This section of the paper gives particular attention to the most advanced efforts to include novel quantitative data in national evaluative schemes, most notably in Australia, the Netherlands and New Zealand. However, we must recognise that these countries do not use these novel data as stand-alone metrics, but rather as information that feeds into a peer-review process. Also, while these exercises are system-wide, in the case of the Australian RQF novel 'quality' indicators will be employed only where disciplines nominate their inclusion, and the post-2008 UK RAE proposes that HASS, mathematics and statistics will receive a light-touch peer-review of research outputs "informed by a range of discipline-specific [that is, novel] indicators" (HM Treasury, 2006b: 57), although no detail is given on what these indicators might be.

Novel quantitative indicators of research quality

Until recently, developments in publication metrics tended to focus on producing ranked lists of journals. However, initiatives such as the UK Arts and Humanities Research Council's attempt to list the top ten "most significant and important" journals in ten subject areas were abandoned in the face of outcry from the academic constituency on the grounds that such an exercise would kill off emerging disciplines and publications, and could ruin developing research careers if subsequent competition meant that only the established elite appeared in these 'top' publications (*THES*, 28 January 2005).

More sophisticated and less contentious efforts have concentrated on producing 'tiers' or 'bands' of journals or book publishers. Norway has developed a national publication-based funding formula that allocates funds to universities on the basis of publisher prestige bands. This is applied to ISI and non-ISI journal articles, and book chapters, with a higher relative weighting given to books (Sivertsen, 2006).[17]

The Australian RQF takes this approach one step further through promising "tiered ranking of discipline-specific research outputs for inclusion in a basket of measures to assist Panels in making their judgements" (DEST, 2006b: 14). The unit of assessment is the "research group", which will report its whole body of work, from which "ranked outputs" will be produced in discipline-specific tiers, for example "refereed journals, professional journals, book publishers, conferences[18] or performance venues" (DEST, 2006b: 20). In this respect, quantitative indicators of research quality may include creative works and performances, thus extending the definition of 'publication' to non-bibliometric research outputs.

Innovations in citation counting have focused on expanding database coverage to include a range of publications beyond ISI-indexed papers. For example, Butler and Visser (2005) constructed a database using DEST and ISI data to calculate "non-standard" citations to books, chapters, non-ISI papers and refereed conference proceedings. The net result is that the visibility of research in the humanities and social sciences increases, as do the citations received. The Australian RQF acknowledges that "standard citation measures are not applicable in all disciplines" and will incorporate this approach where disciplines believe this data will be helpful.[19] However, these citations are mined solely from ISI papers,[20] and so exclude book to book citations. There is evidence that citation cultures vary between journals and books, so that citations derived from journal articles favour quantitative research traditions over qualitative approaches (Clemens *et al*, 1995).

A further area of investigation has been non-bibliometric markers of research quality in the form of esteem indicators, which have come to be identified as particularly important for those fields, particularly in the humanities, arts and social sciences, where bibliometric indicators are difficult to apply (van Raan, 1998: 5). However, while viewed as inherently desirable, little effort has been made to test their actual value.

A recent study (Donovan and Butler, forthcoming 2007) assessed on a comparative basis the following esteem indicators: honours, awards and prizes; election to learned academies and academic professional associations; service to conferences; service to journals; and visiting fellowships. Data were provided for five Australian economics departments and the indicators and findings made subject to expert peer scrutiny. The use of esteem indicators as stand-alone metrics was roundly rejected: peer scrutiny concluded that novel esteem indicators are markers of individual standing and of research-oriented workload, not of actual research quality.

To sum up, the core developments in publication, citation and esteem measures have been to increase the visibility of non-standard research outputs, particularly in the creative arts, humanities and social sciences. Yet do these technical innovations resolve the stock and meta-criticisms of traditional quantitative indicators? Yes, in that much more research output is included, making the indicators far more inclusive and equitable. No, in the fundamental sense that productivity measures, bandings and citations — no matter how well refined — still do not tell us anything about research quality.

Second, any publication weighting would need to be carefully constructed so that: low-'quality' research is not stimulated; caution is exercised to allow appropriate weighting for major works; and, while including non-standard publications allows for broader epistemological representation, care is taken to ensure that journal tiers, for example, do not shore up old biases. Third, novel citation metrics retain epistemological problems: improved database coverage nonetheless relies upon the references made in ISI papers, thereby excluding citations made by non-ISI papers, chapters and books.

Ultimately, the search for novel, HASS-friendly, quantitative measures of research quality obscures the fact that standard metrics, even when applied to STEM disciplines with high ISI coverage (such as physics and chemistry) do not access research quality. No matter the level of refinement achieved, the scientometrics community would advocate that citation and other quality measures should not be used in isolation, but selectively combined and applied to appropriate sub-fields of research within an overall context of expert peer judgement. This points to the need to contain quantitative indicators within qualitative processes.

