evaluating university research performance using metrics

linda butler and ian mcallister*

Research School of Social Sciences, Australian National University, Canberra
ACT 0200, Australia

E-mails: linda.butler@anu.edu.au; ian.mcallister@anu.edu.au

*Corresponding author.

doi:10.1057/eps.2010.13; published online 14 January 2011

Abstract

Evaluations of research quality in universities are now widely used in the advanced economies. The UK's Research Assessment Exercise (RAE), which began in 1986, is the most highly developed of these research evaluations. Based on peer review and involving some sixty-nine panels evaluating the research work of more than 50,000 academic staff, the exercise is expensive and time consuming. In this article, we examine the possibility that a quantitative, metrics-based approach can provide a low-cost alternative to expensive, qualitative peer review. To do this, we build on our previous work on political science by extending a metrics-based model to chemistry, using the results of the 2001 RAE. Our results show that no single model will apply across science and non-science disciplines. Any metrics approach to performance evaluation has to use a discipline-specific suite of indicators.

Keywords research assessment exercise; citations; chemistry; political science; RAE

Over the past quarter of a century, liberal democracies around the world have made strenuous efforts to improve the efficiency of the public sector. Performance evaluation has become a crucial tool in the process of generating better outcomes across all areas of government responsibility, from health and social security, to policing and criminal justice (Heinrich, 2002; Propper and Wilson, 2003). In making government funding dependent

on performance, few areas have been as contentious as higher education. Based on long-established principles of independence and self-government, universities have traditionally resisted government involvement in their affairs, despite relying on the state for the bulk of their funding (Dixon and Suckling, 1996).

With its Research Assessment Exercise (RAE), Britain has moved farther than any country along the path of applying performance evaluation criteria to higher

education funding. Starting with small-scale pilot exercises in 1986 and 1989, by 1992 the RAE had encompassed the whole university sector and the outcome determined the direction of a significant pool of quality-related research funding (Bence and Oppenheim, 2005). Criticisms of the RAE have been many, but one consistent complaint is the cost of the exercise, involving the participation of more than 50,000 academic staff and up to sixty-nine subject panels, each with about twelve members. The high cost of this approach has led many to consider what lower-cost alternatives exist to measuring research performance.

The British government appeared to endorse this criticism, and in 2006 announced its intention to abandon the traditional peer review approach to evaluating the research performance of universities (HM Treasury *et al*, 2006). The reasons given for the change were that the peer review system was too expensive and too cumbersome. In its place, the government proposed to use a measure based on success in attracting competitive grant income to distribute research funds. The widespread opposition to this proposal was not successful in reversing the decision to adopt a metrics-based approach, but the Higher Education Funding Council for England (HEFCE), the government agency responsible for overseeing the introduction of the new system, soon acknowledged that a single measure based on income was not the optimal methodology.

This reassessment opened the way for a more broadly based set of performance indicators. HEFCE launched an extensive series of consultations, and in June 2006 released a consultation document which addressed the possibility of extending the range of metrics to be used in the assessment process (DfES, 2006). But which metrics should be included in this new portfolio? Those based on citation data have widespread backing in most

'We also found that whether or not a department had a member on the assessment panel was a strong predictor of the RAE outcome'.

science, technology, engineering and medical (STEM) disciplines, but are commonly rejected in the humanities, arts and social sciences (HASS) (CHASS, 2005; AHRC and HEFCE, 2006). Proposals to use measures based on research student data can be applicable across all disciplines, but like citation measures, those based on grant income have less currency in the HASS disciplines.

The 2001 UK RAE, with its publicly accessible database of all information provided to assessment panels, provides the ideal opportunity to test the efficacy of a number of proposed metrics, and to develop possible models based on a basket of the more commonly proposed measures. In a previous paper, we developed a model for political science which was based on the measures that appeared to be the best predictors of the actual RAE rank given to departments (Butler and McAllister, 2009).¹ The most important and statistically significant predictive variables were citations (using a target-expanded methodology to identify citations to books and book chapters, in addition to journal articles) and departmental size (represented by student numbers). Research income was not a strong predictor, and nor was departmental size if measured by staff numbers. We also found that whether or not a department had a member on the assessment panel was a strong predictor of the RAE outcome.²

In this article, we extend our analysis to include a STEM discipline, to examine

whether the same metrics are again the major predictors of RAE scores. For this exercise we chose chemistry – a discipline where around 90 per cent of output is in journals, of which over 90 per cent are indexed in Thomson Reuter’s Web of Science (WoS) (Moed, 2005). In theory this should produce a more robust model predicting the RAE outcome, using citations as one of the main indicators. It also provides a more stringent test of our previous analysis of political science, and of our conclusion that metrics provide a reliable, cost-effective alternative to peer review.

