Bibliometric Indicators and the Evaluation of British University Research

Ben R. Martin

1. Introduction

The evaluation of universities and university departments, at least in a formal systematic manner, is a comparatively recent phenomenon in Britain. It is true that, in his attitude surveys of academics in 1964 and 1976, Professor A.H. Halsey did include a question on the location of the best UK departments in the respondent's discipline (for a description, see Halsey, this issue). Apart from this, however, there was little by way of university evaluation, and there was certainly no British equivalent of the comprehensive rankings of academic departments carried out in the United States by Cartter (1966) and by Roose and Andersen (1970).

Yet the situation was to alter quite dramatically during the 1980s, as we shall see in the next section. This first examines the historical background to university research evaluation in the United Kingdom. It then summarises the range of studies by the Science Policy Research Unit (SPRU) on research evaluation more generally, including an analysis of the reasons why such assessments are increasingly needed. There is also a description of the overall approach to research evaluation adopted by the author and his colleagues.

The main part of the paper focuses on two studies, one carried out in 1986-87 and the other currently under way. The former involved an attempt to construct 'bibliometric profiles' of all British universities and polytechnics. After outlining the background to the study and its principal aims, we consider some illustrative results. The empirical findings are then used to explore various topical science policy issues. The section ends with a critical appraisal of the methodology employed in the study.

The second project is exploring the possibilities for constructing a range of academic research performance indicators. Because the study will not be completed until early 1992, the description concentrates on the aims of the research and the approach being employed, although there is also a short

summary of the work to date. Finally, the paper concludes by synthesizing some of the lessons to be drawn from experiences with university research evaluation in the UK, and speculating about the possible implications for its development in Germany.

2. Origins and development of university research evaluation in the UK

2.1 Historical background to university rankings

The election in 1979 of a Conservative Government led by Mrs Thatcher and committed to reducing (or at least constraining the growth of) public expenditure had far-reaching consequences for universities. Higher education was not immune in the Government's search for areas where savings could be made, with overall cuts of 15% being announced in 1981. In addition, concern with Britain's laggardly economic performance over previous decades encouraged an increasingly utilitarian attitude towards academic institutions. It was argued that universities, like other recipients of public spending, should demonstrate that they were making contributions to the economy and society commensurate with the investment in them - in other words, that they represented 'value for money'.

At the same time, for historical reasons many in the Conservative Party were less than sympathetic towards universities, recalling the protests of radical students a decade or so earlier as well as the role of various academics in propounding socialist policies for the economy and society. Matters were not helped by the outright opposition to the new Government's economic plans expressed in a letter signed by over 350 prominent economists, nor by the snubbing of the Prime Minister by Oxford University dons who opposed plans to award her an honorary doctorate. In short, the view of the Government, and one with which many in industry and elsewhere probably had some sympathy, was that universities needed to be subject to more thorough monitoring, evaluation and accountability.

Perhaps not surprisingly, this view initially met with little response from the universities, aside from a ritual restatement of the traditional position that intellectual freedom and academic autonomy are inviolable. However, in December 1982, the **Times Higher Education Supplement** conducted the first of its peer-review surveys of university departments in selected fields. These have been repeated at roughly yearly intervals since then, with the result that most disciplines have now been subject to scrutiny at least once. Initially, the surveys were regarded as a fairly harmless exercise, but later ones provoked

a sharply critical response from those fearful that the results might be used by the Government to impose further cuts (Phillimore, 1989, p.255).

In 1985, the Department of Education and Science published a Green Paper on **The Development of Higher Education into the 1990s**. This discussion document examined the three main types of output from universities - highly qualified manpower, research, and other social benefits - but noted that there was little or no statistical information on the second and third of these. Although recognising the problems involved in evaluating research outputs, it urged that more effort be devoted to this task (ibid., p.259).

That same year, the University Grants Committee (UGC, the body formerly responsible for distributing funds from the Department of Education and Science among universities) bowed to the inevitable and launched an exercise to rank all university 'cost-centres' (these being normally, but not always, equivalent to academic departments). Each cost-centre was invited to submit a two-page description of its research achievements and a list of five of the best publications over the last five years. Information was also compiled on research grants, studentships, income from contracts, prizes and other honorific awards. This information was then considered by UGC subject subcommittees, although how much weight they attached to it compared with their own subjective peer-review judgements of research performance, remains unknown (ibid., p.260).

The results of the ranking exercise, as soon as they were made public in the **Times Higher Education Supplement**, were immediately subject to widespread criticism. The attacks concentrated on three aspects, the first being the process by which the exercise was undertaken, with attention focusing on UGC's secrecy over exactly how the rankings were arrived at (including the lack of information on the precise criteria employed) and the absence of any appeals procedure. Secondly, numerous flaws with the ranking methodology were identified including a systematic bias against smaller departments, against those engaged in interdisciplinary research or work of a more applied orientation, and against newly established departments. The third type of criticism was a more general objection to the very notion that university research could be 'weighed and measured' - a notion that was seen as implying an overly utilitarian view of the role of such research (ibid., pp.261-62). This blanket opposition to university evaluation seems to have cut little ice with those outside the higher education sector, and in 1989 a new research assessment exercise was carried out by UGC and its successor, the Universities Funding Council (UFC). Although this included some methodological improvements (Jones & Sizer, this issue), the results were to prove just as controversial.

2.2 The development of science policy research in Britain

While the 1980s were a far from happy period for universities and scientific research in Britain, science policy research flourished during this time. (This is not as perverse as it might at first seem since, in many respects, science policy research is counter-cyclical. The worse the financial situation confronting science, the more important it is for funding agencies and policy-makers to ensure that the scarce resources which are available are used as effectively as possible. Hence, they may turn increasingly to science policy researchers for advice on how best to achieve that end. This might explain why science policy research grew more rapidly in Britain during the 1980s than in the Federal Republic of Germany, France and especially the United States where the funds available to scientists were not so constrained over this period.) At the University of Manchester, for example, there are now perhaps two dozen researchers working on science and technology policy, two thirds of them at PREST, the Centre for Policy Research in Engineering, Science and Technology. In London, the Royal Society together with the Fellowship of Engineering has set up a Science and Engineering Policy Studies Unit. There are also several smaller groups in universities around the country. However, by far the largest concentration of effort in this field is to be found in the Science Policy Research Unit at the University of Sussex.

Set up in 1966 shortly after the University was founded, SPRU grew gradually in size, reaching around 35 full-time researchers by 1980. Over the next ten years, the number of researchers almost doubled. In addition, the Unit became a university graduate school, and now has 70-80 masters and doctoral students, over half from overseas. This makes it perhaps the largest such centre of science and technology policy research in the world. The staff are divided into ten groups, one of which is the Science Policy and Research Evaluation (SPARE) Group.

2.3 Previous work by the SPARE group

As its name implies, the group is concerned primarily with policies relating to more basic science. (In this, it differs from the other SPRU groups where the focus is more on policies for technological development or for industrial innovation.) Most of the studies by the SPARE group have also involved an element of evaluation. Since the group began work in 1978, assessments have been conducted of the following:

(1) the scientific performance of 'big science' facilities (e.g. optical telescopes and particle accelerators);

(2) the technological 'spin-off' and training benefits from basic research (e.g. from radio astronomy);

(3) scientific performance of research groups working in 'small science' (e.g. condensed-matter physics and protein crystallography);

(4) the output and impact of applied research groups (e.g. in electronics, mechanical engineering and biomass research);

(5) the effectiveness of government R&D programmes or research support mechanisms (e.g. NTNF in Norway, the European Community steel R&D programme and the SERC Biotechnology Directorate in the UK);

(6) national scientific performance (e.g. the relative international standing of British science);

(7) the future scientific prospects for major new research facilities (e.g. accelerators under construction at CERN and elsewhere around the world);

(8) the methods used for looking into the future of science and identifying longer-term priorities for strategic research (e.g. studies in 1984 and 1989 on research 'foresight')

(9) research inputs (e.g. reports in 1986 and 1990 comparing government funding of academic and academically related research in six leading countries);

(10) the factors affecting research performance (e.g. in condensed-matter physics);

(11) research performance of universities and university departments (e.g. the two studies described below);

(12) the links between science and technology (e.g. a current project on the use of academic research by industry).

