2009

Evolving Regimes of Multi-university Research Evaluation

Diana M Hicks, *Georgia Institute of Technology - Main Campus*

Evolving regimes of multi-university research evaluation

Diana Hicks
School of Public Policy
Georgia Institute of Technology
Atlanta, GA 30332
Phone: 404 385 6015
Fax: 404 385 0504
dhicks@gatech.edu
March 10, 2008

Keywords
Composite Index
ERA
NRC ranking
RAE
Ranking
Research
RQF
University

Abstract

Since 1980, national university departmental ranking exercises have developed in several countries. This paper reviews exercises in the U.S., U.K. and Australia to assess the state-of-the-art and to identify common themes and trends. The findings are that the exercises are becoming more elaborate, even unwieldy, and that there is some retreat from complexity. There seems to be a movement towards bibliometric measures. The exercises also seem to be effective in enhancing university focus on research strategy.

Introduction

Early 2008 was a trying time for university research administrators in the Anglo-Saxon world. Research administrators in several countries awaited the results of national scale, departmental level research rankings: UK administrators awaited the results of the current Research Assessment Exercise (RAE); US administrators awaited the release of the latest National Research Council (NRC) rankings; Australian administrators no longer knew what to expect after the cancellation of the Research Quality Framework (RQF) and introduction of the Excellence in Research for Australia initiative (ERA).

The research missions of universities have become the focus of increasing societal attention and performance-based research funding has been increasing around the world. These trends began earlier in the Anglo-Saxon countries, and so their evaluation systems are more mature. In 2007-08, each system underwent revision; thus it is timely to take stock of the lessons learned. Given the decades' long history of multi-university, departmental-level research evaluation in each country and the salience of the exercises for funding and success, there is a rich tradition of critique and discussion of the evaluation exercises. However, this literature is nation-specific and discipline-specific. In contrast this paper is internationally comparative, deriving lessons by finding common themes in the evolution of departmental research ranking in three countries. [i]

Common themes emerge. Each system underwent redesign, suggesting that the state-of-the-art has evolved. Evaluation redesign involved extensive consultation with the academic community which tended to encourage increased complexity. The complexity threatened to become unmanageable however, and simplifications imposed by government loomed. In addition, bibliometric metrics became more important and increased in sophistication. Finally, universities seemed to be very responsive to ranking systems, thus system performance can be increased using these methods.

The United States and non-governmental rankings

In the United States the research quality of university departments is publicly assessed by magazines and independent bodies. When contrasted with the government mandated exercises conducted in the U.K. and Australia, the U.S. efforts might be termed "freelance" ranking projects. Here I focus on the influential rankings of the National Research Council of the National Academies[ii] rather than the possibly better known rankings of *U.S. News and World Report* magazine or the emerging swarm of alternatives produced by think tanks or news media.

The elaborate ranking exercise conducted by the National Research Council (NRC) was undertaken in 1983, 1995 and 2007-08. With its ranking the NRC hoped to provide potential students and the public with accessible information on doctoral programs, and to help universities improve the quality of programs through benchmarking and so enhance the nation's overall research capacity. No funding was allocated as a result of these rankings, but due to the respect afforded the NRC, they promised to attribute prestige to individual departments and to influence the choices of prospective graduate students as well as the desirability of departments as a place to work.

The 1995 rankings were heavily analyzed, usually by scholars discussing their own disciplines. Broadly speaking, the studies seemed to examine the virtues of different ranking systems, or sought to offer advice to those trying to rise in the rankings. In the 1995 exercise departments were ranked based on a reputational survey. Bibliometric information, i.e. departmental level publication and citation counts, were reported in appendix tables but were not incorporated into the rankings.

Methodological question marks marred the NRC 1995 study. Miller et al. examined the political science rankings and noted that given response rate, resulting sample size and sampling error, it was statistically unsound to differentiate the rank ordering of many schools. The NRC report did present confidence intervals in an appendix. Nevertheless, mean ratings were reported "with two decimal places thereby implying more precision than the data warrant." (Miller et al., 1996, p. 716) The exercise did not generate the quality of data required to rank departments with any confidence. Miller et al. also noted that the NRC bibliometric data contained an obvious error: the University of Houston reportedly had no citations. Neither the NRC nor the data provider could explain this because resource limitations precluded checking the accuracy of the publication and citation counts. Even simple validations such as ensuring names were spelled correctly were not undertaken.