RAE BACKGROUND AND PROCEDURES

The main purpose of the RAE is to enable the higher education funding bodies to distribute public funds for research selectively on the basis of quality. Before the 1980s, university research funding in the UK was allocated by subject-based committees using largely subjective criteria on which to base their evaluations. The lack of transparency in this process, together with arguments about the contribution of university research to the national economy, resulted in the design of a more transparent, objective system of research assessment. Table 1 shows the progress of these evaluations, from the first exercise in 1986, to the most recent one in 2008.

The first RAE, conducted in 1986, was a small-scale pilot project which involved only the established universities, with a more comprehensive assessment following in 1989. The 1989 exercise was largely replicated in the following RAE, conducted in 1992. In the 1986 and 1989 exercises, only a small proportion of university funding was dependent on RAE performance, but by 1992 virtually all block grant university research funding was contingent on RAE performance. In 1996 a much more ambitious exercise was mounted, involving over 50,000 academic staff nominated as ‘research active’ by their universities who were evaluated by sixty-nine committees representing discrete subject areas. The 1996 RAE represented the transition from ‘a quality assurance process to a competition for funding’ (UK Parliament, 2002; see also Bence and Oppenheim, 2005). It moved away from collecting full publication lists, focused on ‘research active’ staff in the unit being assessed, and had at its centre the quality assessment of the four submitted publications from each academic. Just as significantly, it was now an important vehicle for establishing or validating the prestige of a department.

The 2001 RAE – the fifth in the series – was substantially based on the 1996 exercise, though it was refined to take account of criticisms, most notably by making better provision for multidisciplinary research and by ensuring the consistency of

Table 1: The UK research assessment exercises, 1986–2008

	Type of evaluation	Assessed units	
		Chemistry	Political science
1986	Small scale pilot	n.a.	n.a.
1989	Universities only (excluded polytechnics)	n.a.	n.a.
1992	43,000 staff, 63 subject committees, 170 HEIs	67	66
1996	> 50,000 staff, 69 subject committees, 192 HEIs	62	66
2001	> 50,000 staff, 68 subject committees, 173 HEIs	45	69
2008	> 50,000 staff, 67 subject committees, 150 HEIs	33	59

scoring across the sixty-eight subject area committees.³ The 2001 exercise considered a research output of over 150,000 publications⁴ from almost 50,000 academic staff in 2,598 submissions from 173 higher education institutions (HEIs). In political science, sixty-nine academic units were assessed from sixty-six HEIs (in the case of three institutions, two assessable units were entered),⁵ while for chemistry forty-five institutions submitted departments for assessment.⁶

The 2001 RAE allowed each assessable unit to nominate their 'research active' staff members, and to provide up to four publications for each nominated staff member 'that reflect their best quality work over the period covered by the RAE' (RAE, 2001a,b). The guidelines allowed for the possibility that a member would not have four assessable publications, particularly in the case of new entrants, those involved in major monographs (particularly in political science), and where there were 'unusual or particular circumstances'.

The sixth RAE was completed in December 2008, and differed from the 2001 exercise in two key ways (RAE, 2008: 3–4). First, the seven-point scale used in the 2001 RAE (which we examine in detail in the next section) was replaced by a 'quality profile' which allocated outputs to one of four grades, plus a fifth 'unclassified' grade. The purpose of the change was to reduce 'the averaging effect of single-point ratings by allowing panels to exercise a finer degree of judgment, especially at grade boundaries' (RAE, 2008: 2). Second, the sixty-seven disciplinary areas were grouped into fifteen 'main panels', in order to ensure consistency in evaluations across cognate disciplines. As with the 2001 exercise, the 2008 RAE was a major undertaking, involving 2,344 submissions from 159 HEIs, listing the work of more than 50,000 researchers. In political science, fifty-nine academic units were assessed, and in chemistry, thirty-three units.

'Critics of the RAE have pointed to ... the tendency to favour safe, mainstream research that will lead to a steady stream of publications over more innovative, riskier research where the publication stream is not assured'.

