Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 36, No. 3 (1996) 343–362

THE USE OF MULTIPLE INDICATORS IN THE ASSESSMENT OF BASIC RESEARCH

B. R. MARTIN

ESRC Centre on Science, Technology, Energy and Environment Policy,
Science Policy Research Unit, University of Sussex, Falmer, Brighton BN1 9RF (UK)

(Received May 24, 1996)

This paper argues that evaluations of basic research are best carried out using a range of indicators. After setting out the reasons why assessments of government-funded basic research are increasingly needed, we examine the multi-dimensional nature of basic research. This is followed by a conceptual analysis of what the different indicators of basic research actually measure. Having discussed the limitations of various indicators, we describe the method of converging partial indicators used in several SPRU evaluations. Yet although most of those who now use science indicators would agree that a combination of indicators is desirable, analysis of a sample of *Scientometrics* articles suggests that in practice many continue to use just one or two indicators. The paper also reports the results of a survey of academic researchers. They, too, are strongly in favour of research evaluations being based on multiple indicators combined with peer review. The paper ends with a discussion as to why multiple indicators are not used more frequently.

Introduction

Early users of bibliometric indicators pursued their development for two reasons: (i) as a library or scientific information search tool (e.g. *Garfield*); and (ii) in historical or sociological studies of science (e.g. *de Solla Price*). Prior to the mid-1970s, only one or two analysts[1] saw the potential of such indicators as a tool for helping to assess the performance of scientists and thus providing an input to science policy. However, in 1977, the Science Policy Research Unit (SPRU) obtained funding for a project to evaluate six of the United Kingdom's 'big science' centres. Those centres accounted for a very high proportion of UK government spending on basic research – approximately 50% of the entire spending on science by the Science and Engineering Research Council.[2,3] The question to be addressed was, 'What had Britain obtained in return for this substantial investment?'

In 1978, *John Irvine* and I began work at SPRU on the 'Big Science Project'. The first task was to devise a methodology for assessing the performance of large

0138–9130/96/US $ 15.00

laboratories engaged in basic research. One starting point was the set of bibliometric tools pioneered by *de Solla Price*, *Garfield*, the *Cole* brothers and others. Those tools had been subject to fierce criticism from scientists[4] and sociologists of science.[5] We therefore began by examining those criticisms and the limitations of the indicators that they revealed. We then attempted to devise an approach that would minimise the effects of those limitations, and applied this to assess the performance of five of the 'big science' centres.[6] As will be seen, this involved examining not only the centres' contribution to scientific knowledge but also their technological and educational contributions. Early results of the 'Big Science Project' were reported in 1980,[7] although publication of the first main paper[8] was delayed until 1983 because of a threat of possible legal action. This represented one of the first uses of multiple indicators of basic research for explicitly science policy-related purposes.

In this paper, we begin by considering why evaluations of government-funded research have become essential. We then analyse the multi-dimensional nature of basic research; each dimension may require a different evaluation approach and indicators. We examine which dimensions of research are reflected in the various indicators, looking at such concepts as scientific 'activity', 'output' and 'progress' and the distinction between 'quality', 'importance' and 'impact'. From this, one can better appreciate which aspect of research each indicator is measuring and the limitations of that indicator. At the same time, we consider the use of peer evaluation for assessing research performance and its limitations. This discussion forms the background and rationale for the 'method of converging partial indicators' first put forward in 1980, a method involving the combined use of multiple indicators with extensive peer evaluation to assess similar laboratories. The paper then analyses what use has been made of evaluation approaches based on multiple indicators in subsequent scientometric studies. It also reports a survey in which academic researchers were questioned about their views on different indicators and evaluation approaches. The paper concludes with a discussion of why evaluations based on multiple indicators are not more common.

The need for assessments of government-funded basic research

There are perhaps four main reasons why assessments of government-funded basic research have become increasingly necessary in recent years. The first relates to the growing costs of the scientific instrumentation, facilities and infrastructure required to conduct frontier research – the so-called 'sophistication factor'. Secondly, virtually all industrialised countries are witnessing increasing constraints on public expenditure,

including spending on research. As a result, it is becoming ever more difficult to find the funds needed to support new areas and new scientists as well as to pay for more sophisticated instrumentation. The UK was one of the first to experience cuts in government spending (during the early 1980s) but most other large industrialised countries now face similar constraints.[9]

A third reason why quantitative assessments are needed concerns the emerging problems with peer review (since 1945 the principal mechanism for determining resource allocation in basic research) as we have moved into an era of level funding. Peer review worked well when government spending was rising by 5-10 % a year (as it was in the 1950s and '60s). Science is inherently dynamic, with new areas and researchers continually emerging, and in the past the annual budget increase could be used to fund those emerging areas and researchers. However, with an essentially level budget, if support for new areas and researchers is to be found, and found promptly, then reductions must first be made in existing commitments. Peer review has proved quite effective in identifying and deciding between new areas and researchers, but it is far less satisfactory when it comes to identifying declining areas and groups.[10]