Given departmental concern to improve in the rankings, the main attributes underpinning results were examined closely. Departmental size was found to be a main driver (see for example Jackman and Siverson, 1996). The NRC acknowledged this and argued that size is an important determinant of quality since bigger programs are broader and have more resources and faculty (alternatively one could argue that size is important for the perception of quality). Jacman and Siverson found that faculty research productivity was not conditional upon faculty size (Jacman & Siverson, 1996), casting some doubt over the contention that size is a measure of research quality. Curiously though, nobody seems to have produced a ranking of departmental quality based solely on number of faculty. In other words, the size variable seems to need "laundering" through a reputational survey to become a legitimate basis for ranking, even for those who argue that equating size with quality is legitimate.

Scholars show far more interest in exploring research productivity as a basis for ranking than size, using bibliometrics to measure productivity. Dusansky and Vernon (1998) compared reputational and bibliometric departmental rankings in economics. They concluded that the reputational ranking and publication productivity ranking seemed to be based on somewhat different information. They also concluded that reputational rankings lag changes in publication productivity. "Established programs

appear to be able to maintain their reputations in the face of declines in their publication productivity, while more aggressive upstart programs must be patient in realizing the full returns from their substantial investments in professorial capital" (Dusansky & Vernon, 1998, p. 170).

The NRC's 2007-08 ranking method emerged from examining the shortcomings of its past rankings. The NRC convened a committee of eminent academics to study past NRC rankings and recommend improvements. The committee analyzed the criticisms of the 1995 report and concluded that the 1995 ranking because it was based on a reputational survey was seen as too "soft". They recommended that the 2007 ranking be based on quantitative variables. A small reputational survey was conducted, and a regression analysis used to identify a weighted mix of quantitative variables that best predicted reputational judgments. Departments would be ranked based on this weighted mix of variables. The perceived legitimacy of this approach presumably rests on the scale of the reputational survey.

The NRC thus required from all departments wishing to be ranked submission of information on the 48 variables to be included in the ranking formula. The 48 variables concern institutional characteristics (i.e. total research expenditure, characteristics of library, childcare and health insurance availability, university housing for PhD students etc.); doctoral program characteristics (i.e. size, time to degree, financial support, facilities for PhD students, test scores, support provided, employment destinations etc.), and program faculty (size, demographics, awards, bibliometrics etc.) (NRC, 2004, Table 4.1). For the bibliometrics, the NRC compiled full bibliographies of *SCI* indexed papers and their citations from Thomson-Scientific and used this information to calculate three bibliometric variables: 1) % of faculty publishing, 2) publications/faculty, 3) citations/faculty.

The 2007-08 NRC method was more elaborate than the 1995 version, and this has a cost. Planning for the exercise began in 2000; it was originally scheduled for 2003-2004 and slated to cost $5 million (direct cost only); it was actually conducted in 2005-2006 for release in 2007; the latest word is that results will be released in summer 2008. It remains to be seen whether the NRC can deliver the promised method and if it does, whether the level of accuracy will be acceptable to the community. The method and results are guaranteed to be subject to endless analysis by academics.

The U.K. Research Assessment Exercise

The Research Assessment Exercise (RAE) was a government-mandated evaluation of research quality in every department in every UK university. Its purpose was to inform the distribution of core research funding to the 160 universities, and it has been conducted by the British government five times since 1986. Approximately 25% of all research support in UK universities was allocated based on the RAE ratings of their departments. These allocations were quite stable. As a result of the 2001 RAE, only one institution saw its total revenues affected by more than 3.7 per cent and the median impact was less than 0.6 per cent (Sastry & Bekhradnia, 2006).