Novel quantitative indicators of research impact

There is an increasing desire within government and funding bodies to evaluate the extra-academic returns of publicly funded research, which tends to translate into the need to formulate quantitative impact indicators. For example, in the USA, while peer review, rather than quantitative indicators, is the basis for research evaluation, in policy circles there is a growing interest in impact indicators, spurred by the wish to evaluate the economic and social outcomes of research (Hicks *et al*, 2004). However, such metrics lack any sophisticated treatment and remain undeveloped. The greatest advances made in reporting research impact have, to date, occurred in New Zealand and the Netherlands, although not in the form of stand-alone metrics.

In the UK, research impact has been conceived as the "relative amount of research universities undertake with business" (HM Treasury, 2006b: 58), and, from 2007–08, HEFCE will allocate £60 million on this basis alone, which suggests using a simple metric of income generated from business. The rationale is to "maximise the economic impact of research", hence "the new system will provide greater rewards for user-focused research", although the user is defined purely as business.

Australia's 2008 RQF will break new ground internationally by defining research impact as "the social, economic, environmental and/or cultural benefit of research to end users in the wider community regionally, nationally, and/or internationally" (DEST, 2006b: 21) and will employ an impact scale that relates to reciprocal engagement with so-called 'end users', the adoption of research, and the significance of the public benefits derived from research. The term 'user' is intended to include industry, practitioners, government, non-government organisations (NGOs), community groups and the general public.[21]

We have seen that popular quantitative indicators of research impact almost exclusively take the form of technometrics that relate to the early stages of commercialisation or technology transfer. Novel technometrics do not deviate from this pattern: for example, number of commercial disclosures and arrangements, number of start-up companies created, number of partnerships and linkages with industry (funds contributed, co-authorship, co-patenting), new technology, new products and research leading to commercial outcomes. Technometrics clearly (although not exclusively) favour STEM research in terms of its value to business, industry and the economy. However, we find that these measures continue to relate to prospective impact and place distinct limitations on how we might conceive of the public benefits derived from linkages, investment, new technology or new products. To restrict impact measurement to technometrics is to ignore the broader benefits that research may bring.

In Australia, some attempts have been made to produce quantitative indicators of the social returns of research, or sociometrics, yet these prove problematic. The Allen Consulting Group (2003) studied the return on investment from Australian Research Council funded research, and explicitly reported the benefits of research solely against economic criteria. In a further attempt (Allen Consulting Group, 2005), social value was defined as "encompassing social attachments, freedom from crime, the level and security of political rights and the extent to which the population engages in political processes", and indicators included "marriage and divorce rates, persons living alone, participation in voluntary work" (Allen Consulting Group, 2005: vii).

These particular benefits and indicators were selected because they matched publicly available data provided by the Australian Bureau of Statistics. The consequence is that the perceived social benefits of research become defined by the limits of the existing technology and data, and social science is understood in technocratic terms as an activity that can provide a 'policy fix'. Yet this pragmatic process is a self-defeating task: when we review the data, we find that it is impossible to posit a causal link between particular research outcomes and macro-level social trends.

There have been attempts to develop novel quantitative impact metrics in two further core areas: linkages with government and policy advice given; and research aimed at, or taken up by, end users. While including these areas broadens the conception of research impact, we find that the proposed novel metrics face similar limitations as those encountered by sociometrics and technometrics. For example, while we may count the instances of policy advice given to government bodies, or the number of government-commissioned reports and consultancies, or the volume of research linkages with government bodies, these relate to prospective outcomes and hence remain low-order impact indicators:[22]

The search for quantitative impact indicators has delivered an array of novel metrics that represent low-order impact, technometrics that privilege private over public interest, and sociometrics that rely on macro-level data with no credible link to the efforts of particular researchers

high-order impact or public value lies in what happens as a consequence of these contributions and linkages.

In terms of contact with end users, the New Zealand Performance Based Research Fund (PBRF) seeks data on a vast array of publications, presentations and displays aimed directly at users (including papers, oral presentations, static displays, web presentations and articles in trade journals or magazines). It also seeks information on reports commissioned by users under contract or formal arrangement, and the number of linkages with organisations or individuals.

Other potential novel indicators include the number of visitors to exhibitions, attendance numbers for performances, quantifying media dissemination and media citations, improved knowledge transfer, increased uptake of advice, stakeholder awareness and satisfaction with advice and information, linkages with community, the number of non-academic visitors who use a unit's facilities, and the number of collaborative projects with non-academics. Again, we find that these are low-order indicators that presage any wider and more significant public benefit that may be derived from research.

Thus we hit an impasse: the urge of agencies to seek quantitative impact indicators has delivered an array of novel metrics that represent low-order impact, technometrics that privilege private over public interest, and sociometrics that rely on macro-level data with no credible link to the efforts of particular researchers. Novel quantitative approaches have concentrated on counting outputs and activities, and not on observing effects; sociometrics rely on statistical datasets rather than the viewpoint of researchers, 'end users' or beneficiaries; and technometrics generate a blinkered vision of the value of research. Yet technometrics, more than any other type of indicator, fit snugly with STI policy imperatives.

