Vol. 8: 83–92, 2008
doi: 10.3354/esep00077

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Printed June, 2008
*Published online January 8, 2008*

*Contribution to the Theme Section 'The use and misuse of bibliometric indices in evaluating scholarly performance'*

OPEN ACCESS

# Using a balanced approach to bibliometrics: quantitative performance measures in the Australian Research Quality Framework

**Linda Butler***

**Research Evaluation and Policy Project, Research School of Social Sciences, Australian National University, Canberra, ACT 0200, Australia**

ABSTRACT: Australia is about to move to a new system of distributing government block grants for research among universities, with the introduction of a process similar to Britain's Research Assessment Exercise (RAE). One of the most significant departures from the current RAE model is that, in the Australian Research Quality Framework, peer judgements will be informed by quantitative performance measures, including bibliometrics. The data will not be used in any formulaic way, but will sit alongside the assessment of other information provided to discipline panels — contextual information provided by the groups being assessed, the full text of the publications they regard as their 'best', and a full list of publications produced in the assessment period. This paper details the metrics to be used in this new framework and outlines some of the reasons why a balanced approach to research assessment was adopted.

KEY WORDS: Bibliometrics · Research assessment · RQF · RAE



## INTRODUCTION

Australia is about to move to a new system for distributing the government's block grant for research among universities, with the introduction of a process similar to Britain's Research Assessment Exercise (RAE) — the Research Quality Framework (RQF). The RQF retains the RAE's expert peer review assessment of the 4 'best' outputs nominated by research active staff, but there are a number of significant innovations in the RQF model.

One major innovation is the attempt to assess not only the quality of the research undertaken within universities, but also the impact of that research outside academia. This broader impact relates to the recognition that research 'has been successfully applied to achieve social, economic, environmental and/or cultural outcomes' (DEST 2006a, p. 10). Another significant difference is the assessment unit — 'groups' of researchers rather than academic organisational units. Universities are free to construct groups in any way they desire, with no requirement for them to conform to the institution's organisational structure. The only constraint is that they are to be organised around Australia's field of research classification, with allowance made in the model for cross-disciplinary groups.

However, in the context of this paper, the most significant variation from the current RAE process is the role of metrics. As well as assessing the quality of the nominated outputs, the deliberations of assessment panels will also be 'assisted by the inclusion of relevant and appropriate quantitative measures of research quality' (DEST 2006a, p. 15). Amongst the suite of indicators to be used, bibliometrics plays a central role. The RAE is now moving to incorporate metrics in their process post-2008; however, they envisage a more central role for quantitative measures than is proposed for the RQF (DfES 2006). The RQF is more balanced with bibliometrics and other quantitative performance measures being used alongside peer review.

*Email: linda.butler@anu.edu.au

## POLICY BACKGROUND

The Australian government has a dual system for funding research in universities. A significant amount of money is distributed by the 2 research councils, the National Health and Medical Research Council (NHMRC) and the Australian Research Council (ARC), via a peer reviewed assessment system. Both agencies distribute the bulk of their funding support in the form of project grants, commonly of 3 years duration. Second, a proportion of the block operating grant to universities (currently in the order of AU$ 1.2 million) is earmarked for research and research training (known as the Research Quantum), with institutions having no restrictions on the internal distribution of these funds.

Since the beginning of the 1990s, this funding has been distributed via a formula. Initially, this formula used institutional success in obtaining national competitive research grant income as the sole basis for allocations, but subsequently student and publication components were added. The weight given to the element relating to grant income decreased from 100% at the start of the funding scheme to 82.5% after additional elements were added in the mid-1990s, and then to 60% after a government review of higher education funding (DETYA 1999). Thus, while there appears to be a dual funding system, it is apparent that success in obtaining grants from the ARC and NHMRC has a flow-through effect, as it directly feeds into the Research Quantum (Marginson & Considine 2000). While the Research Quantum was designed to give universities the capacity to fund long-term, strategic research, the influence of the ARC and NHMRC on the distribution of this money ensures a focus on short to medium term project research.

Other concerns with the Research Quantum, specifically in relation to the publications component, were raised soon after its introduction (Anderson et al. 1996). These issues were considered in a ministerial discussion paper on higher education research and research training issued in June 1999. There were concerns about the reliability of the data and the likelihood that it had 'stimulated an increased volume of publication at the expense of quality' (DETYA 1999, p. 29). Despite the misgivings, the existing allocative mechanism was retained, albeit with an adjustment to the weightings given to each element.