The RAE methodology evolved over the years and grew increasingly complex, but in 2008 it remained a peer review evaluation of departmental research output on a

seven point scale.[iii] The exercise began with relatively simple submission requirements in 1986. Departments described their research achievements in two pages, listed their five best publications and provided data on research income, prizes etc. This method was criticized, as for example favoring large departments with their larger pool of papers from which to choose the top five. In response, the method evolved so that submissions included greater detail on research environment and strategy, and listed four publications per individual and other data. The original four point scale was elaborated to seven – framed as five plus two. In 2008, 68 panels of reviewers were convened to consider departmental submissions and to assign grades.

Although often the subject of comment, the RAE's effect on research performance has not been definitively established largely due to methodological flaws in existing studies. Unfortunately all quantitative analyses were based on UK papers, rather than UK university papers, meaning that trends in publishing by hospitals, firms and non-profits could influence conclusions. The flimsy evidentiary base tended to suggest that the RAE may have increased the research performance of universities: in the 1990s and up to 2005, the number of papers per UK researcher increased (Moed, 2007); the UK's share of world citations rose (Lipsett, 2005); and the share of UK papers that remained uncited decreased (DTI, 2007, 3.05).

Qualitative evidence indicated the mechanisms at work. Researchers and administrators agreed that one effect of the RAE was to create much more focus on how and where to publish (HEFCE, 1997, 132). Researchers and administrators disagreed about whether the quality of research had improved, though a limited sample of journal editors thought that they were seeing better submissions (HEFCE, 1997, 123-124). McNay interviewed administrators and faculty individually and in focus groups finding that at the management level, the RAE had prompted institutions to conduct strategic reviews for the first time (HEFCE, 1997, 50). The RAE generated "awareness of the link between individual performance and the funding of the institution" (HEFCE, 1997, 82). At the faculty level some "said that competition was nevertheless a good and motivating factor. . . .They were more strategic in thinking about their careers, and appraisal was assisting in this" (HEFCE, 1997, 99). UK Vice-Chancellors believed that the benefits arising from the RAE included: the provision of an evidence base for Government to increase research funding, important feedback for university managers and an improved international recognition of the strength of UK research (DEST, 2006, p.1).

Several studies based on questionnaires appeared and reported that 28-64% of faculty, and 81% of department heads agreed that research quality had improved under the RAE (Gläser et al., 2002). But Gläser et al noted that the studies were not scrupulous about reporting sampling procedures, investigating bias due to non-response or constructing questions carefully to avoid passing on negative assumptions about the RAE to the respondents. Gläser et al also argued that the known fact that processes like the RAE reduce researcher autonomy creates in respondents a negative bias in answering questions regarding the effect of the RAE on research performance. This issue was not addressed in these studies. Nor did the studies investigate factors that may shape respondents' responses such as type of university, field, gender or seniority.

Although faculty may have been loath to admit it, UK university research output and the quality of the output seemed likely to have increased in response to the RAE. In

all likelihood, performance increased because the RAE put in place incentives that realized latent capacity in the university system. For example, administrators began conducting strategic reviews, and researchers began to work longer hours (SQW, 1996). Attention was focused on publishing in good journals. U.K. research became more meritocratic and competitive. Mobility increased because there was an institutional payoff to increases in research performance.

Eventually, however, latent capacity is exhausted and more resources must be added to keep increasing performance. Universities will require more money to pay the substantially increased salaries the most eminent scholars now command. Between 2002 and 2006 the number of academics earning more than £100,000 increased by 169%. This increase was fueled by increases in pay in medical and business schools and for administrators (Sanders, 2006). And resources will have to be added to facilitate, for example, reducing teaching loads for promising researchers (Wojtas, 2007)

Although criticism of the RAE for being ineffective seems misplaced, allegations of structural bias have gained more traction (Martin, 2007). The assessment panels were disciplinary and found it difficult to assess interdisciplinary research, which suffered as a result. The panels of academics did not pay equal attention to user-focused research - as requested by the government. Institutions represented on a panel tended to get the highest ranking. When one side of a dispute over appropriate directions for research in a field dominated a panel, it created the sense that there were "insiders" and "outsiders" to the exercise (HEFCE, 1997, 114 & 117).

