Science and Public Policy, 34(8), October 2007, pages 565–574
DOI: 10.3152/030234207X254404; http://www.ingentaconnect.com/content/beech/spp

Assessing university research: a plea for a balanced approach

Linda Butler

The use of quantitative performance measures to assess the quality of university research is being introduced in Australia and the UK. This paper presents the case for maintaining a balanced approach. It argues that 'metrics' have their place, and can make the process more efficient and cost-effective, but that peer review must be retained as a central element in any research assessment exercise. The role of metrics is as 'a trigger to the recognition of anomalies', rather than as a straight replacement for peer review.

RECENT PROPOSED CHANGES in university research assessment are exhibiting a strong cyclical trend. What once was trialled briefly in Australia and abandoned very quickly, has been proposed to replace the UK Research Assessment Exercise (RAE) — distributing research funds on the basis of success in competitive grants. What is about to be abandoned in the UK, has been proposed for Australia — a resource-intensive research assessment exercise relying heavily on peer review.

This oversimplifies the forces at play, but it is of some concern that proposals are being made without any apparent attempt to learn from past experience. This lack of sophistication in the debate has led to a polarising of opinion. Advocates are increasingly aligning themselves in one of two camps. There are the 'red devils' who believe that the application of quantitative measures to assess research is indeed 'the work of the devil' and not to be contemplated under any circumstances. In their view, only peer review can assess the quality of research, and its use in this context is sacrosanct. Proponents for the other side of the debate can be recognised by the 'rose-coloured glasses' that mask their vision. They have an almost total belief in the efficacy of quantitative indicators for assessing research performance, being indifferent to, or ignorant of, any of their shortcomings.

My paper will present a plea for sanity to prevail and for all stakeholders in the debate — researchers, university administrators and bureaucrats alike — to pause and assess the vast wealth of experience that exists from research studies or evaluation exercises around the globe, and to take a more balanced approach to research assessment.

The discussion in this paper focuses on recent developments in the UK and Australia, but the issues it raises, and the model of evaluation proposed as 'best practice', are relevant for any sector-wide research assessment.

Australia

The Australian Government has a dual system for funding research in universities. A significant amount of money is distributed by the two 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 three years' duration.

Second, a proportion of the block-operating grant to universities (in the order of 5%, or approximately Aus$1,039 million) is earmarked for research and research training (known as the Research Quantum), with institutions having no restrictions on the

Dr Linda Butler is with the Research Evaluation and Policy Project, Research School of Social Sciences, The Australian National University, Canberra, ACT 0200, Australia; Email: linda.butler@anu.edu.au; Tel: +61 2 6125 2154; Fax: +61 2 6125 9767.

 0302-3427/07/080565-10 US$08.00

Linda Butler is head of the Research Evaluation and Policy Project at the Australian National University. One of the major findings of her bibliometric research, on the effects of using publication counts to determine the distribution of some research funding to universities, has had significant public policy impact. Her current research is concentrating on the identification and development of novel quantitative indicators for disciplines where standard bibliometric techniques are not appropriate. She is a senior advisor to the Australian Government on the application of metrics to assess the quality of research.

internal distribution of these funds. Since the beginning of the 1990s, this funding has been distributed via a formula. Initially this formula, which came to be known as the Composite Index, used institutional success in obtaining national competitive research grant income as the sole basis for allocations, but subsequently student-related 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, then to 60% after a Government review that encompassed all higher-education funding (DETYA, 1999a).

Thus, while there is the appearance of a dual funding system, as Marginson and Considine (2000) noted, success in obtaining funding from ARC and NHMRC has a flow-through effect, as it directly feeds into the Research Quantum (RQ). While the RQ 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 for which the bulk of their funding is given.

Other concerns with the RQ, specifically in relation to the publications component, which was based on the number of publications produced, were raised soon after its introduction (Anderson *et al,* 1996). These concerns were taken on board in a ministerial discussion paper on higher-education research and research training, issued in June 1999 (DETYA, 1999b):

> The publications component of the Composite Index has been subject to a range of criticisms since its implementation in 1995. These concern the reliability of the information provided by institutions, the costs of data collection and the incentives created by the inclusion of a publications component in the index. It seems likely that the publications component of the Composite Index has stimulated an increased volume of publication at the expense of quality ... on these grounds, the government proposes ... to drop the publications measure in any future indices used to allocate block research funds.

Not all universities were keen to see the removal of the publications element. The notional proportion of the RQ to be distributed via publication counts was 10% in 1999. However, over half the universities, particularly smaller institutions, received more than 10% of their RQ allocation through publications. For one university, the proportion was above 40%, for another five, it was more than 20%. It was predominantly the research-intensive older universities that were at, or even below, the notional 10% level (DEST, 1998).

It was therefore hardly surprising that, in its response to the discussion paper, the Australian Vice Chancellors' Committee (AVCC), representing all 36 institutions that received funds via the RQ, argued for the retention of the publications component: "... of the quality measures that might be utilised, 'publications' is the only measure able to fulfil all the requirements suggested above for a driver of sector-wide funding" (AVCC, 1999). The Government was swayed by the submission of the AVCC and others and in its final policy statement all talk of removing the publications component had disappeared (DETYA, 1999a).