However, the concerns regarding this method of funding research did not disappear, but merely lay dormant for a number of years. In March 2003, an evaluation of the 1999 reforms was published, which included the recommendation that the government and the higher education sector should 'engage in a further discussion on how best to undertake cost-effective research quality assessment' (DEST 2004, p. 53). In so doing, the report's authors urged the government to explore the possibility of designing 'an approach to quality assessment that avoids the RAE's drawbacks' (DEST 2004, p. 53). The greatest concerns were the high implementation cost and the administrative burden on universities. Other issues raised included concerns about game-playing, such as poaching of staff, and the undermining of inter-university and industry collaboration.

In May 2003, the government responded by announcing the establishment of an RQF and appointed Sir Gareth Roberts to chair an Expert Advisory Group, whose remit was to consult widely and develop a model for assessing the quality and impact of research in Australia. Their proposed model was published in February 2006 (DEST 2006a), but a change of minister and further lobbying by the sector led to the establishment of a new Development Advisory Group (DAG) to refine the model. The final model retained most of the key elements of the Roberts proposal, which in turn drew heavily on experiences from the UK's RAE and New Zealand's Performance Based Research Funding scheme.

## THE RQF PROCESS

Under the proposed RQF, the assessment of quality will be undertaken by 13 discipline-based assessment panels, each consisting of 12 members. Because of the need to assess impact, as well as quality, 3 of the panellists will represent the 'end users' of research. The panels are constructed along discipline lines, using the Australian Standard Research Classification scheme devised by the Australian Bureau of Statistics for measuring and analysing research and experimental development undertaken in Australia (ABS 1998). The panels, and the disciplines they cover, are shown in Table 1.

While panels will be asked to judge both the quality and impact of research, this paper focuses solely on the assessment of quality. Panels will be asked to rank the quality of a research group on a scale from 1 (research deemed to fall below the standard of recognised quality work) up to 5 (research that is world leading in its field). In reaching their conclusions about the quality of a group's research, panels will have 2 sets of information available to them. First, each group will submit an Evidence Portfolio (EP), which contains 3 elements: (1) The group's 'context statement', which will provide an overview of the research culture of the group and the institutional context in which it operates, together with any supplementary information they believe demonstrates the quality of their research. These can include esteem measures (e.g. honours, awards, prizes, membership of learned academies, prestigious invited lec-

Table 1. The discipline-based assessment panels established to assess research performance in Australia's Research Quality Framework, detailing the fields of research covered by each panel

| Panel | Title | Discipline coverage |
|---|---|---|
| 1 | Biological sciences | Biochemistry and cell biology, genetics, microbiology, botany, zoology, physiology, ecology and evolution, biotechnology |
| 2 | Physical, chemical and earth sciences | Astronomical sciences, theoretical and condensed matter physics, atomic and molecular physics, nuclear and particle physics, plasma physics, optical physics, classical physics, physical chemistry, inorganic chemistry, organic chemistry, analytical chemistry, macromolecular chemistry, theoretical and computational chemistry, geology, geophysics, geochemistry, oceanography, hydrology, atmospheric sciences |
| 3 | Engineering and technology | Aerospace engineering, manufacturing engineering, automotive engineering, mechanical and industrial engineering, chemical engineering, resources engineering, civil engineering, electrical and electronic engineering, geomatic engineering, environmental engineering, maritime engineering, metallurgy, materials engineering, biomedical engineering, computer hardware, communications technologies, interdisciplinary engineering |
| 4 | Mathematical and information sciences and technology | Mathematics, statistics, information systems, artificial intelligence and signal and image processing, computer software, computation theory and mathematics, data format |
| 5 | Agricultural, veterinary, food and environmental sciences | Industrial biotechnology and food sciences, soil and water sciences, crop and pasture production, horticulture, animal production, veterinary sciences, forestry sciences, fisheries sciences, environmental sciences, and land, parks and agricultural management |
| 6 | Clinical sciences and clinical physiology | Medicine – general, immunology, medical biochemistry and clinical chemistry, medical microbiology, pharmacology and pharmaceutical sciences, medical physiology, dentistry, optometry, clinical sciences (excluding psychiatry), mental health |
| 7 | Public health and health services | Nursing, public health and health services (excluding mental health), complementary/alternative medicine, human movement and sports science |
| 8 | Psychology, etc. | Neurosciences, psychology, psychiatry, cognitive science, linguistics. |
| 9 | Social sciences and politics | Political science, policy and administration, sociology, anthropology, human geography, demography |
| 10 | Economics, commerce and management | Economic theory, applied economics, economic history and history of economic thought, econometrics, accounting, auditing and accountability, business and management, banking, finance and investment, transportation, tourism, services |
| 11 | Law, education and professional practices | Education studies, curriculum studies, professional development of teachers, journalism, communication and media, librarianship, curatorial studies, social work, law, professional development of practitioners, justice and legal studies, law enforcement |
| 12 | Humanities | History and philosophy of science and medicine, art history and appreciation, language studies, literature studies, cultural studies, historical studies, archaeology and prehistory, philosophy, religion and religious traditions |
| 13 | Creative arts, design and built environment | Architecture and urban environment, building, urban environment and building, performing arts, visual arts and crafts, cinema, electronic arts and multimedia, design studies |