The RAE was also burdensome. 70 panels of 10 or more members were convened to work on assessing 180,000 publications, making the exercise expensive. Panels were expected to read papers, though given the impossibility of comprehensive reading, panels varied in their implementation of this (Harman, 2000, p. 115). One author noted that the exercise were conducted as if it were supposed to appraise 50,000 individual researchers and their 180,000 pieces of work in order to make 160 funding decisions (Sastry & Bekhradnia, 2006), which seemed disproportionate. There were also indirect costs born by departments whose effort in preparing submissions increased with each round.

Gläser et al. concluded that as a peer review process, the RAE was subject to the same criticisms as peer review itself, namely that the process discourages unorthodox, new or risky research, encourages a short term focus and disadvantages interdisciplinary research (since peer review panels are constructed along disciplinary lines). Gläser and others expressed the related concern that the variety of topics selected and perspectives applied in research may decrease as a result of the RAE; "homogenization" was a term applied in this context. Gläser et al. concluded that the discussion in the literature can be read as suggesting that the RAE (and indeed other evaluation based methods of research funding) "improve quality to the upper middle level and drive out low quality research but suppress excellence to a certain extent" (Gläser et al., 2002, p. 22).

The question becomes, did the benefits outweigh the costs or vice versa? Geuna and Martin have argued that in the early years benefits probably outweighed costs because resources were shifted away from weaker to stronger performers which encouraged improved performance. However, after several rounds the gains from

initiating departmental research strategies were realized while the cost of the exercise continued to increase because ever more effort was devoted to submissions. Furthermore, over time people learned and responded to the incentives in the RAE and deleterious effects on research behavior appeared, such as avoiding risky research. Eventually, the costs probably outweighed the benefits (Geuna & Martin, 2003).

This argument seems plausible. Early on the requirements of the RAE were simpler, and so the exercise cost less, while the RAE introduced explicit incentives for research performance into a system for the first time no doubt realizing latent capacity. By 2008, the system likely ran at peak capacity while the exercise had become much more elaborate. Rather than dropping the incentives, UK government actions suggested more interest in reducing cost. Costs have been reduced below what might have been expected because the intervals between RAE's increased from 3 years to 4, 5 and then 7 years, reducing the frequency. Cost reduction through simplification has also been discussed by way of substituting a formula based on research grant funding and bibliometric indicators. This method would have the virtue of responding to government concerns that user oriented and interdisciplinary research is undervalued in the RAE process. The 2008 RAE will incorporate a "shadow metrics exercise" in a bid to shape any successor to the RAE (HERO, 2007).

The Australian Composite Index, Research Quality Framework (RQF), and Excellence in Research for Australia (ERA) initiative

The Australian government evaluated the research in its universities using the Composite Index beginning in 1995 using the results to inform the distribution of part of the research portion of general university funds. In 2004, 7% of all research support in Australian universities was allocated based on the Composite Index. [iv]

The Composite Index was a formula at the university level (not at the departmental level of the RAE or NRC rankings). The formula calculated each university's share of total research activity so in essence, it was a ranking of universities (not an assignment of grades like the RAE). The components of the formula were research funding – grants from government, other public sector and industry – and outputs: number of publications and graduate degrees completed (MS and PhD's). Universities submitted lists of publications, which were found to be of questionable accuracy. Audits conducted by KPMG of publication lists submitted by universities found a high error rate (34% in the second audit in 1997); 97% of errors affected final scores and so funding allocations (Harman, 2000, pp. 118-119).

In comparison to the RAE, the Composite Index was a simple thing. That very simplicity elicited a very clear response to its incentives which was analyzed by Butler (Butler, 2003). Over time, the publication portion of the formula became focused on papers indexed in the Web of Science databases such as the *Science Citation Index* (SCI). After a few iterations of funding distributed using the Composite Index, universities could put a dollar value on a paper placed in a journal indexed in the *Science Citation Index*. In the year 2000, such a paper was worth AUS$ 800 to a university, while a book from a recognized publisher was worth AUS$ 4,000. Butler found that Australian

university output increased 8% annually between 1992 and 1996, while the SCI grew at 2% per year. Seemingly the policy had achieved a notable success: greater research output without greater resources, or increased efficiency in the university research system.