The concerns over this method of funding research did not disappear, but merely lay dormant for a number of years. In March 2004, an evaluation of the 1999 reforms was published that 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). In so doing, they urged the Government to explore the possibility of designing "an approach to quality assessment that avoids the RAE's drawbacks". Of most concern 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 2004, the Government responded by announcing the establishment of a Research Quality Framework (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 September 2005 (DEST, 2005), but a change of minister and further lobbying by the sector led to a new Development Advisory Group being established to refine the model. The final model retained most 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 (PBRF) scheme.

The RQF is due to be implemented in 2008, with funding consequences to flow from 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. However, quantitative measures are also being incorporated into the model. They will be used to inform panel deliberations, rather than

being used in any aggregated, formulaic way. Additionally, the measures used will be sensitive to disciplinary characteristics and their different publication practices.

United Kingdom

The UK Government system for funding higher-education research shares many similarities with that operating in Australia. It also is based on a dual funding system, with project grant funding distributed through a number of research councils, and a mechanism for distributing block funding to the universities. The process used for the latter since the mid-1980s has been the RAE. There have been five iterations of the RAE to date, with a sixth to take place in 2008.[1] The RAE relies heavily on peer assessment, with the primary focus of the review being the outputs nominated by researchers as their four 'best' for the census period.

In June 2002, Sir Gareth Roberts was asked to review the UK's research assessment process on behalf of the higher-education funding bodies. The review raised a number of concerns about the RAE process, including the administrative burden, the game-playing that was beginning to undermine the process and the failure to recognise certain research activities (Roberts, 2003).

Roberts proposed some far-reaching changes to the RAE, suggesting different assessment processes according to the intensiveness of the research enterprise. The least research-intensive universities could be assessed separately from those seeking funding via the RAE. The less competitive departments in the remaining institutions could be assessed by proxy measures, or 'metrics' in current terminology. Only the most competitive work of institutions would be submitted for an intensive expert peer assessment. In this way, Roberts sought to reduce the load of the process.

Many of the recommendations of the Roberts Review were not adopted by the funding councils and the Government announced that the 2008 RAE would proceed on much the same basis as previously. There were to be some modifications, such as a continuous, rather than discrete, ranking scale. There were also administrative changes, with the 67 discipline-based sub-panels being grouped under 14 overarching main panels. The aim of this restructure was to address concerns regarding comparability of assessments across disciplines.

In 2006, the UK Department for Education and Skills (DfES) surprised the higher-education sector by announcing its intention to replace the existing RAE process with a metrics-based approach to assessment and funding after the 2008 iteration, and it published a consultation document (DfES, 2006). Perhaps the sector should not have been caught so unawares, as there had been several warning signs. As discussed, Roberts had explicitly addressed the cost and administrative burden of past RAEs in his review, proposing a multi-track assessment process to try to reduce the load (Roberts, 2003).

In 2003, Geuna and Martin had analysed the advantages and disadvantages of performance-based funding in a number of countries and came to the conclusion that "while initial benefits may outweigh the costs, over time such a system seems to produce diminishing returns. This raises important questions about its continued use" (Geuna and Martin, 2003). Their theoretical discussion is supported by what many analysts believe has now happened — the focus on quality engendered by the initial RAEs resulted in an improvement in UK research, with a faster increase in citation rates than for other developed nations and a marked improvement in the proportion of most highly cited papers attributable to the nation (Adams *et al*, 2007). However, it would be extremely difficult, if not impossible, for the UK to sustain such marked improvement once a high status has been reached: subsequent improvement can only be incremental and hence the value of continuing a resource-intensive process needs to be questioned.

Quite valid concerns were raised by the sector when it became clear that the five 'possible' models proposed in the DfES consultation document relied on a single indicator — external research income (DfES, 2006). Not surprisingly, there was apprehension about the impact this would have on the funding schemes whose data would be used in the formula (Sastry and Bekhradnia, 2006). If significant funding and prestige became solely reliant on successes in competitive grants schemes, the funding agencies would find themselves inundated with applications and a disproportionate percentage of effort would be allocated to the type of research that would appeal to these agencies.

The Arts and Humanities Research Council (AHRC) immediately raised concerns about the inappropriateness of external grant income as a proxy for assessing research quality in their disciplines, and indeed this was acknowledged in the consultation document (DfES, 2006; HEFCE and AHRC, 2006). The AHRC subsequently joined with the Higher Education Funding Council for England (HEFCE) to announce the establishment of an expert working group to advise on the potential use of metrics in future RAEs for their disciplines.

Other research councils have voiced similar concerns, and future UK RAEs may yet move closer to the proposed Australian RQF model. The DfES consultation document does list other possible metrics in the appendix, though none were incorporated into their five "possible models". HEFCE is now engaging in extensive consultation with experts on the inclusion of bibliometric indicators in future exercises. During its consultations with the sector, the Roberts Review found that the overwhelming view of academics and research managers was that research should be assessed using a system based on peer review by subject-based panels (Wooding and Grant,

2003). Yet they also supported the use of metrics to inform panel decisions — an ideological shift towards a balanced approach to research assessment.

Fundamental issues in the assessment of research[2]

There are a number of key issues that are fundamental to the use of quantitative measures for the assessment of research performance. An awareness of these is essential to fully appreciate the developments currently occurring in many national systems.