tures), service to journals and conferences, collaborative activities, and competitive grant success. A number of fields will be reported via drop-down menus that will enable them to be routinely summarised for the assessment panels. (2) The full 'body of work' for the group, listing all publications or other forms of research outputs for the 6 year assessment period. The guidelines for each panel will detail the publication types to be included in the body of work, as this will vary from discipline to discipline. (3) The 'best outputs' of the group. Each researcher in a group nominates the 4 publications he/she believes were their 'best' outputs for the 6 year period covered by the RQF, and which demonstrate the quality of their research. It is clear from the size of the panels that panellists will have neither the expertise nor the capacity to assess all the nominated outputs. It is anticipated that many will be assessed using an extensive pool of external experts identified for this purpose. Secondly, panels will be provided with a number of quantitative indicators, which will be centrally collated and analysed and derived from the group's body of work.

In 2006, a number of expert working groups were established by the Australian Department of Education, Science and Training (DEST) to flesh out the recommended assessment model, and I chaired the Qual-

ity Metrics Working Group (QMWG). The QMWG was asked 'to identify the forms and sources of available data that may assist the process of research assessment by expert review' (DEST 2006a). Membership was drawn from across the higher education sector, covering the broad range of universities and disciplines it encompasses. Members discussed in detail 4 issues fundamental to the introduction of metrics in a national research assessment exercise before providing the department with recommendations on the measures to be used. These were the role that metrics were to play, the level of aggregation at which they were to be applied, the source of the data, and the number and range of metrics.

In the following sections I will discuss the debate that surrounded each of the 4 fundamental issues, then detail the final recommendations on metrics made by the QMWG to the department, and conclude with a discussion of the role of quantitative measures in general, and bibliometrics in particular, in the proposed RQF.

## ROLE OF METRICS

Most of the debate surrounding the use of quantitative performance indicators to assess research has focused on bibliometrics. Much of that debate is triggered by concerns about the substantive role they will play. Those that vehemently oppose their use are worried that the proponents of the measures are attempting to replace peer review. Supporters of a fully metrics approach have suggested that, for cost reasons, the peer review component in the British RAE could be largely replaced by citation analyses, given the very high correlations with past rankings (Smith & Eysenck 2002). However, another study of the correlation between RAE scores and bibliometric measures has shown that, while the correlation is high, there are deviant cases (Warner 2000). Their existence raises concerns about the straight replacement of peer review by bibliometrics where funding decisions are coupled with research assessment.

The generally good pattern of correspondence between quantitative indicators and peer judgements has often led to them being characterised as 'objective' measures in contrast to the subjective character of the peer review. However, it should be remembered that the indicators themselves are based in part on peer decisions — journal articles embody the peer evaluations that have led to acceptance for publication, and grant success embodies the peer assessment of applications (Weingart 2003).

Most informed researchers do not see indicators as a replacement for peer evaluation, but rather as a way to make the results of research assessment debatable and to offer experts additional information (van Raan & van Leeuwen 2002). Bibliometric indicators can make peer review more 'transparent' and counterbalance its shortcomings (van Raan & van Leeuwen 2002, Tijssen 2003, Aksnes & Taxt 2004). They are seen as a useful resource in cases of doubt within panel discussions of peers (Moed & van Raan 1988).

In addition, bibliometric indicators can be used to highlight gaps in the knowledge of peers — as 'triggers to the recognition of anomalies' (Bourke et al. 1999, p. 1). Where the indicators do not align with peer evaluation, then the reasons must be sought. It may be due to problems with the indicators, or it may be that the experts have an incomplete knowledge of the research they are assessing. Inconsistencies between quantitative data and peer review are likely to trigger additional, deeper analyses of the performance of units being evaluated by those conducting the assessment.

The QMWG saw the role of metrics as enhancing and complementing the panel assessment process. The measures were seen as just 1 of 3 elements of the panels' deliberations on quality, alongside the assessment of the nominated 'best' outputs and an evaluation of the information in the context statements. There was a strong preference for panels to have access to the data from the beginning of the assessment process.

