Policy Study No. 5

QUANTITATIVE ASSESSMENT OF DEPARTMENTAL RESEARCH

A survey of academics' views

SCIENCE AND ENGINEERING POLICY STUDIES UNIT

The Science and Engineering Policy Studies Unit (SEPSU) is run jointly by the Royal Society and the Fellowship of Engineering. Its staff, based at the Royal Society, assist the Society and the Fellowship in their contributions to policy formulation in science and engineering and, more generally, provide expertise in policy analysis. Much of the Unit's effort is devoted to objective, data-driven studies of policy issues, either of its own choosing or in response to external commissions. The Unit also provides a service to other sections of the Society and Fellowship, and maintains an extensive database of relevant information.

Funding for the Unit is provided privately by the Royal Society and the Fellowship of Engineering, by industrial and commercial sponsors and by contract work. Support from the following sponsors is gratefully acknowledged:

BICC
British Gas plc
Coutts Charitable Trust
Lucas Industries plc
Northern Engineering Industries plc
Prudential Corporation plc
THORN EMI plc

For further information, please write to:

Dr P.M.D. Collins
Director, The Science and Engineering Policy Studies Unit
The Royal Society
6 Carlton House Terrace
London SW1Y 5AG

Tel: 071-839 5561
Fax: 071-930 2170

Peter Collins

26. iv. 91

QUANTITATIVE ASSESSMENT OF DEPARTMENTAL RESEARCH

A survey of academics' views

A report for the CVCP/UFC Subcommittee on Research Indicators

P.M.D. Collins

SEPSU Policy Study No. 5

April 1991

ISBN 0 85403 436 6

SCIENCE AND ENGINEERING POLICY STUDIES UNIT

of

The Royal Society and The Fellowship of Engineering

FOREWORD

The Autumn of 1990 saw the publication of the fourth annual volume of *University Management Statistics and Performance Indicators in the UK*, produced jointly by the Committee of Vice-Chancellors and Principals and the Universities Funding Council. As part of the process of developing this increasingly useful adjunct to the management of universities, a CVCP/UFC subcommittee has for some time been considering what indicators of research performance, if any, might usefully be included in future editions of the publication. The subcommittee has been anxious throughout to ensure that its eventual proposals would be both practicable and broadly acceptable to the academic community; to this end it has consulted widely over a considerable period of time, and I would like to thank most warmly here all those who have responded, often on more than one occasion. One outcome of this process, an agreed basis for the annual recording of research publications, will shortly be promulgated.

A major issue with which the subcommittee has been concerned is the extent to which the regular quantitative assessment of departmental research can go beyond this point. The process of consultation raised a number of questions to which the answers were not immediately apparent, and in the light of this the subcommittee decided to invite Dr Peter Collins, of the Science and Engineering Policy Studies Unit, to review the evidence submitted to it, to point up key issues, and to suggest tentative conclusions. While at this stage neither the CVCP nor UFC is committed to any particular policy, it is already clear that Dr Collins' study will be of great value as we decide how to proceed, and I would like to take this opportunity of thanking him on the subcommittee's behalf. My colleagues and I look forward with great interest to the reactions that it will generate within the university system.

Professor Martin Harris
Chairman, CVCP/UFC Subcommittee on Research Indicators

February 1991

SUMMARY

This report describes what academic staff in UK universities think about quantitative ways of evaluating departmental research. It is based on a uniquely extensive survey of academic opinion, and covers the full range of disciplines from mathematics to music. Written from the perspective of those being evaluated rather than those carrying out the evaluation, it provides an essential element to our *understanding* of research assessment. It should contribute equally to the *practice* of research assessment: decisions based on assessments can be most readily implemented if the assessment methodology takes account of the views of those directly affected.

Respondents typically accepted, more or less grudgingly, the need for research assessment in general and performance indicators in particular, though a few expressed hostility to the whole idea. However, many respondents went out of their way to stress the virtues of peer review, and those who favoured performance indicators saw them as means of strengthening peer review judgements, not as a substitute for them. Acceptance of quantitative indicators was more likely to be found in the sciences than the humanities, but it would be an over-simplification to say either that scientists accepted them or that researchers in the humanities rejected them.

Most respondents accepted that measurement of the volume of research output had some place in assessment of departmental research, and that bibliometric analysis was an important — though by no means the only—route to making such measurements. Citation analysis found a few, guarded, supporters in the natural and social sciences, but the strongly prevailing view was that citation analysis had no role in the assessment of departmental research. The reasons given included the fact that citations *per se* had little or nothing to do with quality; the possibility of citation circles and other abuses; time-lags between carrying out a piece of research and acquiring published citations to it; and the shortcomings of the various Citation Indices as tools for citation analysis (as distinct from their original function as information retrieval tools). The use of a range of esteem indicators found greater support, though it was clear that such indicators could not readily be quantified in any systematic way.

For annually published quantitative data on departmental research, only bibliometric data on the volume of research output found much favour among the respondents to this survey. The degree of favour depended on the extent to which the data were collected in a form that recognized the particular circumstances of individual disciplines. For periodic assessments carried out by relevant experts, as opposed to annually published data that might be misinterpreted by those lacking the necessary knowledge, respondents agreed that esteem indicators had a useful role to play. However, there was virtually no support for the inclusion of citations in annual data series, and little support for their use in periodic assessment. Respondents sent a firm message that, at departmental level, citation analysis added nothing to peer review.

> There are grave dangers of being seduced by the seeming objectivity of a count of publications or citations, which is quite illusory. ... I welcome the apparently cautious approach suggested in the ... document, but fear that we shall have something quantitative but worthless thrust upon us. [Science]
>
> ... considers bibliometric methods to be unreliable performance indicators and is totally opposed to their use in assessing the quantity and quality of research carried out in chemistry departments. [Chemistry]
>
> There is general support for the use of bibliometric methods, if they are components of a wide-ranging process of assessment involving many other factors. [Engineering]
>
> The proposals in the paper for citation analysis and esteem indicators, while not ideal, would at least be an improvement on evaluation simply by volume of publications. [Biology]
>
> The majority feeling is that bibliometric methods are becoming an unwelcome influence on the publishing decisions of academics and trivialize the research process. [Geography]
>
> We have doubts as to whether the whole approach ... is applicable to most of the social sciences, with the possible exception of economics. [Social science]
>
> It is the quality of the process by which the published material is 'accepted' for publication/citation that is crucial to the value of the quantitative measure as an indication of quality. [Social science]
>
> The use of bibliometric methods in the assessment of the quality of research in legal subjects would not only be useless, but positively misleading. [Law]
>
> The drawbacks of bibliometric indicators far outweigh the very few advantages they may offer. We would expect their regular use to have an unwholesome effect on the way in which history would be studied. [History]
>
> Bibliometric measurement is open to very serious objection in the humanities, and it needs not just supplementing but questioning. [English]
>
> This Faculty takes a positive view of proposals to make better use of well-tried research indicators such as lists of publications, membership of learned societies, major prizes ... But ... unanimous condemnation of bibliometric methods. [Theology]
>
> The list of artists, scholars, performers and critics who would have failed any contemporary 'bibliometric' test might well prove the consist of those who have made the greatest contribution to Western civilization in music. [Music]

In addition to these various comments on the acceptability in principle of bibliometric methods, several respondents commented on the cost and the time that would have to be consumed in providing the necessary data for analysis. This was seen as a significant disadvantage of the bibliometric approach.

A further difficulty pointed out by some respondents was that bibliometric data necessarily referred to the past, and not always the

recent past. The past was not self-evidently a clear guide to future performance or long-term status, especially for relatively inexperienced researchers.

> Performance indicators, if used properly, can provide information and can clarify issues about which judgements have to be made. However, they only relate to past performance and cannot judge potential.
>
> ... the unjustified sociological assumption that reputation, in the short term, correlates with quality, as judged by a long-term consensus.

Methodological imperialism

Respondents in the humanities at times saw quantitative assessment as something that had arisen in the sciences and ought to stay there. They objected to an overly scientistic approach being extended to disciplines where more subtle forms of judgement prevailed. For such respondents, one could discern deep-rooted antagonism to the apparent triumphalism of 'scientific method' that ran wider than the present context.

> I do have strong feelings about this nonsense of trying to square the circle of 'devising quantitative measures of quality' and of trying to force the humanities and social sciences into a straitjacket designed for the natural and experimental sciences. [History]
>
> Initiatives on subjects like this often seem Science-based and ill-suited to Arts subjects. [English]
>
> The bibliometric method ... is a method self-confessedly designed to handle scientific publication. No form of juggling or interpretation can adequately adapt this method to produce a meaningful analysis of humanities research, and we strongly object to the continued imposition of scientific criteria on arts subjects. [Drama]
>
> This disturbing paper ... assumes the applicability to all academic disciplines of quantitative measures of quality which are of limited usefulness to a few. ... this unwarranted methodological imperialism ... [Philosophy]

Unit of assessment

The consultative paper focused on the department as the unit to be assessed. 'Department' was not defined: it was assumed to be the same as the unit of assessment employed in the UFC's second research selectivity exercise (1989), which had involved a wide-ranging consultation on which disciplinary groupings were most suitable. The departmental level was chosen partly because that was the level of greatest relevance to the CVCP/UFC, and partly because of the need to generate a critical mass of data points.