Fourthly, there is the requirement from governments for greater public accountability in all areas of public expenditure. With this come demands for evaluation and for performance indicators to assure the government and the public that public money is being well spent. In the case of research, peer review cannot give the necessary assurances (because scientists, as the beneficiaries of public funding, will not be seen as totally unbiased on this issue); a more public form of accountability is therefore required. Again, demands for accountability and evaluations first became prominent in the UK but they have since spread. For example, *Hansen and Jørgensen* describe how in Denmark there is now "a stronger demand for legitimation in the shape of accountability and documentation of results", bringing with it "new forms of research assessment different from classical peer review".[11] Even in the world's largest economy, the United States, accountability has become a major issue. In 1993, Congress passed the Results and Performance Act which requires federal agencies to establish strategic planning and performance measurement. This, in turn, requires the establishment of performance goals and performance indicators to assess output, service level and outcome.[12]

The multi-dimensional nature of basic research

In what follows, we concentrate on basic research. (Applied research requires other forms of evaluation, methodological approaches and indicators.[13]) Few would dispute that basic research is multi-dimensional in terms of its nature and outputs. One

possible classification of those dimensions is the following:[14] (a) scientific – contributions to the stock of knowledge; (b) educational – contributions in terms of skills and trained personnel; (c) technological – contributions to the development of new or improved technologies; and (d) cultural – contributions to the wider society. Each main category can be further subdivided. 'Scientific' contributions to the stock of knowledge occur both in the originating field and in other scientific fields. They can also be theoretical, empirical or methodological, while another subdivision is that between incremental additions and the occasional revolutionary advance. 'Educational' contributions refer to skills and other person-embodied or tacit knowledge and competencies such as the ability to solve complex problems. The 'technological' outputs include (i) new products, processes and services, (ii) new or improved instrumentation, and (iii) new methodologies (simulation techniques, for instance) applied outside of basic research in the development of innovations. These technological outputs, along with the educational ones, may result in either economic or social benefits, while the fourth main category of 'cultural outputs' also represents a form of social benefit.[15]

Because of the multi-faceted nature of basic research, no single indicator of research output or performance will ever reveal more than a small part of the multi-dimensional picture. This point was emphasised in the first main paper from the 'Big Science Project',[16] and has been stressed many times since. For example, a principal conclusion emerging from *Kostoff*'s thorough review of research assessment approaches is that, "Since research impact has many facets, its assessment must use as many methods and as many types of experts as required to address as many of these components as possible."[17] Besides needing different evaluation approaches and indicators to assess the various forms of output, one must also recognise that no absolute quantification of basic research is possible. One can only make comparisons. Furthermore those comparisons will only be valid if they focus on reasonably similar research entities – i.e. one can only legitimately compare 'like with like'.[18] As *Miller*, in a recent analysis of 53 laboratories of various types, concluded, "comparisons of scientific impacts should be made only with laboratories that are comparable in their primary task and research outputs".[19]

For reasons of space, we shall concentrate in what follows on methods for evaluating the scientific contributions from basic research. For more basic research, these often correspond to the primary reason for government funding of such research. (This is not to imply, however, that the educational and technological contributions are unimportant and can therefore be ignored.[20] Details of how they might be assessed can be found elsewhere.[21]) In the 'Big Science Project', we assessed the scientific

contributions of the centres using a combination of (i) a wide range of bibliometric indicators and (ii) extensive peer evaluation. It should be stressed that the latter was not the same as conventional peer review (i.e. the use of two or three peers to referee a paper or a grant proposal); it involved conducting interviews with large numbers of peers at the British 'big science' centres and at their equivalents overseas.[22]

What do the different indicators actually 'measure'?[23]

In this section, we consider certain conceptual distinctions which may help in understanding what the various indicators of basic research actually 'measure'.

Scientific activity, production and progress

The first of these three categories, scientific activity, is concerned with the consumption of the inputs to basic research, and is related to such factors as the number of scientists involved, the level of funding, the number of support staff and the scientific equipment. The second, scientific production, refers to the extent to which this consumption of resources creates a body of scientific results. Those results are embodied both in research publications and in other types of less formal communication between scientists. The third, scientific progress, refers to the extent to which scientific activity results in substantive contributions to scientific knowledge. As we shall see below, although some output indicators are fairly closely linked with scientific production, their relationship to scientific progress is more complex. However, indicators of scientific progress are most relevant to assessing scientists' success in fulfilling the primary goal of basic research, the production of new scientific knowledge.

Publications

Numbers of scientific publications – that is, articles reporting substantive research results published in peer-reviewed learned journals – are a reasonable measure of scientific production. However, they are a much less adequate indicator of contributions to scientific progress. One essential problem is that most publications make only a very modest incremental addition to knowledge, while only a very few make a major contribution. Yet publication counts cannot distinguish between these. Nevertheless, the analysis reported below reveals that scientometric studies continue to make far more use of this indicator than any other, often on its own or at least without any indicator that relates to the varying magnitudes of the contributions to knowledge represented by different papers.

Publication counts are, at best, only a partial indicator of contributions to knowledge - that is, a variable reflecting (a) the level of scientific progress made by an individual or group and (b) a number of other factors such as social and political pressures. These include the publication practices of the employing institution, the country and the research area, as well as the emphasis placed on publications for obtaining promotion or grants. It cannot be assumed that the effects of (b) are relatively small compared with (a), nor that the effects are randomly distributed and therefore cancel out for large aggregations or long periods of time. The relative importance of (a) and (b) can only be established empirically.