It was acknowledged that quantitative measures have the potential to exert undue influence on all panel decision-making. The QMWG proposed 2 strategies to lessen these concerns. The first was to recommend that no attempt be made to aggregate the indicators to produce a single score. The second strategy was to propose that the role of each of the various elements of a research group's EP be made transparent by stipulating the weight given to each in the overall assessment. This proposal is in line with the UK's 2008 RAE practice, where a minimum weighting for each RAE element is specified (research outputs 50%, research environment 5%, and esteem indicators 5%), but panels have the flexibility to determine the actual weighting for each element in their discipline within these broad limits. The QMWG left open for further discussion a recommendation on what the 3 weightings should be in the RQF context.

The formal recommendation of the QMWG was that 'metrics should be used to inform decision making' (DEST 2006b, p. 2). To make the process for the interaction of metrics and the discipline panel assessment process more explicit, this recommendation was elaborated on as follows: (1) metrics should not be used in isolation, such as the 'shadow exercise' proposed for the 2008 UK RAE; (2) the data should be available to discipline panels from the beginning of the assessment process; (3) no attempt should be made to aggregate indicators to produce a single 'quality score'.

### Level of aggregation

It was clear from the outset of QMWG deliberations that bibliometric indicators were one of the favoured metrics. Discussion about the validity of citation indicators has shown that it is important to allow for the highly skewed nature of the distribution of citations. Most publications receive relatively few citations, with only a tiny minority being heavily cited (Garfield 1979). It is possible that the average citation rate of a research unit is high because 1 article of the group is highly cited, with other publications receiving very few. Concerns relating to this skewed distribution are most critical if the number of publications is small — less than 50 publications (Moed et al. 1995). van Raan proposes 10 or 20 publications per year — the usual output of a research group in the sciences — as a sufficient basis for bibliometric calculations, while rejecting those based on a few publications per year (van Raan 2000). The RQF assessment period covers 6 years, so for a research group (also the unit of assessment for the RQF), van Raan's productivity threshold suggests a minimum level of 100 publications.

The primary concern of the QMWG was ensuring that the proposed metrics would be based on a sufficient body of work in order to be robust, thus giving confidence in their use in the RQF process. Preference was stated for bibliometric measures to rest on sets of at least 100 publications. It was obvious that limiting the measures to those publications nominated as 'best' outputs by members of the group was unlikely to provide a sufficient number, particularly for small groups (the minimum size of a group is to be 5 researchers). There was also concern about the ability of measures to discriminate between the performances of research groups when only a small fraction of their output was being assessed.

The QMWG therefore made the recommendation that the measures should be applied to the total 'body of work' of the Research Grouping for the RQF census period (DEST 2006b, p. 3).

### Source of data

As I have indicated, it was assumed in the deliberations of the QMWG that bibliometrics were likely to be recommended, at least for science disciplines, and there was considerable discussion on how the data could be extracted for such an extensive research assessment process, and who would undertake the required analyses. At the time of these deliberations, citation analysis rested primarily on data from Thomson's *Web of Knowledge* indexes. The QMWG noted that Australian universities had varying capacity to undertake in-house analysis of this data. At that point, only the Australian National University had access to the raw data files from Thomson, allowing them to undertake sophisticated analyses. A number of the research-intensive universities had access to a range of Thomson products that they could use for national and international comparisons, while other universities had more limited options and data sources available to them.

If metrics are based solely on data available to all universities, their range would be extremely limited. If the RQF sought to use more sophisticated metrics, few universities would have access to the necessary data sources. If universities were given free reign to provide whatever metrics they were able to construct, panels were likely to be faced with a conglomeration of incompatible measures. Therefore, the only viable option was for the analysis to be undertaken centrally.

In contrast, data for other possible metrics considered (grant income and ranked outputs) could not be obtained from a central source and would have to be provided by the research groups.

The QMWG therefore made the recommendation that 'citation analysis should be undertaken centrally ... other measures are to be constructed from data supplied by the Research Groupings' (DEST 2006b, p. 5).

### Number and range of metrics

The character of research 'quality' is complex and multidimensional. No single quantitative measure can address all its facets. In addition, since each indicator has different strengths and weaknesses, it has been suggested that evaluations should always incorporate more than 1 indicator (Martin & Irvine 1983), and that indicators should never be used in isolation, especially if applied to individual groups (van Raan 1996). This is also the proposed standard practice for OECD surveys of R&D activities (Godin 2002). The Centre for Science and Technology Studies (CWTS) puts this into practice by always using a set of indicators, their 'crown indicators', in evaluative studies (described in van Leeuwen et al. 2003).

The selection of a suitable suite of indicators for a given evaluation task is by no means clear-cut. In a number of studies, Australian researchers have been sent questionnaires asking which indicators best reflect the work in their field. For example, in a study conducted by Hattie and colleagues (Hattie et al. 1991, Tognolini et al. 1994, Print & Hattie 1997), scientists rated a large list of indicators divided into 6 groups. Similar questionnaires were used in a study by Grigg & Sheehan (1989) and by a research group chaired by Linke (NBEET 1993). While the lists were comprehensive, none of the studies came up with a preferred set.