These assumptions attracted some comments. Many departments would appear to be smaller than estimates of the minimum threshold needed for reliable statistical analysis. Moreover, it was not clear that departmental performance could be regarded as the sum of individual performance:

> One cannot talk about a 'group' in an Arts subject—each scholar is an individual. [Languages]
>
> If a department consists of three Nobel prizewinners and 20 people who do no research at all, what does a research performance indicator for that department as a whole mean, and could it justify withholding support for the research of the three Nobel prizewinners? [Mathematics]

A sense of alienation from the assessment process is apparent in one comment on this issue:

> It is not accepted that performance indicators operate best at the institutional rather than the individual level. Bibliometrics are already operating at an unacceptable distance from the object that is being measured; operating at the institutional or departmental level increases that distance and multiplies inaccuracies. [Arts and Social sciences]

Care would also have to be paid to defining eligible departmental staff. Should visiting staff be included? How should staff moving from one department to another be handled—should the second department be credited with the career performance of the new member or only with work done after the move?

Inter-disciplinary comparisons

Some respondents reported 'widespread concern about the notion that a single set of indicators can be used regardless of discipline', though others recognized that the consultative paper did not suggest that the various measures should be used in the same way for each discipline. The impossibility of using performance indicators to make direct comparisons between disciplines was said to apply also to comparisons between specialisms within a single discipline. This could lead to problems where a single department combined a range of specialisms having distinct patterns of behaviour: to combine results from the various specialisms into a single departmental result could give a misleading impression.

Additional measures

Several respondents proposed measures of performance additional to those discussed in the consultative paper, for example:

- regularity of research grants [Geology]
- research income per member of staff [Engineering]
- grant value per paper and/or per citation [Business studies]
- number of postdoctoral staff employed on competitive external grants [Biology]
- number of competitive studentships held by the department [Biology, Engineering]
- bibliometric impact and/or influence measures
- the quality of journals in which most scientific work was published [Biology]
- published reviews of the books produced by the department [History]

Some of these attracted criticism from other disciplines:

> The amount of outside funding obtained by an individual historian gives no indication of his industry or of the quality of his work. [History]

Other suggestions were reading the reports of external examiners of graduate theses [History], inviting researchers to present all the information they considered relevant against a definition of quality agreed to be pertinent to their department [Science], and making use of the reports of visiting accreditation teams from the professional institutions [Engineering].

Institutional experience

Some universities reported work they had already undertaken to develop their own performance indicators. In one, a paper commenting favourably on the usefulness of bibliometric indicators had been written for the university's Research Committee and subsequently circulated to all academic staff. In another, there had been a university-wide citation survey and a detailed analysis in four sociology departments of citation performance against rating in the UFC research selectivity exercise. A third reported:

> We assiduously collect data on the publication record of staff which is incorporated into the University's database and analysed by type of publication.

A fourth commented:

> Esteem indicators are also welcomed and have been collected here for some years.

It would appear, then, that at least in some institutions the notion of systematically collecting performance data is becoming internalized and is not regarded simply as an externally imposed chore. This is in addition to the practice of many departments of including lists of publications in annual reports.

Summary

Respondents discussed some broad methodological issues relating to assessment of departmental research as well as dealing with the detailed points raised in the consultative paper. There was a strong consensus, across all disciplines, in favour of peer review, though many respondents allowed quantitative techniques a subsidiary role as an input to the peer review process. Some respondents argued that quantification of research performance, and especially of the quality of research, was inherently impossible. Those who saw a role for quantitative measures argued that it was preferable to use a broad portfolio of such measures rather than just one or two; that raised the question of how to distil a single message from the outcome, especially when the measures did not all point in the same direction.

On bibliometrics in general, views ranged from cautious acceptance to outright rejection. Acceptance was more likely to be found in the sciences than the humanities, but it would be an over-simplification to say that scientists accepted bibliometrics while researchers in the humanities rejected them. Some humanities respondents, though, objected vociferously to what they saw as a scientific paradigm being thrust upon them.

Many respondents pointed to the dangers of trying to use quantitative measures to draw comparisons between departments in different disciplines. Some took this further, arguing that the particular mix of subdisciplines represented in a given department would affect its overall performance on certain measures and that comparisons between departments in the same general discipline therefore needed to be approached with care.

Some institutions reported that they already collected, systematically, some of the sorts of data mentioned in the consultative paper. A move towards regular (if limited) use of quantitative measures in some degree would therefore not fall on wholly unprepared ground.

CHAPTER 4: MEASURES OF VOLUME OF OUTPUT

Acceptability of measuring volume

If we leave aside for the moment the questions of what constitutes the output of research and how it might be measured, most respondents were ready to go along with the suggestion that there was some merit in trying to measure systematically the volume of academic research output.

> The volume of research output is clearly crucial and probably not difficult to measure. [Modern languages]
>
> There can be no objection to assessment of the quantity of research publication generated by a department ... there is some correlation between the amount a person publishes and the liveliness of his/her engagement in the subject. [English]

But there were, as ever, some dissenting voices.

> There is a danger of emphasizing short-term output at the expense of long-term output and quality. [Social science]
>
> Volume of output is one of the most misleading criteria that can be adopted. ... Harvard University ... when considering people for promotion or tenure permits them to cite only five papers in support of their claim. ... This decision has completely changed the publication practices of the junior staff. [Medicine]
>
> I always considered that 'publish or perish' was more an American failing than a British one. Clearly, the CVCP/UFC are determined to rectify this. [Mathematics]

Publications as research output

Most respondents agreed that publications constituted at least one element of the output of academic research, and that collecting data on the volume of published output had a place in the assessment of research.

> With the exception of commercial and other confidential material, we accept that volume of published research output is as reliable an indicator as is likely to be attainable of the volume of research output.
>
> Some departments, but not all, would accept that the quantification of numbers of articles in refereed journals was a useful measure.
>
> Each university should establish its own database of publications.
>
> We intend to put our publication list on a database this year. [Chemistry]

And, again, some dissent:

> The plethora of scientific publications really means that almost anything can get published somewhere; unless a review body has some feel for the quality of journals involved, it might as well weigh the papers for all the information a simple list will give. [Medicine]
>
> The naive assumption that the fact of publication ensures novelty or worthiness is hardly worth refutation. [History]

Journal papers

The great bulk of bibliometric analysis has focused on papers in journals. Quite a few respondents advocated making some attempt to recognize that some journals were more equal than others.

> If publication counts are to become the accepted method by which research output is quantified, some weighting should be given to the number of publications in the major journals for each discipline. Each discipline would be asked to agree 6 major respected journals and these would be weighted equally at 5 or 6 times all other journals.
>
> The quality of research within a department is better measured by counting the number of papers published in prestigious journals. [Physical sciences]
>
> Any measure of the volume of published output must be tied to a quality indicator (e.g. X full papers in a journal with the highest international standard, Y letters in some lesser category journal). [Engineering]
>
> It is very important that journals are graded and not just all accepted as being equal. [Engineering]

Two further caveats were highlighted in the context of counting papers. One was that, in a range of technical disciplines from medicine to architecture, the target audience for research results was not only fellow researchers but also practitioners of the relevant technical skill; results were therefore often published in technical rather than research journals, and this should be taken into account in any selective approach to counting published output. The second was the importance of sensitivity to external factors that could influence the rate of output:

> How do you compare the productivity of London-based historians of England ... with those pursuing scholarship under difficulties in foreign lands with inadequate infrastructure and funding? [History]

Other published output

The finding reported in the consultative paper, that 44% of the published output of Earth science departments was of a form other than that of papers in journals scanned by the Science Citation Index, attracted much interest. Respondents in disciplines outside the mainstream natural sciences stressed particularly the importance of including forms of publication other than journal papers.

> A considerable amount of publication in many humanities and some social science subjects takes the form of books or of papers that appear as chapters in books.
>
> A large proportion of published work, especially by established scholars, is in book or book-chapter form. [Social & political sciences]
>
> Valuable work in English can consist of (i) the publication of new knowledge, e.g. of previously unknown documents; (ii) the re-presentation of existing but inadequate knowledge, e.g. the editing of texts; (iii) the interpretation of works of literature; (iv) the promulgation of theory about the subject. [English]

Conference proceedings were regarded as an important form of publication in many disciplines, and one that did not necessarily imply inferior standards of quality control to refereed journals. Some, but not all, respondents thought conference proceedings should be restricted to invited papers, other papers not counting towards total published output.

> Much published material appears in conference proceedings and probably some journals not covered by the major indices. [Computer science]

Relevance of citations to quality

A fundamental objection lodged against the use of citation analysis as a surrogate for quality in the assessment of research was that citations had little or nothing to do with quality.

