Ben R. Martin

Research assessment in the United Kingdom and how it might be improved

1 Introduction

I am very grateful to the organizers of this conference for the opportunity to write about experiences with research assessment in the UK. Research assessment is one area where perhaps other countries have something to learn, if only because in Britain we have longer experience of research assessment than most. As regards the structure of my paper, it is useful to look first at the last ten to fifteen years to see how research assessment in Britain has evolved.[1] Then I shall discuss a study that Social Policy Research Unit (SPRU) carried out in the early 1990s which looked at the approach being adopted in the Research Assessment Exercise (RAE), how well it worked, and how, if at all, it might be improved. On the basis of that, I shall draw some conclusions about research assessment.

2 Science Policy in the UK over the last 20 Years

2.1 Evolution of UK Science Policy

To understand how research assessment emerged and evolved in Britain, let us begin by considering science policy under the last three Prime Ministers, beginning with Mrs Thatcher. As you will recall, she pursued a tight monetarist policy of trying to reduce public expenditure. Her goal was, in her favourite phrase, "to roll back the state", to subject the public sector to the discipline of the market place. She was also determined to ensure value for money with the emphasis on the so-called three E's – economy, efficiency and effectiveness.

---

[1] This overlaps a little with some of what John Krebs said about the UK Research Assessment Exercises, where we come to broadly similar conclusions.

What did this mean for science and technology policy? Those in the public science sector were encouraged to move closer to industry, to seek other sources of funding and to rely less on the state. They were encouraged to focus on exploiting the benefits from research, particularly the economic benefits and to ensure that Britain reaped those economic returns. However, there was a change in policy around 1987 whereby the Government stopped funding "near market" research. In my view, that introduced something of a contradiction: how could you focus on economic returns if you were not being funded to do research which was somewhere between traditional curiosity-oriented research and applied research. There was also increasing emphasis on accountability and hence on monitoring and evaluation. From about 1986 onwards, there was increasing evaluation both by the Research Councils[2] and what was then called the University Grants Committee.[3] In 1986, we had the first Research Assessment Exercise (RAE) of universities.

What were the policies under John Major? The key event was the 1993 White Paper on science, engineering and technology. Its title "Realising Our Potential" summarised what it was about – exploiting the UK science base for economic benefits in particular but also for benefits in terms of improved quality of life. The Research Councils were re-organized, given new missions and told that "users" had to be more directly involved. What that White Paper essentially did was to set out a new "social contract" between science and technology, on the one hand, and society and the state on the other. Under this, researchers who received money from the public purse had a responsibility, even a duty, to identify who might be the eventual users or beneficiaries of their research, and then to go to them, to help them identify their longer-term research needs, and to work with them in trying to meet those needs. In other words, under the revised social contract, if you receive public money for your research, you are accountable to society for that. One of the key mechanisms set up to achieve that goal was the Foresight Programme, the aim of which was to link science and technology more closely to national needs in relation to wealth creation and improved quality of life.

How has research policy changed under the Blair administration? The short answer is, "Not much!" There was a significant increase in funding earlier in the year but, apart from that, the policies and the mechanisms have not changed a great deal. However, there was one key development which started before the new Labour Government took over – namely the completion of the work by the Dearing Committee which was looking at the future of higher education in Britain. This advocated

---

2 The UK then had five Research Councils. In 1993, there were re-organized so that there are now six (e.g. Medical, Engineering and Physical Sciences etc.).

3 It later became the University Funding Council and subsequently the Higher Education Funding Council (HEFC) for England (with similar bodies for Scotland, Wales and Northern Ireland).

a more market-oriented approach to student education and more teaching quality assessment, as well as encouraging universities to continue seeking more funds from non-traditional sources – i.e. from users of various types.

2.2 The Research Assessment Exercise

In the first Research Assessment Exercise in 1986, there were 37 fields or units of assessment, and the method used (as it has been in all the subsequent exercises) was peer review by panels. At that stage, there were only four grades – below average, average, above average, and international excellence – and each unit or department submitted their five best publications.[4] However, it was quickly pointed out that this can introduce a bias in favour of large departments. If you assume that the quality of publications is perhaps distributed approximately on a normal curve, then for a larger department the five best papers are more likely to be further along that spectrum of excellence than for a small department. So that approach was dropped subsequently in the 1989 and 1992 exercises where there were further refinements in the approach. (In parallel with this, in 1994 we had the first Teaching Quality Assessments, initially with three grades – unsatisfactory, satisfactory and excellent. Subsequently that changed to assessment in terms of six different dimensions, each of which was ranked on a 4-point scale so a department can earn up to a maximum of 24 points. However, unlike in the Research Assessment Exercise, there is no extra money attached to doing well in the Teaching Quality Assessment, at least not directly.[5])