How to define quality

The intention of any application of performance indicators is to either identify 'high quality' or to find out which research is 'better'. It is therefore not surprising that any proposed application of quantitative indicators triggers a debate about our understanding of what is meant by research quality, and how the various indicators are related to it.

The character of research quality is complex and multidimensional. Trying to assess it is difficult enough at the level of the individual (for promotions, appointments or awards) or project (research funding assessment). The problems encountered when trying to identify it within a whole higher-education sector are significantly increased. To clarify what is being measured by research assessments generally, and performance indicators in particular, a number of attempts have been made to define what is meant by research quality'.

Cole and Cole suggested that quality might be defined in two ways. First: "[a] traditional historian of science might apply a set of absolute criteria in assessing the quality of a paper. Those papers which embody scientific truth and enable us to better understand empirical phenomena are high-quality papers" (Cole and Cole, 1973: 23). Measuring the quality of work in historical retrospect has the advantage that it does not exclude work that is momentarily fashionable or temporarily ignored.

Secondly, Cole and Cole suggest using a social, as opposed to an absolute, definition of quality. They say that, since absolute truth does not exist, high-quality work can be defined: "as that which is currently thought useful by one's colleagues. If scientists in their everyday behaviour find a particular idea useful in their work, that idea is a valuable one, and we shall call it a high-quality idea" (Cole and Cole, 1973: 24).

The relative and social character of research quality is also discussed by other authors (see, for example, Martin and Irvine, 1983; Herbertz and Müller-Hill, 1995). Martin and Irvine define quality as "a property of the publication and the research described in it. It describes how well the research is done, whether it is free from obvious 'error', how aesthetically pleasing the mathematical formulations are, and so on". They claim that "quality is relative, it is socially and cognitively determined; it is not just intrinsic to research but judged by others with differing research interests and social goals".

Similarly, Moed *et al* (1985) refer to the multidimensional character of research quality and differentiate among a cognitive (importance of the specific content of scientific ideas), methodological (accuracy of methods and techniques) and aesthetic dimension of quality. They contend that these three dimensions of quality cannot be measured by quantitative indicators, but can only be judged by peers.

Van Raan (1996) suggests the following definition:

> Quality is a measure of the extent to which a group or an individual scientist contributes to the progress of our knowledge. In other words, the capacity to solve problems, to provide new insights into 'reality', or to make new technology possible. Ultimately, it is always the scientific community ('the peers', but now as a much broader group of colleague-scientists than only the peers in a review committee) who will have to decide in an inter-subjective way about quality.

While the introductory sentence of this quote appears to indicate that research quality is a quantitatively measurable property, van Raan makes clear immediately afterwards that he, too, regards quality as something that in its complexity must incorporate the opinions of peers. Quantitative indicators may be related to quality and measure certain aspects of it, but cannot stand alone in any assessment.

Performance indicators are used either to identify 'high quality' or to find out which research is 'better', so any proposed application of quantitative indicators triggers a debate about what is meant by research quality, and how the various indicators are related to it

Focusing on impact

A way to circumvent the problem of defining quality is to treat it as one characteristic of research among others. Martin and Irvine (1983) introduce a differentiation among quality, importance and impact in relation to publications. Making a distinction from their definition of quality (see above), they contend the importance of a publication refers to its potential influence on surrounding research activities if communication channels in science were flawless.

Finally, they define the impact of a publication as "the actual influence on surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language and availability of the publishing journal".

Because of the difficulty in defining a concept of quality that is appropriate when analysts seek to apply quantitative measures, discussions have concentrated on clarifying what impact is. On this there is much more agreement. In addition to Martin and Irvine's definition, other authors have defined it along similar lines: "the actual importance of a paper judged by the scientific community" (Dieks and Chang, 1976), "the reception of research work by other scientists" (Weingart *et al*, 1988), or "the effect which a published research finding has on its audience" (Phillimore, 1989: 263).

Moed *et al* (1985) distinguish between "short-term impact" and "long-term impact". The former is the impact of research at the cutting-edge of its discipline within a few years of completion, and is the focus of assessment exercises such as the RQF and RAE. The long-term impact indicates whether, and to what degree, the research has made a more permanent contribution to scientific advance.

Research quality as a whole cannot be assessed without input from peers. However, impact is now regarded as one aspect of research quality that can be measured by quantitative indicators (see for example Nederhof and van Raan, 1987; van Raan, 1996).

This differentiation between quality and impact when discussing the use of quantitative indicators raises problems in current UK and Australian developments. There is an emerging trend in science policy to regard impact, the measurable part of quality, as a proxy measure for quality *in total*, and disregard the need for peer input to assess those aspects not well covered by quantitative measures. Additionally, in Australia, the Government seeks to distinguish between the *quality* of research ("its intrinsic merit and academic impact") and its *impact* ("the use of original research outside the peer community"). In relation to the RQF, proponents of quantitative measures are being forced to use 'metrics' in the same breath as 'quality', something that many, particularly experienced bibliometricians, are uncomfortable with.