Novel metrics as palliatives

It is a logical step to produce more sensitive and inclusive quantitative indicators of research quality and impact, yet such metrics fail to provide the solution to the pitfalls exposed by stock and meta-criticisms of popular measures because they serve the same logic: accounting for the worth of research according to a mechanistic rating of outputs, or as an instrumental end-product aimed at specific users, or through a crude economic rationalisation of its commercial value. This overlooks the essence of HASS research and its major benefits to society at large; the returns of STEM research are similarly constrained, thus sustaining a false distinction between the public value of STEM and HASS research. Indeed, we have seen that, in terms of innovation and analysis, criticisms and proposed improvements tend to polarise HASS and STEM, treating HASS as a special case and seeking greater quantitative parity, and not addressing the possibility that STEM may similarly be the captive of badly chosen indicators.

This paper maintains that, while clearly desirable, the search for novel quantitative metrics is a palliative for the deficiencies of an outmoded STI policy framework. We are witnessing a shift in public policy away from an economic rationalist agenda premised on utility and profit maximisation, and towards principles of sustainability embodied in social, environmental and economic returns (what is known as the 'triple bottom line'). Gallop (2007: 7) characterises this as a response by governments on the left and right to the contradictions produced by the New Public Management (NPM) and its methods of accounting for policy outcomes. Such a shift in government philosophy entails that the imperatives underpinning STI policy must similarly evolve.

An economic rationalist STI model employs accounting techniques and indicators that complement the goals of productivity, wealth creation and competitiveness: that is, quantitative techniques that privilege quantity over quality, income generated over value generated, and commercial activity over public benefit. Extending this accounting logic to HASS detects that which we assume is most like STEM.

However, we must acknowledge that these quantitative indicators are infused with human values, and that STI policy is contingent and need not necessarily be premised on economic rationalism. Can it be that STI reporting defines what we believe to be the characteristics of STEM? Do we believe quantitative indicators are 'science-friendly' because we assume the reported data neatly matches STI principles? Or could it be that STEM is similar to HASS in that a significant remainder of research outcomes is rendered invisible by the economic rationalism and NPM accounting techniques currently enshrined in STI policy?

Qualitative future of research evaluation

A re-imagined STI policy framework requires a re-imagined evaluative framework. Following post-NPM logic, accounting for the triple bottom line of social, environmental and economic returns will fail if it relies purely on quantitative indicators: this is why we find that novel impact metrics, such as sociometrics and technometrics, ultimately disappoint us by not genuinely accessing the public value of research.

This section of the paper therefore reviews advances made towards post-NPM quality and impact measurement, and advocates the use of qualitative impact modelling for research evaluation and STI foresight planning purposes. This approach has the necessary flexibility to allow for a fair assessment of the distinctive qualities of all disciplines and subfields, and captures a diverse range of public value (intellectual, social, cultural, environmental *and* economic) thereby dissolving the contingent HASS/STEM divide perpetuated by purely quantitative approaches.

Towards holistic research evaluation

The evolution of post-NPM attempts at capturing the public value of scientific research consists of three waves. The first was conceptually bound by the continued desire to quantify, hence the emergence of studies with sociometric aspirations (Bozeman, 2003: 34; Allen Consulting Group, 2005).

The second wave broke new ground by employing models of research quality and impact that attempted to combine quantitative and qualitative data with peer assessment. For example, in 2002, the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) created the Social and Economic Integration (SEI) Emerging Science Initiative, which integrated social and economic considerations into the design, conduct and delivery of scientific research with the aim of increasing the positive impact of scientific research and development on people's lives.

This is apparent in how CSIRO accounted for "Delivering impact from our science" where an Outcome–Outputs Framework reported the triple bottom line of economic, social and environmental benefits from CSIRO's four main arms of research in sum (presented in Figure 1) and individually (CSIRO, 2004: 38–41).[23]

Although the precise methodology is not defined by CSIRO, we can nonetheless see that impact 'outputs' cover new knowledge, technology and products, and policy advice, while benefits accrue to Australian business, the technology society, and environment and lifestyle. However, the reporting style remains tied to counting outputs that contribute to outcomes, rather than describing the outcomes. The conceptual leap needed to make the *y* axis in Figure 1 redundant is the key feature of the third, and most qualitative, wave.