2.4 Why are research evaluations needed?

In view of the large amount of effort that has been devoted at SPRU and elsewhere to evaluation, the reader might be tempted to ask: Why do we need research assessments or performance indicators for academic science? Why can't we continue to rely on traditional peer review - in other words, leaving it to scientists to decide which research activities have been more or less successful in the past, and which new ones should be funded in the future?

There can be little doubt that for two or three decades after the Second World War the peer-review process worked very successfully in determining the distribution of resources among research areas. During this time, science was characterised by four main conditions: (a) substantial annual increases in funding (typically of 5-10% per annum in real terms); (b) a 'free market' for scientific ideas - in most specialties, there were large numbers of small groups competing for funds, and it was therefore relatively easy to find 'neutral' peers able to judge proposals objectively (i.e. scientists whose own funding prospects would not be significantly affected by the decision to support or not to support a particular project); (c) most research proposals fell within existing scientific disciplines; and (d) in deciding which projects to fund, much greater emphasis was given to 'internal' scientific criteria than to 'external' criteria such as the likely technological or economic impact of the proposed research.

Now, however, the 'boundary conditions' confronting science have changed considerably, posing serious challenges for the peer-review system (Irvine and Martin, 1984a, pp.71-78). First, the rapid growth rates witnessed in earlier decades have given way in many countries to approximately level budgets. As Ziman (1987; 1989) and others have argued, we are now faced with the phenomenon of 'steady-state science' where growth in one part of the system is only possible at the expense of contraction in another. Peers may find it relatively easy to determine which groups should be given additional funds (for example, to hire a new researcher or to purchase a piece of equipment), the main type of decision in more affluent times. However, the decision to cut back the level of existing support, jeopardising as it may the careers of professional colleagues, is of a qualitatively different type, and one which peers, for institutional or psychological reasons, find exceedingly difficult to take. Hence, the peer-review system is far less satisfactory when there are no new funds available.

Secondly, the policy of increased selectivity and concentration pursued by many funding agencies has resulted in research efforts in numerous areas

being focused on an ever smaller number of laboratories. The situation is most pronounced in 'big sciences' like high-energy physics where the researchers in each country may have access to only one or two central facilities. There has thus been a change from a situation approximating to the classical 'free market' (characterised by many small competing groups) to one of 'oligopoly' (Irvine and Martin, 1984a, p.76). Under such conditions, it is virtually impossible to find neutral peers because all researchers will have a direct interest in a given proposal being funded or rejected. (Either they will be users of the proposed new scientific facility, or they will be associated with a rival laboratory which is likely to benefit through more funds being available if the proposal is turned down.)

A third factor affecting peer review is the growing importance of cross-disciplinary research. Proposals for work in emerging interdisciplinary areas are often handled inadequately by review committees based on traditional scientific disciplines. For example, a proposal for a science policy research project may be judged less sympathetically by an economics committee or a political science committee than one of equal merit in the disciplinary mainstream. (At the very least, it is much less likely to have an 'advocate' on the committee to argue in favour of support.)

Fourthly, the 1980s have witnessed the growing importance of 'strategic' research - that is, basic research carried out in the reasonable expectation that in perhaps five, ten or fifteen years it will produce a broad base of knowledge likely to form the background to the solution of current or future practical problems (Irvine and Martin, 1984, p.4; Martin and Irvine, 1989, p.7). Strategic research thus falls somewhere between the traditional categories of pure (or curiosity-oriented) research and applied research. In determining whether to fund strategic research, consideration must be given not only to internal scientific criteria but also to external technological, economic or social criteria. In assessing the latter, one clearly cannot rely solely on the views of scientific peers.

Finally, in several countries governments have been demanding that there should be greater public accountability in relation to the distribution of resources to science. Peer-review has tended to operate in the past in a fairly secretive manner, or at least in a way such that its workings are not very transparent to 'non-peers' (i.e. to everyone except the scientists engaged in that particular specialty). It is therefore seen by some as more a mechanism for special pleading than one for demonstrating to the public that 'value for money' has been obtained from the resources invested.

Under these changed boundary conditions facing science, continued reliance on traditional peer review tends to have a number of adverse consequences (Irvine and Martin, 1984a, pp.71-78):

(1) It encourages the reproduction of past priorities - those groups or areas funded generously in the past, because they have become well represented on decision-making bodies, may continue to be supported generously irrespective of the scientific merits of their work. This gives rise to what the sociologist, Robert Merton, has termed the 'Matthew effect' ("Unto every one that hath shall be given, and he shall have abundance").

(2) It increases the degree of politicisation of decision-making, this being especially pronounced in 'big science'. In US high-energy physics, for example, the three main laboratories involved during the 1960s and '70s operated an informal policy of supporting each other's proposals in public (regardless of what they privately thought of them), taking it in turn to approach the government for funding. To the outside world, therefore, this appeared to be a field where the researchers were agreed on their priorities and one which therefore merited preferential support over areas where the scientists did not seem to have such a clear idea as to where they were heading. This system finally broke down when the costs of one project (to build the ISABELLE collider at Brookhaven Laboratory) escalated to such an extent that the future of the whole field was thrown into jeopardy. Only then did scientists at the other laboratories finally come out into the open to express their doubts (which many had harboured in private from an early stage), but by the time that the project was eventually cancelled around $200 million had already been spent.

(3) New areas of science are not picked up sufficiently promptly because there is no-one to argue their case on peer-review committees. Interdisciplinary subjects are particularly vulnerable, either being ignored because they fall between the 'cracks' of the existing committee structure, or being held back because they are subject to inter-committee wrangling as to which body should be responsible (biotechnology has suffered this particular fate in several countries).

(4) The likely economic or social benefits from strategic research are appraised inadequately or, at best, unsystematically by scientific peers, some of whom have little or no interest in such external spin-offs, while others may exaggerate their probable impact as a ploy to improve their funding prospects.

The rationale for conducting research assessments or constructing scientific performance indicators is to offset the effects of some of these problems. More specifically, the aim of research evaluation is to provide **accessible** information on scientific performance in a **systematic** form to **feed into** (but not supplant) the peer-review process, **opening it up** to enable non-peers to participate and rendering decision-making more **transparent** to the outside world.

2.5 The SPRU approach to research evaluation

Although the exact methodological details vary widely, the approach adopted in most SPRU evaluations exhibits a number of common characteristics:

(1) Its starting point is an input-output model of science - that is, it involves identifying and assessing the various inputs (funding, personnel and so on) and outputs, and then relating them to each other. This does not mean that other influences on performance (sociological, cultural or whatever) are ignored; rather the assumption here is that the effects of such influences cannot be investigated empirically in the absence of reliable data on inputs and outputs.

(2) It is institutionally focused. The unit of analysis is not the individual scientist, but the research group, department, laboratory or institution. The rationale for this is that modern science is essentially a social activity conducted largely by groups (the day of the individual scientist constructing all his or her own equipment, carrying out the experimental work and analysing the results entirely alone, is virtually over). Furthermore, most funding decisions, and certainly all the larger ones, focus on research groupings of one form or another rather than individuals.