> Citation analysis measures the number of times a paper is cited by others. That is all. It may give some indication of how widely it has been read, but bears little relation to quality.
>
> There is widespread disquiet concerning the use of citation analysis, fundamentally because no such analysis can distinguish which research has significance as opposed to short-term impact.
>
> There are grave doubts as to what precisely citation analysis purports to measure. ... whilst citation analysis may be used as a 'surrogate measure of quality' it does not actually measure any qualitative aspects of the work.

Of particular concern was the observation that many publications were cited in order to be criticized rather than praised. Nearly every respondent who commented in detail on citation analysis raised this point. It was clearly influential in undermining the proposition that citations could be used as a surrogate measure of quality, and in stimulating opposition to citation analysis as a whole. Similarly, respondents pointed to seminal papers that quickly passed into the common currency of the discipline and ceased to be explicitly cited, and to review articles and papers reporting new techniques which tended to be cited out of proportion to their originality or creativity.

Discipline-specific considerations

In certain disciplines, particular citing (or non-citing) practices were held to vitiate the use of citation analysis in assessment.

> In many, perhaps most, areas of English studies citation is not a standard part of academic discourse to quite the extent it seems to be in Science disciplines. [English]
>
> Citation of earlier work by legal writers performs a role different from that in the sciences. ... It is not the usual practice ... for academic writings which have been used by counsel to be cited in court either in argument or in the judgement. [Law]

The characteristic choices of certain disciplines about where to publish were said to diminish the suitability of those disciplines for citation analysis.

> The citation indices themselves would seem to use mainly Anglo-Saxon sources. Clearly this would put people likely to be cited in non-English speaking countries at a great disadvantage. ... If all that is being said about 1992 is to be taken seriously, we have to come to terms with its multi-lingualism. Hence to assess people on the basis of essentially Anglo-Saxon sources and standing would be politically inept and damaging. [Modern languages]
>
> It would be impossible to scan a sufficient number of foreign and interdisciplinary periodicals to make the citation analysis meaningful in our discipline. [Modern languages]
>
> Where one publishes a paper depends on the intended audience as much as likely acceptance due to quality. [Regional studies]

> In my own smallish department, colleagues have recently written in Spanish, French, Italian, Polish, Russian, Finnish and Japanese, and will doubtless be cited in due course in books in these and other countries. There is little likelihood of many of these citations being recorded in the standard citation indices. [History]

Disciplines where research results were said to be published commonly in professional rather than scholarly journals, and were therefore likely to be misrepresented in citation analyses, included accountancy, architecture, engineering, medicine and planning. The practice of reporting results in confidential reports and various types of grey literature beyond the scan of the citation indices has been discussed above in the context of measuring the volume of research output.

Fashion

There was considerable concern that citation analysis reflected, and in turn would (adversely) influence, social behaviour rather than academic merit. For example, several respondents thought citation analysis measured chiefly the transient swings of fashion, and feared that it would encourage researchers to flit from one fashionable research topic to the next instead of making a sustained effort in a particular area.

> There is a danger that citation measurements can actually harm research by stifling creativity. Younger and less established scholars may be reluctant to explore new or unusual areas of enquiry if this means they get fewer citations.

> Citation is, in many cases, a question of fashion and not necessarily of importance or quality. [Physical science]

> Citations often measure popularity, not necessarily quality. [Engineering]

> People doing original research, not in a fashionable field, are bound to be disadvantaged. [Plant science]

> The system could lead to an overconcentration of effort in some areas of research. Local diseases would suffer from loss of interest, because of lack of 'international interest' in them, thus leading to a University ... neglecting its responsibilities to the local community which helps to support it. [Medicine]

Citation circles and other abuses

A large number of respondents were concerned about possible abuses of citation practice. Many reported anecdotal evidence of citation circles operating in the USA, and feared that similar behaviour would occur in the UK if citation analysis became a normal port of the assessment of academic research.

> We should wish to discourage the development of a habit, to be observed among some younger scholars in other countries, of citing teachers, colleagues and friends merely to improve their citation count. We should abhor a situation which discourages scholars from freely exchanging ideas without thought of enhancement of their own reputation. [Modern languages]

> The widespread anxiety about citation in America has produced results in English studies which this country ought to avoid. Everyone cites everyone else, in the hope of trade-offs in reviews and further articles. ... I don't think it cynical to think that the American author recognizes that

there is little professional advancement to be gained by citing authors who are dead. [English]

There are also some small groups of historians who frequently cite each other's work as they seek to gain currency for their idiosyncratic interpretations of historical problems. [History]

It is unclear whether the proposed citation analysis recognizes and has ways of coping with 'citation circles' (you scratch my back, I'll scratch yours) which are now reputedly common in the USA.

One respondent reported what might be called a non-citation circle:

Competition exists between groups: I know of two major research groups who *never* refer to the work of one another. [Earth science]

Self-citation may be regarded as a special case of a citation circle. The consultative paper suggested that it could be eliminated if desired. Many respondents stated or implied that it should be eliminated, but pointed out that this would be a major task and even so might not greatly improve the value of citation analysis as a performance indicator.

Self-citation or own group citation should be eliminated: but further checks are required which can detect and eliminate cartels of citation.

It would be exceedingly time-consuming to remove self-citation and own group citation; own group citation could never be completely eliminated as people move jobs and it would be impossible to eliminate citation circuses.

We are not sure how self-citation could be eliminated without labour-intensive work, and believe that such citation might be expected to increase if citation analysis was widely used as an indicator.

The possibility of the elimination of self-citation, etc is not an 'advantage' but the mitigation of a disadvantage, and it would be interesting to see the methodology that would be used to achieve it reliably and systematically. Furthermore, has the Subcommittee reflected that more sophisticated methods of increasing one's citations will be developed if there is an advantage to be gained?

There is some scepticism about whether the system can be rendered free of abuse, even if self-citation is excluded.

Not everyone, however, regarded self-citation as necessarily mischievous:

Self-citation may be open to abuse but, equally, it may (and normally will) be a legitimate practice, reflecting continuity in an author's or group's work. Apart from the particular practical difficulties of discounting self-citation, to disregard such instances in toto penalizes researchers who have secured a dominant position in a particular area.

Time lags

The consultative paper suggested that one had to wait at least three years after publication before assessing the citation performance of a publication. This was presented as a disadvantage of citation analysis, because the data then reflected projects initiated five or more years previously, sometimes by people no longer in the department. The time-lag arose because it was often only after three years that the citation performance of typical publications began to be apparent.

This question of the time-lag gave rise to some misunderstanding. One respondent objected to not being able to include citations until three years had lapsed, while another objected to not being able to include citations after three years had lapsed.

> The cut-off point of three years after which citations would cease to be counted is extremely unfair in a subject like mine where there are long publication delays and where important papers continue to have influence for a long period. [Mathematics]

As one respondent pointed out, it is difficult to select a time-lag that is long enough to give a fair representation of the publication's eventual citation performance over its whole life, yet short enough to be relevant to assessment of the current strength of the department. It is also unrealistic to suppose that the time-lag necessary to achieve the first of these objectives will be the same in all disciplines.

> Too long a time scale will fail to pick up changes in the quality of research, but a short-term view, particularly in arts subjects, can omit the very important work which takes a long time to come to fruition.

Respondents provided a range of views as to what time-lag would fairly represent the circumstances of their own disciplines.

> In mathematics, the period from submission of a manuscript to its appearance in print may typically vary between one and two years. ... A three-year citation index may really be measuring citations of a paper in research done within a few months of its publication. [Mathematics]

> The three year time limit would exclude many seminal papers whose importance does not become apparent until some years after publication. Any paper establishing a new field is likely to fall into this category. [Physics]

> A period of 7 to 10 years for citations would be more reasonable. [Engineering]

> ... the microbiological folk whose method of communicating results had to be rapid or it was useless and as a consequence the time-lag was little more than a year. I expect that the high temperature super-conductivity people work similarly.

> Three years ... may reflect an under-estimate of the time required for a fair measure to be taken. [Zoology]

> Recognition of work published after long research projects could be missed by use of the citation index over a period of one or two years. [Medicine]

> It is futile to try to measure the impact of research until several years after publication (i.e. generally a *minimum* of 5 years after the research has been undertaken). [Geography]

> Three years would be too short a time for most papers in historical fields: even five years might be too short. If measures of the quality of current, and not past, research are wanted, citation counts can scarcely provide them. [History]

> There is no necessary reason why the first three years after publication should be particularly significant as far as citations of humanities publications are concerned. One would expect a much longer period in our field. [Modern languages]

> In our field publication of a journal article often takes more than two years. Hence the citation performance of a publication may reflect research work initiated far earlier than the 'five or more' years noted in the paper. [Theology]
>
> A few moments' reflection on the history of music suggests that the whole exercise would be utterly misconceived, particularly if subject to the proposed three-year cut-off point: after all, if we are really talking about quality we are talking about enduring value. [Music]
>
> The proposed period would be much too short. We would suggest a minimum of ten years but would prefer no time limit.