The evolution of the RAE shows its origins in a small, pilot scheme focused on the traditional universities, to a major exercise determining the funding of the whole higher education sector. That has necessarily shaped the behaviour of universities, as they have sought to maximise their financial returns under the scheme's rules. Critics of the RAE have pointed to the emergence of a football-style transfer market to buy in research stars, and the tendency to favour safe, mainstream research that will lead to a steady stream of publications over more innovative, riskier research where the publication stream is not assured. Some evidence of the way in which institutions have adapted to the system is shown in Table 1, with a declining number of HEIs who participated in each RAE since the peak in 1996. This reflects the declining returns from poorly ranked departments, so that the effort required to make a submission often outweighs the potential return.⁷

MODELLING THE 2001 RAE OUTCOME

Our aim is to see how far we can replicate the peer review-based outcome of the 2001 RAE in chemistry and political science, relying instead on a metrics approach. The dependent variable in our

Figure 1 The distribution of RAE grades for chemistry and political science (number of departments).

model is the rank given to each of the academic departments submitted for assessment to the RAE. The scoring adopted by the RAE was on a seven-point scale running from 1 to 5, with 3 being divided into 3a and 3b, and 5 into 5 and 5*. A department was awarded the top 5* ranking if, in the opinion of the panel members, more than half of its research activity was of international excellence, and the remainder demonstrated it had attained levels of national excellence. At the other end of the scale, a department was awarded a ranking of 1 if virtually none of the research activity it submitted was judged to have attained national excellence.

Figure 1 shows the distribution of the RAE scores across the forty-five chemistry and fifty-nine political science departments submitted to the RAE. No department received the lowest grade of 1, and only five departments were awarded grades of 2 or 3b. However, there were significant differences in overall outcomes between the two disciplines. Chemistry departments were awarded 5* scores at almost double the rate of their political science counterparts and were given lower scores (2, 3a, 3b) much less frequently. In both disciplines, the median outcome was a score of 4.

Each of the sixty-nine subject area panels laid down clear rules in advance for evaluating their respective fields. The

panel for political science identified three evaluation criteria (RAE, 2001b):

1. the quality of research outputs;
2. evidence of the vitality of the research culture; and
3. evidence of the quality of research activity as reflected by a number of other factors such as research post-graduates, esteem indicators and contributions to academic public goods.

The chemistry panel identified an expanded list of evaluation criteria, highlighting the importance of external income in their assessment (RAE, 2001a):

1. the quality of research as judged by publications and other output. There will be no *prima facie* discrimination between forms of research output;
2. the extent of research and its quality as indicated by the activity of research students, research assistants, post-doctoral workers, visiting academic and industrial researchers and others who are deemed to add value to the research output;
3. evidence of support by external funders, as indicated by research income from research councils, industry, international agencies etc.;
4. evidence of the vitality of a department, as demonstrated by research strategy and achievement over the assessment period;

5. evidence of national and international peer group recognition.

In the case of publications, both panels committed to collectively reading at least one quarter of the submitted output of a department, with political science panelists additionally undertaking to examine a greater proportion of publications if the outcome turned out to be borderline. The procedure for political science was to evaluate the publications using 'the standards of research that are employed in a peer-review process' (RAE, 2001b), with authored books judged most important. In chemistry, the panel undertook to take into account 'the quality of the concept, the methodology employed and the importance of the results' (RAE, 2001a). No mention was made of the use (either implicit or explicit) of quantitative indicators such as citations or journal impact factors. For both disciplines, the quality of the publications is measured here by the mean citations received for the publications entered by each staff member.⁸

At this point, it is worth emphasising an important and novel aspect of our citation analysis for political science. Our counts were obtained by interrogating Thomson Reuter's Web of Science (WoS) database for citations to all 4,400 publications submitted for assessment to the RAE. The analysis is not restricted to articles in journals indexed in the WoS, and hence our measure is far more comprehensive than standard citation measures. This method is more defensible for a discipline such as political science, where only a quarter of submitted outputs were articles in WoS journals. In contrast, 98 per cent of publications submitted for assessment by chemistry departments were indexed in the WoS, so there were no concerns about employing standard bibliometric analyses based on journal articles only.