By 1996 and the fourth Research Assessment Exercise, the methodology had begun to settle down with peer review by nearly 70 panels (which by then included a few users) and a classification based on seven grades.[6] Each active researcher now listed four publications (or other forms of public output in the case of an artist, for example). No bibliometric statistics were used, however.[7] Units could include all researchers who were in employment on a particular census date, something which perhaps encouraged the development of an academic "transfer market" between universities. Lastly, by 1996 there were quite wide differentials in funding for the

---

4 As we read in the presentation by Dr Barend van der Meulen, the same approach is currently used in the Netherlands.

5 With a high Teaching Quality Assessment, a department may attract more students and thus generate more income from their fees.

6 What had been the 3 grade in 1992 was split into 3A and 3B, and a 5* grade was added at the top end based on the proportion of research that was judged to be of international excellence.

7 This was tried in the 1992 exercise, using data on the total number of publications produced by each department, but it was dropped in the 1996 exercise.

various grades; if you got a grade 1 or 2, you got no research funding; a 3B yielded one unit, a 3A yielded 50 % more, and so on up to five, with a 5* ("five star") earning 20 % more than a 5.

The results of the Research Assessment Exercise now influence large sums of money – 95 % of the research money from the Higher Education Funding Council. This is in contrast to the Teaching Quality Assessment where there is no direct financial consequences at present, although there has been some discussion as to whether there should be in the future.[8] My personal assessment of the RAE is that it has probably improved the overall quality of research, particularly in lower ranked universities. Most universities now have clearer and arguably more effective research strategies. However, there are also several negative aspects. For example, there is a lot of "game playing" that now goes on – a lot of emphasis on how best to present yourself.[9] There has also been increasing concentration of resources, although one can argue whether that is a good or a bad thing. What I am more concerned about, however, is the contradiction inherent in government policy. On the one hand, publicly funded researchers are encouraged to seek out their "users", to get close to them, and to help them address their research needs. Yet those needs tend to come in interdisciplinary form – to require research drawing on several disciplines as well as being of a more strategic or applied nature. Then, every four or five years those researchers are assessed in a Research Assessment Exercise which is carried out on a disciplinary basis and which gives more emphasis to basic and mainstream research rather than more applied or less conventional research.[10] So there is a contradiction between policies encouraging us to engage with users which

---

8 One obvious consequence of the current financial system is that many academics and departments inevitably give more emphasis to their research compared with improving the quality of their teaching.

9 There is probably even more of this game-playing in the Teaching Quality Assessments – in other words, these assessments have a lot of effect on how people present their teaching when subject to the visits by a group of peers, but if you look at the quality of the teaching actually delivered, I do not think that the assessments have had a lot of impact on the quality of the teaching received by students. Indeed, I would argue in the case of Teaching Quality Assessment that the costs especially in terms of people's time are probably greater than any benefits.

10 Certainly in SPRU where the research is intrinsically interdisciplinary, this is a continuous dilemma. We have to choose which pigeon-hole to screw ourselves up into every four or five years for the purpose of the Research Assessment Exercise. We could go in politics, in economics, in management, or perhaps even in sociology. We choose to go into politics (because it represents the least bad option) but it does mean that the great majority of our work which is not political science is being assessed by about six political scientists who obviously find some difficulty in ascertaining whether the majority of our work is of international excellence or not.

inevitably draws you into interdisciplinary research, and than being subject to an assessment system based on traditional disciplines in which interdisciplinary research and more applied research is not regarded so favourably.[11]

3 SPRU Study on the Assessment of Academic Research

In the early 1990s, SPRU carried out a study, the aims of which were to evaluate the approach adopted in the UK Research Assessment Exercise, to explore whether peer review might be complemented by performance indicators, to assess the feasibility of using bibliometric indicators for this purpose, and to investigate the potential of a range of other indicators. There were two main components of the study: the construction of a very large database on all publications and citations for UK university science over a 10-year period; and four casestudies in four different fields based on interviewing around 120 academics in some two dozen university departments. I won't go into detail on the first part where we concluded that it was feasible to construct bibliometric indicators at the level of departments but it is extremely labour intensive to clean up and unify all the addresses to the appropriate degree of accuracy at the level of the department.[12]

In the second part of the study, we looked at how well peer review works for assessing entire departments. As you will recall, peer review was first introduced several centuries ago for assessing papers submitted to journals. Later during the 20th century, it was applied to assessing proposals for grants. Now, we have a new application – to assess a whole department. How well does it work for that new task? From the interviews with 120 academics in four fields of science and engineering,[13] we found that the typical academic is familiar with research in between six and ten other British university departments. However, that knowledge is generally confined to their own subfield. For example, a solid-state physicist would know about solid-state physics in six to ten other departments. One must therefore ask whether a panel of about six peers can truly assess *all* the university departments in the UK and all the

---

[11] Recent stories in the *Times Higher Education Supplement* suggest that this may be one reason why some thought is now being given as to whether the Research Assessment Exercise should be continued.