Unfortunately, Butler also found that the impact of Australia's research fell over the same period. Between 1988 and 1993 Australia's citation impact dropped from 6th to 11th among OECD countries. Analysis revealed that Australian researchers were publishing more papers, but in journals with lower average citation impact (impact factor). This suggests that while the apparent volume of research produced increased, the apparent quality of Australian research suffered. Butler's analysis provided a very clear demonstration of a response to a policy's incentives that was ironically detrimental to the overall goals of the policy.

Butler's point was accepted by the Australian government under John Howard and a new system was devised – the Research Quality Framework or RQF. The rationale was that the Composite Index did not reward research excellence or encourage the wider community to increase its investment in research and so a broader assessment of quality and impact was required (Australian Government DEST, 2006. p. 9). The design of the RQF was notable for the extensive consultation behind it; the massively increased complexity of the exercise in comparison to the Composite Index, and the correspondingly increased sophistication in the metrics to be used.

The recommended RQF methodology was developed by an advisory group in 2006 who built on the work of a prior advisory group. The group developed a set of guiding principles for the RQF, solicited feedback from every university, talked to senior UK academics and consulted widely in Australia with groups representing business and education. The details of the methodology were fleshed out by four working groups covering: quality metrics, research impact, information technology and exploratory modeling.

Not unrelated to this wide consultative exercise was the increased complexity of the RQF. The RQF was RAE-like in that 13 subject area panels of 12 members, at least three foreign and three end users, would be convened to consider the submissions and metrics of each research group in the country. The submissions were to comprise staff lists, evidence of collaboration, awards won, students and their employment destinations, grant income, the four best outputs for each researcher, a full list of outputs, evidence, including indicators, of impact against generic and panel specific impact criteria; up to four case studies illustrating impact, and end user referees who might be contacted to verify impact claims. For each group, the metrics were to report: the distribution of output across (unweighted) discipline-specific tiers of output; the citations per publication; the proportion of work that falls into the top citation percentiles in its field. After considering this information the panels were to assign each group scores on two five point scales representing research quality and research impact. The exercise would be conducted every six years.

The increased complexity in comparison to the Composite Index was obvious. The RQF was also more complex than the RAE. Scores on two scales would be assigned, not one as in the RAE. And most importantly, the RQF moved assessment to

the level of the research group, where the Composite Index assessed at the university level, and the RAE at the departmental level. Given that research groups do not have the stability of officially recognized legal or administrative entities and can be quite fluid in their makeup, and that the assessment cycle was to be 6 years long, problems could be anticipated. Notably, the performance of groups may not be independently measurable even given directives such as: if a collaborative paper is submitted by researchers belonging to more than one research group, it must be "with the respective contributions duly reflected" (Gläser et al., 2004; DEST, 2007, p. 17). Whereas cost reduction is under discussion in the U.K., the RQF prompted requests for increased support (DEST, 2007, p. 1).

In December 2007, the government of Prime Minister Kevin Rudd replaced that of Prime Minister John Howard. On February 26, 2008, the Minister for Innovation, Industry, Science and Research announced a new research quality and evaluation system to be called ERA – Excellence in Research for Australia, to replace the "now defunct" RQF. The ERA announcement describes the new system as workable, streamlined and transparent. The system has yet to be finalized, so details are not available. The proposal is for a progressive (rather than simultaneous) examination of discipline clusters by institution to identify internationally competitive and emerging areas. Research quality will be assessed "using a combination of metrics and expert review by committees comprising experienced, internationally-recognized experts" (Carr, 2008).

The idea behind ERA seemed to be to add expert review and international comparison to the Composite Index's focus on departmental comparison using metrics only. Carried over from the RQF process was a sophisticated appreciation that appropriate metrics vary by discipline and will need to be tailored in consultation with disciplinary experts. However, ERA jettisoned the complexity of evaluation at the research group level and detailed submission requirements.