We may note that the second and third waves share an important element in common: the attempt to embed quality and impact assessment within a holistic reporting framework. The most advanced

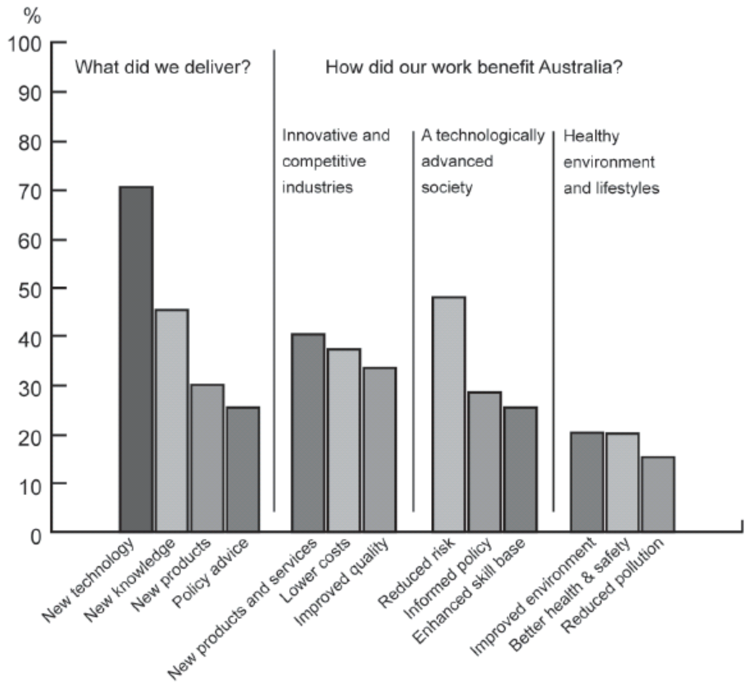

Figure 2. Overview of selected achievements, 2003–2004
Source: Reproduced with permission of CSIRO (2004: 39)

exemplar of the third phase[24] is an impact modelling exercise conducted jointly by Brunel University's Health Economics Research Group (HERG) and RAND Europe, for the Arthritis Research Campaign, in order to evaluate the outcomes and outputs (or 'payback') resulting from research funding (Wooding *et al*, 2004). The Payback model combines measures of research quality and impact with peer and user evaluation, to produce a logic model of the research process (see Figure 2).

Category A in Table 1, Knowledge, includes bibliometric analyses of publications resulting from research funding, and Categories A to E are considered by panels of expert peers and research users who rate each category on a scale of 1 to 9. This is mapped, resulting in various profiles such as those illustrated in Figure 2.

The other categories are benefits to future research and research users (which is largely forward looking and about capacity building), Political and administrative benefits, Health sector benefits (including cost reduction and qualitative improvements in service delivery, improved equity and intellectual property revenues), and Broader economic benefits (including wider economic benefits from commercialisation and a healthy workforce).[25] While this approach has been utilised within biomedical research at the project level, it can be applied to the research group, discipline, institutional and national level.[26]

We find, therefore, that the Payback approach is a form of comparative qualitative impact modelling, which is framed by the views of various stakeholders, and which subsumes quantitative quality

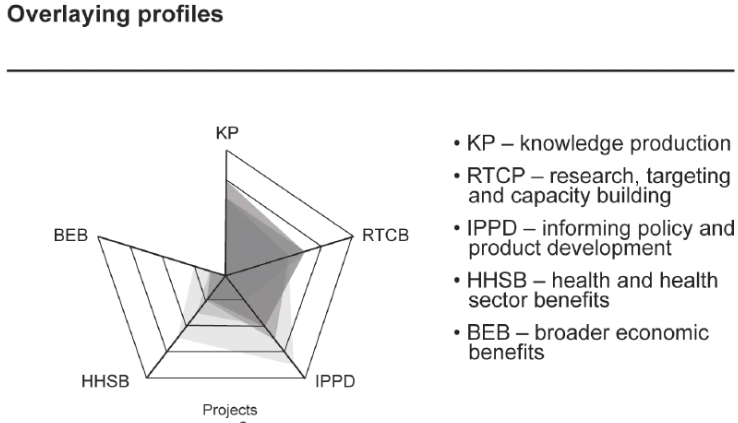

Figure 1. Example of 'payback profile' impact assessment
Source: Reproduced with permission of the HERG/RAND group (Wooding *et al*, 2004:54)

Table 1. Payback categories for arthritis research

A	**Knowledge**	
B	**Benefits to future research and research use**	
	i	better targeting of future research
	ii	development of research skills, personnel and overall research capacity
	iii	critical capability to utilize appropriately existing research, including that from overseas
	iv	staff development and educational benefits
C	**Political and administrative benefits**	
	i	improved information bases on which to take political and executive decisions
	ii	other political benefits from undertaking research
D	**Health sector benefits**	
	i	cost reduction in the delivery of existing services
	ii	qualitative improvements in the process of service delivery
	iii	increased effectiveness of services, eg increased health
	iv	equity, eg improved allocation of resources at an area level better targeting and accessibility
	v	revenues gained from intellectual property rights
E	**Broader economic benefits**	
	i	wider economic benefits from commercial exploitation of innovations arising from r&d
	ii	economic benefits from a healthy workforce and reduction in working days lost

Source: Reproduced with permission of the HERG/RAND group (Wooding *et al*, 2004: 10)

and impact data within a broader qualitative framework. Payback categories may be constructed that are appropriate to any research field, and that fit the appropriate kinds of quantitative and qualitative indicators, and other contextual information, which can inform a review panel's assessment. Indeed, for many research fields, novel quality metrics would provide the ideal quantitative data to inform the evaluation of knowledge production.