Citations

The aim of citation analysis is to allow for and to estimate the varying contributions to scientific progress made by different publications. The use of citation counts is beset with technical problems,[24] but there are also substantive conceptual problems such as critical citations of 'mistaken' work, the failure to cite early scientific 'classics', variations in citation rates across fields and with type of paper (e.g. methodological versus empirical or theoretical papers), and the 'halo effect'. These problems arise because scientific authors are not completely logical or consistent in their referencing habits. As *Luukkonen* has recently pointed out, over thirty years after analysts first started to use citation indicators, we still lack an adequate theory of citation,[25] although there have been some useful empirical studies. For example, in one of the latest studies, *Shadish et al.* show that, of the four main reasons why authors cite particular publications, "at least three ... have the flavour of describing Kuhnian exemplars, classic works that show how something is done or thought of in a field".[26]

Quality, importance and impact of publications

In order to understand what citation counts actually measure, we need to make a conceptual distinction between the quality, importance and impact of publications.[27] In an earlier paper, we defined 'quality' as follows:

> a property of the publication and the research described in it. It describes how well the research has been done, whether it is free from obvious 'error', how aesthetically pleasing the mathematical formulations are, how original the conclusions are, and so on.[28]

This is not an entirely satisfactory definition, especially in the light of subsequent work by sociologists of science. However, we did go on to stress that

> quality is still relative rather than absolute, and it is socially as well as cognitively determined; it is not just intrinsic to the research, but is something judged by others who, with differing research interests and social and political goals (i.e. different cognitive and social 'locations' ...) may not place the same estimates on the quality of a given paper.[29]

The 'importance' of a publication was defined as

> its *potential* influence on surrounding research activities – that is, the influence on the advance of scientific knowledge it would have if there were perfect communication in science ... However, there are 'imperfections in the scientific communications system, the result of which is that the *importance* of a paper may not be identical with its *impact*.[30]

In contrast, the 'impact' of a publication describes

> its *actual* influence on surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language and availability of the publishing journal.[31]

Of these three concepts, it is the third, scientific impact, that is most closely related to the notion of contributions to scientific progress. Furthermore, from these definitions, it should be apparent that citation counts are an indicator more of impact than of quality or importance. Even so, the number of citations is but a partial indicator of impact – that is, influenced partly by the impact of a paper (or group of papers), but also influenced by communication practices, the visibility of authors, their previous work and employing institution, and so on. As with publication counts, the effects of those other factors cannot be assumed to be small nor necessarily randomly distributed; their relative importance can only be established empirically.

If citations are seen as merely a partial indicator of scientific impact, then some of the problems are diminished. For example, a 'mistaken' paper can nonetheless have a significant impact in terms of stimulating other work.[32] A high quality paper in a small unpopular field or published in a low-circulation journal may have a relatively low impact. Conversely, a paper by an eminent scientist may be more visible and therefore have more impact, earning more citations, even if its quality is no greater than those by less well known authors.

Peer evaluation

Peer evaluation is the method of assessing scientific progress most favoured by scientists but it is by no means perfect. It is based on scientists' *perceptions* of contributions by others and is influenced partly by the magnitude of those contributions and partly by other factors. There are at least three main problems here.

First, political and social pressures within the scientific community will affect the way scientists assess the contributions by their peers. Peer review depends on finding neutral peers whose material or other prospects will be unaffected by the judgements they give. Modern research is, however, increasingly competitive and this, together with the trend towards oligopoly (i.e. the concentration of research resources in an smaller number of large centres), makes it ever more difficult to find truly neutral peers in more capital-intensive areas of research. Secondly, different peers in different cognitive and social locations may evaluate a given scientific contribution rather differently. Thirdly, no peer will have perfect information on the contribution being evaluated and will therefore base their assessment on limited or imperfect information. As a result of these and other problems, peer evaluation is no more than a partial indicator of contributions to scientific progress.

Other indicators

Kuhn pointed out that the great mass of research results in only minor incremental contributions to scientific progress ('normal science') with only the occasional major discovery or radical advance ('revolutionary science').[33] Publication and citation counts may not reveal which groups have been responsible for those major advances. For these, there are two possible approaches. One is to ask peers to identify which groups have been responsible for those crucial advances and to assess their impact. The other is to construct an indicator based on highly cited papers. This dual approach was adopted in the 'Big Science Project' and the study of CERN and found to yield broadly consistent results.[34]

Another possible evaluation approach is based on the recognition accorded to scientists through the awarding of medals, prizes, invitations to give prestigious lectures and the like. Although such an 'esteem' indicator might seem attractive in theory, there are a number of problems. Some of these are practical (e.g. obtaining the necessary data in a systematic and comparable form) and some more conceptual. For example, the allocation of such awards is based not only on the magnitude of a given scientist's contributions (or, more accurately, on other scientists' *perceptions* of those) but also on other factors perhaps reflecting other contributions he or she may have made (e.g. editing a leading journal). Hence, this approach again offers only a partial indicator of contributions to scientific progress.