As we would expect, the distribution of citations across departments is not linear, and for that reason we use a log

transformation in the multivariate analyses. although the mean citations per submitted work is much higher for chemistry departments than for political science units, the two distributions exhibit very similar curvilinear distributions. Chemistry means range from a low of 9.83 (for de Montfort University) to a high of 68.49 (for the University of York), whereas political science ranges from a low of 0.78 (for Middlesex University) to a high of 32.48 (for the Open University), with an overall mean of 29.55 for chemistry and 5.3 for political science. For both disciplines, the exceptional citation performance for the departments with the highest mean was predicated on a small number of very highly cited publications. In York, two of the submitted works received 3,007 citations and 1,472 citations,⁹ whereas in the Open University, the high score was substantially attributable to the 711 citations received by one book, Paul Hirst and Grahame Thompson's *Globalization in Question* (Polity, first edition 1996), and the 610 citations received by a second book, David Held, Anthony McGrew, David Goldblatt and Jonathan Perraton's *Global Transformations* (Stanford University Press, 1999).¹⁰

Several other variables can be identified to represent 'the vitality of the research culture' and the 'quality of the research activity' as outlined in the panels' assessment criteria. The model includes the annual research income from research councils covered by the Office of Science and Technology and the British Academy (OST/BA). Research income is a particularly important inclusion in the model as initial proposals in the UK suggested that funding decisions might be made solely on the basis of success in obtaining external funds (DfES, 2006). Demonstrating the importance of OST/BA income in chemistry, the mean across all 44 departments is £144,240 compared to just £8,630 for political science.

Indeed, the smallest amount reported for a chemistry department (£96,225 for Northumbria) was close to the highest level in political science (£119,930 for Sussex-SPRU) (see Table 2). The highest levels reported for chemistry were around £20 million (£20.72 million for Oxford, and £19.44 million for Cambridge). At the other end of the scale, twelve political science departments reported receiving no OST/BA income at all.

Research 'esteem', always a difficult concept to quantify, is measured here by membership of national academies – the British Academy and the Royal Society of Edinburgh for political science and the Royal Society of London and the Royal Society of Edinburgh for chemistry. Academy membership was far more prevalent in chemistry departments with over half having at least one member among their staff, whereas only thirteen of the sixty-nine departments in political science had a member. Highest numbers

in the respective disciplines were nine for political science (Oxford) and fourteen for chemistry (Oxford and Cambridge).

The size of a particular department is often regarded as an indicator of research activity in its own right, with larger departments capitalising on their size to generate an active research culture and to sustain research groupings that further attract interest and resources (Jackman and Siverson, 1996). The model includes the number of academic staff submitted in the RAE – in other words, the number considered 'research active'.¹¹ On average, chemistry departments were double the size of political science departments – a mean of 31 versus a mean of 17. This difference was due primarily to the existence of only one chemistry department with less than ten research active staff submitted, while nineteen submissions were received from political science departments with less than ten nominated staff. The largest submissions

Table 2: Variables, definitions and means

Variable	Scoring	Means	
		Chemistry	Political science
RAE ranking	1 = 2, 2 = 3a, 3 = 3b, 4 = 4, 5 = 5, 6 = 5 star	4.30	3.97
Citations (log)	Mean citations per submitted work, logged	7.84	1.47
<i>Research culture</i>			
External annual income	£1,000s	144.24	8.63
Academy membership	1 = member(s) in department, 0 = none	0.52	0.19
<i>Department size</i>			
Academic staff	Number	31.41	17.00
Graduate students	Mean students per staff member	3.72	6.02
Member on RAE panel (N)	1 = yes, 0 = no	0.16 (44)	0.17 (69)

Note: Academy membership is the Royal Societies of London and Edinburgh for chemistry, the British Academy and the Royal Society of Edinburgh for political science. Academic staff is full-time equivalents.

were similar for both disciplines – eighty in political science (LSE) and seventy-eight in chemistry (Oxford). The number of graduate students is also a good measure of research activity, since high profile researchers will attract students who want to work with them. The mean number of graduate students, measured as full-time equivalents per research active staff, was 6.02 for political science and 3.61 for chemistry.

Finally, the model controls for whether or not the academic department had a member on the RAE panel. Panels consisted of twelve members, predominantly drawn from the top ranked departments in the previous RAE. Of the four departments ranked as 5* in political science in 1996, three were represented on the 2001 panel, as were three of the six departments ranked 5 in 1996; both chemistry departments ranked 5* in 1996 were represented on the panel, as were three of the seven departments ranked 5.

How effective is the model in predicting the outcome of the RAE for the departments in the two disciplines? The results of the OLS regression analyses¹² in Tables 3 and 4 suggest that the model works well for political science, explaining 60 per cent of the variance in the RAE outcome, and even better for chemistry where it accounts for 86 per cent of the variance. However, there are major differences between the two disciplines in which variables are the most important predictors of RAE outcomes. For political science, the mean number of citations a work attracts significantly improves the RAE outcome for a department and is the most important predictor in the model. However, for chemistry it is external income that is the most significant variable.