Beyond economic rationalism

Various approaches to qualitative impact modelling demonstrate that the value of publicly funded research may be conceived of in meaningful terms broader than "utility", "commercialisation" and "wealth creation" to which it is currently confined. What might additional alternative values underpinning STI policy be? A starting point may be drawn from two sources attempting to break out of NPM economic rationalism and thus reframe the value of research for the nation. The first is from an Australian Research Council review of the Humanities and Creative Arts. A survey of leading researchers was combined with researcher and end user descriptions of the impact of research taken from end of grant reports (ARC, 2004: 15; 145–152). The findings point the way towards developing payback categories or qualitative impact modelling informed by academic and user perspectives:

- Understanding ourselves and our community;
- Understanding other societies and cultures;
- Informed public debate, improved policy-making and economic impact;
- Knowledge/education;
- Cultural/economic.

Similarly, The British Academy (2004: vii–ix) examines in detail five key functions of research in the arts, humanities and social sciences, which are summarised as:

- Contributing to cultural and intellectual enrichment;
- Contributing to economic prosperity and well-being;
- Contributing new knowledge and understanding of major challenges facing both the UK and beyond;
- Contributing to public policy and debate;
- Providing a rigorous, beneficial and fulfilling education.

We can see striking similarities between these examples, and there is also considerable overlap with the biomedical objectives illustrated in Figures 1 and 2 and Table 1. It follows that a re-imagined STI policy may embrace wider intellectual, social, cultural, environmental and economic returns, using qualitative measures and processes to capture research outcomes, and thus dissolve the HASS/STEM divide supported by NPM accounting techniques.

Conclusion

This paper maintains that quantitative indicators are as infused with human values as are qualitative approaches. Indeed, advances in the use of quality and impact metrics have followed a trajectory away from the unreflexive use of standardised quantitative metrics divorced from expert peer interpretation, towards triangulation of quantitative data, contextual analysis and placing a renewed and greater value on peer judgement combined with stakeholder perspectives.

The post-2008 UK RAE attempts to resolve the dilemmas of applying traditional quality metrics by introducing a dual assessment system: traditional metrics for STEM; novel metrics and peer review for HASS. However, this paper concludes that the notion that STEM is well served by standard metrics alone is false — it just so happens that the failings are more apparent in HASS disciplines. The more an evaluation system moves towards privileging quantitative indicators (even if these are novel measures), and polarising HASS and STEM, the further away it will move from measuring genuine research quality and impact.

In the course of measuring the quadruple bottom line of social, economic, environmental and cultural returns, the Australian RQF has broken new ground by recognising the limitations of novel technometrics and sociometrics: "Impact metrics are underdeveloped, and hence are neither robust surrogates for research impact, nor stand alone tools to inform the allocation of research funds" (DEST, 2006a: 6). The RQF avoids this impasse by adopting a post-NPM, qualitative, contextual process to evaluate the extra-academic impact of publicly funded research. This process promises to combine impact statements, case studies, appropriate quantitative and qualitative data, and stakeholder testimony; and impact evaluations are made by panels of peers and end users.

A shift in public policy towards the triple bottom line of social, economic and environmental returns will necessarily entail a re-imagined STI policy framework and new breed of post-NPM evaluative metrics. To avoid the pitfalls produced by traditional and novel quantitative indicators of research quality and impact, this paper advocates the use of qualitative impact modelling for research evaluation purposes. This approach provides a fair assessment of the diverse range of research outcomes appropriate to all disciplines and subfields, and allows quality and impact to be framed by stakeholder concerns.

In this light, for research evaluation and foresight planning purposes, we can demonstrate that research may be conceived of in meaningful terms broader than utility, commercialisation and wealth creation. Research evaluation should therefore no longer aspire to the standardised use of blunt quantitative metrics: the future of STI policy and research evaluation is qualitative.