(3) It is comparative. Because there are no absolute measures of scientific output (it cannot, for example, be added up and 'weighed' in a single unit such as dollars), the only approach is to compare the outputs from a number of research groups. Here, an important proviso is that one can only legitimately compare 'like' with 'like' - in other words, groups working on similar problems in the same specialty, supported with broadly equivalent levels of resources, publishing in the same set of international journals (where they are subject to the same refereeing procedures and engage in similar referencing behaviour), attending the same scientific conferences and so on.

(4) Again because there are no perfect measures of scientific output but only a number of imperfect or partial indicators, the approach involves the combined use of several such partial indicators, each reflecting different facets of research performance (e.g. total output, output per unit input, overall impact, number of 'discoveries' or major advances), but each also subject to other institutional, social, political and psychological factors (see Martin and Irvine, 1983, for further details).

Nevertheless, despite these common characteristics, it is vital to recognise that, for different types of research, one needs different assessment approaches and indicators. The starting point of any research evaluation consists of two related questions: What is the primary form of output for this research? And who is the main audience or 'customer' for that output? From the answers to these questions, one must then devise an approach and, if possible, a set of indicators which can be used to assess that output and its impact on the target audience.

For more basic research, the primary output consists of contributions to scientific knowledge, and these are generally written up and published in the form of articles in learned journals or books. The intended audience is normally other researchers, often those working in the same specialty. Hence, indicators based on numbers of publications may provide a reasonable assessment of the comparative output of groups working in a given area, while the number of times that those publications are referred to or cited by fellow scientists should give some indication as to which groups have had most impact on the research community. To take one example, in our evaluation of the scientific performance of the particle accelerators at CERN, the European laboratory at Geneva, compared with that for accelerators at various other high-energy physics centres around the world, we used a combination of two approaches: (a) various bibliometric indicators (numbers of journal articles, citation totals, numbers of highly cited papers etc.); and (b) peer-rankings based on the results of structured interviews with 200 particle physicists from a dozen countries in the East and West who were asked to rank accelerators in terms of major 'discoveries', on the one hand, and incremental advances (e.g. better statistics on known particles or their properties), on the other. It was found that the rankings in terms of discoveries were consistent with the data on highly cited papers, while the rankings in terms of incremental advances were in agreement with the citation totals (Martin and Irvine, 1984).

For applied research, by contrast, the primary output consists of contributions to technology, industry, society, welfare or health, defence and

so on. The output can take a much wider variety of forms - for example, a new product or process, an influence on social or economic policy, improved health care or a new weapons system. The nature of the intended audience or 'customer' for that product also varies considerably. Consequently, there is a wide range of possible methodological approaches and output or impact indicators. In a study of mechanical engineering in Norway, for example, bibliometric indicators proved almost useless since researchers in this area see their primary output as a new device or process rather than an advance in scientific knowledge. Hence, little emphasis is attached to preparing journal articles. Moreover, when researchers do write up the results in papers, these tend to be in Norwegian and to appear in trade or professional journals rather than academic ones. They are therefore hardly ever referred to by researchers outside Norway, with the result that they receive no citations from the international scientific community. In the light of this, the evaluation approach instead involved a combination of (a) peer review (in which researchers in the area were asked to compare the performance of various groups), and (b) 'customer review' whereby some 40 mechanical engineering firms were approached for their views on the quality and utility of the work by different groups. By and large, it was found that the results from peer review and from customer review converged, although not in a few individual cases (for details, see Schwarz et al., 1982).

3. Study to develop bibliometric profiles of UK academic institutions

3.1 Aims and approach

Earlier SPRU evaluation studies of 'big science' facilities proved relatively labour-intensive, not least because of the effort involved in exact citation matching - that is, in looking up every publication individually in the **Science Citation Index**(SCI) and counting the citations received in each year. (In the case of the CERN study referred to earlier, 10,000 papers had to be looked up in the **SCI** in the year of publication and in each of the next ten years. This took approximately eight person-months of effort.) Hence, one of the objectives of the project described here was to explore a 'short cut' to full citation analysis whereby, rather than weighting publications by the actual number of citations that they earned, they were instead weighted by the average influence of the journal in which they appeared. Thus, articles in **Nature**, for example, were scored more highly than those appearing in low status journals. There are two main ways to do this. Either one can use the journal impact factors calculated by the Institute for Scientific Information

(ISI, the producers of the SCI), these being equivalent to the average number of citations per paper for each journal. However, the impact factors vary considerably between fields like molecular biology (where each paper on average contains perhaps 30 references and where the journal impact factors are correspondingly high), and engineering or mathematics (where the average number of references per paper is typically only half a dozen or so). The alternative developed by CHI Research (a US consultancy company) is to normalise the journal impact factors across fields (essentially by dividing by the average number of references contained in each paper) in order to arrive at a 'journal influence weight'.

As was noted above, in 1985 the UGC began its controversial ranking of the research performance of all the departments (or 'cost-centres') in British universities. A year later, the UK Advisory Board for the Research Councils (ABRC, the body which advises the Secretary of State for Education and Science on the distribution of the 'Science Budget' between the five Research Councils) commissioned an experimental study by SPRU to compile bibliometric profiles of all British universities, polytechnics and research council institutes. The aim was to provide an overview of research performance across the entire academic sector.

The study offered an opportunity to test the influence-weight methodology developed by CHI Research. In principle, this offers three advantages over exact citation counting: (a) with citation counting, it is necessary to wait two or three years to establish whether papers are going to be highly cited or not, whereas no such time-lag is involved in calculating the total influence for the papers produced by different groups; (b) exact citation matching is rather labourintensive and/or expensive, while the calculation of influence statistics can be largely computerised once the publication data (especially the institutional addresses) have been 'cleaned up'; and (c) influence indicators are less subject to the variations in referencing practices across fields described above. However, offsetting these advantages was the recognition that the influence-weight approach is an approximate one which is suitable only for larger groups or sets of papers. One of the aims of the project was to establish exactly how large a number of publications is needed before the results can be regarded as reliable. In addition, the study provided a chance to explore current science policy issues with the aid of systematic empirical evidence.

Although details of the methodology will not be given here (for these, see Carpenter et al., 1988), the approach involved identifying all the science and engineering papers published by each of the UK institutions during 1983-84 in the 3000 or so journals scanned by the SCI. Next, each paper was clas-

sified into one of eight fields and around 100 subfields on the basis of the journal of publication (rather than the departmental address). This was done using the journal categorisation scheme developed by CHI Research. Then, for each institution and each field or subfield, the following indicators were calculated:

(1) total influence =

(number of papers in journal i) x (influence weight of journal i)

(2) average influence per paper

(3) influence score = number of standard deviations the average influence per paper for that institution is above or below the mean for UK institutions

(4) average research level (calculated using the CHI categorisation of journals on a 4-point scale from applied to basic)

3.2 Illustrative results

A specimen set of results for University A is shown in Table 1. In the case of biology (one of the eight main fields), it can be seen from the first column that this university published 46.5 papers in biology journals in 1983-84 (after fractionating collaborative articles). The figure in the third column shows that those represented 9.7% of the institution's total published output. For all UK universities, biology papers accounted for 12.7% of the total output. If we define an 'activity index' as the percentage of University A's papers in a given field divided by the percentage of all UK academic papers in that field, this gives a value for biology of 0.8 (i.e. 9.7/12.7), as can be seen from the second column. In other words, University A published fewer biology papers in relation to its total output than the average for Britain. The fourth column shows that University A's 46.5 papers represented 1.5% of the total UK effort in biology, compared with a figure of 2.0% for all science and engineering fields combined.