The second important predictor of outcome for a political science department is having a member on the RAE panel, whereas for chemistry it is mean number

'The number of graduate students is also a good measure of research activity, since high profile researchers will attract students who want to work with them'.

of citations. It seems intuitively surprising for a discipline well served by citation databases, where 98 per cent of the output submitted for assessment is indexed in the WoS, that citation data do not play a more significant role. We conducted a decomposition of effects which is essentially an exercise to see why citations in chemistry are so much less important than in political science, which on face value appears counter-intuitive. The answer is the strong correlation between income and RAE rank in chemistry, which is not found in political science, and demonstrates a clear and important disciplinary difference.

Ordinary least squares regression equations showing partial and standardised coefficients and standard errors predicting RAE outcome. See Table 2 for details of variables and scoring. $N=44$ chemistry departments, 69 political science departments.

Although limiting the citation analysis to the 'best four' publications for each submitted staff member was the most valid choice in the first instance, as we sought to replicate the political science analysis and mirror the information provided to panelists, it is unlikely that any metrics-based assessment system would be so restrictive. We therefore modelled the chemistry results using two alternative citation measures – the log of citations to all publications from each university department for the full

Table 3: Predicting RAE outcome in chemistry and political science (OLS estimates)

	Chemistry			Political science		
	Partial	(SE)	Standard	Partial	(SE)	Standard
Citations (log)	0.491**	(0.170)	0.370**	0.410**	(0.136)	0.273**
Department member on RAE panel	0.216	(0.184)	0.075	0.903**	(0.248)	0.318**
<i>Department size</i>						
Academic staff	0.005	(0.008)	0.064	0.020*	(0.008)	0.245*
Graduate students	0.076	(0.068)	0.087	0.033**	(0.012)	0.221**
<i>Research culture</i>						
OST/BA annual income	0.006**	(0.002)	0.414**	0.009	(0.005)	0.147
National academy membership	0.240	(0.164)	0.113	0.203	(0.278)	0.074
Constant	-0.984	(1.134)		2.551	(0.206)	
Adj R-squared	0.862			0.599		
(N)	(44)			(69)		

**statistically significant at $p < 0.01$, * $p < 0.05$.

Table 4: Predicted RAE outcome in chemistry and political science using metrics (OLS estimates)

	Over-prediction			Under-prediction	
	Actual	Predicted		Actual	Predicted
<i>Chemistry</i>					
Huddersfield	3b	3a	Imperial	5*	5
Surrey	3a	4	Loughborough	4	3a
Northumbria	3b	3a	Bristol	5*	5
			Durham	5*	5
			Heriot-Watt	4	3a
			East Anglia	5	4
<i>Political science</i>					
Southampton Institute	2	3a	Kings College (War)	5*	5
Open University	3a	4	Queens Univ Belfast	5	4
LSE	5	5*	Bradford	5	4
Sussex (Politics, IR)	4	5	St Andrews	5	4
Cambridge	4	5	De Montefort	5	4
Univ College London	3a	4	Reading	5	4
Westminster	4	5	Bristol	5	4
Kent	3a	4	Exeter	5	4
Derby	3b	3a	Sheffield	5*	5
Middlesex	3b	3a	Aberystwyth	5*	5
Swansea	3a	4	Essex	5*	5
Sussex (SPRU)			Birmingham	5	4
			Coventry	4	3a
			Dundee	4	3a
			Keele	5	4
			Durham	4	3a
			Hull	5	4
			Queen Mary	4	3a

six-year period; and the h-index for the department. The two alternatives both marginally reduced the predictive power of the model, suggesting that the panel did focus on the publications submitted for assessment, rather than taking into account the wider output of departments.

METRICS AND RESEARCH QUALITY

The results presented so far have shown that quantitative indicators of research quality predict a large proportion of the variance in the 2001 RAE outcome in both chemistry and political science, which was arrived at through peer evaluation. The results also show that, at least for political science, membership of the RAE panel

had a significant influence on the outcome, net of other things. In this section, we examine the extent to which the use of peer evaluation to measure research quality could be dispensed with, relying instead solely on objective, quantitative measures such as citations, income and student and staff numbers. If the 2001 RAE in chemistry and political science had adopted this metrics-only approach, how would the results have differed from what was actually decided upon, using peer evaluation?