Notes

1. This paper is a revised version of a report by the author, which appears without acknowledgement as Appendix A in CHASS (2005: 56–81), which took von Tunzelmann and Mbula (2003) as its starting point. The author was Chair of the Australian Research Quality Framework Development Advisory Group (RQFDAG) Technical Working Group on Research Impact, so has had some influence on the Australian policy process. However, the purpose of this paper is to offer an independent analytical discussion of the future of quality and impact assessment in the RQF and beyond.
2. In the UK, debate about use of the terms STEM and non-STEM includes discomfort about the label 'non-STEM' defining this sector in opposition to, and as inferior to, the sciences. The acronyms STEM and HASS are popular in Australia, although an amused colleague has recently explained that this may not be an improvement because 'HASS' means 'hate' in German.
3. See REPP (2005) for a comprehensive review of the scientometric literature about the pros and cons of standard quantitative indicators.
4. For example see Butler and Visser (2005: 239), who demonstrate that, in Australia, there is a vast gulf in ISI coverage of all outputs among the physical sciences, social sciences and humanities: chemistry at 90%, psychology 54%, philosophy 32%, economics 27%, history 18%, the arts 13% and law 6%.
5. For example, see Strand (1998: 53–55).
6. Note that the policy discourse talks of 'quality' and 'impact' indicators to refer respectively to the academic and extra-academic impact of research, and this is the terminology the paper will now adopt.
7. The Higher Education Funding Council for England (HEFCE) was asked to consult with the university sector about appropriate bibliometric systems in time for the 2007 Pre-Budget Report. However, the Report does not cover this ground (HM Treasury, 2007).
8. See Donovan (2006) for a critique of applying a 'Rolls-Royce' system to HASS and not STEM.
9. HM Treasury has requested that HEFCE oversee this process.
10. The panel specifications were published in September 2007 and are light on detail about specific impact metrics (DEST, 2007).
11. See Jasanoff (2007) for an account of the mismatch between the grand visions of science policy and the practice of everyday 'mundane' science.
12. See Bozeman and Sarewitz (2005) for an account of how economic thinking and economic metaphors shape science policy in the USA.
13. See HM Treasury (2006a: 33–38) for the original Government proposal, and Sastry and Bekhradnia (2006) for the models and potential effects.
14. Luwel *et al* (1999: 30–31; 156) present an extensive range of scholarly and 'enlightenment' publications aimed at practitioners and the public that academics routinely produce in law and linguistics, which are excluded from standard publication counts.
15. A phenomenon Espeland and Stevens (1998) refer to as "commensuration". See Donovan (2007) for a discussion of social science, citation metrics and commensuration.
16. It is interesting to note that perceived limitations of quantitative indicators and proposed solutions often proceed on the basis of finding a better deal for HASS subjects. However, this paper contends that the limitations of quantitative indicators are simply more visible in the case of HASS. Indeed, the paper maintains that the shortcomings of quantitative measures exposed by investigating the case of HASS acts as a foil to reveal systemic weaknesses that apply, albeit less obviously and perhaps counter-intuitively, to STEM also.
17. See Donovan and Butler (forthcoming) for details of a trial of a similar metric applied to Australian economics departments.
18. For example, conference publications are the primary mode of communication in computing science. A tiered ranking of information and computing technology (ICT) conferences has been produced by Computing Research and Education Australasia (CORE): see <http://www.core.edu.au/>, last accessed 14 September 2007. Also see Butler (2006).
19. For the results of a trial of this novel citation measure applied to history and political science see: <http://www.chass.org.au/papers/bibliometrics/CHASS_Methodology.pdf>, last accessed 14 September 2007; and <http://www.chass.org.au/papers/bibliometrics/CHASS_Report.pdf>, last accessed 14 September 2007.
20. See also Butler and McAllister (2006) who mine citations to non-ISI journals, books and chapters produced by Australian political science departments.
21. For a detailed explanation and critique of the RQF impact model see Donovan (forthcoming).
22. A public servant suggested to me that a suitable impact indicator would be the number of citations to one's research in *Hansard*, and whether these citations were positive or negative. When I asked how we would distinguish between positive or negative citations made by government and opposition, the idea was swiftly abandoned.
23. CSIRO's four research groups were: information technology, manufacturing and services; sustainable minerals and energy; environment and natural resources; and agribusiness and health.
24. In Australia, the National Medical and Health Research Council's (NHMRC) Research Committee on Measures of Research Impact and Achievement (MORIA) has for several years been developing its own impact assessment methodology, and has independently arrived at a process for assessing research impact that is very similar to the Payback model. NHMRC plans to allocate research funds through its Programs Grants Scheme on the basis of previous research impact, especially to humanity, through gains in knowledge, health and commercial development. The idea is to use peer review informed by quality and impact data relating to each of these areas on scales ranging from "activity", "recognition" and "acclaim" to "global impact". However, this methodology has not yet been released into the public domain, but did inform the Australian RQF's approach to measuring research impact.
25. A report by the Royal Academy of Engineering (2000: 17 and

following) attempts to measure excellence in engineering research and similarly uses logic or 'footprint' modelling to visualise categories such as "strategy", "vitality and sustainability" and "scholarship", which are generated through mostly quantitative and some qualitative judgements about research group activities.

26. For example, Seares and Schippers (2005) apply the model to research in the creative and performing arts in Australia, and RAND Europe has used this methodology to assess the impact of an Economic and Social Research Council funding programme on the Future of Work (Nason *et al*, 2007).

References

Allen Consulting Group 2003. *A Wealth of Knowledge: the Return on Investment from ARC-funded Research.* Report to the Australian Research Council. Available at <http://www.arc.gov.au/pdf/ARC_wealth_of_knowledge.pdf>, last accessed 23 March 2007.