Table 1
Specimen bibliometric profile for University A, 1983-84

Subject	Number of papers	Activity index	% internal effort	% total UK effort	% papers with influence	Average influence	Influence score	Average research level
Biology	46.5	0.8	9.7	1.5	85.8	23.6	2.6	3.6
Biomedical research	89.8	1.2	18.8	2.5	93.9	63.3	2.7	3.9
Chemistry	122.4	1.3	25.6	2.5	98.1	26.9	2.2	3.9
Clinical medicine	21.3	0.5	4.4	0.9	78.8	16.0	0.3	2.9
Earth & space science	40.4	1.1	8.4	2.1	97.3	26.1	0.2	4.0
Engineering	26.0	0.5	5.4	1.0	76.1	5.9	-1.1	2.0
Mathematics	11.0	0.5	2.3	1.1	95.5	4.0	-1.2	3.3
Physics	121.3	1.4	25.3	2.8	93.8	21.4	0.9	3.7
All fields combined	478.7	1.0	100.0	2.0	92.8	30.1	5.1	3.7

The fifth column of Table 1 relates to the fact that, for a few comparatively small or new journals, CHI Research was unable to calculate a meaningful 'influence-weight' figure. However, 85.8% of University A's biology papers were in journals for which an influence weight was estimated, and these had an average influence per paper of 23.6. This compares favourably with the figure of 17.5 for all biology papers produced by UK academic institutions. University A's average influence corresponds to 2.6 standard deviations above the UK mean, so it has an 'influence score' of +2.6 (see seventh column). The final column, 'average research level', describes how applied or basic are the journals in which University A published. CHI Research has classified the 3000 journals scanned by SCI into four categories, with level 1 corresponding to journals of an applied technological nature, and level 4 representing very basic scientific literature. A research level of 3.6 means that most of the biological research by University A is relatively basic while, as one might expect, research in the field of engineering, for example, is more applied.

3.3 Relevance of results to current policy issues

A major objective of this experimental study was to see if the bibliometric profile data could cast any light on topical science policy issues. The first such issue to be considered is that of selectivity and concentration. Over recent years, many in Britain (and indeed elsewhere) have argued that, because the nation's scientific resources are limited, a policy of increased selectivity and concentration is required. First, however, it would seem desirable to know the degree of concentration of effort that we already have.

Table 2 shows how publications and influence were distributed across the UK university sector in 1983-84. For example, the top five institutions (i.e. the top decile) published 25.7% of all papers in biology journals and the top 12 (the top quartile) 48.8%. Conversely, the bottom quartile and decile produced only 4.6% and 1.1% of publications respectively (universities without departments of biology have been excluded here). The figures in the second column reveal that, in terms of total influence, there is a slightly higher degree of concentration in the leading universities, with the top five, for example, accounting for 26.8% of influence compared with 25.7% of papers. As one moves across the table through the progressively more capital-intensive areas of chemistry, engineering and physics, one finds the leading institutions in each field obtaining an increasing share of publications and influence. In physics, where the costs of carrying out frontier research are generally greatest and 'critical mass' effects might therefore be expected to be most pronounced, the top five institutions published no

Table 2
Distribution of papers and influence across UK university sector, 1983-84

Universities		Biology (n = 50)		Chemistry (n = 54)		Engineering (n = 48)		Physics (n = 54)	
		% of papers	% of influence	% of papers	% of influence	% of papers	% of influence	% of papers	% of influence
Top	10%	25.7	26.8	26.8	28.8	31.2	37.9	38.2	43.1
	25%	48.8	49.0	48.4	51.5	53.8	58.4	57.8	62.6
	50%	77.2	77.4	75.8	77.8	79.0	82.2	79.9	81.0
	75%	95.4	96.4	91.7	92.8	93.0	94.4	93.4	94.2
	90%	98.9	99.0	97.7	98.1	98.1	98.4	98.4	98.5
	100%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

less than 38.2% of papers which accounted for 43.1% of influence, while for the bottom 27 universities (i.e. the bottom 50%) with physics departments the corresponding figures were only about half this (20.1% and 19.0% respectively). In short, for some scientific subjects there is already quite a heavy degree of concentration of research efforts in the United Kingdom.

A second policy question upon which these bibliometric data might shed some illumination is the nature of the UK higher education system and in particular the relative research performance of polytechnics compared with universities. In Britain, there has traditionally been a binary or bipolar higher education system. On the one hand, universities are expected to be active in teaching and research across the broad range of arts and science subjects (with the exception of a few specialised institutions). In recognition of this, university faculty are paid both to teach and to conduct research (typically dividing their time in the ratio 60% to 40%). On the other side of the higher education sector are the polytechnics which have been treated as primarily teaching institutions. Although their staff are free to apply to Research Councils and other agencies for project grants, they are not paid to do research. Yet over time, many polytechnic departments have become quite active in research. This has given rise to much debate, with some advocating that the polytechnic and university sectors should be merged, coming under the same funding body and with identical financial conditions for the employment of staff. Others, however, have pointed to the dangers of such a centralised unitary system, and have instead argued in favour of a transition towards a much more differentiated higher education sector along the lines of that to be found in the United States with its wide spectrum of institutions ranging from research universities to four-year teaching colleges.

From the figures in Table 3, it is clear that, for the two fields shown, any gap between universities and polytechnics in relation to their research efforts had disappeared by 1983-84. In the case of chemistry, there is little difference in terms of numbers of publications and influence between the bottom five universities and the top five polytechnics. For biology, the overlap is more pronounced with the output of the leading polytechnics (including equivalent Scottish 'central institutions') being appreciably better than that of the bottom five universities with biology departments. Thus, although the research profiles of polytechnics are still on average markedly inferior to those of universities, these figures suggest that the binary nature of the UK higher education sector has given way to a continuous spectrum in relation to research output.

Table 3
Publication and influence data for leading universities and polytechnics and for the lowest placed universities, 1983-84

University/ Polytechnic	Rank	Biology		University/ Polytechnic	Rank	Chemistry	
		No of papers	Total influence			No of papers	Total influence
University	1	160.2	1800	University	1	330.5	8316
	2	150.9	2869		2	325.8	8309
	3	147.3	3165		3	227.2	4777
	4	136.0	2087		4	214.2	4929
	5	130.2	1620		5	165.5	3398
	46	9.7	131		50	27.7	601
	47	8.9	122		51	22.4	473
	48	5.6	55		52	21.8	193
	49	4.1	60		53	19.5	308
	50	3.5	58		54	18.3	414
Polytechnic	1	31.5	372	Polytechnic	1	27.0	591
	2	27.4	271		2	26.6	474
	3	25.3	216		3	22.7	41[a]
	4	16.5	236		4	16.4	100
	5	15.3	144		5	12.0	250

[a] Only 26% of the chemistry papers from the third placed polytechnic were published in journals for which CHI has calculated an influence weight.

The third policy question concerns the reliability of the results from the UGC exercise in 1985-86 to rank all university department on a four-point scale ('outstanding', 'above average', 'average' and 'below average'). As we saw earlier, that exercise was subject to widespread methodological criticism (for example, the uncertainty over whether the notion of 'average' referred to a British or a world average). One particular ambiguity was whether departments had been assessed in terms of their total research output, or whether the rankings were based on a size-adjusted notion of output (i.e. their 'productivity' or output per unit input). At least some of the subject committees responsible for producing the rankings claimed that they had taken size into account but certain critics argued that the methodology was seriously biased in favour of larger departments (e.g. Gillett, 1987).