Size-adjusted indicators

For many evaluations (for example of laboratories, departments or countries), one needs to allow for the differing size of research activity – that is, for the differing scale of inputs. One therefore requires input as well as output indicators – e.g. numbers of scientists and support staff, recurrent funding and capital or equipment funding. From such data, one can then calculate size-adjusted output indicators such as publications per person or per unit of funding, citations per unit of funding or per paper, and so on. Such size-adjusted indicators are essential if smaller research units or entities are to be compared on a fair basis with larger ones. Yet the analysis of scientometric studies reported below suggests that something like two thirds of them rely solely on size-dependent indicators such as publication or citation totals.

The methodology of converging partial indicators

As the previous discussion has illustrated, all quantitative measures of research are, at best, only partial indicators – indicators influenced partly by the magnitude of the contribution to scientific progress and partly by other factors. Nevertheless, selective and careful use of such indicators is surely better than none at all. Furthermore, the most fruitful approach is likely to involve the combined use of multiple indicators. However, because each is influenced by a number of 'other factors', one needs to try and control for those by matching the groups to be compared and assessed as closely as one can. Clearly, no matching will be perfect. However, the hypothesis that we set out to test in the 'Big Science Project' was that, if one succeeded in obtaining a reasonable match, one would expect to find convergence both between the various bibliometric indicators employed and with the peer evaluation results. This hypothesis was confirmed in our studies of radio astronomy observatories,[35] large optical telescopes[36] and electron accelerators,[37] and was also confirmed in the later and more extensive study of the world's leading proton accelerators.[38] This is consistent with the conclusion reached by *Baird* and *Oppenheim* in a recent review of citation-based studies:

> [T]here is not, and never can be, one single measure of the value of information that will be universally acceptable. However, there are a number of measures that might, in combination, lead to some sort of index of the value of a piece of information, an individual's research contribution, or a collection of information.[39]

Similarly, *Kostoff*, in his comprehensive review of research assessment studies, drew the following conclusion:

> The concluding hypothesis of this Handbook ... is that the greater the variety of measures and qualitative processes used to evaluate research impact, the greater is the likelihood of converging to an accurate understanding of the knowledge produced by research.[40]

It could be argued that such convergence is misleading because of at least some of the indicators are related – that is, the dimensions that they measure are not orthogonal but overlap to some extent. However, the response to this is that the various indicators are not all measuring the same thing – or measuring research along a single dimension. Hence, a result based on the convergence of several indicators (preferably including extensive peer evaluation) is likely to be more reliable than one based on a single bibliometric indicator or on peer review alone. To take an example, the total number of citations earned by a research group depends partly on the number of publications it produces so these two indicators are not orthogonal. However, neither do they measure performance on exactly the same dimension. The citation total depends on the average impact of the published papers as well as on their numbers. More generally, many of the indicators used in research assessments are based to some extent on peer review (the number of articles reflects peer-review decisions to accept papers for publication in a journal, the number of citations reflects the assessments of subsequent authors as to which papers to cite, and so on). Nevertheless, an assessment based on multiple indicators should be more reliable than one based on a single indicator.

The use of multiple indicators in practice

According to *Kostoff*, "Much of the research evaluation community has come to believe that simultaneous use of many techniques is the preferred approach"[41] if one wishes to capture the different dimensions of research. Yet has the scientometric community acted accordingly? To analyse this, a selection of articles published in *Scientometrics* was examined to establish how many employed multiple indicators. The sample consisted of 12 recent issues of *Scientometrics* (Volumes 31 to 34 published in 1994-95) and 12 issues published a few years earlier (Volumes 14 and 15 which appeared in 1988-89). The analysis focused only on empirical studies reporting indicators in tables, graphs or figures. Indicators were classified into a number of main categories – for example, publications, citation totals, and so on. Where there was some ambiguity as to whether a paper used two separate types of indicators, that paper was given the benefit of the doubt. For example, a paper reporting both absolute publication counts and percentage shares of the world total was classified as being based on two indicators rather than one. The results of this analysis are shown in Table 1 below.

Table 1
Numbers of indicators reported in a sample of *Scientometrics* papers

Number of distinct indicators	Empirical papers published in *Scientometrics* in		TOTAL
	1988-89	1994-95	
1	24	25	49
2	14	18	32
3	8	15	23
4	6	3	9
5 or more	2	6	8
TOTAL	54	67	121

Of the 54 papers published in 1988-89 employing indicators of one form or another, 38 (i.e. 70%) used only one or two indicators. Only 2 (or 4%) used five or more indicators, and less than a third used three or more. The situation in 1994-95 was little different: most of the papers analysed (43 out of 67 or 64%) were based on only one or two indicators. Out of the entire sample of 121 papers, only one employed eight distinct indicators, the same number used in our original assessment of radio astronomy observatories. In short, there is little evidence here that the scientometric community is acting on the basis of *Kostoff*'s conclusion that "simultaneous use of many techniques is the preferred approach" (if we interpret 'many' as three or more distinct types of indicator used in combination). Nor do we find very convincing evidence to support the claim of *Rubenstein* and *Geisler*[42] that the use of multiple indicators has been growing; the proportion of papers based on three or more indicators in 1994-95 is not significantly greater than that in 1988-89 (36% compared with 30%).