Allen Consulting Group 2005. *Measuring the Impact of Publicly Funded Research*. Report to the Australian Government Department of Education, Science and Training. Available at <http://www.dest.gov.au/NR/rdonlyres/6FB9F35B-FADD-44C2-A15F-B145FF9416F6/1392/report.pdf>, last accessed 23 March 2007.

ARC, Australian Research Council 2004. *Draft of Review of the Humanities and Creative Arts Discipline Grouping 2004*. Canberra: ARC, 14 October.

Bourke, Paul, Linda Butler and Beverley Biglia 1996. Monitoring research in the periphery: Australia and the ISI indices. Canberra: REPP Monograph Series no 3, Research School of Social Sciences, ANU.

Bozeman, Barry 2003. Public mapping of science outcomes: theory and method. Monograph of the Public Value Mapping Project. Washington DC: Center for Science, Policy and Outcomes.

Bozeman, Barry and Daniel Sarewitz 2005. Public values and public failure in US science. *Science and Public Policy*, **32**(2), April, 119–136.

British Academy 2004. *That Full Complement of Riches: the Contributions of the Arts, Humanities and Social Sciences to the Nation's Wealth*. London: The British Academy.

Butler, Linda 2003. Modifying publication practices in response to funding formulas. *Research Evaluation*, **12**(1), April, 39–46.

Butler, Linda 2004. What happens when funding is linked to publication counts? In *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. H F Moed, W Glänzel and U Schmoch, pp. 389–340. Dordrecht/Boston/London: Kluwer Academic Publishers.

Butler, Linda 2006. ICT assessment: moving beyond journal outputs. In *Proceedings of the 9th International Conference on Science and Technology Indicators, 7–9 September 2006, Leuven, Belgium*, pp. 29–30. Leuven: Katholieke Universiteit.

Butler, Linda and Ian McAllister 2006. The Hirsch index: is it applicable to the social sciences? In *Proceedings of the 9th International Conference on Science & Technology Indicators, 7-9 September 2006, Leuven, Belgium*, pp. 31–32. Leuven: Katholieke Universiteit.

Butler, Linda and Martijn Visser 2005. Extending citation analysis to non-source items. *Scientometrics*, **66**(2), 327–343.

CHASS, Council for the Humanities, Arts and Social Sciences 2005. Measures of quality and impact of publicly funded research in the humanities, arts and social sciences. Occasional Paper 2. Canberra: Department of Education, Science and Training. Available at <http://www.chass.org.au/op2.pdf>, last accessed 23 March 2007.

Clemens, E S, W P Walter, K McIlwaine and D Okamoto 1995. Careers in print: books, journals and scholarly reputations. *American Journal of Sociology*, **101**(2), 433–494.

CSIRO, Commonwealth Scientific and Industrial Research Organisation 2004. *Annual Report 2003–-04.* Canberra: CSIRO. Available at <http://www.csiro.au/files/files/p2kj.pdf>, last accessed 23 March 2007.

DEST, Department of Education, Science and Training 2004. *Evaluation of Knowledge and Innovation Reforms Consultation Report*. Canberra: Commonwealth of Australia.

DEST, Department of Education, Science and Training 2006a. *Research Quality Framework: Assessing the Quality and Impact of Research in Australia — Impact*. Research Quality Framework Development Advisory Group. Canberra: Commonwealth of Australia. Available at <http://www.dest.gov.au/NR/rdonlyres/D51500F4-A864-4227-B67A-1A5B983A924F/14100/rqf_research_impact.pdf>, last accessed 23 March 2007.

DEST, Department of Education, Science and Training 2006b. *Research Quality Framework: Assessing the Quality and Impact of Research in Australia — the Recommended RQF*. Canberra: Commonwealth of Australia. Available at <http://www.dest.gov.au/NR/rdonlyres/7E5FDEBD-3663-4144-8FBE-AE5E6EE47D29/14867/Recommended_RQF_Dec2006.pdf>, last accessed 23 March 2007.

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. Available at <http://www.dest.gov.au/NR/rdonlyres/6720E340-D0D0-4B8F-B069-D0BCC88CA478/18505/RQFSubmissionSpecifications.pdf>, last accessed 15 November 2007.

Donovan, Claire 2005. The governance of social science and everyday epistemology. *Public Administration*, **83**(3), 597–615.

Donovan, Claire 2006. An instrument too blunt to judge sharp minds. *Times Higher Education Supplement*, 11 August, 14.

Donovan, Claire 2007. The hidden perils of citation counting for Australasian political science. *Australian Journal of Political Science*, **42**(4), December, 655–678.

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.

Donovan, Claire and Linda Butler forthcoming. Testing novel quantitative indicators of research 'quality', esteem and 'user engagement': an economics pilot study. *Research Evaluation*.