Practical difficulties

Respondents also raised a range of practical difficulties that would have to be resolved if citation analysis was to be a viable component of research assessment. Particularly, but by no means only, in the social sciences and humanities, the focus in citation work on journals to the exclusion of books, reports, conference papers and various forms of grey literature was regarded as unacceptable. For the same reasons as explored above in the context of using publications as a measure of the volume of research output, it was emphasized that citation analysis had to be sensitive to specific dissemination practices if it was to be acceptable to individual disciplines.

Even within the context of journal publications, there was concern about the coverage of the various citation indices.

> Architectural journals are not covered by any citation index that we know. [Architecture]
>
> The Social Science Citation Index is regarded as unrepresentative of much published work, unreliable and difficult to use.
>
> A brief examination of the Science and Social Science Citation Indices confirms that many influential journals are not scanned. [Business studies]
>
> The coverage of the Citation Index declines as one moves away from science. [Social science]
>
> Law is not covered by any of the existing citation indexes. ... Legal research is not simply consumed and assessed within the academic legal journals. [Law]
>
> The large number of publications *not* scanned by the Institute of Scientific Information means that published citation indices have little value in a discipline such as ours. [History]
>
> The Arts and Humanities Citation Index provides even less comprehensive coverage of publications in the field of English than the figure of 44% quoted for citation indices applicable to Earth sciences. [English]
>
> The Arts and Humanities Citation Index is unlikely to be comprehensive in fields such as the theory of literary criticism, where new journals are continually appearing. [French]

The coverage of the citation indices was not the only practical problem impeding achievement of the minimum standards of comprehensiveness and accuracy necessary for citation analysis to be

a viable element in research assessment. Elementary accuracy was highlighted by several respondents:

> Two respondents to my call for comments (named Jones and Smith) draw attention to the risk of misattribution. [Modern languages]

The 'first-named author' problem gave rise to a fair amount of disquiet.

> Citation indices are prepared by reference to the first-named author. If all the authors are from the same research group this may not matter too much, but for any paper produced by collaboration across different research groups this produces serious inaccuracies.

In many disciplines respondents reported that departments were typically too small to produce a statistically analysable number of publications or citations. That was not to say that they lacked the critical mass to discharge their teaching and research roles—merely that they lacked the critical mass for a particular form of assessment. This caused considerable concern.

A further cause for concern was the cost of carrying out citation analysis at a worthwhile level of accuracy and sensitivity. This was almost universally regarded as high, both in cash terms and in terms of the time that would be lost to research.

Summary

Overall, then, there was not much support for using citation analysis in the assessment of departmental research. As one respondent delicately put it:

> From certain viewpoints citation indexes appear to offer attractive possibilities. Nevertheless, we do not feel that citation analysis is sufficiently sophisticated to bear the weight of interpretation which analysts, in the absence of a numerical index for quality, might be tempted to place upon it.

The discussion of citation analysis in the consultative paper concluded: 'It is therefore open to discussion whether the insights generated by citation analyses are worth the effort and cost of collecting and analysing the data'. Many respondents picked this up and answered it unequivocally.

> The general consensus in the University, ranging from science and technology to humanities, is that the negative points in the Subcommittee's report on citation analysis outweigh the positive.
>
> It is most unlikely that the degree of acceptance of citation counts as a valid measure of research quality will justify the vast resources needed to produce them. [Mathematics]
>
> Overall, citation analysis has very little relevance to engineering and will certainly not quantify accurately what it is trying to measure. [Engineering]
>
> The disadvantages listed ... far outweigh any useful meaning that could possibly be obtained. [Biochemistry]
>
> The disadvantages of citation analysis are clearly recognized in the paper and ... are *not* worth the effort and cost of collection, which would be better spent on research time and resources. [Geography]

> ... extremely cautious about any extended use of citation analysis. In inexperienced hands it could be an extremely damaging technique, and even in very professional ones its utility is likely to be rather marginal. [Government]
>
> The academic legal community is uniformly hostile to citation analysis and sees no place for it in assessing the quality of research in university law schools. [Law]
>
> Citation analysis ... is flawed to the point of being both misleading and inherently absurd. [History]
>
> The paper recognizes the limitations of citation analysis. ... But these disadvantages do not merely *reduce* the value of such analysis; they *destroy* it. It's not as if 40% of the apple crop is infected, and we can use the other part of it. We know that each and every apple is up to 40% infected, so the whole crop must be rejected. [Philosophy]
>
> The 'advantages' listed are questionable, the 'disadvantages' decisive. [Arts]

Alternative strategies

Despite all these difficulties, some respondents recognized that the pressures to develop ways of looking at the output of research, and of going beyond merely trying to measure its volume, had nevertheless to be faced. Two broad strategies were proposed, based respectively on bibliometric indicators other than basic citation analysis and on variants of peer review.

The bibliometric approaches that were proposed were, implicitly or explicitly, based for the most part on the notion summarized in the term 'journal impact factor'. This starts from the premise that one can tell something about the quality of a paper by the company it keeps, i.e. by the repute of the journal in which it is published. The 'impact factor' approach, and its more elaborate descendant the 'journal influence measure', attempt to quantify this repute by reference to the typical citation performance of papers in that journal. A less mechanistic variant involves expert judgement of which are the key journals. These notions found some support, though it should be noted that they can be applied to assessment only of papers in journals and not of other forms of research output.

> We would like to see more weight given to the type of journal in which publications appear, since this is in practice the measure that scientists, at least, apply in assessing quality.
>
> The quality of journals is important. It would not be difficult for the medical sciences to develop a simple classification system for the quality of journals. [Medicine]
>
> Classification of journals into leagues (on the basis of their selectivity in what they publish, the esteem in which they are held and their impact factor) would be a better tool than citation analysis. The classification would have to be revised every 2/3 years. [Medicine]
>
> The particular journal in which a paper is published must be seen as an important measure of the quality of the paper. [Psychiatry]

One respondent suggested that the way forward lay in developing conventional citation analysis.

> ... generally in favour of the use of citation analysis ... hoped that the UFC would seek an improved, more European based, system which could include chapters from books in addition to journal articles. [Psychology]

The second broad strategy was to develop peer review. Respondents' views on the central place of peer review in assessment of research have been discussed in chapter 3 above: the strong consensus was that no amount of bibliometric endeavour could replace peer review. There was no acceptable substitute to consulting experts able at first hand to judge the quality of the output of a department.

> It is not easy to quantify quality. But if there has to be such an attempt, the judges and/or their referees will have to *read* and evaluate the publications. Subjective judgements are inescapable. [Biotechnology]

> Quantitative methods cannot replace qualitative ones. Gains in facility or in objectivity are illusory. If the job is to be done it should be done properly, which requires specialist qualitative assessment. [Philosophy]

> To be of any value as an indicator, the citations would have to be interpreted by specialists, in which case they may as well be abandoned in favour of the present system of periodic peer review. [History]

> Peer review is not perfect, but it seems preferable to citation indices. [Social administration]

CHAPTER 6: ESTEEM INDICATORS

Purpose of esteem indicators

The consultative paper suggested that an additional source of data relevant to assessment of departmental research might be found in esteem indicators—a set of variables that could be interpreted as measures of the esteem in which a department as a whole or its individual members were held. Four illustrative examples of esteem indicators were given: election to learned societies, major prizes, visiting scholars and journal editorships.

Reaction to this suggestion was mixed, with a majority of respondents expressing support to a greater or lesser extent and a significant minority being opposed. Instances of support and of opposition were to be found across all disciplines.

Support in principle for esteem indicators

Those who favoured the use of esteem indicators commonly regarded them as variants of peer review—measures, or at least concrete expressions, of the esteem that individuals were accorded by their peers. Support for esteem indicators sometimes reflected negative responses to the alternative approaches to quantitative assessment.

> To include measures of esteem that effectively quantify the judgements of the peer group seems entirely reasonable.
>
> Esteem indicators may perform a useful function to guide and challenge informal opinion but could represent a hazardous and ambiguous basis for untutored assessment. ... Nevertheless, as an adjunct to peer review and in the context of detailed examinations of particular subjects or institutions, esteem indicators could represent a significant set of data.
>
> The use of citation indices is problematic ... I am more in favour of esteem indicators. [Computer science]
>
> In our field esteem indicators would be more appropriate than citation analysis. [Social science]
>
> It is on the category of esteem indicators that we would wish to place much more emphasis (than on bibliometrics) when historians are to be assessed. [History]
>
> We are in sympathy with the broad aims behind the paper. ... In particular, we welcome the proposed use of esteem indicators. [Humanities]
>
> We wish to urge that the section on esteem indicators be extended to the point where it takes over in significance for the Humanities from the section on citation analysis. [English]
>
> The only true way in which our standing as scholars can be assessed is by the esteem of our peers. The indicators listed in the consultative document must be regarded as one means of measuring this esteem. [Modern languages]

Many respondents implicitly indicated their support for the principle of esteem indicators by suggesting additions to the list of possible indicators. These suggestions will be discussed below.

Opposition in principle

A significant number of respondents, however, expressed opposition in principle to the construction and use of esteem

indicators. Two who already had experience of handling esteem indicators were not enthusiastic.