Before considering how the results of our study compare with the UGC rankings, we should point out two problems in attempting to draw such comparisons. First, as mentioned above, the breakdown by field for the profile data is based on journal subject classification rather than the departmental affiliation of authors. Secondly, many of the UGC 'cost centres' correspond only loosely with the CHI field categories. Nonetheless, two fields where there is a reasonably good correspondence are chemistry and physics.

Table 4 gives research profile data for the 54 UK universities with physics departments, together with their respective UGC ratings. Four bibliometric indicators are shown, two of which (number of papers and influence) relate to total output while the other two (average influence per paper and influence score) represent size-adjusted measures of performance. In the table, universities have been ranked in terms of the total influence of their physics publications. It can be seen that the four institutions with the greatest influence were the only four to be judged 'outstanding' by UGC, while four of the next five were classified as 'above average'. At the other extreme, 10 of the bottom 11 universities with least influence received a 'below average' rating. Overall, there is a correlation of 0.63 between total influence and UGC ranking, while that between numbers of papers and UGC rating is very similar (0.65).

Table 4
Comparison of research profile data with UGC rankings for physics

University	Number of papers	Average influence	Influence score	Influence[a]	UGC ranking[b]
1	552.4	22.9	3.1	11,739	4
2	469.2	23.3	3.5	10,251	4
3	296.8	20.1	0.6	5,697	4
4	144.7	18.7	-0.1	2,679	4
5	165.2	15.5	-1.5	2,499	3
6	121.3	21.4	0.9	2,435	2
7	120.3	20.7	0.6	2,254	3
8	105.3	21.6	0.9	2 086	3
9	100.4	20.4	0.5	1,993	3
10	99.2	17.2	-0.6	1,664	2
11	95.2	17.2	-0.6	1 637	2
12	65.0	24.5	1.5	1,519	1
13	44.3	34.1	3.4	1,508	2
14	94.8	16.7	-0.7	1,466	2
15	72.6	21.0	0.5	1,426	2
16	98.2	14.4	-1.5	1,342	3
17	75.2	17.0	-0.6	1,278	N/A
18	62.4	21.6	0.6	1,125	2
19	76.1	14.8	-1.2	1,088	2
20	81.1	15.2	-1.1	1,081	1
21	69.7	15.7	-0.9	982	3
22	62.4	15.5	-0.9	947	1
23	66.1	17.3	-0.4	943	3
24	52.9	19.8	0.2	908	1
25	44.8	21.2	0.5	860	1
26	39.1	23.1	0.8	857	2
27	68.3	15.2	-1.0	842	3
28	77.3	11.8	-2.1	837	1
29	60.7	14.0	-1.3	829	N/A
30	42.1	17.9	-0.2	718	3
31	49.5	17.5	-0.3	718	2
32	36.4	19.4	0.1	706	2
33	35.2	19.4	0.1	673	3
34	53.5	14.1	-1.1	670	1
35	51.3	14.9	-0.9	650	2
36	30.3	21.1	0.4	637	1
37	38.8	17.3	-0.3	591	N/A
38	41.7	13.5	-1.2	522	1

(continued on next page)

Table 4
(continued from previous page)

University	Number of papers	Average influence	Influence score	Influence[a]	UGC ranking[b]
39	30.5	17.2	-0.3	521	3
40	30.8	16.4	-0.5	505	1
41	37.7	13.4	-1.1	496	2
42	30.8	16.4	-0.5	494	1
43	28.0	16.7	-0.4	447	1
44	21.7	20.7	0.3	445	1
45	31.5	13.5	-1.0	412	1
46	54.2	8.7	-2.6	408	2
47	11.0	28.4	1.0	312	1
48	28.9	10.5	-1.5	288	1
49	20.4	14.2	-0.7	287	1
50	19.0	16.9	-0.3	270	1
51	17.3	15.8	-0.4	262	N/A
52	13.1	16.4	-0.3	215	1
53	19.4	10.4	-1.0	113	1
54	9.6	9.4	-1.0	90	N/A

[a] Universities ranked in terms of total influence.
[b] UGC rankings: 4=outstanding 3=above average
2=average 1=below average
N/A=not available

In contrast, the correlations between the UGC rankings and the two size-independent indicators of average influence per paper and influence score are much smaller (0.22 and 0.34 respectively). Thus, it would seem that the UGC assessments were more influenced by the total output of physics departments than by their 'productivity'. This points to a further problem with peer-review judgements of university departments - namely, that big groups are more 'visible' (because a peer is more likely to know of at least one leading researcher or one piece of outstanding work by the group). It therefore tends to result in a systematic bias in favour of larger departments. Hence from the UGC results, it is impossible to be certain whether the top-ranked departments are better because they are bigger, or only appear to be better because they are more visible.

The bibliometric statistics also cast doubt on the rankings of certain individual institutions. For example, Table 4 shows that Universities 39 and 46 received higher rankings than their publication and influence records might indicate, while Universities 6 and 12 seem to have been comparatively harshly treated. This raises the question of whether these institutions would have been given the same ranking if such data had been available to the subject committees. In some cases, there may have been particular 'extenuating circumstances' known of by the peers which would have led them to set the bibliometric figures on one side; but in others one suspects that the peers might well have reconsidered their rankings of a few departments and made adjustments to bring them closer into alignment with the quantitative evidence.

3.4 Assessment of the influence methodology

The study reported here was essentially an experiment to investigate the validity and potential policy utility of bibliometric research profile data. As we have seen, it succeeded in generating empirical results of direct relevance to important science policy issues. Nevertheless, given that this was a pioneering exercise, it was perhaps inevitable that a number of methodological problems were encountered.

First, the classification of some journals into the various CHI subfields was criticised by researchers during the process of validation. The categorisation scheme therefore needs to be checked, updated and fully validated if it is to be used with greater confidence in the future.

Secondly, in certain cases the influence weights of journals appeared not to be in accord with researchers' perceptions as to which are the more important journals. Again, further validation is required here.

A third problem is that the influence indicators were found to be significant only when considering institutions producing more than 30-40 papers in a given area. Hence, they are generally only useful at the field rather than the subfield level.

Fourthly, the breakdown into fields (and subfields) is, as we have seen, based on the journals in which authors publish rather than the departments in which they work. Consequently, the results for 'chemistry', for example, do not necessarily relate entirely to a university's chemistry department. A few of the papers in chemistry journals may have been written by, say, members of a physics department, while some staff in the chemistry department may

publish in journals classified by CHI as 'chemical engineering'. As such, the institutional profile data are less useful for policy purposes than if the classification solely reflected departmental affiliation.

Lastly, if research profile data were to be adopted for routine use in policy-making, this might very well affect scientists' choices as to the journals in which they elected to publish their research results. Some institutions might pursue a strategy of publishing wherever possible in journals with the highest influence, while others might, for reasons of loyalty perhaps, refuse to alter their publication habits. It would then require considerable efforts by peer-review committees to allow for such 'manipulation' of the bibliometric profile data. (In fairness, it should be pointed out that traditional peer review conducted in the absence of research output indicators is just as likely to be subject to 'manipulation' by certain peers - perhaps even more so when it is conducted behind closed doors.)

Because of these various methodological difficulties, in the new study currently under way it was decided to revert to full citation counting (which can now be completely computerised, making it rather less labour-intensive). In addition, publications are being sorted by departmental address, even though this involves a very large amount of 'cleaning up' of the data.