In addition, we analysed the papers to see which indicators were most commonly employed. Table 2 below reveals that by far the most common indicator is publication counts. This was employed in 72 of the 121 papers in the sample (60%), almost double the next most common indicators – citation counts (38 or 31%) and citations per paper or impact factors (26%). These, in turn, appeared several times more frequently than any other indicator; for example, co-citations were used in only 5% of the total. In short, scientometric analysts would seem to rely rather heavily on the simplest of indicators, despite their well known limitations, and they mostly use only one or two indicators rather than a larger combination.[43]

Table 2
Numbers of *Scientometrics* papers based on particular indicators

Indicators	Empirical papers published in *Scientometrics* in		TOTAL
	1988-89	1994-95	
Publications	29	43	72
Citation totals	16	22	38
Citations per paper/ impact factor	14	18	32
Publication % share	8	6	14
Peer review	8	4	12
Collaborations	2	9	11
Journals	4	5	9
Co-citation	3	3	6
Citation % share	2	3	5
Patents	2	3	5
Immediacy/Price index	2	2	4
Co-word	1	2	3
Funding	2	0	2
Highly cited papers	1	1	2
Students	0	2	2
Total number of papers	54	67	121

Views of university scientists on the use of multiple indicators

What are the views of scientists on the use of multiple indicators to assess their research? This question was investigated by the author in a study conducted in 1990-92. First, however, we need to consider the background to that study. In 1986, the University Grants Committee (UGC) ranked the research performance of all British university departments or 'cost centres'. The evaluations were carried out by UGC subject groups of half a dozen or so experts who ranked all UK university departments in their field on a 4-point scale. The approach was based almost entirely on peer-review and the results were extremely controversial. The exercise was repeated in 1989 and again in 1992 by the Universities Funding Council. In 1992, rather more performance-related information was collected (for example, on departmental publication counts) but again the approach relied primarily on peer-review judgements to classify departments, this time on a 5-point scale. A broadly similar approach was adopted in the 1996 Research Assessment Exercise by the Higher Education Funding Council, although the scale was further expanded to seven points.

The aim of the SPRU study was to explore the feasibility of constructing research performance indicators for science and engineering departments, and to establish whether such indicators might be used to complement conventional peer-review procedures for assessing university departments in these research assessment exercises. The approach adopted had two main components. One involved the construction of a large database on the inputs and outputs for all UK university departments. From this, a range of indicators was constructed, and the rankings of departments obtained with these various indicators were compared with the ratings given by UGC/UFC in 1986 and 1989.

The other component consisted of case-studies focusing on four selected fields – mathematics, physics, biochemistry and chemical engineering. In interviews at a range of universities, academics were asked for their views on the strengths and weaknesses of different approaches to the evaluation of university departments and on how the approach adopted by UFC might be improved. Approximately 120 researchers in 25 university departments were interviewed. Among the questions addressed were the following:

(a) How well does peer-review work in practice as a means of evaluating entire departments?

(c) What are the relative strengths and weaknesses of each approach or indicator? And how would academics most like future research assessment exercises to be carried out?

The first question involved an empirical investigation of peer review. Traditionally, peer review has been used to assess individuals or, at most, relatively small groups, not entire departments. The aim here was to determine the extent of knowledge possessed by academics about research at other departments in the same field. Interviewees were asked to rank the research performance of a number of departments in their field on the UFC five-point scale. The results were then compared with those of the UFC expert panels. In general, there was broad agreement between the UFC rankings and those we obtained, but in approximately 10% of cases the rankings differed by at least one point on the 5-point scale.

Interviewees were also asked how many departments they were sufficiently familiar with to rank their research performance with some confidence. A typical academic is reasonably familiar with the work of six to ten UK departments. However, that knowledge is almost entirely confined to his or her own subfield. Even if one chooses half a dozen singularly well informed researchers to serve on a panel, it unlikely that they will have direct knowledge of research in all the subfields (there might be six or eight in total) for all university departments in the country. This may explain why the

results of our peer-review exercise to rate departments suggest that up to 10% of the UFC rankings could be wrong by at least one unit.

As noted above, another objective of the study was to obtain the views of academics on the strengths and weaknesses of different approaches to assessing the research performance of university departments. They were asked about the approach adopted in the 1986 and 1989 exercises and about a number of possible indicators. Their responses were subsequently classified into five categories: (1) strongly in favour – no major weaknesses; (2) some weaknesses but outweighed by strengths – on balance in favour; (3) mixed views on strengths and weaknesses; (4) some strengths but outweighed by weaknesses – on balance against; and (5) strongly against – no major strengths. The results are given in Table 3 below.