Martin & Irvine (1983) suggest identifying the combination of indicators that provides the strongest correlations and thereby the best combination. However, important information may be lost if indicators are chosen on the basis of their convergence — contradictory results could enhance, rather than detract from, an assessment of performance.

While many indicators have a common starting point — a particular data source — their final form may be quite dissimilar. There is considerable room for 'manipulation by selection, weighting and aggregating indicators' (Grupp & Mogee 2004, p. 87). These concerns have been specifically raised in relation to bibliometric indicators, where a special session of the major international conference in the discipline was devoted to the issue.[1]

Taking note of the strong evidence provided in the literature on quantitative performance indicators, the QMWG therefore considered it essential that a 'basket of measures' should be assembled. By not relying on a single metric, the possibility of unintended and undesired responses to the measures would be reduced. The group recommended that some generic measures, applicable in all discipline panels, should be used to ensure confidence in cross-panel comparability. However, it was acknowledged that standard citation measures were not applicable in all disciplines, being primarily restricted to the sciences, and that an attempt should be made to develop and test alternative equivalent metrics for the applied sciences (particularly computer science), the arts, the humanities and many of the social science disciplines.

The QMWG therefore made the following recommendations: (1) discipline panels should employ a basket of measures; and (2) from the list of proposed measures, each discipline panel should be free to choose the combination of indicators most appropriate for their disciplines, with some generic measures across all panels (DEST 2006b).

[1]Proceedings of the Workshop on 'Bibliometric Standards' at the 5th International Conference of the International Society for Scientometrics and Informetrics (ISSI) are published in Vol 35(2) of *Scientometrics*. This volume contains only papers taken from the workshop and is accordingly named: Proceedings of the Workshop on 'Bibliometric Standards'

## PROPOSED METRICS

When these fundamental issues had been addressed by the QMWG, the choice of metrics became relatively straightforward, once the desirable characteristics of such measures were taken into account. The QMWG asserted that the indicators used should (1) measure some aspect of research quality (and not, for example, refer solely to productivity), (2) be transparent, (3) be reliable when applied to a 6 year time frame, (4) not involve an excessive financial or time burden to the sector, (5) avoid undue complexity, and (6) encourage desirable responses from researchers and institutions. This last characteristic was regarded as of paramount importance.

The QMWG also took note of the varying publication practices among disciplines when making their recommendations. Data that Australian universities report each year to the government was analysed to determine, for each field of research, the proportion of output that appeared in Institute for Scientific Information (ISI) journals. These data are reported in Table 2.

It is clear from Table 2 that the use of bibliometric analyses is defensible for the sciences, but for most disciplines in the social sciences and humanities, the use of standard bibliometric measures cannot be sup-

Table 2. Proportion of Australian university publications appearing in Institute for Scientific Information (ISI)-indexed journals, by field of research (classified according to the standard Australian research classification scheme).*

| Field of research | Total publications[a] | No. ISI journal publications[b] | % publications in ISI |
|---|---|---|---|
| Chemistry | 2430 | 3234 | 83 |
| Physics | 2506 | 2964 | 74 |
| Biology | 4571 | 4626 | 72 |
| Medicine | 22 631 | 18 075 | 65 |
| Agric, Veterinary, Environ | 5157 | 3487 | 61 |
| Earth Sciences | 2060 | 2256 | 60 |
| Mathematics | 2078 | 2735 | 55 |
| Psychology | 2294 | 2040 | 52 |
| Engineering | 8819 | 9650 | 35 |
| Philosophy | 659 | 613 | 26 |
| Economics | 1903 | 1917 | 24 |
| Studies in Human Society | 1678 | 1070 | 18 |
| Politics and Policy | 1195 | 993 | 15 |
| Computing | 2237 | 2904 | 15 |
| History | 1095 | 1160 | 14 |
| Management | 4788 | 4826 | 11 |
| Language | 1940 | 1167 | 10 |
| Education | 4524 | 3165 | 9 |
| The arts | 2272 | 446 | 7 |
| Architecture | 1340 | 936 | 5 |
| Communication | 393 | 334 | 4 |
| Law | 3196 | 1925 | 4 |
| Total | 86 720 | 78 709 | 43 |

[a]Total publications include counts of books, book chapters, refereed journal articles and refereed conference publications
[b]Number of articles published in journals indexed by ISI

*Errors were found in this table after publication. Please see corrected table in Erratum

ported, even when they are not the sole indicators of performance as is the case with the RQF. The QMWG strongly supported further investigation into alternative citation metrics for those disciplines where Thomson databases covered less than half their output. The following 3 metrics were recommended for the RQF.