Espeland, Wendy N and Mitchell L Stevens 1998. Commensuration as a social process. *Annual Review of Sociology*, **24**, 313–343.

Gallop, Geoff 2007. Towards a new era of strategic government. In *A Passion for Policy: Essays in Public Sector Reform*, ed. John Wanna, pp. 75–89. Canberra: ANU E Press. Available at <http://epress.anu.edu.au/anzsog/policy/pdf/whole_book.pdf>, last accessed 15 November 2007.

Glänzel, Wolfgang and Urs Schoepflin 1999. A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management*, **35**, 31–44.

Hicks, Diana 2004. The four literatures of social science. In *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. H F Moed, W Glänzel and U Schmoch, pp. 473–495. Dordrecht/Boston/London: Kluwer Academic Publishers.

Hicks, Diana, Hiroyuki Tomizawa, Yoshiko Saitoh and Shinichi Kobayashi 2004. Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation*, **13**(2), August, 78–86.

HM Treasury 2006a. *Science and Innovation Investment Framework 2004–2014: Next Steps*. Norwich: HMSO.

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

HM Treasury 2007. *Meeting the Aspirations of British People: 2007 Pre-Budget Report and Comprehensive Spending Review*. Cmnd. 7227. London: The Stationery Office.

Jasanoff, S 2007. Frontier of dreams: basis science in society. Paper presented at the Austrian Science Fund/European Science Foundation Conference on Science Impact: Rethinking the Impact of Basic Research on Society and the Economy, Vienna, 10–11 May. Available at <http://www.science-impact.ac.at/documentation/pdf/Keynote_Jasanoff.pdf>, last assessed 14 September 2007.

Katz, Sylvan J 1999. *Bibliometric Indicators and the Social Sciences*. Report prepared for UK Economic and Social Research Council. Brighton: SPRU.

Luwel, M, H F Moed, A J Nederhof, V De Samblanx, K Verbrugghen and L J van der Wurff 1999. *Towards Indicators of Research Performance in the Social Sciences and Humanities*. Leiden: CWTS.

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

Nason, Edward, Lisa Klautzer, Jennifer Rubin, Stephen Hanney, Steven Wooding and Jonathan Grant 2007. *Policy and Practice Impacts of Research Funded by the Economic and Social Research Council: a Case Study of the Future of Work Programme, Supporting Data*. Report prepared for the Economic and Social Research Council. Cambridge: RAND Europe. Available at <http://www.rand.org/pubs/technical_reports/2007/RAND_TR438.pdf>, last accessed 15 November 2007.

Nederhof, Anton J 2006. Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientometrics*, **66**(1), 81–100.

Nederhof, A J, R A Zwaan, R E De Bruin and P J Dekker 1989. Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: a comparative study. *Scientometrics*, **15**(5–6), 423–435.

REPP, Research Evaluation and Policy Project 2005. Quantitative indicators for research assessment — a literature review. Discussion paper 05/1. Canberra: REPP, Research School of Social Sciences, The Australian National University. Available at <http://repp.anu.edu.au/Literature%20Review3.pdf>, last accessed 23 March 2007.

Royal Academy of Engineering 2000. *Measuring Excellence in Engineering Research*. London: Royal Academy of Engineering.

Sastry, Tom and Bahram Bekhradnia 2006. *Using Metrics to Allocate Research Funds: a Short Evaluation of Alternatives to the Research Assessment Exercise*. Oxford: Higher Education Policy Institute. Available at <http://www.hepi.ac.uk/downloads/23RAEandmetricsfullreport.pdf>, last accessed 23 March 2007.

Seares, Margaret and Schippers, Huib 2005. *The Payback Model and Research in the Performing and Creative Arts*. Canberra: Council for the Humanities, Arts and Social Sciences.

Sivertsen, Gunnar 2006. A bibliometric model for performance based budgeting of research institutions. In *Proceedings of the 9th International Conference on Science & Technology Indicators, 7-9 September 2006, Leuven, Belgium*, pp. 133–135. Leuven: Katholieke Universiteit.

Strand, Dennis 1998, *Research in the Creative Arts*. Canberra: Department of Employment, Education Training and Youth Affairs.

THES, *Times Higher Education Supplement* 2005. Journals ‘top ten’ sparks a rebellion. *Times Higher Education Supplement*, 28 January.

van Leeuwen, Thed 2006. The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics*, **66**(1), 133–154.

van Raan, Anthony F J 1998. Assessment of social sciences: the use of advanced bibliometric methods as a necessary complement of peer review. *Research Evaluation*, **7**(1), April, 1–6.

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

Wooding, Steve, Steve Hanney, Martin Buxton and Jonathan Grant 2004. *The Returns from Arthritis Research Volume 1: Approach, Analysis and Recommendations*. Cambridge: RAND Europe. Available at <http://www.rand.org/pubs/monographs/2004/RAND_MG251.pdf>, last accessed 23 March 2007.