[12] However, the second part of this study did raise a severe question mark as to whether the department is actually the most appropriate unit for this type of assessment or whether one should focus instead on subfield-based groups within departments.

[13] The social sciences and humanities were not included in this study. The findings obtained in this study for science and engineering should not necessarily be assumed to hold in social sciences or humanities.

research within them extending across all the subfields. Our conclusion was that it was somewhat unlikely that the panel would have direct knowledge of research in all subfields in all departments.

To take the example of physics again, a field where one can identify perhaps eight or ten subfields – low temperature physics, solid state physics, particle physics and so on. Even if you have particularly knowledgeable peers, each of whose knowledge extends to rather more than six to ten departments, it is unlikely that, between all six of them, they are going to have direct knowledge of the research in all subfields in all university physics departments in the country. Therefore in some cases, they will be ranking departments perhaps on the basis of extrapolation from the parts of the department that they are familiar with to the rest of that department, or attempting to get the information from these long complicated assessment submissions. For a department of forty researchers, for example, the submission will list 160 publications. The panel is most unlikely to read 160 publications, so they will probably look instead at the journals (or at the publishers of books) and they will come to some conclusion about the appropriate ranking. If it is borderline between two grades, then someone may be asked to read a sample of the published work from the department in question.

The results from the peer assessment that we conducted in our study agreed with those from the Research Assessment Exercise in about 90 % of the cases but they disagreed in the remaining 10 % by one or more grades. One possible explanation is that perhaps up to 10 % of the RAE rankings are wrong by one or more unit. This may be size related since there is some evidence that peer review is intrinsically biased in favour of larger (and hence more visible) departments; we found that the best correlations between the Research Assessment Exercise rankings and the various indicators we constructed were with indicators based on total output or total citations and not those size-adjusted indicators such as the output of publications per member of staff or per pound or the average number of citations per paper.

So what are the problems with peer review when applied to whole departments? We asked the 120 academics for their views and, as can be seen from Table 1, those that they identified included the following: a tendency for peers to rank more highly those departments and subfields they know well (an almost inevitable psychological affect); a concern that the field or unit of assessment is often too broad to be ranked by a small panel, familiar in each case with only their own subfields; a bias perhaps against small departments, perhaps stemming from definitions of the rankings;[14] a bias against departments specialising in non-mainstream subfields; and inadequate

[14] The definitions in 1992 centred on whether the work in a majority of subfields was of international or national quality. For a small department of perhaps a dozen researchers in which there are just two subfield-based groups, one of which is internationally excellent while the other is not, does that constitute a "majority" or not?

- tendency for peers to rank more highly departments and subfields they know well
- field/unit of assessment often too broad to be ranked by a small panel familiar with only some of the subfield components
- a bias against small departments, perhaps stemming from definitions of rankings
- problems in ranking departments with interdisciplinary interests not falling within a single field/unit of assessment
- a bias against departments specialising in non-mainstream subfields
- inadequate normalisation across fields with consequent adverse financial consequences for fields obtaining lower average rankings
- period between early RAEs too short – needs around 5 years rather than 3 to improve significantly
- absence of foreign peers from panels even though the definitions of the top rankings are based on international excellence

Table 1
Weaknesses in RAE Approach

normalisation across fields.[15] Another problem is that initially the period between RAEs was too short – it was three years (the first three exercises were in 1986, 1989 and 1992), then it became four years (the next was in 1996) and now it has become five years (the next is due in 2001). With a cycle of three years, if a department does poorly and the university decides to do something about it, it might take a year to recruit some good new people; it may take them a year or so to raise some research funds; then it will take at least another one or two years to produce some good published research outputs. By then, the next assessment has come and gone, and that department still has not done very well, so the people who have been hired become demoralised and they may go off to another university. Five years is probably a more sensible time-scale. Lastly, in the earlier exercises, there was an absence of foreign peers, despite the fact that the definitions of the top rankings are based on international excellence.