> We gather this data for our internal assessment exercise. Its interpretation is even more problematic than for publications data.

> The University already uses certain esteem indicators in its departmental profiles. ... In practice, we feel that this information is of relatively little value, and that certain problems of definition exist. Nonetheless, we propose for the time being to continue to include it in the profiles.

One strong objection raised against esteem indicators was that they reflected past performance—sometimes distant past—rather than present endeavour or, still less, future potential. They were thus of limited value in decision-making.

> The esteem indicators mentioned in the paper did not commend themselves to us as acceptable—often such accolades come only when an individual is past the peak of his or her research career.

> Esteem indicators seem to us as the worst possible way of assessing an institution. ... Such awards tend to go to people who have reached at least the mid-point in their career. [Earth science]

> All such esteem relates to past reputation rather than current activity. [Science]

A second strand of objection was that the social allocation of esteem was controlled by those who had themselves already benefited from the process: there was no open market to which all had equal opportunity of access.

> Less difficult to quantify is the undoubted feeling that FRS's tend to be awarded to some extent on the basis of who you know so that pockets develop in certain Universities which are not necessarily better than others. In this as in many other things there appears to be a bias towards the south of England. [Chemistry]

> In general esteem indicators act much as peer review, and may be subject to the same abuses, in particular the possibility of 'patronage'.

> It would be disturbing if 'membership of learned societies where admission is by competitive election' were to be counted unless these societies were to operate a system where anyone is free to apply for election. [Mathematics]

A respondent from one of the ancient universities argued, as an objection, that esteem indicators would be likely to work in favour of the ancient universities.

Although esteem indicators look quantitative at first sight, the reality is more complex. Several respondents pointed to the difficulties this would pose to any attempt to use esteem indicators systematically as a performance indicator.

> Esteem indicators can never be adequately quantified and we support the practice of the UFC in the recent research selectivity exercise to provide a separate section in which each unit of assessment was able to indicate what it thought were the measures of esteem in which the department as a whole and its individual members were held.

> How many journal editors are the equivalent of one FRS? [Physical science]
>
> Esteem indicators of the kind mentioned would run into problems in our disciplines. In order for this sort of analysis to be at all comprehensive, a wide range of indicators would have to be used and the problems of weighting different indicators would be considerable. [Social sciences]
>
> There would remain, of course, the problem of attaching weightings to these activities. [History]

One respondent argued that, if esteem indicators were effectively a quantification of peer review, they constituted double counting and were redundant.

> By esteem indicators do we mean—is one a member of the establishment? Why should it matter if one is a member of a learned society or a journal editor? Being a member is a reflection of professional activities and presumably one therefore has an esteem which can be formulated into resource allocation by independent peer review. Why multiply this effect? [Biochemistry]

Several respondents commented that one man's esteem indicator might be another man's chore.

> Some 'esteem indicators', such as journal editorship or committee membership, may be seen as chores by some scholars, to be undertaken in rotation under pressure.
>
> I myself have been offered two major journal editorships in the last two years and have turned these down because of pressure of other work. [Mathematics]
>
> Editorships are subject to many variables and are subject to individual choices about how time is best spent. Many would view an editorship as a chore rather than a distinction. [Social studies]

One respondent warned:

> There has been little or no systematic assessment of the reliability and validity of esteem indicators, nor indeed of whether one can compile the necessary information in a form that is truly comparable across departments and institutions. [Social science]

A number of respondents were bluntly dismissive of the whole notion of esteem indicators.

> Esteem indicators are not worthwhile as these are crude measures and could result in few individuals in the academic world as a whole being quoted.
>
> The esteem measures were felt to be implausible, and of little value. [Computer science]
>
> I do not consider that the esteem indicators give a reliable measure of a department's work. [Physics]
>
> There was particular dismay expressed at 'esteem indicators'. The criteria seemed to bear no relationship whatever to the reality of arts research. [Arts]

Practical considerations

Some specific points in addition to those already mentioned were identified with particular candidate esteem indicators.

> There are too few FRSs and prizes for the statistical uncertainties to have averaged out. [Physics]
>
> Titles such as FBA are still too quirkish in many areas to be an acceptable guide. [Social studies]
>
> Prizes are few in the field of humanities and social science and in the former tend to be restricted to creative work. [Social studies]
>
> Election to learned societies is a valid yardstick, but all societies where election depends on scholarly excellence should be included. [History]
>
> There are few literary prizes and no learned society to which members of the English departments might aspire. [English]
>
> These are of dubious relevance to arts-area Schools; an FRAS is a very different from an FRS; major prizes for research are so rare as to be an endangered species; editorships frequently measure the good fortune of academics employed by those institutions whose finances enable them to sustain journals. [Arts]

Additional measures

Whatever the misgivings that some respondents felt about the principle or the practicality of esteem indicators, many brought forward their own suggestions for possible indicators. Those mentioned by several independent respondents included the following, though it was recognized that no given indicator was necessarily relevant to all disciplines.

- involvement in various international organizations
- appointment as coordinators for international research studies
- active involvement in national organizations such as advisory committees, boards of management, working groups, learned societies, government bodies
- participation in peer review and the commissioning and funding of research
- refereeing for funding agencies, journals or publishers
- external assessing for academic promotion
- external examining
- consultancies and other links with industry
- measures of the satisfaction that senior customers felt about the work they commissioned from individuals and groups
- success in securing research grants and other external research income
- numbers of research students
- editorship of book series
- reviewing for the quality media
- invited papers and chairmanship of panels at national and international conferences
- invitations to give distinguished lectures and to be a visiting professor in the UK or abroad

– award of honorary degrees

Summary

The prevailing attitude to esteem indicators was summarized by one respondent in the following terms.

> Esteem indicators seem peripheral and likely to relate to individuals rather than the sum of research produced by a unit of assessment. They need to be taken into account, all the same. [Modern languages]

> ... extremely cautious about any extended use of citation analysis. In inexperienced hands it could be an extremely damaging technique, and even in very professional ones its utility is likely to be rather marginal. [Government]

> The academic legal community is uniformly hostile to citation analysis and sees no place for it in assessing the quality of research in university law schools. [Law]

> Citation analysis ... is flawed to the point of being both misleading and inherently absurd. [History]

> The paper recognizes the limitations of citation analysis. ... But these disadvantages do not merely *reduce* the value of such analysis; they *destroy* it. It's not as if 40% of the apple crop is infected, and we can use the other part of it. We know that each and every apple is up to 40% infected, so the whole crop must be rejected. [Philosophy]

> The 'advantages' listed are questionable, the 'disadvantages' decisive. [Arts]

Alternative strategies

Despite all these difficulties, some respondents recognized that the pressures to develop ways of looking at the output of research, and of going beyond merely trying to measure its volume, had nevertheless to be faced. Two broad strategies were proposed, based respectively on bibliometric indicators other than basic citation analysis and on variants of peer review.

The bibliometric approaches that were proposed were, implicitly or explicitly, based for the most part on the notion summarized in the term 'journal impact factor'. This starts from the premise that one can tell something about the quality of a paper by the company it keeps, i.e. by the repute of the journal in which it is published. The 'impact factor' approach, and its more elaborate descendant the 'journal influence measure', attempt to quantify this repute by reference to the typical citation performance of papers in that journal. A less mechanistic variant involves expert judgement of which are the key journals. These notions found some support, though it should be noted that they can be applied to assessment only of papers in journals and not of other forms of research output.

> We would like to see more weight given to the type of journal in which publications appear, since this is in practice the measure that scientists, at least, apply in assessing quality.

> The quality of journals is important. It would not be difficult for the medical sciences to develop a simple classification system for the quality of journals. [Medicine]

> Classification of journals into leagues (on the basis of their selectivity in what they publish, the esteem in which they are held and their impact factor) would be a better tool than citation analysis. The classification would have to be revised every 2/3 years. [Medicine]

> The particular journal in which a paper is published must be seen as an important measure of the quality of the paper. [Psychiatry]

One respondent suggested that the way forward lay in developing conventional citation analysis.

> ... generally in favour of the use of citation analysis ... hoped that the UFC would seek an improved, more European based, system which could include chapters from books in addition to journal articles. [Psychology]

The second broad strategy was to develop peer review. Respondents' views on the central place of peer review in assessment of research have been discussed in chapter 3 above: the strong consensus was that no amount of bibliometric endeavour could replace peer review. There was no acceptable substitute to consulting experts able at first hand to judge the quality of the output of a department.

> It is not easy to quantify quality. But if there has to be such an attempt, the judges and/or their referees will have to *read* and evaluate the publications. Subjective judgements are inescapable. [Biotechnology]

> Quantitative methods cannot replace qualitative ones. Gains in facility or in objectivity are illusory. If the job is to be done it should be done properly, which requires specialist qualitative assessment. [Philosophy]

> To be of any value as an indicator, the citations would have to be interpreted by specialists, in which case they may as well be abandoned in favour of the present system of periodic peer review. [History]

> Peer review is not perfect, but it seems preferable to citation indices. [Social administration]

CHAPTER 6: ESTEEM INDICATORS

Purpose of esteem indicators

The consultative paper suggested that an additional source of data relevant to assessment of departmental research might be found in esteem indicators—a set of variables that could be interpreted as measures of the esteem in which a department as a whole or its individual members were held. Four illustrative examples of esteem indicators were given: election to learned societies, major prizes, visiting scholars and journal editorships.