Table 3
Views of academics on assessment approaches and performance indicators

Assessment approach or performance indicator	% holding particular view					% in favour –% against
	Strongly favour		Mixed views		Strongly against	
	1	2	3	4	5	
UFC 1989 exercise	12	74	8	5	2	79
UGC 1986 exercise	1	41	41	11	5	26
International peers	25	37	17	17	4	41
Opinion poll	7	45	24	14	12	26
Research income	7	59	19	9	7	49
Publications	11	61	20	3	5	64
Weighted publications	42	37	12	9	0	70
Citations	9	57	20	9	4	54
Esteem indicators	5	53	15	18	9	31
Trained researchers	17	57	10	16	1	57

Table 3 shows that the great majority of academics favoured the peer-review-based approach adopted in the 1989 Research Assessment Exercise: 86% (i.e. 12 + 74%) were in favour to a greater or lesser extent compared with only 7% against, a net balance of 79% in favour. Nevertheless, interviewees pointed to many weaknesses in the 1989 UFC approach. These included: (a) the tendency for peers to rank more highly departments and subfields they know well; (b) the cost-centre often being too broad to be ranked by a small panel familiar with only some of the component subfields; (c) a bias against small departments, perhaps stemming from the UFC definitions for the five rankings; (d) problems in ranking departments with

interdisciplinary interests that do not fall neatly within a single cost-centre; (e) a bias against departments specialising in non-mainstream subfields; (f) inadequate normalisation across cost-centres, leading to adverse financial consequences for some fields; and (g) the absence of international (foreign) peers on UFC assessment panels.[44]

The interviewees were much more critical of the comparatively primitive approach adopted in the 1986 Research Assessment Exercise (the net majority in favour was only 26%). The same percentage favoured ranking departments on the basis of an opinion poll of researchers, while rather more were in favour of including international peers in the assessment. A research income indicator was favoured by two thirds of interviewees but they pointed to substantial problems, the main one being the wide variation in cost across subfields within a broad field. In addition, the funding data may sometimes be incomplete and therefore misleading. Publication indicators were favoured by 72% of those questioned. One worry here was the variation in the importance of papers. As a result, the great majority (nearly 80%) would like to see a system for 'weighting' publications according to the status of the journals in which they appeared. To construct such a weighting scheme, most favoured a peer-review survey to identify the leading journals in the field.

Citation indicators were certainly seen as problematic but two-thirds of interviewees were in favour of their inclusion with only 13 % against. Worries were expressed about departments earning large numbers of citations by producing 'bad' papers or through citation circles, yet no-one had first-hand knowledge of such cases. The belief that citation circles are already at work seems to be more of a modern legend than an established phenomenon. There was also concern about the variation in citation rates among subfields, with larger and more fashionable subfields being at an advantage. Despite this, the overall view was that citation data were worth including, at least for science and engineering. Esteem indicators were also on balance seen as worth including, but the difference between those in favour and those against was smaller than for most other indicators, and several problems were foreseen. Many academics were concerned that honours and prizes are allocated by the scientific 'establishment' on grounds that are not exclusively scientific and often for work performed many years earlier.

It can be argued that trained researchers are just as an important output from university departments as scientific advances. Certainly, a key function of universities is to educate young scientists, and an indicator based on numbers of PhDs awarded each year will reflect this aspect of a department's contribution to the economy or society. The inclusion of such an indicator in departmental assessments was favoured

by three quarters of interviewees. Again, there are limitations such as variations in the quality of PhD training and the dependence of the indicator on other factors (such as how many studentships the department was awarded).

When we asked academics to compare all the different approaches to departmental assessment directly, the 1989 UFC approach was rated first equal with international peer review and weighted publications. Next came four indicators - citations, (unweighted) publications, PhD numbers, and research income - together with the 1986 UGC approach. Opinion poll ratings and esteem indicators were ranked some way behind these.[45]

The above results concern the use of different approaches in isolation. Another question put to university staff was whether departmental research performance was best assessed using peer review alone, performance indicators alone, or some combination of the two. For those choosing the last of these three options, we also asked whether equal weight should be given to the peer review and performance indicator components, or whether rather more emphasis should be attached to one of them. The results are contained in Table 4 below.

Table 4
Relative importance of peer review (PR) versus performance indicators (PI)

	Number of interviewees expressing a particular view				
	Physics	Biochemistry	Mathematics	Chemical engineering	TOTAL
Peer review (PR) only	0	0	1	0	1
PR and PI - more weight to PR	11	7	13	2	33
PR and PI - equal weight to each	9	11	5	6	31
PR and PI - more weight to PI	6	13	7	2	28
Performance indicators (PI) only	0	0	1	2	3

The first point to note is that only one person (out of 96 interviewees who addressed this issue) suggested that peer review should be used without any recourse to performance indicators. Likewise, only three argued that performance indicators completely supplanted the need for any element of peer review when assessing departments. The vast majority (96%) believed that peer review should be combined with performance indicators. They were split approximately equally between those who favoured giving similar weight to the two elements (32%), those who advocated

more weight to peer review (34%) and those pressing for more emphasis on performance indicators (29%). Significantly, several interviewees stressed that as wide a range of indicators as possible should be employed.

To sum up: university researchers certainly have many criticisms of bibliometric indicators and especially citations. However, if one adopts a symmetrical approach, asking about the strengths and weaknesses of each possible means for evaluating the research performance of university departments, one finds approximately a similar level of criticism of peer review as of publications and citations. Consequently, the great majority of academics favour an approach based on combining peer review with a range of research performance indicators.

Why is there so little use of multiple indicators?