To make these estimates, we use the regression models in Table 3, but omitting the variable measuring membership of the RAE panel. To calculate the predicted RAE outcome using this model, the regression equation is re-estimated for each

of the forty-four chemistry departments and sixty-nine political science departments, substituting the means for each department into the equation and calculating the predicted RAE outcome. The predicted outcome is then compared with the actual outcome to see which departments have improved their grade and which would have dropped in grade as a result of moving to a solely metrics-based evaluation. The results of this exercise are shown in Table 4.

The 'actual' figure is the 2001 RAE outcome for each department. The 'predicted' figure is the prediction based on the OLS regression model in Table 3, minus membership of the RAE panel; see text for details of estimation. Only departments that changed in grade because of the metrics model are shown, and the predicted figure has been rounded to the nearest grade. The departments are ranked by the size of the difference between the actual and the predicted outcome.

Overall, only one of the 113 departments would have received an outcome on the metrics model which would have differed by more than one unit on the peer evaluation grading used by the RAE. This is Southampton Institute, which would have moved from a 2 to a 3a grade in political science. Of the remaining departments, Table 4 shows that fourteen departments would have improved their position by one grade, whereas twenty-four departments would have moved down one grade. In general, the departments in the left-hand side of Table 4 that would have improved their position were lower ranking departments, the only exception being the LSE, which would have moved from a 5 to a 5-star grade in political science.

The twenty-four departments that would have dropped one grade on the right-hand side of Table 4 are, conversely, disproportionately drawn from the higher ranking departments. As predicted by

'... only one of the 113 departments would have received an outcome on the metrics model which would have differed by more than one unit on the peer evaluation grading used by the RAE'.

the adjusted *r*-squared figures, the model has a much better fit to outcomes in chemistry than in political science – the rankings of 80 per cent of departments remained unchanged for the former, whereas only 60 per cent remained the same in the latter.

These results, while affecting the grades of thirty-nine of the 113 departments, suggest that a metrics approach to measuring research quality would produce results which are close to those that come from peer evaluation of research, particularly in the natural sciences. Moreover, the metrics results are more defensible, since they are achieved without recourse to a panel of assessors drawn from departments that are themselves being assessed, and the accompanying risk of accusations of real or imagined conflicts of interest. The robustness of the metrics-based results suggests to us that such evaluations are a much more efficient and transparent approach to measuring research quality.

DISCUSSION

Britain has been a pioneer in the use of metrics to assess research performance and to use the resulting evaluations to determine university funding. However, in the rush to find a more cost effective evaluation process, there has been less acknowledgement of the differences

between the research cultures of the various disciplines. This has implications beyond Britain, since many countries are moving in a similar direction and are watching British developments closely. The findings presented here suggest, unequivocally, that a metrics-based model, using objective, transparent indicators drawn from a range of readily available measures, will yield results which are close to those of a peer-based evaluation model, and can be used with confidence. However, they also point to strong differences between HASS disciplines (represented in our analysis by political science) and STEM disciplines (represented by chemistry).

Our analysis demonstrates empirically the differences that exist between STEM and HASS disciplines in terms of the choice of measures that are most appropriate in a metrics-based system. Briefly, these differences are as follows.

1. For chemistry, a core STEM discipline, there was a very close association between RAE rank and research income. In political science, the predictive power of research income was almost non-existent.
2. Our target-expanded citation metric was by far the strongest predictor of RAE outcomes in political science, but in chemistry citations ranked second behind research income. The latter finding was because of the very close association between RAE rank and research income in each department, which was even closer than between citations and RAE rank.
3. Our finding on the significance of panel membership in political science was not replicated in chemistry. There are a number of explanations for this. Tacit knowledge of the high correlation between quantitative measures in chemistry may have influenced the deliberations of the panel members, even though there was no formal use of metrics in the process. In political

science, this may not have been the case and the opinions and assessments of panel members may have had a much stronger role to play.¹³

4. The size of a department does matter, but not to a great extent. Again we see a difference between the two disciplines – in chemistry the effect is much smaller compared to political science.
5. In both disciplines, a measure based on esteem (membership of a national academy) has no predictive power and is unnecessary for any metrics-based system.
6. Our model explains over 86 per cent of the variance in the RAE outcome for chemistry, but only 60 per cent of the variance for political science.