In the study, we also looked at a range of other possible indicators and asked academics whether they would like to see them used. The results are summarised in Table 2. As can be seen, an indicator based on research income was favoured by 66 % and opposed by 16 %. The main problem with such an indicator is the wide variation in costs across subfields. Take the example of physics again: some sub-

[15] There was no attempt to guide the panels as to what percentage of departments should be ranked as 5*, 5, 4 or so on. Some panels were more generous and gave out large numbers of 5s and 5*s, while other panels were tougher. These grades were then translated into financial resources with the result some fields suffered compared with others.

Research Income
Favoured by 66 % cf. 16 % opposed
But (i) wide variations in cost across subfields (ii) data incomplete

Publication indicators
72 % favoured cf. 8 % opposed
Most felt position of department based on publications about right
But problem of variation in importance of papers and in publication practices across subfields
Weight publications by importance of journals (as assessed by peer review)?

Citation indicators
Favoured by 66 % cf. 13 % opposed (NB Only scientists & engineers interviewed)
When shown positions based on citations, 60 % agreed with position and only 3 % expected to be a lot higher (or lower)
Worries about departments earning citations through "citation circles" (but no-one able to quote specific instance – modern legend?)
Problem of variation in citation rates across subfields

Esteem indicators
58 % favoured cf. 27 % opposed
Problems with (i) availability of reliable data, (ii) time lag, (iii) influence of 'non-scientific' factors on awards

Numbers of PhD students trained
74 % favoured cf. 17 % opposed
Reflect output of trained people (as opposed to scientific advances/knowledge)
But problems with (i) variation in quality (ii) influence of other factors (e.g. availability of studentships, general prestige of university, facilities of local city)

Table 2
Academics' Views on Different Research Performance Indicators

fields are very expensive, others less so. In addition, when we asked departments for such information, most of them did not have the data in an appropriate form so such an indicator might be difficult or time-consuming to operationalise.
Another possible indicator is one based on numbers of publications. 72 % of those interviewed favoured this being used. When academics were shown the position of their department in a table based on numbers of publications, most felt that the ranking was about right. However, they pointed to the problem of variations in the importance of papers and in publication practices across subfields.

Citations have been subject to much criticism. However, when we approached this in a symmetrical way, asking for each indicator what were the pros and cons, there was just as much enthusiasm, or just as little opposition, to this as with any of the other indicators – although one is likely to get a very different answer in social sciences and humanities. Of the interviewees, 66 % favoured it being used in university research assessment and only 13 % opposed it. There were worries about certain problems with citations such those citations earned through citation circles – "you cite me and I'll cite you". However, when we asked for direct evidence of this, nobody could provide any. (They might claim to know of someone in another country who was reportedly engaging in this but it always appeared to be more of a modern legend than something for which there was specific evidence.)

Esteem indicators – winning prizes, medals and so on – were favoured by a slightly smaller percentage and opposed by rather more. There are problems with the availability of reliable data (it is just not collected systematically in departments); there is often a long time-lag between the research and the recognition in the form of a prize; and there is the influence of non-scientific factors as well as scientific ones on the allocation of such prizes – whether you have given good service in the scientific community by editing a journal, organizing conferences and so on.

Another possible indicator is the number of doctoral students produced. It could be argued that this is a more important output from research, that trained people and the skills they embody are more beneficial than new knowledge per se. This indicator was favoured by 74 % of those questioned. Again, there are problems arising from the variation in quality of those students and from the influence of other factors; some universities may attract lots of students because of the general prestige of the university or the attractions of the local city.

We then asked the sample of 120 academics how they would like Research Assessment Exercises to be carried out – would they prefer peer review on its own, as was being done in most of the RAEs (with the exception of the 1992 one), or would they prefer some combination of peer review and performance indicators. No less than 96 % favoured some combination of the two. When they were asked whether more weight should be given to peer review or to performance indicators, the responses were fairly evenly divided with about one third (31 %) arguing that peer review and performance indicators should be given equal weight, similar numbers (33 %) saying that more weight should be given to peer review, and slightly fewer (28 %) saying that more weight should be given to performance indicators. However, what they all agreed on was one should use as wide a range of performance indicators as possible, endeavouring to develop a multi-dimensional profile of research performance (as is apparently done in the Netherlands) rather than trying to conflate everything on a single dimension as is done in the UK at present.