Discussion and Conclusions

Comparing the NRC rankings, RAE, Composite Index, RQF, and ERA identified common themes in the evolution of national university system research evaluation. Notable was a tension between increasing complexity and practicality. Complexity was reflected in methodological choices that in some cases seem rash. Neither the RQF goal of assessing at the research group level, nor the NRC goal of compiling a full (and presumably correct) bibliography and citation count for every U.S. academic had been accomplished on anywhere near the proposed scale and accuracy before. Not surprisingly, the RQF is gone, and the NRC ranking process threatens to consume a decade. The complexity of submissions required by the RAE increased over the years; as well departments elaborated their submissions over time in an effort to become more competitive. This raised questions about the cost/benefit ratio of the exercise, and the UK government proposed a metrics-only future for the RAE.

Complexity emerged in these systems as a response to consultation which produced pressures for fairness across heterogeneous academic disciplines. Presumably, complexity increased easily in the absence of any accounting of the full cost. Estimates of the full cost of these exercises were not found, which precludes systematic analysis of cost/benefit ratio. Any estimate of full cost would need to account both for the work at

the center, that is the framing then gathering of submissions and the work of the panels, as well as the work embedded in the system, that is the time and effort spent compiling the submissions. Perhaps even the cost of time spent in consultation and argument in the design phase should be incorporated.

The role of peer evaluation versus quantitative metrics, in particular bibliometrics (paper and citation counts, impact factors), is also worth considering. To someone with a background in bibliometrics, such as this author, the techniques seem well suited to the task of large scale evaluations of research. In addition, throughout the 1980s when these evaluation systems were initiated, each country housed world renowned specialists in the techniques, offering expertise that could be drawn upon in designing evaluation systems.[v] It was rather surprising therefore that bibliometrics has played little or no role in the British or U.S. evaluation systems.

The RAE has been the canonical exercise in peer evaluation of a nation's university departments, with no formal metrics component. However, in the U.K. there was a desire to move to a quantitative formula and eliminate the peer panels. If this happens, the move from peer panels to formula would contrast strikingly with the move being implemented in Australia from a very simple formula to peer panels informed by metrics. The NRC 1995 ranking was strongly driven by peer evaluation, though not the informed peer review of the panel exercises; rather the NRC used an opinion survey to elicit peer rankings. NRC then judged this inadequate and aimed to rank departments on quantitative variables. If, as seems likely, the opinion survey results correlate with a few mostly size related variables the basis for the final NRC ranking could in fact be fairly similar to the Australian Composite index or the "shadow metrics exercise." Thus we may see a convergence of method towards peer informed, metrics based, departmental level evaluation.

Metrics invariably include bibliometric variables. Any "shadow metrics exercise" in the RAE will include bibliometrics. The ERA seems likely to include a quite sophisticated suite of bibliometric indicators compiled centrally. The NRC ranking variables include three bibliometric measures compiled centrally. Table 1 records notes on the inclusion of bibliometrics and its role in the evaluation judgments.

Table 1 – Bibliometrics in university evaluation exercises

Evaluation exercises

	US	UK	Australia
Old	NRC 1995	RAE	Composite Index
New	NRC 2000-2008	Shadow metrics exercise	ERA

Does bibliometrics play a role?

	US	UK	Australia
Old	No, reputational ranking, though bibliometric data present in appendix	No, 4 papers per faculty member submitted for "reading"	Yes, paper counts, along with funding
New	Yes, full departmental bibliography with citation count	Yes, along with funding indicators	Yes, now internationally comparative

What is the primary basis for the evaluation?

	US	UK	Australia
Old	Peer judgment, survey based	Peer judgment	Indicator based, bibliometrics prominent
New	Indicator based, weight of bibliometrics unknown at present	Indicator based, bibliometrics prominent	Indicator based, discipline-specific indicators developed with peer involvement

Again there is convergence in that three exercises that initially eschewed or simplified bibliometrics moved to incorporate more sophisticated bibliometrics. Perhaps over the past decade advances in computing have made system level metrics seem more cost effective and achievable. It may also be that when the systems were initially implemented, academics were loath to accept the idea of evaluation, and the inclusion of bibliometrics would have made evaluations even more controversial. After a decade or so, academics have adapted to evaluation and bibliometrics may seem like just another methodological tweak, with advantages and disadvantages like the rest.