The implications to be drawn from our analysis is that the 'basket of metrics' chosen for the assessment of STEM disciplines may look different to that assembled for HASS disciplines. A STEM 'basket' might rely almost exclusively on citation and external income data, with little need for other indicators. For these disciplines, the focus would be on selecting citation measures that are sensitive to varying citation practices within broad disciplines, that do not mitigate against the inclusion of recent publications that have had little time to attract citations, and that do not prove to be a disincentive for collaborative and interdisciplinary research activities. There would be little requirement for peer input except to scrutinize the data for any errors and to determine the cause and, potentially, the treatment of obvious outliers. Any suite of metrics would also need to be chosen with an eye to minimizing their potential for 'game playing' by those who seek to manipulate the system. However, citation metrics have powerful 'gatekeepers' – journal editors – who are unlikely to countenance blatant perverse practices.

For HASS disciplines, filling the 'basket' may prove more problematic. There is

little of value to be gained from using grant income for many disciplines, but collecting student and staff data and undertaking more comprehensive citation analysis using the target-expanded approach may prove to be a valid method. One lingering doubt is over the cost of adopting the latter approach to bibliometric analysis, which is significantly more expensive and time-consuming than the traditional approach used for STEM disciplines. It is also likely that peer input may still have a necessary role to play in the assessment of HASS disciplines, even in a streamlined metrics-based approach. However, our data have demonstrated, at least for political science, the influence of peers on the assessment process. Care must therefore be taken to ensure that incorporating peer judgements to overcome some of the shortfalls of a metrics-based system does not introduce bias.

Linking performance to government funding is now accepted in the public sector, but it remains controversial in higher education. The British experience of using qualitative assessment has shown that, over several iterations of the RAE, a transparent and effective process of research performance evaluation is feasible. The drawback is that such evaluations are hugely resource-intensive and disruptive for the universities concerned. Our research has shown that a metrics approach has the potential to arrive at similar evaluations, but at a much reduced cost – in time, resources and disruption. The key to such an exercise is to be sensitive to the different research cultures between disciplines, and to accommodate these differences in the basket of metrics that is eventually decided upon. Metrics do work, but such an approach has to use different metrics for different disciplines.

Notes

1 Our earlier paper was part of a symposium on research evaluation in political science, and our contribution was the subject of comment and criticism from, among others, Albert Weale, Andrew Russell and Claire Donovan. We do not seek to address or report on all of the points made in these debates as they can be found in the January 2009 issue of *Political Studies Review*.

2 Parts of the discussion and analysis relating to political science draw on Butler and McAllister (2009).

3 In 2001 the subject areas were reduced by one, to sixty-eight.

4 Research outputs are usually, but not always, publications. Other types of output – including films, exhibitions or contributions to television programmes – are increasingly common (UK Parliament, 2002).

5 The units were: King's College London (War Studies and Defence Studies); Nottingham Trent University (International Relations, and Politics and International Studies); and the University of Sussex (SPRU – Science and Technology Policy Research, and Politics and International Studies).

6 For one unit of assessment, from the North East Wales Institute for Higher Education, it proved impossible to identify data within the Web of Science for use in some of our analyses, and this unit was omitted.

7 For example, in 1992 no funds were allocated to departments ranked lowest in five grades; in 1996 this was extended to the lowest two of seven grades; and in 2001, to the lowest three of seven grades.

8 Other ways of measuring citations include: total citations, relative citation impact, and the Hirsch index (or h-index as it is now commonly known, see Hirsch, 2006). The measure used here – mean citations per submitted work, across the sixty-nine departments – is the most straightforward. Early analyses controlled for the year in which publication of the submitted work took place, since the period of time that has elapsed might influence the opportunities for the work to be cited. Only works published between 1996 and 2000 could be submitted to the 2001 RAE. This control made no substantive difference to the results and has been omitted in the interests of presenting as parsimonious a model as possible.

9 Namely: Buzan *et al* (1998), Mouffe (2000).

10 Our earlier paper omitted Thompson from the database because of the problem of having to search by first author for publications indexed in the ISI. That has been corrected, and makes no substantive difference to the results of our analysis. The only changes are slight variations in the size of the

coefficients reported for political science in Table 3, and a slight variation in the ordering of the political science departments reported in Table 4.

11 A valuable measure would of course be the total number of academic staff in a department, but that was unfortunately not available.

12 The estimates were also made using logistic regression (to take into account the skewed distribution of the dependent variable). Since the results were substantially the same as those obtained using the OLS model, we have relied on the latter here, as the results are more easily interpretable.