What were the conclusions to the SPRU study? The first was that evaluations are here to stay. As we heard in discussion at the conference, the need for public accountability in all areas where public spending is involved is inescapable, as scien-

tists themselves have come to recognise. As Stolte-Heiskanen found in a 1991 survey of Finnish academics, "the common assumption that academic scientists are against evaluations seems to be based more on a myth than a reality. A favourable attitude toward evaluations seems to be widespread especially among productive scientists."

Secondly, peer review must remain central in the assessment of university research. However, peer review complemented with performance indicators is arguably better than peer review on its own, at least in science and engineering. Furthermore, if one uses indicators, then it is better to employ a range of indicators rather than just one or two because that enables one to capture a wider range of aspects of research performance. It also means that it becomes harder to manipulate the system; if, say, four or five indicators are used, then to improve one's performance in relation to all of these will almost certainly require that one does better research – one is not able to "cheat" such a system. (As some of those interviewed remarked, there are similarities here between evaluation and the Heisenberg principle in that, once you start measuring a system, you influence or disturb it in a somewhat unpredictable manner.)

Another conclusion was that performance indicators designed for science and engineering should not be uncritically applied to social sciences and humanities. For example, although indicators based on publications and citations in journals scanned in the *Science Citation Index* may work reasonably for science and engineering, they work much less well for social sciences and arguably not at all for humanities.

The study also raised a fundamental question as to whether the university department is the right unit of analysis for such assessments. For example, when interviewees were asked whether it makes any difference to them as to whether they are based in a big department of say 30 or 50 researchers or in a small department of 15 or so,[16] they answered that the department is almost irrelevant here; what is important is whether a researcher in a given subfield has around him or her half a dozen researchers in that department working in the same subfield. If you do, then you can do world class research, whether you are embedded in a big department of 50 or a small department of 15.

This is particularly true in this age of cheap, fast, easy communication. It may have been different in earlier decades when, if you ran into a problem, say, with your equipment, you might wander down the corridor and find someone from another group who knew about equipment. Arguably in those days, department size did make more of a difference because you were more likely to find someone who could help you in a bigger department. However, these days in your subfield you

[16] At this time in British science policy, there was a strong belief that 20 academics was a critical size for a university department and that those below 20 were sub-critical in size.

will know who is the expert on that piece of equipment and you will email them; you do not need to wander the corridor trying to see if a colleague from another subfield group can help.

So the argument made by interviewees was that it is actually the subfield-based group that is the important unit for the purposes of research and its assessment. The department may be the appropriate unit for the organisation of teaching but not for research since you can do international quality research in a small department or in one where the groups around you are not of international or even national excellence. Yet if the latter is the case, your department will get a low ranking and you will not receive much money so you are penalised unfairly.

Lastly, one needs more research on the long-term effects of Research Assessment Exercises and whether, for example, they are discouraging interdisciplinary research and disadvantaging teaching, to pick up two of the questions which John Krebs mentioned earlier.

4 Conclusions

The first broad conclusion for the UK is that we have entered the phase of a revised social contract in which acute financial and political pressures have resulted in an emphasis on accountability and obtaining value for money and hence on evaluating performance and results. We are witnessing something similar in the United States with the Federal Government Performance and Results Act (GPRA). The UK, perhaps along with the Netherlands, may be at the forefront of the development of performance assessment in research, experimenting with bibliometric and other indicators and with the evolving approach adopted in the Research Assessment Exercise.

Secondly, the Research Assessment Exercise could be improved through combining peer review with a range of indicators of the type described above to yield multi-dimensional profiles for departments. On the other hand, that would entail much greater costs to do it in a more thorough way and, as in all evaluations, one needs to balance costs against benefits, to pick up one of the issues raised earlier in the conference. The benefits of an evaluation exercise must be greater than the costs – if the costs escalate too much, then you have to ask whether it is worth doing the evaluation at all.

In addition, one needs to determine the most appropriate unit of analysis for the Research Assessment Exercise. It may be that for research activities it is the subfield-based group rather than the department. If so, the costs of carrying out the evaluation will become considerably greater and that may again bring one up against the problem of the costs being in danger of exceeding the benefits. Lastly, as noted earlier, we need more research on the longer-term affects of assessments.

In the UK, an improved approach to evaluation is gradually evolving, with more peers including some users and some foreign peers. One could also complement peer review with a range of indicators to generate multi-dimensional profiles of research performance, ideally focusing perhaps on the group rather than the department. However, this is only worth doing if the benefits continue to outweigh the costs. Up to now, my personal assessment in relation to the British Research Assessment Exercise is that the benefits have been greater than the costs. However, if pressed, my conclusion about the Teaching Quality Assessment would be rather different.