Reaction to this suggestion was mixed, with a majority of respondents expressing support to a greater or lesser extent and a significant minority being opposed. Instances of support and of opposition were to be found across all disciplines.

Support in principle for esteem indicators

Those who favoured the use of esteem indicators commonly regarded them as variants of peer review—measures, or at least concrete expressions, of the esteem that individuals were accorded by their peers. Support for esteem indicators sometimes reflected negative responses to the alternative approaches to quantitative assessment.

> To include measures of esteem that effectively quantify the judgements of the peer group seems entirely reasonable.
>
> Esteem indicators may perform a useful function to guide and challenge informal opinion but could represent a hazardous and ambiguous basis for untutored assessment. ... Nevertheless, as an adjunct to peer review and in the context of detailed examinations of particular subjects or institutions, esteem indicators could represent a significant set of data.
>
> The use of citation indices is problematic ... I am more in favour of esteem indicators. [Computer science]
>
> In our field esteem indicators would be more appropriate than citation analysis. [Social science]
>
> It is on the category of esteem indicators that we would wish to place much more emphasis (than on bibliometrics) when historians are to be assessed. [History]
>
> We are in sympathy with the broad aims behind the paper. ... In particular, we welcome the proposed use of esteem indicators. [Humanities]
>
> We wish to urge that the section on esteem indicators be extended to the point where it takes over in significance for the Humanities from the section on citation analysis. [English]
>
> The only true way in which our standing as scholars can be assessed is by the esteem of our peers. The indicators listed in the consultative document must be regarded as one means of measuring this esteem. [Modern languages]

Many respondents implicitly indicated their support for the principle of esteem indicators by suggesting additions to the list of possible indicators. These suggestions will be discussed below.

Opposition in principle

A significant number of respondents, however, expressed opposition in principle to the construction and use of esteem

indicators. Two who already had experience of handling esteem indicators were not enthusiastic.

> We gather this data for our internal assessment exercise. Its interpretation is even more problematic than for publications data.
>
> The University already uses certain esteem indicators in its departmental profiles. ... In practice, we feel that this information is of relatively little value, and that certain problems of definition exist. Nonetheless, we propose for the time being to continue to include it in the profiles.

One strong objection raised against esteem indicators was that they reflected past performance—sometimes distant past—rather than present endeavour or, still less, future potential. They were thus of limited value in decision-making.

> The esteem indicators mentioned in the paper did not commend themselves to us as acceptable—often such accolades come only when an individual is past the peak of his or her research career.
>
> Esteem indicators seem to us as the worst possible way of assessing an institution. ... Such awards tend to go to people who have reached at least the mid-point in their career. [Earth science]
>
> All such esteem relates to past reputation rather than current activity. [Science]

A second strand of objection was that the social allocation of esteem was controlled by those who had themselves already benefited from the process: there was no open market to which all had equal opportunity of access.

> Less difficult to quantify is the undoubted feeling that FRS's tend to be awarded to some extent on the basis of who you know so that pockets develop in certain Universities which are not necessarily better than others. In this as in many other things there appears to be a bias towards the south of England. [Chemistry]
>
> In general esteem indicators act much as peer review, and may be subject to the same abuses, in particular the possibility of 'patronage'.
>
> It would be disturbing if 'membership of learned societies where admission is by competitive election' were to be counted unless these societies were to operate a system where anyone is free to apply for election. [Mathematics]

A respondent from one of the ancient universities argued, as an objection, that esteem indicators would be likely to work in favour of the ancient universities.

Although esteem indicators look quantitative at first sight, the reality is more complex. Several respondents pointed to the difficulties this would pose to any attempt to use esteem indicators systematically as a performance indicator.

> Esteem indicators can never be adequately quantified and we support the practice of the UFC in the recent research selectivity exercise to provide a separate section in which each unit of assessment was able to indicate what it thought were the measures of esteem in which the department as a whole and its individual members were held.

> How many journal editors are the equivalent of one FRS? [Physical science]
>
> Esteem indicators of the kind mentioned would run into problems in our disciplines. In order for this sort of analysis to be at all comprehensive, a wide range of indicators would have to be used and the problems of weighting different indicators would be considerable. [Social sciences]
>
> There would remain, of course, the problem of attaching weightings to these activities. [History]

One respondent argued that, if esteem indicators were effectively a quantification of peer review, they constituted double counting and were redundant.

> By esteem indicators do we mean—is one a member of the establishment? Why should it matter if one is a member of a learned society or a journal editor? Being a member is a reflection of professional activities and presumably one therefore has an esteem which can be formulated into resource allocation by independent peer review. Why multiply this effect? [Biochemistry]

Several respondents commented that one man's esteem indicator might be another man's chore.

> Some 'esteem indicators', such as journal editorship or committee membership, may be seen as chores by some scholars, to be undertaken in rotation under pressure.
>
> I myself have been offered two major journal editorships in the last two years and have turned these down because of pressure of other work. [Mathematics]
>
> Editorships are subject to many variables and are subject to individual choices about how time is best spent. Many would view an editorship as a chore rather than a distinction. [Social studies]

One respondent warned:

> There has been little or no systematic assessment of the reliability and validity of esteem indicators, nor indeed of whether one can compile the necessary information in a form that is truly comparable across departments and institutions. [Social science]

A number of respondents were bluntly dismissive of the whole notion of esteem indicators.

> Esteem indicators are not worthwhile as these are crude measures and could result in few individuals in the academic world as a whole being quoted.
>
> The esteem measures were felt to be implausible, and of little value. [Computer science]
>
> I do not consider that the esteem indicators give a reliable measure of a department's work. [Physics]
>
> There was particular dismay expressed at 'esteem indicators'. The criteria seemed to bear no relationship whatever to the reality of arts research. [Arts]

Practical considerations

Some specific points in addition to those already mentioned were identified with particular candidate esteem indicators.

There are too few FRSs and prizes for the statistical uncertainties to have averaged out. [Physics]

Titles such as FBA are still too quirkish in many areas to be an acceptable guide. [Social studies]

Prizes are few in the field of humanities and social science and in the former tend to be restricted to creative work. [Social studies]

Election to learned societies is a valid yardstick, but all societies where election depends on scholarly excellence should be included. [History]

There are few literary prizes and no learned society to which members of the English departments might aspire. [English]

These are of dubious relevance to arts-area Schools; an FRAS is a very different from an FRS; major prizes for research are so rare as to be an endangered species; editorships frequently measure the good fortune of academics employed by those institutions whose finances enable them to sustain journals. [Arts]

Additional measures

Whatever the misgivings that some respondents felt about the principle or the practicality of esteem indicators, many brought forward their own suggestions for possible indicators. Those mentioned by several independent respondents included the following, though it was recognized that no given indicator was necessarily relevant to all disciplines.

- involvement in various international organizations
- appointment as coordinators for international research studies
- active involvement in national organizations such as advisory committees, boards of management, working groups, learned societies, government bodies
- participation in peer review and the commissioning and funding of research
- refereeing for funding agencies, journals or publishers
- external assessing for academic promotion
- external examining
- consultancies and other links with industry
- measures of the satisfaction that senior customers felt about the work they commissioned from individuals and groups
- success in securing research grants and other external research income
- numbers of research students
- editorship of book series
- reviewing for the quality media
- invited papers and chairmanship of panels at national and international conferences
- invitations to give distinguished lectures and to be a visiting professor in the UK or abroad

– award of honorary degrees

Summary

The prevailing attitude to esteem indicators was summarized by one respondent in the following terms.

> Esteem indicators seem peripheral and likely to relate to individuals rather than the sum of research produced by a unit of assessment. They need to be taken into account, all the same. [Modern languages]

CHAPTER 7: CONCLUSIONS

(i) Introduction

Many questions can be asked about quantitative methods of assessing departmental research performance. This concluding chapter will focus on two. In the light of the responses received from a wide segment of the UK academic community, are quantitative methods valid? And are they useful?

These questions are important. There is little point in investing time and effort in assessing performance unless the results of the assessment are to play some role in decision-making processes, for example about structures or about resource allocation. If the decisions are to secure the general assent of the community of researchers as a whole, it is prudent to seek, and to take at least some account of, their views on the processes by which the decisions are taken. That is the objective of the consultation analysed in this paper.

That said, neither the CVCP/UFC joint committee nor any other body is bound by the results of this consultation as presented here. A different author might have produced a different analysis; external circumstances may alter the pertinence of respondents' comments; and, as every researcher knows, the prevailing majority view is not always right. However, the following conclusions do deserve to be taken seriously.