13 A further possible explanation, put forward by one of the reviewers for this journal, was the concentration of chemistry within the older universities by 2001, so that tacit knowledge of the RAE rules and procedures was evenly spread across chemistry departments. By contrast, political science had a larger number of departments spread across the old and new universities, whereas the panel membership was mainly drawn from the old universities. Tacit knowledge was therefore not as evenly spread across the political science discipline, making panel membership a more important resource for understanding the RAE process.

References

- Arts and Humanities Research Council and the Higher Education Funding Council of England (AHRC and HEFCE). (2006) 'Use of research metrics in the arts and humanities', available at <http://www.hefce.ac.uk/research/ref/group/>, accessed 3 December 2008.
- Bence, V. and Oppenheim, C. (2005) 'The evolution of the UK's research assessment exercise: publications, performance and perceptions', *Journal of Educational Administration and History* 37: 137–155.
- Butler, L. and McAllister, I. (2009) 'Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science', *Political Studies Review* 7: 3–17.
- Buzan, B., Wæver, O. and de Wilde, J. (1998) *Security: A New Framework for Analysis*, Boulder, Co/London: Lynne Reiner.
- Council for the Humanities, Arts and Social Sciences (CHASS). (2005) 'Measures of quality and impact of publicly-funded research in the humanities, arts and social sciences', available at <http://www.chass.org.au/papers/PAP20051101JP.php>, accessed 3 December 2008.
- Department for Education and Skills (DfES). (2006) 'DfES consultation on the reform of higher education research assessment and funding: summary of responses', available at <http://www.dfes.gov.uk/consultations/conResults.cfm?consultationId=1404>, accessed 3 December 2008.
- Dixon, H. and Suckling, J. (1996) 'Measuring Outcomes in Higher Education', in P. Smith (ed.) *Measuring Outcome in the Public Sector*, London: Taylor and Francis, pp. 47–56.
- Heinrich, C.J. (2002) 'Outcomes-based performance management in the public sector: implications for government accountability and effectiveness', *Public Administration Review* 62: 712–725.
- Hirsch, J.E. (2006) 'An index to quantify an individual's scientific research output', *Proceedings of the National Academy of Science* 102(46): 16569–16572.
- HM Treasury, Department of Trade and Industry, Department for Education and Skills, and Department of Health. (2006) 'Science and innovation investment framework 2004–2014: next steps', available at http://www.berr.gov.uk/dius/science/science-funding/framework/next_steps/page28988.html, accessed 3 December 2008.
- Jackman, R.W. and Siverson, R.M. (1996) 'Rating the rating: an analysis of the national research council's appraisal of political science Ph.D. programs', *PS: Political Science and Politics* 29: 155–160.
- Moed, H.F. (2005) *Citation Analysis in Research Evaluation*, Heidelberg: Springer.
- Mouffe, C. (2000) *The Democratic Paradox*, London: Verso.
- Propper, C. and Wilson, D. (2003) 'The use and usefulness of performance measures in the public sector', *Oxford Review of Economic Policy* 19: 250–267.
- Research Assessment Exercise (RAE). (2001a) 'Panels' criteria and working methods. 3.12 chemistry UoA 18', available at http://www.rae.ac.uk/2001/Pubs/5_99/ByUoA/crit18.htm.
- Research Assessment Exercise (RAE). (2001b) 'Panels' criteria and working methods. 3.31 politics and international studies, UoA 39', available at http://www.rae.ac.uk/2001/Pubs/5_99/ByUoA/crit39.htm.
- Research Assessment Exercise (RAE). (2008) 'Research assessment exercise 2008: The outcome', available at <http://www.rae.ac.uk/pubs/2008/01/>, accessed 25 December 2008.
- UK Parliament. (2002) 'Memorandum submitted by the higher education funding council for England (HEFCE)', available at <http://www.parliament.the-stationery-office.co.uk/pa/cm200102/cmselect/cmsctech/507/2012302.htm#n7>, accessed 16 January, 2009.

About the Authors

Linda Butler is a visiting fellow in the Research Evaluation and Policy Project at the Australian National University. She has published widely on bibliometrics and advised government agencies on the quantitative assessment of research. Her current research focuses on developing indicators for disciplines where standard journal citation analysis is inappropriate.

Ian McAllister is professor of political science at The Australian National University. He is the co-author of *The Australian Electoral System* (UNSW Press, 2006) and *Dimensions of Australian Society* (Palgrave Macmillan, forthcoming 2010). He has been director of the Australian Election Study since 1987 and is editor of the *Australian Journal of Political Science*.