Role of quantitative indicators

Much of the debate surrounding the use of quantitative indicators is triggered by concerns about the way in which they will be deployed. Those that vehemently oppose their use are worried that the proponents of the measures are attempting to replace peer review. Several years before DfES proposed a metrics-based approach, Smith and Eysenck (2002) had suggested that, for reasons of cost, the peer-review component in the British RAE could be largely replaced by citation analyses given the very high correlations with past rankings. However, Warner's (2000) study of the correlation between RAE scores and bibliometric measures has shown that, while the correlation is high, there are deviant cases. He cautions that 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 sometimes led to the tendency to point to quantitative indicators as objective measures in contrast to the subjective character of the peer review. However, it should be remembered that indicators themselves are based in part on peer decisions, for instance, 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 see indicators being used, not to replace the peer evaluation, but rather to make the results of research assessment debatable and to offer experts additional information (van Raan and van Leeuwen, 2002). Peer review can become more 'transparent' by using bibliometric indicators, which can also counterbalance peer review's shortcomings (see van Raan, 1996: 401; Tijssen, 2003; Aksnes and Taxt, 2004). For example, the indicators can control for the rapid decline of research fields, and protect against the operation of 'old boys' networks' in peer reviews (Weingart, 2003).

Quantitative indicators are also seen by scientists as a useful resource in cases of doubt within panel discussions (Moed and van Raan, 1988). In addition, they can be used to highlight gaps in the knowledge of peers – as "triggers to the recognition of anomalies" (Bourke *et al*, 1999). Where the indicators do not align with peer evaluation, the reasons must be sought. It may be because of 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 entities being evaluated by those conducting the assessment.

Inconsistencies between quantitative data and peer review are likely to trigger additional, deeper analyses of the performance of entities being evaluated by those conducting the assessment

Even the role of quantitative indicators as an additional source of information needs to be more closely defined. What weight are the assessors to place on the information? How should conflicting evidence be handled? The haphazard 'feeding in' of bibliometric indicators to a peer review can distort the evaluation when the existence of anomalies is not recognised and investigated. For Gläser (2004), the main danger of this practice is that "amateur bibliometricians" do not understand all the methodological issues addressed in the construction of such measures, and are likely to trust them blindly: such practices may lead to a situation where "bibliometrics involuntarily takes over because peers do not judge content anymore" (Gläser and Laudel, 2005). Care needs to be taken to ensure those incorporating the data into their deliberations understand the methodological limitations and have all the necessary information to interpret the data accurately and sensitively.

Selection of indicators

The challenge facing policy-makers is to identify robust indicators, particularly for those disciplines not well served by standard measures. Since each indicator has different strengths and weaknesses, researchers suggest that evaluations should always incorporate more than one indicator (Martin and Irvine, 1983). This is also the proposed standard practice for Organisation for Economic Co-operation and Development (OECD) surveys of R&D activities (Godin, 2002: 7). The Centre for Science and Technology Studies (CWTS) puts this into practice by always using a set of "crown indicators" in their evaluative studies (van Leeuwen *et al*, 2003).

The treatment of this issue to date demonstrates one of the major differences between the UK and Australian approaches. Whereas in Australia the use of a "basket of indicators" is quite explicitly mandated, the stance of DfES is ambiguous (DEST, 2006; DfES, 2006). Numerous indicators are listed, but whether the aim is to construct a suitable suite of indicators, or to provide a list from which each discipline is to choose, remains unclear.

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, department, and so on: for example, a study was conducted by Hattie *et al* (1991), where scientists rated a long list of indicators divided into six groups; similar questionnaires were used in a study by Grigg and Sheehan (1989) and by a research group chaired by Russell Linke (NBEET, 1993). While the lists were comprehensive, none of the studies came up with a preferred set.

Martin and 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, the analysis.

While many indicators have a common starting point — a particular data source — their final form may bear few similarities. There is considerable room for "manipulation by selection, weighting and aggregating indicators" (Grupp and Mogee, 2004: 86–87). These concerns have been specifically raised in relation to bibliometric indicators; a special session of the major international conference in the discipline was devoted to the issue.[3]

A final selection of a suite of indicators for a metrics-based RAE remains some way off, though developments already in place with the RQF can provide a concrete example. The preferred model has settled on a basket of three indicators. Two of these —ranking outputs and competitive grant income — will be applicable across all disciplines.

The benchmarks against which quality will be assessed will vary. For example, grant income will be much higher in the experimental sciences than in the humanities. The type of output to be ranked will also vary — conferences for computer science, journals for the experimental sciences, publishers for the humanities, venues for the performing arts, to name just some possibilities. Equivalent measures can be developed for both these indicators. It is only the third indicator — measures based on standard citation analysis — that is not applicable across all disciplines. Even in this case, research is underway to identify novel citations measures that may have a wider application (CHASS, 2006).

Validation

Many of the concerns raised about the use of quantitative measures in the evaluation of science centre on the question of whether their use for such purposes is valid. There is a large body of literature on the topic, though it is limited in that most references focus on bibliometrics, rather than quantitative indicators more generally.

Most have been conducted comparing the results of quantitative analyses against peer review, but validation attempts have proved difficult: "The fundamental problem in any discussion of the validity of indicators of scientific productivity is the fact that there is no absolute standard measure of such productivity" (Narin, 1976: 82).

Some of the problems of using peer judgements for validation purposes were also discussed in detail by Martin and Irvine (1983). Nederhof (1988: 209) claims that, since the reliability of peer judgements is weak, the correlations between the peer-review results and bibliometric indicators can only be moderately strong. In addition, peer-review judgements and bibliometric evaluations are not unrelated: many peers use bibliometric measures in their deliberations,

such as the number of publications in top journals (van Raan, 2004: 38). Nevertheless, peer review remains the main criterion against which the validity of quantitative measures is assessed.