Respondents were generally willing to recognize that some form of assessment was necessary. With varying degrees of enthusiasm or resignation, most were prepared at least to explore what quantitative methods could contribute to assessment. Hostility (as opposed to reasoned opposition) against quantitative methods, where it was found, arose less from antagonism towards assessment as such than from fear that quantitative methods would displace peer review, in comparison with which they were regarded as a very blunt instrument. Nearly all respondents stressed that quantitative methods should be regarded as an input to peer review, not a substitute for it. Indeed, one of the most striking outcomes of the consultation was the spontaneous affirmation of the value of peer review.

(ii) Validity

Definition

Are quantitative methods valid? For this paper, that question translates into whether quantitative methods can give consistent, reproducible and accurate (i.e. reasonably compatible with the judgement of most informed observers) information about the research performance of university departments. The question can be approached in terms of specific quantitative methods and the specific aspects of research performance to which they are claimed to relate. There are, first, some more general considerations to address.

Departments ...

For example, what is a department? This is not as trite as it sounds. Any evaluation exercise must, of course, define its terms, for example whether research students, visiting academics or retired but still active staff should be included in the definition of 'department'. One must also consider how to deal with staff joining the department during the

period being reviewed: should their earlier work be credited to their new department, which did not contribute to it but now has a stake in the total career history of the new member of staff? This last point is particularly relevant to esteem indicators, which tend to reflect cumulative career achievements rather than immediate history.

... or individuals

But beyond these points of detail lies the issue of whether it is valid to talk about departments at all. Departments are administrative and financial units (cost centres or parts of cost centres), and act collectively to provide research infrastructure and to discharge their teaching responsibilities. However, to a degree that varies from one discipline or subdiscipline to another, research is done by individual researchers, not by departments. Many of the concerns expressed by respondents about the validity of quantitative methods stemmed from the problems attached to their use for evaluating individuals. The normal response is that the risks of inaccuracy are much reduced when dealing with aggregations of individuals. It may be that the validity of this response depends on the extent to which it is appropriate to talk of research as a departmental rather than an individual activity: a certain level of teamwork within the department may be necessary before it is valid to apply at the departmental level assessment methodologies that one would not apply at the individual level.

The degree of teamwork also affects perceptions of how one should assess collaborative research, especially if departments have different traditions of favouring internal or external collaboration.

Overall departmental performance, as measured by the standard quantitative techniques, is the sum or average of the performances of individual members of the department. Several respondents questioned the validity of such an averaging process in cases where individual performances varied widely, for example where significant numbers of staff had chosen to focus their energies on activities other than research. Quantitative techniques should, ideally, be able to distinguish non-researchers from poor researchers.

Critical mass

If it is, in principle, valid to apply quantitative methods to departments that one would not apply to individuals, the degree of that validity will increase with the size of the department. It is far from clear what minimum size of department is needed to give a reasonable degree of validity, and this will anyway vary with discipline and with the particular indicator under consideration. But for bibliometric methods in the basic sciences (the most thoroughly explored case), it is likely that many departments in UK universities currently fall below the critical threshold of size.

Publications

The validity of any given quantitative technique depends crucially on the discipline concerned. All quantitative techniques of research assessment involve measurement of surrogates: validity depends on how closely the surrogate matches the relevant aspect of performance. Most respondents accepted, for example, that data on the volume of published output had a place in assessment of research, i.e. that it was valid to count publications. But what forms of publications should be counted, how they should be counted and what else should be counted varied greatly from one discipline to another.

Citation analysis

The validity of citation analysis in assessment of departmental research was much more controversial: only a few respondents, even in the basic sciences, gave it much credence. The reasons for rejecting citation analysis as invalid included the specific referencing traditions of individual disciplines, the many non-approbationary functions of citations and the inadequacy of the databases available for counting citations. Citation analysis may have a role in analysis of comparative trends in national performance in basic science, but its use at departmental level found little favour among respondents. Even the more modest claim for citation analysis as a surrogate measure of impact (rather than quality) had relatively few supporters.

Esteem indicators

Esteem indicators proved to be more widely acceptable as valid, provided individual disciplines were allowed to specify what should be included as an indicator of esteem. Systematic use of such indicators would not be straightforward, partly because one man's esteem indicator was another man's chore (e.g. journal editorships) and partly because aggregation of an inevitably disparate set of measures would be close to impossible.

(iii) Usefulness

Context

The Research Indicators Subcommittee has been examining two distinct issues. One is the production of annual data on the outputs of university research, to go with existing annual statistical series on other facets of university life such as finance, staff and students and first destinations of graduates. The other is the production of data to assist in detailed periodic reviews such as the UFC selectivity exercises, carried out at intervals of several years. The usefulness of quantitative methods has to be considered in these two separate contexts. As with the question of validity, issues both of principle and of detail can be raised.

Peer review

It is nowhere seriously suggested that quantitative methods can or should replace peer review. Their claimed role is two-fold: to strengthen the peer review process by making available to peer reviewers as much relevant information as possible, and to render the peer review process more open to scrutiny by outside groups, including those being reviewed. Respondents were lukewarm in their estimates of how usefully quantitative methods could fulfil these roles.

Some respondents saw little use for quantitative methods in this context. At best they gave the same message as the traditional peer review process (was this not the test of their validity?), in which case they were redundant; at worst they gave a conflicting message, in which case they would be ignored.

Other respondents examined in greater detail the claim that quantitative methods were useful because they injected an element of objectivity into peer review. There was some sympathy for this claim, but it was not unlimited. It was pointed out that subjectivity entered into the selection of which quantitative measures to use, into the aggregation of the messages coming from different indicators and, particularly, into the way these messages were fed into the judgement-forming process. The claim for objectivity is central to the

attractiveness of performance indicators; respondents generally thought that claim should be handled with some care.

The same goes for the claim that quantitative methods can help to democratize peer review by making it more open. If peer review ultimately depended on the judgement of informed peers—and respondents would not have it any other way—then there were inherent limits on how open it could be. Indeed, it could be argued that quantitative methods are more useful in defending judgements than in influencing them.

The experience of the Earth sciences exercise was that bibliometrics clearly identified the strongest, and the weakest, 15% or so of departments—which were already well known anyway. Bibliometrics proved much less incisive in ordering the middle 70% of departments, which is where the difficult and controversial decisions have to be made.

Time lags

One characteristic limiting the usefulness of all quantitative methods to a greater or lesser extent is their lack of timeliness. Quantitative methods of research assessment tend to measure things that have already happened. There can be delays of many months between completion of a piece of research and its appearance in print; further delays (notwithstanding oral dissemination, conferences and preprints) before results are incorporated in other research and publicly cited; and yet further delays before measurable esteem accrues to the original researcher. These time lags mean that quantitative methods may be measuring the performance of a group of people very different from those now in the department. Many respondents argued that the problem of time lags severely diminished the usefulness of quantitative methods. Some discussed ways of reducing the time lags, for example by using journal impact factors instead of counting actual citations, but this did not greatly alleviate the problem.

Opportunity cost

A more practical, and to respondents equally visible, problem with quantitative methods was the time taken to produce them—the opportunity cost in terms of time taken from research. A key feature of the quantitative approach is that, by involving the academics concerned in generating the data, it enables them to have an input to the assessment process. Many respondents, however, were clearly worried about the demands this would impose on their time. Some institutions, on the other hand, already routinely collected publication and esteem (though not citation) data, for example for use in annual reports, so this problem may be more perceived than real.

Non-verbal outputs

Several respondents, often in engineering but also in other disciplines, stressed the need to take account of non-verbal outputs of research. This was important for the departments themselves; it was also important for actual or potential customers for the departments' work. Exclusive emphasis on verbal output would be detrimental to departments that saw their research mission as focused on the needs of, for example, industrialists wanting a functioning model or audiences wanting a live performance. Even where the initial output was in a verbal form, account should be taken of any subsequent non-verbal outputs.

Annual and periodic use

Usefulness, as mentioned earlier, depends on context. For periodic assessments (such as the UFC selectivity exercises, carried out at intervals of several years), relatively subtle and complex methods may be used. The infrequency of these exercises, coupled with the fact that they lead directly to strategic decisions of long-term importance, justifies a relatively high investment of time and effort; and because periodic assessments are generally carried out by groups of peers with relevant first-hand experience, data requiring expert interpretation may safely be used. Annually published data, on the other hand, must be reasonably cheap to collect, must be fairly stable from year to year and, above all, must be interpretable by outsiders. Respondents were greatly concerned that annually published data could be used by those without a proper understanding of the complexities of the situation to substantiate simplistic—and by implication detrimental—judgements. They were not unconcerned about the possible misuse of periodic data, but felt that it might be easier to control.

Publications

The Research Indicators Subcommittee has been conducting trials across a variety of disciplines to assess the feasibility, and usefulness, of collecting annual publications data. The responses to the present consultation exercise suggest that such an exercise might be useful, provided individual disciplines could determine which sorts of publications were included and provided any published outcome included, loud and clear, all necessary caveats about interpretation. Once the series was established, the use of three-year rolling averages would help to smooth out misleading year-to-year fluctuations. Cumulative totals would provide a useful input to periodic assessments.