4. Current SPRU study on academic research performance indicators

4.1 Aims and approach

In 1989, the Science Policy Research Unit embarked upon a new three-year project to explore the possibilities for constructing performance indicators for academic scientific research. (It is important to stress that the work involves research on, and not the operationalisation of, such indicators. The SPRU view is that several more years of research and development are required before these indicators may be safely used for policy purposes.) The study is funded jointly by the Advisory Board for the Research Councils and the Economic and Social Research Council. It is being conducted by SPRU in collaboration with the Research School of Social Sciences at the Australian National University. Discussions are also under way about the possibility of developing collaborative links with researchers in Finland, the Federal Republic of Germany, France and the Netherlands.

The SPRU study has five main objectives:

(1) to develop input, output and performance indicators for scientific and engineering research in British universities and polytechnics, and in their departments or cost-centres;

(2) to arrive at a better theoretical understanding of the significance and validity of different research performance indicators;

(3) to study institutional determinants of research performance;

(4) to identify policy implications at the national and institutional level;

(5) to examine more specific implications regarding the exploitability of academic research, its industrial relevance and collaboration with companies.

The project can be divided into three main stages. In the first, the emphasis is on the exploration of existing data-bases. As regards the inputs to British academic research, use is being made of the figures compiled by the Universities Statistical Record (USR), the Polytechnics Finance Officers Group and by SPRU in earlier studies of government funding of academic research. Here, the central problem is ascertaining what proportion of UGC/UFC resources should be attributed to research. Until around 1986, it was traditionally assumed that the figure was approximately 30%. However, re-examination of the time-budget data from which that figure was derived (see Martin and Irvine, 1986; Irvine et al., 1990), together with the results of a study by Clayton (1987), suggests that the actual figure is probably closer to 40%.

In looking at the outputs from academic research, we are focusing initially on the **Science Citation Index** data-base, with ten years of publication and citation statistics (covering the period 1981-90) having been purchased (together with equivalent information for the **Social Sciences Citation Index** and the **Arts and Humanities Citation Index**). For these output data, the most difficult task is to link all the papers to individual departments, which involves 'cleaning up' the institutional addresses given by authors (there can be many different variants for the same departmental or institutional address which all have to be merged on the computer). Once this task has been completed, it will then be possible to construct indicators based on publications, total citations, average citations per paper, numbers of highly cited papers, average number of publications per researcher and so on for

each department. Besides the **Science Citation Index**, the study is also investigating other output-related data bases (e.g. abstract services such as **Chemical Abstracts**, and patent data-bases) and findings from previous reputational surveys of UK universities (such as those conducted by Professor Halsey, the **Times Higher Education Supplement** and the UGC/UFC).

The second phase of the project will consist of in-depth case studies of perhaps four fields and a sample of higher education institutions (including a few from overseas to see where British universities stand in international terms). The case studies will seek more detailed statistics on:

(1) the inputs to academic research - funds, staff (including estimates of the percentage of time devoted by faculty to research) and instrumentation;

(2) outputs - (a) departmental publication lists (to check the comprehensiveness of the **Science Citation Index**), (b) recognition-based indicators (prizes, journal editorships, officers of learned societies etc.), (c) other research-related indicators (e.g. numbers of doctoral degrees awarded) and (d) peer-ranking data (where academics will be asked to rank the performance of other groups in their field).

The views of researchers and university officials will also be sought on the strengths and limitations of different indicators, and on the determinants of successful research performance.

At least one of the four fields chosen for more detailed study will be a more applied area. For this, additional information will be sought on the following:

(1) the extent and nature of industrial linkages (e.g. commissioned research, consultancy);

(2) the industrial relevance and potential exploitability of the research conducted (including any information available on patents, licensing, royalties etc.);

(3) the views of researchers and others on (a) the utility of data on patents, contract income and the like as indicators of performance in applied academic research and (b) factors encouraging effective applied research and its exploitation;

(4) 'customer review' - that is, the views of potential users of the research results (e.g. industrial R&D managers) on the comparative performance of different university departments and the 'marketability' of their students.

The third and final phase will involve processing all the interview and other case-study material, and completion of the analysis of the input, output and performance indicators. A preliminary version of the conclusions will be drafted and circulated for comment, criticism and validation. The final results will then be disseminated through publications, presentations at seminars or conferences, and very importantly through the training of research students and perhaps of agency officials on secondment to SPRU.

4.2 Likely outputs

The probable outputs from the project fall under three headings. The first will consist of theoretical contributions, helping to advance our understanding of two sets of questions:

(1) What constitutes and determines 'successful' research performance? What is the relationship of this to theories of scientific and technological progress?

(2) What are the most useful types of research performance indicator? What aspects of performance (and what higher education functions) does each relate to?

The second output will take the form of contributions to methodologies for evaluating university research performance. Here, the questions to be addressed include the following:

(1) How can one best apply existing indicators (e.g. publications and citations) to the assessment of academic departments?

(2) What other indicators (e.g. recognition-based ones) could be operationalised without undue effort or cost?

(3) What is the most effective way of employing performance indicators on a continuous and routine basis in the future?

The third and perhaps most important output will be to provide empirical evidence on a number of important science policy issues:

(1) Is there a 'critical mass' effect in academic research? What level of effort is needed for (a) conducting frontier research, and (b) pursuing a 'watching brief' role (i.e. maintaining a small amount of research and closely monitoring overseas advances while retaining the option of quickly building up the level of effort in the field to that needed for frontier research if a new scientific discovery or a sudden change in the demand for the results of such research makes this desirable)? How do these vary across fields?

(2) Is there evidence for dis-economies as well as economies of scale? What are the implications for policies aimed at increased selectivity and concentration as opposed to those emphasising pluralism and dispersion?

(3) Should a policy of selectivity and concentration focus on university departments (i.e. choosing the best departments in the country in a given field regardless of the institution in which they are located) or on entire universities? (In Britain, the former option was favoured by UGC for several years, while ABRC advocated the latter.)

(4) Is the relationship between teaching and research truly an essential symbiosis, as many academics maintain (largely in the absence of any reliable evidence up till now)? Does research excellence depend upon access to first-rate students (and vice-versa)? Should Britain follow the United States in encouraging the development of a more differentiated higher-education sector with some universities specialising only in teaching?

(5) Is there an optimum size and form of organisation for academic research teams? How does this vary across fields? What are the implications for Britain's newly created Interdisciplinary Research Centres?

(6) What are the advantages and disadvantages of different research support mechanisms - of UGC/UFC general institutional support versus research council project grants versus longer-term or rolling programme grants versus selected centres of excellence?

(7) How crucial is access to 'state of the art' instrumentation for conducting pioneering research, and again how does this vary across fields?

(8) What are the factors promoting and inhibiting the 'exploitability' and industrial relevance of academic research?

These policy questions are extremely wide-ranging and together comprise a research agenda for all those intending to carry out work in the area of policies for university science. Clearly, this project will not resolve all these difficult issues, but at the very least it should begin to furnish some systematic information in an area where this has previously been rather lacking.

4.3 Early results

Although the project is still in its early stages, some progress has been made in relation to at least one of policy issues listed above. During the 1980s, much science policy in the UK has been predicated on the seemingly commonsense notion that there are economies of scale in research - in other words, that 'bigger is better'. For example, a review of the future of earth sciences took certain US data (very selectively and not very accurately - see Hicks and Skea, 1989, p.31) and argued that a minimum size for university departments in this field is around 20 faculty. This threshold figure of 20 scientists was subsequently taken up and repeated in similar reviews of physics and chemistry.