A general theme of the literature is that discrimination in relation to research performance in the middle range can prove difficult. Nederhof and van Raan (1989) sought to determine if this could be done using bibliometric techniques. Their analysis of *cum laude* and non-*cum laude* doctorates showed that those with *cum laude* publish more and were more highly cited than those without before graduation, but this difference rapidly vanished. Lindsey (1989) also found that the differences between articles that attract no citations and those that attract two or three citations were not substantial and concludes: "Thus, in the heavily populated middle range of the continuum of quality, citation counts are of doubtful utility".

In non-scientific disciplines, validation continues to rely on peer judgements, but the quantitative data used stretches beyond assessing journal-based publications. In the humanities, Lewison (2001) uses surveyed peer opinion alongside bibliometric data to define what counts as a 'quality' book in medical history, and obtains a high degree of qualitative agreement between citation outcomes and separately generated peer evaluation. For the humanities and social sciences Luwell *et al* (1999: 22) rely on peer judgements to operationalise a concept of academic quality covering a myriad range of standard and non-standard publications, and they assess the connection between the two methods of assessment. Moed *et al* (2002) go on to champion a reflexive approach to research evaluation that routinely feeds back peer judgments about bibliometric findings. Meho and Sonnenwald (2000) demonstrate a close relationship between journal-based citation analysis and peer opinion in the form of book reviews and solicited rankings of individuals in the social sciences.

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

Discipline differences

Research fields differ in their publication practices. The most obvious examples are the differences that exist between the natural and physical sciences on the one hand, and the humanities, arts and social sciences on the other. There is general agreement that the research methods and orientations in the social sciences and humanities are distinct from those of the experimental sciences, that their communication practices or literatures are thereby differently structured and that this has consequences for quantitative measures (see, for example, Hicks, 2004: 473; Glänzel and Schoepflin, 1999; Luwel *et al*, 1999: 13; Nederhof and van Raan, 1989). While quantitative indicators are relatively straightforward for the evaluation of scientific research, and sophisticated methods are available for dealing with interdisciplinarity, we find that "[w]hen challenged to evaluate scholarly work in the social sciences and humanities, we are rudely forced to work outside this comfort zone" (Hicks, 2004: 474).

Even within seemingly coherent fields, subdiscipline variations in publication and citation practices can occur. For example, in high-energy physics, theorists tend to publish more frequently than experimentalists (Irvine and Martin, 1985). Small fields encompassing a limited number of researchers have lower citation rates than those with many researchers (van Raan, 1996: 403). Different types of publication are relevant in different research fields. While in some disciplines journal articles are the most important channel of communication, in others books, book chapters and conference papers play a crucial role (for example, see Small and Crane, 1979; Butler and Visser, 2006).

Given these differences, both the RQF and RAE follow 'best practice' as both make explicit allowance for field-specific characteristics by establishing discipline-based panels and allowing for variations in assessment methodologies within an overall framework.

Influence on behaviour

> ... I urge the funding councils to remember that all evaluation mechanisms distort the processes they purport to evaluate. (Roberts, 2003)

Any system used to assess research that affects money and/or prestige is likely to affect the behaviour of researchers and administrators. This applies equally to qualitative and quantitative assessments. Two major classes of unintended consequences, which can occur as the result of any evaluation, have been identified.

One almost inevitable consequence of repeatedly applying the same type of measure is goal displacement: high scores in the measures become the goal rather than a means of measuring whether an objective (or performance level) has been attained (Perrin, 1998). The re-assessment of the RAE by the House of Commons Science and Technology Committee (2004) devoted considerable time and space to the subject of 'game-playing' by universities in response to the assessment criteria. A number of submissions to their enquiry go to the nub of this issue. One senior academic commented: "... the improvement in results represented a 'morass of fiddling, finagling

and horse trading' (House of Commons Science and Technology Committee, 2004: 21).

A second effect may be that the research process itself is modified: researchers adapt their behaviour in response to the method of evaluation. This occurs in ways that are more complex and more difficult to observe than goal displacement. The type of response that can be observed most easily is a change in publication behaviour. Scientists reported changing their publication strategy by placing more of their work in international journals making it difficult for nationally oriented journals to attract sufficient manuscripts (de Bruin *et al*, 1993: 40). Although a stronger international orientation in publishing seems desirable, it could lead to the neglect of nationally important topics. Mojon-Azzi *et al* (2003) investigated publications in opthalmological journals and found some work published twice: they speculated that it may be because of a desire to increase the author's performance in assessment measures based on simple publication counts.

Other behavioural changes have been hypothesised, such as risk avoidance and clinging to the mainstream, which can lead to a reduced diversity of approaches in research. Marginson and Considine claim that Australia's previous formula-based funding favours research quantity rather than research quality, short-term rather than long-term research, and mitigates against new researchers and emergent approaches (Marginson and Considine, 2000: 141–171). The authors' conclusions were largely reached on the basis of anecdotal evidence rather than resulting from systematic analysis.