Notable in each system was evidence that universities were extremely responsive to hierarchical ranking. One effect of the RAE was to create what McNay termed assured, aspiring and anxious universities (HEFCE, 1997, 47). Attention devoted to RAE submissions did not decrease, even though as mentioned above, Sastry and Bekhradnia calculated that the median impact on total university revenue of the last exercise was 0.6%. A clear response to the Australian exercise was elicited even though it effected only 7% of university research revenue, and so an even smaller portion of total university revenue. And the rankings in the United States have shaped university strategy even though no money was attached to them at all.

Marginson noted in relation to the introduction of a university assessment in Australia in 1993:

> *Nothing less than the positional status of every institution were at stake; the process of competitive ranking had a compelling effect, leading to the rapid spread of a reflective culture of continuous improvement.* (Marginson, 1997, p. 74)

Harman related that in Australia allocation of funding based on the Composite Index had become "an important vehicle for developing status hierarchies" as data are published in newspapers and widely used (Harman, 2000, p. 116). Perhaps most tellingly, many UK universities may now be choosing high ranking over more money. RAE 2008 allows selective inclusion of faculty members.

> *. . . research-intensive institutions indicated that they would seek the best ratings rather than the financial rewards that could be won by entering more staff.* (Lipsett, 2007)

Even without an explicit tie to funding distribution, universities will seek to rise in rankings over time. Thus, greater scientific productivity is achieved for the cost of the evaluation, which is presumably less than the cost of increasing research funding.

Far from being a fad or passing preoccupation of one or another government, over the past two decades national university departmental ranking exercises have become embedded in several systems. The exercises effectively focused universities' attention on improving their research enterprises. The methods have evolved to include more metrics as well as a peer component. However, exercises designed in consultation with the academic community tended to become increasingly elaborate, to the point of becoming unwieldy. Going forward the challenge will be to find a balanced system that is fair to different disciplines, with costs that are fully accounted for and controlled, that takes advantage of sophisticated bibliometric and computational techniques to ensure accuracy, and that can be conducted swiftly enough that results reflect current departmental configurations at their release. Finally, it might be best if the systems are continually tweaked because this will make it more difficult for universities to focus simply on improving scores on specific indicators and more likely that the only sure route to success will be a long term focus on improving the research enterprise.

Acknowledgements

Consultations with Ben R. Martin, Linda Butler and Phil Shapira were most helpful in the writing of this paper and the author is grateful for their insights. There remains a distinct possibility that misinterpretations remain however, and for those the author is solely responsible.

References

Australian Bureau of Statistics (2006). *Research and Experimental Development: Higher Education Organisations, 2004 Reissue*, 8111.0, July, http://www.ausstats.abs.gov.au/ausstats/

Australian Government, Department of Education, Science and Training (2006). *Research Quality Framework: Assessing the quality and impact of research in Australia, The Recommended RQF.* Commonwealth of Australia: October

Australian Vice Chancellors Committee (AVCC) (2005). *University Funding and Expenditure*, January, http://www.universitiesaustralia.edu.au/documents/publications/stats/Funding&Expenditure.pdf

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

Carr, Senator the Hon. Kim (2008). *New Era for Research Quality*: Announcement of Excellence in Research for Australia Initiative. Press release, February 26, 2008. Retrieved March 6, 2008 from: http://minister.industry.gov.au/SenatortheHonKimCarr/Pages/NEWERAFORRESEARCHQUALITY.aspx

Department of Trade and Industry (DTI), Office of Science and Innovation (2007). *PSA Target Metrics 2006*. (London: HMSO)

Dusansky, R. & Vernon, C.J. (1998). Rankings of U.S. Economics Departments. *The Journal of Economic Perspectives*, 12, 1, 157-170

Geuna, A. & Martin, B.R. (2003). University Research Evaluation and Funding: An International Comparison. *Minerva*, 41, 277-304

Gläser, J., G. Laudel, S. Hinze & Butler, L. (2002). *Impact of evaluation-based funding on the production of scientific knowledge: What to worry about and how to find out*, Expertise for the German Ministry for Education and Research, May, http://repp.anu.edu.au/expertise-glae-lau-hin-but.pdf