Given the policy importance of the issue, two colleagues at SPRU, Diana Hicks and Jim Skea, decided to examine the evidence for economies of scale. They looked first at the available literature. They found firstly that there had been surprisingly little published on this topic, and secondly that the few results which had been reported were contradictory or, at best, very ambiguous.

Their next step was to carry out an analysis of the relationship between research output statistics (numbers of publications listed in the SCI) and an input variable, the number of scientists. If there are indeed economies of scale, then a plot of 'productivity' - that is, numbers of papers per person - against size (in terms of numbers of researchers) should yield a line with a significant upward slope. Furthermore, if scale effects do exist, they might be expected to feature most prominently in relatively expensive areas of research, so attention is focussed on physics and chemistry. The results on the variation of productivity with university department size for physics are shown in Figure 1, with the continuous line representing the best-fit regression and the broken lines delineating the 95% confidence level of the regression. It can be seen that there is a wide spread of points and that the

two variables are only very weakly correlated (r squared = 0.08). True, the best-fit regression line does have a small positive slope of 0.0041 but this is only slightly greater than the standard error of 0.0020 (ibid., p.33).

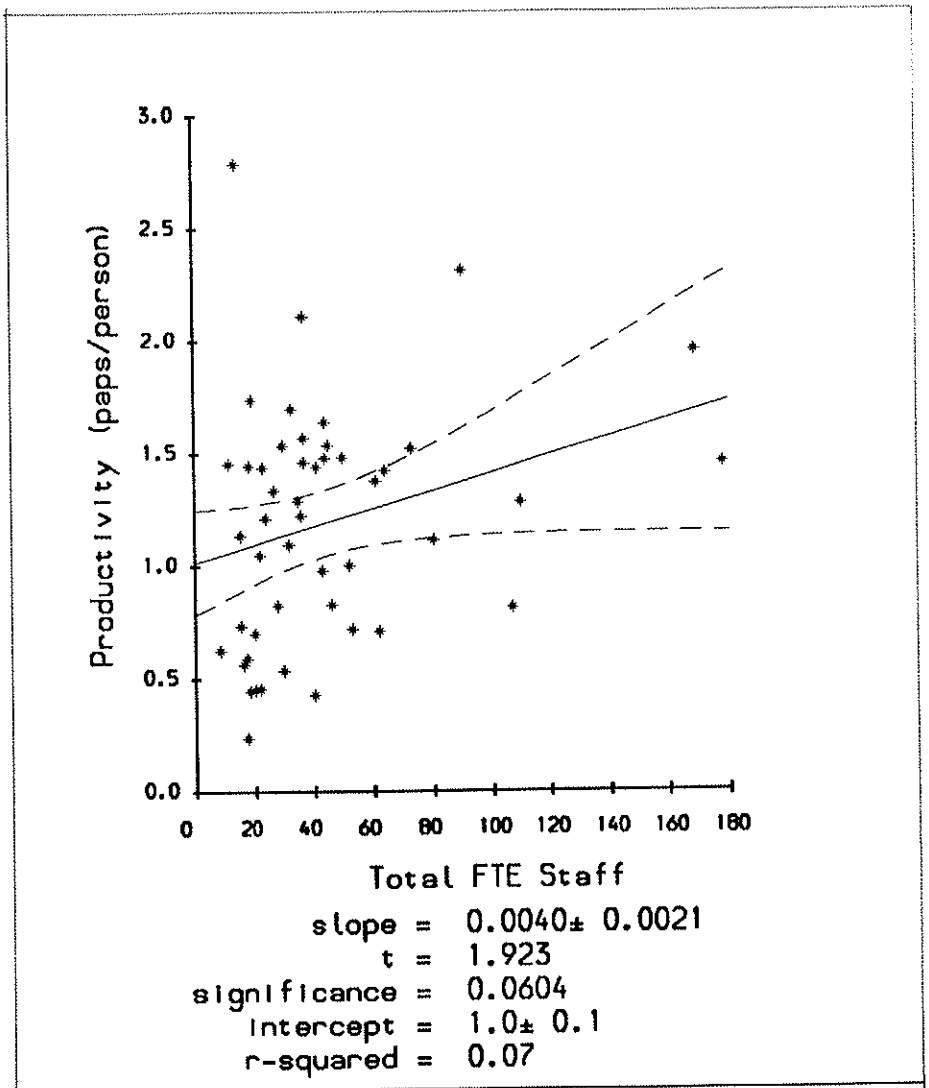

Figure 1: Variation of Productivity with Departmental Size (Subject: Physics)

Furthermore, closer investigation reveals that the upward slope is almost entirely due to two outlier points corresponding to Oxford and Cambridge Universities. If these two are excluded (and there are various reasons why one might expect Oxbridge to have certain advantages in terms of the quality of its staff and students), the slope of the regression line drops to 0.0023 with a standard error of 0.0024 - in other words, it is not significantly different from zero. Hence there is no evidence here that larger departments produce greater numbers of papers per person, although it would seem that Oxbridge physicists are rather more productive than their counterparts at other universities (ibid., p.34).

Nor are these results confined to physics. Analysis of the statistics for chemistry yields virtually identical results with a regression line slope of -0.0006 (and a standard error of 0.0045) once Oxford and Cambridge chemistry departments have been excluded. While it is conceivable that an examination of other output indicators such as numbers of citations may yet reveal the existence of economies of scale, at present it would seem unwise to base policies on the assumption that such effects exist.

5. Conclusions

What general conclusions can be drawn from the work on university research evaluation in Britain described above? And what implications, if any, can be identified for Germany? One conclusion has already been mentioned but is so important that it warrants repeating, if only to avoid any possible misunderstanding. This is that research evaluations and performance indicators are merely a complement to, not a replacement for, peer review. Peer review must remain at the heart of decision-making in science, not least because performance indicators need careful interpretation and this, in turn, often requires the assistance of scientific peers. The aims of research evaluation are rather more modest - to furnish some assurance to government and others that resources have been used effectively in the past, and to provide systematic information to peer-reviewers that may have some bearing on future scientific performance. Thus, evaluation is a management information tool, not a panacea for science policy.

Secondly, we have seen how research evaluation techniques are still in a fairly preliminary state. The influence-weight methodology, for example, although it succeeded in yielding results of direct policy relevance, faces several major problems. Although the approach adopted in the current SPRU study avoids most of these, it is nevertheless proving very

time-consuming. In short, a lot of work remains to be done before quantitative evaluation techniques can be applied on a routine basis.

A third and closely related conclusion is that, if university research evaluation is to be successful, especially over the long-term, the methods and results must be accepted as valid and reliable by those being assessed. Part of this means ensuring that the approach adopted is consistent with prevailing views on the functions or goals of academic institutions. (For example, an assessment methodology suitable in a country where university research is viewed primarily as a long-term investment in technological and economic development is hardly likely to be appropriate in a situation where a less utilitarian view of academia predominates.) In addition, success depends on developing an 'evaluation culture' whereby academics come to see why such assessments are necessary or, better still, why they are actually in their own interests. In countries where that culture has yet to become established, the strategy towards the introduction of evaluation must be to proceed gradually and on an experimental basis, determining which approaches are successful in that particular environment and which are not. Furthermore, the academic community must be fully involved at all stages. One advantage of assessment approaches that rely on interviews and not just quantitative indicators is that these provide an opportunity to explain to sceptical scientists why the evaluation is needed and to answer any criticisms of the assessment techniques being employed. Considerable effort must also be devoted to validation of the evaluation results by the scientific community and to effective widespread dissemination (through publications, seminars and so on).