However, a bibliometric study of Australia's scientific output (Butler, 2003) provided some direct evidence, at least for their first response. It showed a significant increase in the country's journal output in the mid-1990s. A number of the characteristics of the data, in particular the timing of this productivity increase in relation to the introduction of funding formulas, suggests a causal relationship.

Conclusions

The character of research quality is complex and multidimensional. Both the UK's RAE and Australia's RQF seek to assess it across the breadth of a national higher-education system in all its disciplines in order to inform funding. A robust process is imperative. No single quantitative measure, or even a 'basket' of indicators, can address all its facets. Nor can a small panel of peers be expected to combine sufficient knowledge of the performance of all a nation's institutions and all a nation's researchers active in their discipline to enable them to arrive at error-free judgements. The most sensible approach is to combine the two methods, by assembling a group of highly qualified experts in the discipline and arm them with reliable, discipline-specific data to assist their deliberations.

In a balanced approach to research assessment, the data is viewed as triggers to the recognition of anomalies. As has been demonstrated by many studies, the two methods will usually produce similar results. Reaching the same conclusion from two perspectives will increase confidence in the assessments. Then the bulk of the time panel members have available to them can be productively used to determine the reasons for discrepancies in those cases in which the two methods result in different outcomes (whether this is a result of problems with the data or gaps in the knowledge of panel members).

The model proposed for Australia's RQF encompasses a balanced approach, with a range of quantitative measures, sympathetic to discipline differences, informing the more traditional peer-review process. The UK's current RAE makes no explicit routine use of metrics, though a number of panels are known to use them informally. New proposals for a metrics-based assessment process are a move to the opposite extreme, though the lobbying and consultative processes currently in train may yet move the final model to a more balanced approach.

While disciplines have their distinctive characteristics, it is not practical or desirable to develop a discrete set of quantitative measures for every distinct discipline or group of similar disciplines. There must be consistency in assessment among disciplines, while allowing for sensible adjustments to generic indicators. The Australian model shows the possibilities that exist. One measure, classifying research into bands based on the prestige of their outlet, can be modified to encompass the most important outlets in each discipline, whether it be journal articles, books, conferences or performance venues. Differences for standard indicators, such as external grant income, can be catered for by establishing discipline-specific benchmarks (at either national and/or international levels) against which performance can be judged. By incorporating a set of common or equivalent indicators, the addition of a limited number of discipline-specific measures will not compromise the consistency of assessment.

Any research assessment process, particularly one with significant funding consequences, will affect the way people behave. The balanced approach, using both peer review and a suite of indicators, lessens the likelihood that those responses will be too perverse. Researchers and administrators alike will be confronted by an array of signals to respond to, which means a range of possible responses. Stakeholders appear to be much more aware of the possibilities of game-playing, featuring strongly in all consultations surrounding the RQF and debate on changes to the RAE. Yet there is often no simple assessment of likely responses. Much is said in the UK about staff poaching in the lead-up to each RAE, with the implicit assumption in the rhetoric that it is undesirable. However, if one of the aims of any assessment exercise is to concentrate high-quality research in any given field in a limited number of

institutions, then this is precisely the response to be sought.

Most of the worst perversions of any assessment process can be avoided, or at least minimised, if a balanced approach is employed, using both peer review and quantitative approaches. It is the most robust methodology, and also the most cost-effective. I return to my opening plea for sanity to prevail and for all stakeholders in the debate — researchers, university administrators and bureaucrats alike — to pause and assess the vast wealth of experience that exists from research studies or evaluation exercises around the globe, and to take a more balanced approach to research assessment.

Notes

1. RAEs were conducted in 1986, 1989, 1992, 1996 and 2001.
2. Some of the discussion in this section is drawn from a REPP working paper co-authored with Claire Donovan and Grit Laudel, which encompassed an extensive literature review undertaken at the commencement of a research project in 2003 (REPP, 2005).
3. Fifth International Conference of the International Society for Scientometrics and Informetrics, River Forest, Illinois in 1995. Selected papers from the session were published in *Scientometrics*, **35** (1996).

References

Adams, Jonathan, Karen Gurney and Stuart Marshall 2007. Profiling citation impact: a new methodology. *Scientometrics*, **72**, 325–344.

Aksnes, Dag and Randi Taxt 2004. Peer review and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation*, **13**(1), April, 33–41.

Anderson, Don, Richard Johnson and Bruce Milligan 1996. *Performance-based Funding of Universities: National Board of Employment Education and Training, Commissioned Report No. 51*. Canberra: NBEET.

AVCC, Australian Vice-Chancellors' Committee 1999. *Discussion Paper on Higher Education Research and Research Training. Canberra*: AVCC. Available at <http://www.avcc.edu.au/archive/news/speeches/1999/newknowledge.htm>, last accessed 14 September 2007.

Bourke, Paul, Linda Butler and Beverley Biglia 1999. *A Bibliometric Analysis of Biological Sciences Research in Australia*. Department of Education, Training and Youth Affairs Report no 6307HERC99A. Canberra: DETYA.

Butler, Linda 2003. Explaining Australia's increased share of ISI publications: the effects of a funding formula based on publication counts. *Research Policy*, **32**, 143–155.

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

CHASS, Council for the Humanities, Arts and Social Sciences 2006. *History and Political Science: the Case for Bibliometrics*. Available at <http://www.chass.org.au/papers/bibliometrics/CHASS_Report.pdf>, last accessed 14 September 2007.