We have seen how most scientists favour the combined use of multiple indicators while in practice most evaluations continue to rely either on peer review alone (in evaluations by funding agencies) or on just one or two indicators (in *Scientometrics* articles). What are the reasons for this? To answer this, we need to contrast the benefits and the costs associated with using multiple indicators. The benefits are relatively easy to identify. As argued earlier, the use of multiple indicators, preferably in conjunction with peer review, is the only way to capture the multi-dimensional nature of basic research. It is also the only effective way to meet the evaluation needs set out at the start of the paper. And as *Kostoff* has argued, better evaluations can contribute to improved organisational efficiency and increased communication between researchers and potential research users (leading to more effective exploitation of the results of research). The end consequence is a better means to justify research funding.[46]

As regards the costs, the first point to note is that publication counts, citation counts and journal impact factors are all relatively cheap and easy to construct. Consequently, these indicators tend to be used a lot "because they are there". Often, there is little regard for precisely which aspects of research they are capturing and which they are neglecting. Other indicators may be harder or more expensive to obtain. For example, size-adjusted indicators require input data which may take much effort to produce. Peer evaluation can be even more time-consuming. (In the SPRU evaluation of proton accelerator laboratories, for instance, we interviewed 200 particle physicists in a dozen countries.) It is therefore perhaps not surprising that peer review features infrequently in the empirical studies reported in *Scientometrics*. Similarly, to evaluate the technological and educational contributions from basic research, one needs to conduct case-studies, interviews or surveys – again, comparatively time-consuming activities compared with generating some simple bibliometric data.

Conclusion

In this paper, we have argued that using a range of performance indicators is better than just one or two. This is partly because any single indicator can at best capture only one aspect of performance. However, there is another important reason for favouring the use of multiple indicators, namely that it minimises the risk that scientists will in some way 'play the game' and manipulate the indicators to their advantage. There is an interesting philosophical point here: any attempt to assess scientific research will change the research system in some way.[47] In other words, there is a form of Heisenberg Principle at work here – if you measure a research system, you disturb it.[48] However, with a number of indicators being applied, it then becomes much more difficult, if not impossible, to manipulate all the indicators without at the same time improving one's research.

In short, rather than attempting to measure research performance on a one-dimensional scale, what one needs in evaluations is a multidimensional set of ratings or a profile.[49] Instead, many evaluators have taken the lazy way out and used only one or two indicators that are cheap or easy to obtain. The scientific community is naturally unhappy with these simple-minded approaches to evaluation and is still somewhat reluctant to adopt bibliometric and other indicators as a complement to peer-review in decision-making in science. Only when those involved in constructing and using research performance indicators routinely use multiple indicators reflecting the different facets of basic research in combination with systematic peer evaluation[50] are we likely to see more acceptance of the results.[51]

*

Ben Martin is Professor of Science and Technology Policy Studies at the Science Policy Research Unit, University of Sussex. He is grateful to the Economic and Social Research Council for supporting his work through the Centre on Science, Technology, Energy and Environment Policy (STEEP). He would also like to acknowledge the immense contributions made by his colleague and collaborator, *John Irvine*, over a period of a dozen years of fruitful partnership. The study on academic research performance indicators reported here benefited from inputs from *Jim Skea* and from collaboration with *Paul Bourke* and *Linda Butler* at the Australian National University. Lastly, *Diana Hicks* and *Ron Kostoff* provided helpful comments on a preliminary draft of this paper.