Gläser, J., T. S. Spurling & Butler, L. (2004). Intraorganisational evaluation: are there "least evaluable units"? *Research Evaluation*, 13, 1

Harman, G. (2000). Allocating Research Infrastructure Grants in Post-binary Higher Education Systems: British and Australian Approaches. *Journal of Higher Education Policy and Management*, 22, 2, 111-126

Higher Education Research Opportunities in the United Kingdom (HERO), [online] Accessed June 8, 2008. Available from World Wide Web: <http://www.hero.ac.uk/uk/research/research_quality_and_evaluation/research_assessment_exercise_rae_2001.cfm>

National Research Council (2004). *Assessing Research Doctorate Programs: A Methodology Study (2004)*. (Washington DC: National Academies Press) http://www.nap.edu/openbook/030909058X/html/28.html

U.K. Department of Trade and Industry (DTI) (2007). *Science, Engineering and Technology Statistics*, Table 5.1

Higher Education Funding Council for England (HEFCE) (1997). *The impact of the 1992 Research Assessment Exercise on higher education institutions in England*. (Bristol: Higher Education Funding Council for England, M6/97) http://www.hefce.ac.uk/pubs/hefce/1997/m6_97.htm

Jackman, R.W. & Siverson, R.M. (1996). Rating the Ratings: An Analysis of the National Research Council's Rating of Political Science PhD Programs. *PS: Political Science & Politics*, 29, 2, 155-160

Lipsett, A. (2005). RAE Raises UK Journal Activity. *Times Higher Education Supplement*, July 1, section 1698, 4

Lipsett, A. (2007). RAE Selection Gets Brutal. *Times Higher Education Supplement*, February 2, section 1779, 1

Marginson, S. (1997). Steering from a distance: Power relations in Australian higher education. *Higher Education*, 34, 1, 63-80

Martin, B.R. (2007, January). *Replacing the RAE with Metrics – is this the way forward?* (Seminar presented at Imperial College)

Miller, A.H., C. Tien & Peebler, A.A. (1996). Department Rankings: An Alternative Approach. *PS: Political Science & Politics*, 29, 4, 704-717

Moed, H.F. (2007). *UK Research Assessment Exercises: Informed judgments on research quality or quantity? Scientometrics,* Forthcoming

Sanders, C. (2006). Boom Time for High-Flyers. *Times Higher Education Supplement*, March 17, section 1734, 6

Sastry, T. & Bekhradnia, B. (2006). *Using Metrics to Allocate Research Funds: A short evaluation of alternatives to the Research Assessment Exercise*. (May, Oxford: Higher Education Policy Institute)

Segal, Quince Wickstead (SQW) (1996). *A Study of Selectivity*. (Bristol: HEFCE, M 20/96) http://www.hefce.ac.uk/pubs/hefce/1996/m20_96_2.htm

Wojtas, O. (2007). RAE fuels trend for research only time. *Times Higher Education Supplement*, January 19, section 1777, 64

---

[i] Only research evaluation is considered here. The teaching and social or economic development missions of universities are not discussed.

[ii] The National Research Council (NRC) functions under the auspices of the National Academy of Sciences (NAS), the National Academy of Engineering (NAE), and the Institute of Medicine (IOM). The NAS, NAE, IOM, and NRC are part of a private, nonprofit institution that provides science, technology and health policy advice under a congressional charter.

[iii] Note the difference with the NRC exercise. The NRC will issue a rank ordering of departments. The RAE issues "grades" to each department.

[iv] The portion of research funding based on the evaluation results was called the "research quantum" until 2001 and the "institutional grants scheme" thereafter. In 2004, the Institutional Grants Scheme accounted for AU $285 million of AU $4,283 in R&D funding in universities (HERD) (Australian Vice Chancellors Committee, 2005, Table A.1; Australian Bureau of Statistics, 2006, p. 3).

[v] In the UK were found Ben Martin and John Irvine at SPRU, University of Sussex; in Australia Paul Bourke and Linda Butler at the Australian National University in Canberra; and in the U.S., Francis Narin of CHI Research and also ISI, now Thomson Scientific, provider of the Science Citation Index, the database most used for bibliometric analysis.