### Ranked outputs

The publications of research groups will be classified into 4 prestige tiers according to where they appear. Journals will be ranked in all disciplines. In addition, other outlets will be ranked where they carry publications that are important for a particular discipline, e.g. book publishers for the social sciences and humanities, conferences for computer science, and venues for the performing arts. These rankings are being undertaken by discipline workgroups and involve comprehensive consultation throughout the higher education sector. Disciplines aim to classify journals, publishers, conferences and/or venues into tiers according to the following distribution: Tier A* (5%), Tier A (15%), Tier B (30%) and Tier C (50%). The central RQF information system will produce an analysis of the relevant outputs from the full body of work for each research group, allocating outputs into the 4 prestige tiers based on these rankings. It will then produce a summary report to panels showing the number and distribution of publications across tiers for each research group being assessed by that panel.

### Citation data

The QMWG discussed an extensive range of possible measures based on citation data. These covered 2 types of analysis, and both were supported for use in the RQF.

#### Standard bibliometrics

These are measures based on the indexed journal literature. I have labelled them as 'standard' because they encompass indicators that are routinely used in citation analyses. The proposal is to apply these measures for disciplines where at least half their publications appear in the indexed journal literature. The following are the 2 standard measures to be used in the RQF in this category:

*Citations per publication.* An analysis will be undertaken of journal articles from the full body of work for each research group, obtaining total publication and citation counts and calculating a citation per publication rate for the group's œuvre. This data will be provided in summary form to the panels, together with relevant world and Australian benchmark data for the disciplines they cover.

*Centile distribution of a group's output.* A second analysis will be undertaken of journal articles from the full body of work for each research group, producing a distribution of all articles across centile bands. This will show the number and proportion of each group's articles judged to be among the top 1%, 10%, 20% and 50% most highly cited publications for their discipline in any given year. The benchmark data on which this analysis is based will be obtained from Thomson and will be provided to the sector prior to the submission of research groups.

#### Non-standard bibliometric

Where support exists for the methodology, non-standard measures may be applied in some disciplines in the social sciences and humanities. The data will be centrally collated and extracted, and the extraction of citation counts will be extended to books, book chapters, and journal articles not traditionally covered by the major citation indices (Butler & Visser 2006). Because this novel approach has not previously been used in an extensive research assessment exercise, particularly one on which significant funding implications rest — and because the process is labour intensive — it is only likely to be applied to a limited number of disciplines.

### Grant income

As outlined above, grant income will be provided by groups in their EPs. It is anticipated that the 4 categories of income judged relevant to this exercise will be entered by groups into defined fields. The categories to be reported include competitive grant income (category 1), other public sector income, industry and other income, and funding for competitive research centres. Category 1 income will be used as a quality metric for all disciplines. Most disciplines will restrict data on other research income for use as an indicator of impact, rather than quality.

The QMWG also provided further rationale for the choice of particular indicators, and the reasons for rejecting others that had been proposed.

### Rationale for citation data

The QMWG believed that the overall thrust of the citation measures was to encourage researchers to

achieve highly cited publications — aiming for quality rather than focusing on quantity — and believed this was a desirable behavioural outcome.

Some concerns were raised that relying solely on citation per publication rates as a measure could tempt some research groups to limit subsequent output if they had produced a highly cited article, thus enhancing their citation average. To counteract this, a second measure identifying the citation percentile to which each publication belonged (based on discipline-specific yearly citation benchmarks) should be used. While citation rates allow comparison between research groups of different sizes, it was also felt that it would be essential for the discipline panels to be provided with the data that underpinned these averages (i.e. total publication and citation counts and the number of publications in each percentile).

### Rationale for grant income data

While some members of the QMWG regarded grant income as an input to research, others supported it as an indicator of 'quality', given the assessment of researcher's track record embodied in the grant application process. A decision was made to include it as a metric for the RQF, as it could contribute to the holistic picture of the group for the assessors. The QMWG recommended limiting allowable grants to those listed as 'Category 1' by DEST, but expanding the coverage to include significant international agencies, e.g. National Institutes of Health (NIH), European Union (EU) — the qualifier being that they are peer reviewed funding programs. Panels should decide the international funding sources relevant to their own disciplines.

The measure was regarded as generic in that it could be applied in all discipline panels, though benchmarks (e.g. average income per staff member) would obviously vary significantly across panels.

### Rationale for ranked outputs

This indicator is also generic, though its construction would be discipline-specific, as the type of output/outlet to be ranked would vary — e.g. journals, conferences, publishers and exhibition venues. The QMWG determined that because the measure is to be used to inform the assessment process, there is no need to weight the tiers and attempt to derive a score for each research group.