Fourthly, many current science policies are based on implicit assumptions about research performance. These urgently need to be investigated empirically. We have seen, for example, how the belief that science benefits from economies of scale has little evidence to support it at present. Another example is the assumption that collaboration - whether with other university departments, industrial laboratories or scientific groups abroad - is beneficial for academic research performance, with enthusiasm for European collaboration being particularly pronounced at present. Yet much more work remains to be done, looking not just at the benefits but also at the 'costs' of collaboration (in particular, the time which it takes to initiate, organise, finance and maintain collaborative partnerships).

Finally, research evaluation is somewhat like technological innovation in that it is only likely to be successful where there is a clearly defined 'customer' for the end-product with well specified needs. In Britain, we have seen how, under a Conservative Government which has limited the growth of university

funding and at the same time sought to achieve better 'value for money', there has been an obvious demand for evaluations of academic research performance. In the Federal Republic of Germany, the situation has been different for at least three reasons. One is that, because universities are funded by individual states rather than the Federal Government, there is less need for nation-wide comparisons of research performance. Another is the prevailing ideology that all universities are equally excellent, or at least should have equal access to resources. The third is that the financial constraints have probably been less tight than in the UK, and certainly there has not been the same political concern in the Federal Republic with 'value for money' in relation to spending on higher education.

Now, however, the situation is changing quite dramatically as Germany embarks upon re-unification. Clearly, vast sums will have to be spent on improving the infrastructure of universities in the five Eastern states, with the result that the resources available to research groups in the West are likely to be much more limited than in the past. In addition, the incorporation of universities from the former Democratic Republic into the overall German higher education sector will mean that the country is faced by a new situation in which all universities are manifestly not equally excellent. As a starting point, it will be important to establish just how great are the differences in research capacity between institutions in the East and West (and how these vary across fields) in order to estimate how much needs to be invested in the former. This, combined with the tightened pressure on funding, means that the demand for university research evaluations will surely grow.

If such evaluations are carried out in Germany, this is likely to pose an interesting dilemma for science policy-makers. The results will most probably reveal a wide range of research performance across the higher education sector. The question is how to allocate resources in the light of this. There are perhaps two principal options. The first would be to pursue a 'laissez faire' policy of allowing research groups to compete for funds strictly on their scientific merits. In such a competition, the stronger departments will tend to be more successful, reinforcing the differences between the leading groups and the rest. Here, the European dimension is likely to become important. If, as many people anticipate, 1992 leads to greater mobility, this may well affect universities, with the best researchers and students migrating to the leading European universities. The race will then be on between individual member states of the Community to build up selected universities as international centres of excellence. The 'laissez faire' policy might then

give way to a more deliberate strategy of increased selectivity and concentration.

Yet, at the same time, the process of re-unification will bring political pressures in the opposite direction, with demands that East German universities should be brought up to the same level as their counterparts in the West. Hence, the main alternative policy option is one of 'positive discrimination' whereby resources are allocated preferentially to universities with a weaker research standing. Given the shortage of resources available to the governments of the five Eastern states, this will almost certainly mean that the Federal Government will have to intervene, something which has been fiercely resisted in the past on the grounds that universities should be free from such 'interference'. However, the counter-argument will be that this is the only way to ensure the continued strength of the German higher education system as a whole (as opposed to the success of a few elite universities).

(There is an analogy here with the sporting world. In relation to European soccer teams, a 'free market' supposedly pertains. The ideology is that the clubs are all equal, at least in terms of having the same access to footballing talent. In practice, however, the star players are lured to the more successful teams, which by and large are those already with most resources, further accentuating the differences in performance. In American football, by contrast, a more 'socialist' policy is pursued - somewhat ironically, given that the United States is normally regarded as the home of the 'free market'. There, it is the weaker teams who are given first choice of the new recruits. The question is which of these two models will be preferred in relation to German universities.)

The above discussion provides a clear example of how research evaluations never lead unambiguously to a specific policy decision. Although they may reveal which institutions are stronger and which weaker, there is then always a **political** decision to be made - whether to reward the more successful or to build up the weaker institutions. How this dilemma will be resolved for universities in a newly unified Germany situated in post-1992 Europe remains to be seen.

Acknowledgements

The first study reported here was funded by the UK Advisory Board for the Research Councils (ABRC) while the second is being jointly supported by ABRC and the Economic and Social Research Council (ESRC). An

preliminary version of the paper was given at ESRC, Swindon in December 1988. The third section draws heavily on an article written jointly by the author and others (Carpenter et al., 1988). Preparation of this paper for the Augsburg symposium was carried out as part of the programme at SPRU on 'Academic Research Performance Indicators' funded by ESRC, to whom the author is grateful for support. He would also like to thank all his collaborators at SPRU who have participated in the work described here, especially John Irvine, Diana Hicks and Jim Skea.

References

Carpenter, M.P., Gibb, F., Harris, M., Irvine, J., Martin, B.R. and Narin, F. (1988), 'Bibliometric profiles for British academic institutions: an experiment to develop research output indicators', **Scientometrics** 14, 213-233.

Cartter, A.M. (1966), **An Assessment of Quality in Graduate Education**, American Council on Education, Washington, DC.

Clayton, K. (1987), **The Measurement of Research Expenditure in Higher Education**, University of East Anglia, Norwich.

Gillett, R. (1987), 'Serious anomalies in the UGC comparative evaluation of the research performance of psychology departments', **Bulletin of the British Psychological Society** 40, 42-49.

Hicks, D. and Skea, J. (1989), 'Is big really better?', **Physics World** 2 (12), 31-34.

Irvine, J. and Martin, B.R. (1984), **Foresight in Science: Picking the Winners**, Pinter Publishers, London and Dover, NH.

Irvine, J. and Martin, B.R. (1984a), 'What direction for basic scientific research?', in Gibbons, M., Gummett, P. and Udgaonkar, B.M. (eds), **Science and Technology Policy in the 1980s and Beyond**, Longman, London and New York.

Irvine, J., Martin, B.R. and Isard, P.A. (1990), **Investing in the Future: An International Comparison of Government Funding of Academic and Related Research,** Edward Elgar, Aldershot.

Martin, B.R. and Irvine, J. (1983), 'Assessing basic research; some partial indicators of scientific progress in radio astronomy', **Research Policy** 12, 61-90.

Martin, B.R. and Irvine, J. (1984), 'CERN: past performance and future prospects', **Research Policy** 13, 183-210, 247-84 and 311-42.

Martin, B.R. and Irvine, J. (1986), **An International Comparison of Government Funding of Academic and Academically Related Research**, Advisory Board for the Research Councils, London.

Martin, B.R. and Irvine, J. (1989), **Research Foresight: Priority-Setting in** Publishers, London and New York.

Phillimore, A.J. (1989), 'University research performance indicators in practice: the University Grants Committee's evaluation of British universities, 1985-86', **Research Policy** 18, 255-71.

Roose, K.D. and Andersen, C.J (1970), **A Rating of Graduate Programs,** American Council on Education, Washington, DC.

Schwarz, M., Irvine, J., Martin, B.R., Pavitt, K. and Rothwell, R. (1982), 'The assessment of government support for industrial research; lessons from a study of Norway', **R&D Management** 12, 155-67.

Ziman, J. (1987), **Science in a Steady State: The Research System in Transition,** Science Policy Support Group, London.

Ziman, J. (1989), **Restructuring Academic Science: A New Framework for UK Policy,** Science Policy Support Group, London.

Author's address:

Ben R. Martin
Science Policy and Research Evaluation Group
Science Policy Research Unit
University of Sussex

Falmer, Brighton BN1 9RF