Notes and references

1. E.g. J. H. WESTBROOK, Identifying significant research, *Science*, 132 (1960) 1229–1234; N. WADE, Citation analysis: a new tool for science administrators, *Science*, 188 (1975) 429–432.
2. J. IRVINE, B. R. MARTIN, What direction for basic scientific research?, Chapter 5 in M. GIBBONS, P. GUMMETT, B. M. UDGAONKAR (eds.), *Science and Technology Policy in the 1980s and Beyond*, London, Longman, 1984, pp. 67–98.
3. B. R. MARTIN, J. IRVINE, Assessing basic research: some partial indicators of scientific progress in radio astronomy, *Research Policy*, 12 (1983) 61–90.
4. E.g. ANON, Is your lab well cited?, *Nature*, 227 (1970) 219; ANON, More games with numbers, *Nature*, 228 (1970) 698–699.
5. E.g. D. LINDSEY, Production and citation measures in the sociology of science: the problem of multiple authorship, *Social Studies of Science*, 10 (1980) 145–162.
6. The sixth, CERN, was left to a subsequent study two years later. The results were published in a series of three articles: B. R. MARTIN, J. IRVINE, CERN: past performance and future prospects – I – CERN's position in World High-Energy Physics, *Research Policy*, 13 (1984) 183–210; J. IRVINE, B. R. MARTIN, CERN: past performance and future prospects – II – The scientific performance of the CERN accelerators, *Research Policy*, 13 (1984) 247–284; and B. R. MARTIN, J. IRVINE, CERN: Past performance and future prospects – III – CERN and the future of world high-energy physics, *Research Policy*, 13 (1984) 311–342.
7. J. IRVINE, B.R. MARTIN, A methodology for assessing the scientific performance of research groups, *Scientia Yugoslavia*, 6 (1980) 83–95.
8. MARTIN, IRVINE, *op. cit.*, note 3. This article appeared in April 1983 even though it had been accepted for publication in September 1980.
9. J. IRVINE, B. R. MARTIN, P. A. ISARD, *Investing in the Future: An International Comparison of Government Funding of Academic and Related Research*, Aldershot and Brookfield, Vermont, Edward Elgar, 1990.
10. IRVINE, MARTIN, *op. cit.*, note 2.
11. H. F. HANSEN, B. H. JØRGENSEN, *Science Policy & Research Management: Can Research Indicators Be Used?*, Institute of Political Science, University of Copenhagen, Copenhagen, 1995, p. 1.
12. R. N. KOSTOFF, *The Handbook of Research Impact Assessment* (Fifth Edition), DTIC Report Number ADA296021, 1995.
13. For example, the evaluation of government-funded applied research in Norway employed a combination of peer review and 'customer review' – see J. IRVINE, B. R. MARTIN, M. SCHWARZ, K. PAVITT, R. ROTHWELL, *Government Support for Industrial Research in Norway: A SPRU Report*, Oslo: Universitetsforlaget Norwegian Official Publication NOU 30B, 1981.
14. See Fig. 1 on p.64 in MARTIN, IRVINE, *op. cit.*, note 3.
15. A good example here would be popular books by scientists such as *Stephen Hawking*.
16. *Ibid.*, note 3, p. 64.
17. KOSTOFF, *op. cit.*, note 12, p. 8.
18. MARTIN, IRVINE, *op. cit.*, note 3, p. 75.
19. R. MILLER, The influence of primary task on R&D laboratory evaluation: a comparative bibliometric analysis, *R&D Management*, 22 (1992) 3–20.
20. For an example of how the educational technological outputs from basic research may be assessed, see the references cited in note 21.
21. J. IRVINE, B. R. MARTIN, The Economic Effects of Big Science: The Case of Radio Astronomy, *Proceedings of the International Colloquium on* the *Economic Effects of Space and Other Advanced Technologies*, Strasbourg, 28–30 April 1980, Paris, European Space Agency, ESA SP-151, 1980; and B. R. MARTIN, J. IRVINE, Spin-Off from Basic Science: The Case of Radio Astronomy, *Physics in Technology*, 12 (1981) 204–212.

22. Over 150 scientists were interviewed in the 'Big Science Project'.
23. This section draws heavily on MARTIN, IRVINE, *op. cit.*, note 3.
24. See, for example, the discussion in *ibid.*, p. 67.
25. T. LUUKKONEN, The cognitive and social foundation of citation studies - why we still lack a theory of citation, submitted to *Science, Technology and Human Values* (1995).
26. W. R. SHADISH, D. TOLLIVER, M. GRAY, S. K. SEN GUPTA, Author judgements about works they cite: three studies from psychology journals, *Social Studies of Science*, 25 (1995) 477-498 - quote on p.481.
27. See also the related distinction between 'quality' and 'relevance' in HANSEN, JØRGENSEN, *op. cit.*, note 11 , p. 3.
28. MARTIN, IRVINE, *op. cit.*, note 3, p.70.
29. *Ibid.*
30. *Ibid.*
31. *Ibid.*
32. Examples of this in the field of experimental high-energy physics can be found in MARTIN, IRVINE, *op. cit.*, note 6.
33. T. S. KUHN, *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press, 1970.
34. MARTIN, IRVINE, *op. cit.*, note 3 ; *idem.*, *op. cit.*, note 6.
35. MARTIN, IRVINE, *op. cit.*, note 3.
36. J. IRVINE, B. R. MARTIN, Assessing basic research: The case of the Isaac Newton Telescope, *Social Studies of Science*, 13 (1983) 49-86.
37. B. R. MARTIN, J. IRVINE, Internal criteria for scientific choice: an evaluation of the research performance of electron high-energy physics accelerators, *Minerva*, XIX (1981) 408-432.
38. MARTIN, IRVINE, *op. cit.*, note 6.
39. L. M. BAIRD, C. OPPENHEIM, Do citations matter?, *Journal of Information Science*, 20 (1994) 2-15 (quote on p. 13).
40. KOSTOFF, *op. cit.*, note 12, p. 37.
41. *Ibid.*, p. 118.
42. A. H. RUBENSTEIN, E. GEISLER, Evaluating the outputs and impacts of R&D/innovation, *International Journal of Technology Management*, 6 (1991).
43. Not all scientometric analysts are guilty of this. For example, the ISI analysts who periodically publish lists of leading research institutes in *Science Watch* normally use three indicators - papers, citations and citations per paper.
44. Full details of the study and the results can be found in B. R. MARTIN, J. E. F. SKEA, *Academic Research Performance Indicators: An Assessment of the Possibilities*, Brighton, SPRU, 1992.
45. See Table 12 in *ibid.*
46. KOSTOFF, *op. cit.*, note 12, p. 8.
47. HANSEN, JØRGENSEN, *op. cit.*, note 11, p.5.
48. MARTIN, SKEA, *op. cit.*, note 44, p. 75.
49. *Ibid.*, p. 75.
50. J. P. DE GREVE, A. FRIJDAL, Evaluation of scientific research profile analysis - a mixed method, *Higher Education Management*, 1 (1989) 83-90.
51. KOSTOFF, *op. cit.*, note 12, p. 9.