As with the citation measures, the main thrust of this indicator is to encourage researchers to publish in the most prestigious outlets for their discipline, a response that the QMWG believed was a desirable outcome.

It was noted that many disciplines, including computer science, education and the creative arts, had already commenced developing output rankings. However, this work would have to be validated, and other disciplines would need to develop rankings relevant to their own outputs through committees or workshops. Discipline peak bodies and the learned academies were identified as potential drivers of these developments.

### Rationale for rejecting measures

Many additional measures were discussed as potential metrics but subsequently rejected. The 4 most seriously considered and discussed in-depth were as follows

- **Webmetrics.** At this point, measures based on these statistics have not been fully developed as assessment tools. It is anticipated that for future RQF rounds they may be more robust, though there was concern over how readily such data could be audited, and a belief that they might be easily manipulated
- **Collaborations.** With the exception of jointly-authored publications, the effort required to collect data on collaborations is a time consuming process, and its use as a formal metric was rejected
- **Contextual metrics.** The QMWG identified a number of measures that could not be used as stand-alone metrics, but which might be reported in a research group's context statements. These included measures that related more to a researcher's whole career (e.g. esteem measures, service to journals) or were more related to identifying the capacity for generational change (such as research student data)
- **ISI Impact Factor.** While ISI's Impact Factor is used extensively throughout the scientific community, it was rejected for a number of reasons. It was believed that actual citation counts are a far better citation measure for judging the performance of groups than surrogates based on the average citation rates of the journals which carry that work. There were also concerns about the way in which the indicator is calculated and anecdotal evidence of increasing manipulation of the indicator by a few journal editors. Even when ranking journals, some disciplines had already made it clear that they wished to look beyond the Impact Factor and undertake a more detailed assessment of the quality of journals.

## DISCUSSION

The character of research 'quality' is complex and multidimensional. No single quantitative measure, or even a 'basket' of indicators, can always provide an

'accurate' and unambiguous result. Nor can a small panel of peers be expected to combine sufficient knowledge of the performance of all of a nation's institutions and researchers active in their discipline to enable them to arrive at error-free judgements. The most sensible approach is to combine the 2 methods — assemble a group of highly qualified experts in the discipline and arm them with reliable, discipline-specific data to assist their deliberations. The data should be viewed as triggers for recognising anomalies. As has been demonstrated by many studies, the 2 methods will usually produce similar results. Reaching the same conclusion from 2 perspectives will increase confidence in the assessments. The bulk of the time panel members have available to them can be productively used to determine the reasons for discrepancies in those cases where the 2 methods result in different outcomes — whether this is due to problems with the data or to gaps in the knowledge of panel members.

The challenge facing policy makers is to identify robust indicators, particularly for those disciplines not well-served by standard citation analysis. However, considerable progress is being made by a small number of units worldwide, particularly the Centre for Science and Technology Studies at the University of Leiden and my own unit at the Australian National University. This work is providing a path to using citation data in novel ways more sensitive to the output of the humanities and social sciences, and is demonstrating effective methods of ranking outputs into prestige bands in a way developed and supported by researchers in the discipline.

The RQF is due to be implemented in 2008, with funding to be allocated, based on the results, starting in 2009. It has the potential to follow the 'balanced approach' methodology if all the recommendations are carried through to fruition. Peer review is an essential component of the scheme — assessing both the quality of nominated research outputs and the claims of research impact beyond academia. But quantitative measures, and specifically bibliometrics, are also being incorporated into the model. They will inform panel deliberations, rather than being used in any aggregated, formulaic way. Additionally, these measures will be sensitive to disciplinary characteristics and their different publication practices.

The discussion of the validity of using quantitative data to assess research performance must necessarily return to a reflection on the role of quantitative indicators in the assessment of research. The significance of many of the concerns on validity is reduced when the indicators are used as an aid to peer review where differences between values can be interpreted and exceptions can be discussed. They are, however, at the forefront of concerns related to their use in isolation from informed peer input.