Cole, Jonathan and Stephen Cole 1973. *Social Stratification in Science*. Chicago: The University of Chicago Press.

de Bruin, R, A Kint, M Luwel and H Moed 1993. A study of research evaluation and planning: the University of Ghent. *Research Evaluation*, **3**(1), April, 25–42.

DEST, Department of Education, Science and Training 1998. *Characteristics and Performance Indicators of Higher Education Institutions*. Canberra: DEST. Available at <http://www.dest.gov.au/archive/highered/otherpub/characteristics.pdf>, last accessed 14 September 2007.

DEST, Department of Education, Science and Training 2004. *Evaluation of Knowledge and Innovation Reforms Consultation Report*. Canberra: DEST Available at <http://www.dest.gov.au/NR/rdonlyres/654E1226-6F91-44C5-BDEA-FE8FCB228E88/2788/pub.pdf>, last accessed 14 September 2007.

DEST, Department of Education, Science and Training 2005. *Research Quality Framework: Assessing the Quality and Impact of Research in Australia. The Preferred Model*. Canberra: DEST. Available at <http://www.dest.gov.au/NR/rdonlyres/AF74E4A9-C7DD-48A4-8D94-847FF35C6B97/7845/RQFPreferredModelPaper.pdf>, last accessed 14 September 2007.

DEST, Department of Education, Science and Training 2006. *Research Quality Framework: Assessing the Quality and Impact of Research in Australia. The recommended RQF*. Canberra: DEST.

DETYA, Department of Employment, Training and Youth Affairs 1999a. *Knowledge and Innovation*. Canberra: DETYA. Available at <http://www.dest.gov.au/sectors/higher_education/publications_resources/profiles/archives/knowledge_and_innovation_policy_statement.htm>, last accessed 14 September 2007.

DETYA, Department of Employment, Training and Youth Affairs 1999b. *New Knowledge, New Opportunities*. Canberra: DETYA. Available at <http://www.dest.gov.au/sectors/higher_education/publications_resources/profiles/archives/new_knowledge_new_opportunities.htm>, last accessed 14 September 2007.

DfES, Department for Education and Skills 2006. *Reform of Higher Education Research Assessment and Funding*. London: HMSO. Available at <http://www.dfes.gov.uk/consultations/downloadableDocs/consultationDocument%20jcutshall2.doc>, last accessed 14 September 2007.

Dieks, D and H Chang 1976. Differences in impact of scientific publications: some indices derived from a citation analysis. *Social Studies of Science*, **6**, 247–267.

Geuna, Aldo and Ben Martin 2003. University research evaluation and funding: an international comparison. *Minerva*, **41**, 277–304.

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

Gläser, Jochen 2004. Why are the most influential books in Australian sociology not necessarily the most highly cited ones? *Journal of Sociology*, **40**(3), 261–282.

Gläser, Jochen and Grit Laudel 2005. Advantages and dangers of 'remote' peer evaluation. *Research Evaluation*, **14**(3), December, 186–198.

Godin, Benoit 2002. Outline for a history of science measurement. *Science Technology & Human Values*, **27**, 3–27.

Grigg, Lyn and Peter Sheehan 1989. *Evaluating Research: the Role of Performance Indicators*. Brisbane: Office of the Academic Director of Research, The University of Queensland.

Grupp, Hariolf and Mary Ellen Mogee 2004. Indicators for national science and technology policy. In *Handbook of Quantitative Science and Technology Research*, eds. Henk Moed, Wolfgang Glanzel and Ullrich Schmoch, pp. 75–94. Dordrecht: Kluwer Academic Publishers.

Hattie, J, J Tognolini, K Adams and P Curtis 1991. *An Evaluation of a Model for Allocating Funds Across Departments Within a University Using Selected Indicators of Performance*. Canberra: Department of Education, Employment and Training.

HEFCE and AHRC, Higher Education Funding Council for England and the Arts and Humanities Research Council 2006. *HEFCE and AHRC Announce Expert Group on Research Metrics*. Bristol: HEFCE. Available at <http://www.hefce.ac.uk/news/hefce/2006/metrics.htm>, last accessed 14 September 2007.

Herbertz, Heinrich and Benno Müller-Hill 1995. Quality and efficiency of basic research in molecular biology: a bibliometric analysis of thirteen excellent research institutes. *Research Policy*, **24**, 959–979.

Hicks, Diana 2004. The four literatures of social science. in *Handbook of Quantitative Science and Technology Research*, eds. Henk Moed, Wolfgang Glanzel and Ullrich Schmoch, pp. 473–496. Dordrecht: Kluwer Academic Publishers.

House of Commons Science and Technology Committee 2004. *Research Assessment Exercise: a Re-assessment*. London: The Stationery Office Limited.

Irvine, John and Ben Martin 1985. Evaluating big science: CERN's past performance and future prospects. *Scientometrics*, **7**, 281–308.

Lewison, Grant 2001. Evaluation of books as research outputs in history. *Research Evaluation*, **10**(2), 89–95.

Lindsey, Duncan 1989. Using citation counts as a measure of quality in science: measuring what's measurable rather than what's valid. *Scientometrics*, **15**, 189–203.