## LITERATURE CITED


Aksnes D, Taxt R (2004) Peer review and bibliometric indicators: a comparative study at a Norwegian university. Res Eval 13:33–41

Anderson D, Johnson R, Milligan B (1996) Performance-based funding of universities. National Board of Employment Education and Training, Canberra, Commissioned Report No. 51

ABS (Australian Bureau of Statistics) (1998). Australian Standard Research Classification. ABS Catalogue No. 1297.0

Bourke P, Butler L, Biglia B (1999). A bibliometric analysis of biological sciences research in Australia. Department of Education, Training and Youth Affairs, Canberra, DETYA No. 6307HERC99A

Butler L, Visser M (2006) Extending citation analysis to non-source items. Scientometrics 66:327–343

DEST (Department of Education Science and Training) (2004) Evaluation of knowledge and innovation reforms consultation report. DEST, Canberra. Available at: www.dest.gov.au/NR/rdonlyres/654E1226-6F91-44C5-BDEA-FE8FCB228E88/2788/pub.pdf

DEST (Department of Education Science and Training) (2006a) Research quality framework: assessing the quality and impact of research in Australia. Available at: www.dest.gov.au/NR/rdonlyres/7E5FDEBD-3663-4144-8FBE-AE5E6EE47D29/17543/RecommendedRQF2.pdf

DEST (Department of Education, Science and Training) (2006b) Research quality framework assessing the quality and impact of research in Australia: quality metrics. Available at: www.dest.gov.au/NR/rdonlyres/EC11695D-B59D-4879-A84D-87004AA22FD2/14099/rqf_quality_metrics.pdf

DETYA (Department of Employment Training and Youth Affairs) (1999). New knowledge, new opportunities. DETYA, Canberra. Available at: www.dest.gov.au/archive/highered/otherpub/greenpaper/fullpaper.pdf

DfES (Department for Education and Skills) (2006) Reform of research assessment and funding (Letter from DfES to HEFCE). Available at: www.hefce.ac.uk/News/HEFCE/2006/LetterAJtoDY.pdf

Garfield E (1979) Perspectives on citation analysis of scientists. In: Garfield E (ed) Citation indexing — its theory and application in science, technology, and humanities. John Wiley & Sons, New York

Godin B (2002) Outline for a history of science measurement. Sci Technol Human Values 27:3–27

Grigg L, Sheehan P (1989) Evaluating research: the role of performance indicators. Office of the Academic Director of Research, The University of Queensland, Brisbane

Grupp H, Mogee ME (2004) Indicators for national science and technology policy. In: Moed H, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research. Kluwer Academic Publishers, Dordrecht

Hattie J, Tognolini J, Adams K, Curtis P (1991) An evaluation of a model for allocating funds across departments within a university using selected indicators of performance. Department of Employment, Education and Training, Canberra

Marginson S, Considine M (2000). The enterprise university: power, governance, and reinvention. Cambridge University Press, Cambridge

Martin BR, Irvine J (1983) Assessing basic research: some partial indicators of scientific progress in radio astronomy. Res Policy 12:61–90

Moed H, van Raan AJ (1988) Indicators of research performance: applications in university research policy. In: van Raan AJ (ed) Handbook of quantitative studies of science and technology. Elsevier, Amsterdam

Moed HF, De Bruin RE, van Leeuwen TN (1995) New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. Scientometrics 33:381–422

NBEET (National Board of Employment, Education and Training) (1993) Research performance indicators survey. NBEET, Canberra, Commissioned Report No. 21

Print M, Hattie J (1997) Measuring quality in universities: an approach to weighting research productivity. High Educ 33:453–469

Smith A, Eysenck M (2002) The correlation between RAE ratings and citation counts in psychology. Department of Psychology, University of London, London, Technical report. Available at www.pc.rhul.ac.uk/vision/citations.pdf

Tijssen R (2003) Scoreboards of research excellence. Res Eval 12:91–103

Tognolini J, Adams K, Hattie J (1994) A methodology to choose performance indicators of research attainment in universities. Aust J Educ 38:105–117

van Leeuwen TN, Visser MS, Moed HF, Nederhof TJ, van Raan AFJ (2003) The holy grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. Scientometrics 57:257–280

van Raan AJF (1996) Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics 36:397–420

van Raan AJF (2000) The Pandora's box of citation analysis: measuring scientific excellence — the last evil? In: Cronin B, Atkins HB (eds) The web of knowledge. Information Today, Medford, NJ

van Raan A, van Leeuwen T (2002) Assessment of the scientific basis of interdisciplinary, applied research — application of bibliometric methods in nutrition and food research. Res Policy 31:611–632

Warner J (2000) A critical review of the application of citation studies to the research assessment exercises. J Inf Sci 26:453–460

Weingart P (2003) Evaluation of research performance: the danger of numbers. In: Bibliometric analysis in science and research: applications, benefits and limitations. Schriften des Forschungszentrums Jülich, Vol 11 (Proc 2nd Conf Central Library, Jülich, 5–7 November 2003): 7–19

*Editorial responsibility: Howard Browman, Storebø, Norway and Konstantinos Stergiou, Thessaloniki, Greece*

*Submitted: September 2, 2007; Accepted: November 10, 2007*
*Proofs received from author(s): December 7, 2007*