Luwel, Mark, Henk Moed, Anton Nederhof, V de Samblanx, K Verbrugghen and L van der Wurff 1999. *Towards Indicators of Research Performance in the Social Sciences and Humanities*. Leiden: CWTS.

Marginson, Simon and Mark Considine 2000. *The Enterprise University: Power, Governance, and Reinvention*. Cambridge: Cambridge University Press.

Martin, Ben and John Irvine 1983. Assessing basic research: some partial indicators of scientific progress in radio astronomy. *Research Policy*, **12**, 61–90.

Meho, L and D Sonnewald 2000. Citation ranking versus peer evaluation of senior faculty research performance: a case study of Kurdish scholarship. *Journal of the American Society for Information Science*, **51**(2), 123–138.

Moed, Henk and Antony van Raan 1988. Indicators of research performance: applications in university research policy. In *Handbook of Quantitative Studies of Science and Technology*, ed. Anthony van Raan, pp. 177–192. North-Holland: Elsevier.

Moed, Henk, W Burger, J Frankfort and Anthony van Raan 1985. The application of bibliometric indicators: important field- and time-dependent factors to be considered. *Scientometrics*, **8**, 177–203.

Moed, Henk, Mark Luwel and Anton Nederhof 2002. Towards performance indicators in the humanities. *Library Trends*, **50**(3), 498–520.

Mojon-Azzi, Stefania, Xiaoyi Jiang, Ulrich Wagner and Daniel Mojon 2003. Journals: redundant publications are bad news — publishing the same work twice is unethical and casts doubt on the integrity of research. *Nature*, **421**, 209.

Narin, Francis 1976. *Evaluative Bibliometrics: the Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Cherry Hill NJ: Computer Horizons Inc.

NBEET, National Board of Employment, Education and Training 1993. *Research Performance Indicators Survey*. National Board of Employment, Education and Training, Commissioned Report no 21. Canberra: NBEET.

Nederhof, Anton 1988. The validity and reliability of evaluation of scholarly performance. In *Handbook of Quantitative Studies of Science and Technology*, ed. Antony van Raan, pp. 193–228. North-Holland: Elsevier.

Nederhof, Anton and Antony van Raan 1989. A validation study of bibliometric indicators: the comparative performance of cum laude doctorates in chemistry. *Scientometrics*, **17**, 427–435.

Nederhof, Anton and Anthony van Raan 1987. Peer review and bibliometric indicators of scientific performance: a comparison of cum laude doctorates with ordinary doctorates in physics. *Scientometrics*, **11**, 333–350.

Perrin, Burt 1998. Effective use and misuse of performance measurement. *American Journal of Evaluation*, **19**, 367–379.

Phillimore, A J 1989. University research performance indicators in practice: the University Grants Committee's evaluation of British universities, 1985-86. *Research Policy*, **18**, 255–271.

REPP, Research Evaluation and Policy Project 2005. *Quantitative Indicators for Research Assessment: a Literature Review*. REPP Working 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 14 September 2007.

Roberts, Gareth 2003. *Review of Research Assessment*. Available at <http://www.ra-review.ac.uk/reports/roberts.asp>, last accessed 14 September 2007.

Sastry, Tom and Bahram Bekhradnia 2006. *Using Metrics to Allocate Research Funds: A Short Evaluation of Alternatives to the Research Assessment Exercise*. UK: Higher Education Policy Institute.

Small, Henry and D Crane 1979. Specialties and disciplines in science and social science: an examination of their structure using citation indices. *Scientometrics*, **1**(5–6), 445–461.

Smith, Andy and Mike Eysenck 2002. The correlation between RAE ratings and citation counts in psychology. Technical Report, Psychology. London: University of London.

Tijssen, Robert 2003. Scoreboards of research excellence. *Research Evaluation*, **12**(2), August, 91–103.

van Leeuwen, Thed, Martijn Visser, Henk Moed , Anton Nederhof and Antony van Raan 2003. The holy grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, **57**, 257–280.

van Raan, Antony 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, **36**, 397–420.

van Raan, Antony 2004. Measuring science. In *Handbook of Quantitative Science and Technology Research*, eds. Henk Moed, Wolfgang Glanzel and Ullrich Schmoch, pp. 19–50. Dordrecht: Kluwer Academic Publishers.

van Raan, Antony and Thed van Leeuwen 2002. Assessment of the scientific basis of interdisciplinary, applied research – application of bibliometric methods in nutrition and food research. *Research Policy*, **31**, 611–632.

Warner, Julian 2000. A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science*, **26**, 453–460.

Weingart, Peter 2003. Evaluation of research performance: the danger of numbers. Proceedings of the 2nd Conference of the Forschungszentrum Jülich GmbH Zentralbibliothe, Bibliometric Analysis in Science and Research: Applications, Benefits and Limitations, Jülich, 5–7 November 2003.

Weingart, P, R Sehringer and M Winterhager 1988. Bibliometric indicators for assessing strength and weaknesses of West German science. In *Handbook of Quantitative Studies of Science and Technology*, ed. Anthony van Raan, pp. 391–430. North-Holland: Elsevier.

Wooding, Steven and Jonathan Grant 2003. *Assessing Research: The Researchers' View*. RAND Europe Report for HEFCE. Bristol: HEFCE. Available at <http://www.ra-review.ac.uk/reports/assess/AssessResearchReport.pdf>, last accessed 14 September 2007.