Citation analysis

There was virtually no support for the inclusion of citations in annual data series and, in view of the earlier comments about validity, little support for citation analysis as a useful input to periodic assessments. Respondents sent a firm message that, at departmental level, citation analysis added nothing to peer review. The only way to assess the quality of published departmental output was to invite the department to select representative items and then for the reviewing panel to read them.

Esteem indicators

The disparate nature of esteem indicators militated against the usefulness of their inclusion in annual data series. However, they were regarded as a useful input to periodic assessments. They should therefore be collected systematically (i.e. annually) so as to be readily available when needed.

(iv) Conclusion

Usefulness

One outcome of this consultation, then, is to dispel any notion that quantitative methods provide a quick fix to problems of assessment of departmental research. Under certain circumstances some quantitative methods can be useful; but they are not particularly quick and they do not provide an easy fix. They are most useful, and most acceptable to academic researchers, when employed as an input to periodic peer review.

Cross-disciplinary comparisons

Nor can quantitative methods be used directly to compare different disciplines. Many respondents, and not only in the humanities, were worried that they would be faced with a set of indicators inappropriate to their own disciplines. It is evident that no set of indicators of research performance can be validly applied across all disciplines. Sensitivity to the specific circumstances of individual disciplines is more important than superficial uniformity.

Impact on academic behaviour

Finally, many respondents voiced anxieties that the introduction of performance indicators would distort academic behaviour, for example by encouraging excessive publication at the expense of quality or short-term fashion at the expense of long-term importance. A key aim of quality control mechanisms is, of course, precisely that of influencing behaviour, by clarifying what is most likely to reap rewards. In introducing new, additional methods of assessing departmental research it is therefore important to consider carefully whether the behaviour likely to be encouraged by those methods is consistent with the academic mission of advancing and disseminating knowledge.

ANNEX A

MEMBERSHIP OF THE CVCP/UFC SUBCOMMITTEE ON RESEARCH INDICATORS

Chairman: Professor M.B. Harris, Vice-Chancellor, University of Essex

Members: Professor E.W. Abel, Professor of Inorganic Chemistry, University of Exeter
Mr S.R. Bosworth, Registrar, University of Salford
Dr P.M.D. Collins, Director, Science & Engineering Policy Studies Unit
Mr B.D. Cullen, DES
Mr K.S. Davies, Principal Assistant Secretary, CVCP
Professor M. Hart, FRS, Professor of Physics, University of Manchester
Mr M. Markus, DES
Ms A. Frost, Advisory Board for the Research Councils
Mr J. Irvine, Science Policy Research Unit, University of Sussex
Ms J. King, Cancer Research Campaign
Professor J.R. Quayle, FRS, Vice-Chancellor, University of Bath
Professor J. Sizer, CBE, Professor of Financial Management, University of Loughborough
Mr D. Tupman, Senior Administrative Officer, CVCP

Secretary: Mr C.R. Doherty, Principal, UFC

CHAPTER 7: CONCLUSIONS

(i) Introduction

Many questions can be asked about quantitative methods of assessing departmental research performance. This concluding chapter will focus on two. In the light of the responses received from a wide segment of the UK academic community, are quantitative methods valid? And are they useful?

These questions are important. There is little point in investing time and effort in assessing performance unless the results of the assessment are to play some role in decision-making processes, for example about structures or about resource allocation. If the decisions are to secure the general assent of the community of researchers as a whole, it is prudent to seek, and to take at least some account of, their views on the processes by which the decisions are taken. That is the objective of the consultation analysed in this paper.

That said, neither the CVCP/UFC joint committee nor any other body is bound by the results of this consultation as presented here. A different author might have produced a different analysis; external circumstances may alter the pertinence of respondents' comments; and, as every researcher knows, the prevailing majority view is not always right. However, the following conclusions do deserve to be taken seriously.

Respondents were generally willing to recognize that some form of assessment was necessary. With varying degrees of enthusiasm or resignation, most were prepared at least to explore what quantitative methods could contribute to assessment. Hostility (as opposed to reasoned opposition) against quantitative methods, where it was found, arose less from antagonism towards assessment as such than from fear that quantitative methods would displace peer review, in comparison with which they were regarded as a very blunt instrument. Nearly all respondents stressed that quantitative methods should be regarded as an input to peer review, not a substitute for it. Indeed, one of the most striking outcomes of the consultation was the spontaneous affirmation of the value of peer review.

(ii) Validity

Definition

Are quantitative methods valid? For this paper, that question translates into whether quantitative methods can give consistent, reproducible and accurate (i.e. reasonably compatible with the judgement of most informed observers) information about the research performance of university departments. The question can be approached in terms of specific quantitative methods and the specific aspects of research performance to which they are claimed to relate. There are, first, some more general considerations to address.

Departments ...

For example, what is a department? This is not as trite as it sounds. Any evaluation exercise must, of course, define its terms, for example whether research students, visiting academics or retired but still active staff should be included in the definition of 'department'. One must also consider how to deal with staff joining the department during the

period being reviewed: should their earlier work be credited to their new department, which did not contribute to it but now has a stake in the total career history of the new member of staff? This last point is particularly relevant to esteem indicators, which tend to reflect cumulative career achievements rather than immediate history.

... or individuals

But beyond these points of detail lies the issue of whether it is valid to talk about departments at all. Departments are administrative and financial units (cost centres or parts of cost centres), and act collectively to provide research infrastructure and to discharge their teaching responsibilities. However, to a degree that varies from one discipline or subdiscipline to another, research is done by individual researchers, not by departments. Many of the concerns expressed by respondents about the validity of quantitative methods stemmed from the problems attached to their use for evaluating individuals. The normal response is that the risks of inaccuracy are much reduced when dealing with aggregations of individuals. It may be that the validity of this response depends on the extent to which it is appropriate to talk of research as a departmental rather than an individual activity: a certain level of teamwork within the department may be necessary before it is valid to apply at the departmental level assessment methodologies that one would not apply at the individual level.

The degree of teamwork also affects perceptions of how one should assess collaborative research, especially if departments have different traditions of favouring internal or external collaboration.

Overall departmental performance, as measured by the standard quantitative techniques, is the sum or average of the performances of individual members of the department. Several respondents questioned the validity of such an averaging process in cases where individual performances varied widely, for example where significant numbers of staff had chosen to focus their energies on activities other than research. Quantitative techniques should, ideally, be able to distinguish non-researchers from poor researchers.

Critical mass

If it is, in principle, valid to apply quantitative methods to departments that one would not apply to individuals, the degree of that validity will increase with the size of the department. It is far from clear what minimum size of department is needed to give a reasonable degree of validity, and this will anyway vary with discipline and with the particular indicator under consideration. But for bibliometric methods in the basic sciences (the most thoroughly explored case), it is likely that many departments in UK universities currently fall below the critical threshold of size.

Publications

The validity of any given quantitative technique depends crucially on the discipline concerned. All quantitative techniques of research assessment involve measurement of surrogates: validity depends on how closely the surrogate matches the relevant aspect of performance. Most respondents accepted, for example, that data on the volume of published output had a place in assessment of research, i.e. that it was valid to count publications. But what forms of publications should be counted, how they should be counted and what else should be counted varied greatly from one discipline to another.

Citation analysis

The validity of citation analysis in assessment of departmental research was much more controversial: only a few respondents, even in the basic sciences, gave it much credence. The reasons for rejecting citation analysis as invalid included the specific referencing traditions of individual disciplines, the many non-approbationary functions of citations and the inadequacy of the databases available for counting citations. Citation analysis may have a role in analysis of comparative trends in national performance in basic science, but its use at departmental level found little favour among respondents. Even the more modest claim for citation analysis as a surrogate measure of impact (rather than quality) had relatively few supporters.

Esteem indicators

Esteem indicators proved to be more widely acceptable as valid, provided individual disciplines were allowed to specify what should be included as an indicator of esteem. Systematic use of such indicators would not be straightforward, partly because one man's esteem indicator was another man's chore (e.g. journal editorships) and partly because aggregation of an inevitably disparate set of measures would be close to impossible.

(iii) Usefulness

Context

The Research Indicators Subcommittee has been examining two distinct issues. One is the production of annual data on the outputs of university research, to go with existing annual statistical series on other facets of university life such as finance, staff and students and first destinations of graduates. The other is the production of data to assist in detailed periodic reviews such as the UFC selectivity exercises, carried out at intervals of several years. The usefulness of quantitative methods has to be considered in these two separate contexts. As with the question of validity, issues both of principle and of detail can be raised.

Peer review

It is nowhere seriously suggested that quantitative methods can or should replace peer review. Their claimed role is two-fold: to strengthen the peer review process by making available to peer reviewers as much relevant information as possible, and to render the peer review process more open to scrutiny by outside groups, including those being reviewed. Respondents were lukewarm in their estimates of how usefully quantitative methods could fulfil these roles.

Some respondents saw little use for quantitative methods in this context. At best they gave the same message as the traditional peer review process (was this not the test of their validity?), in which case they were redundant; at worst they gave a conflicting message, in which case they would be ignored.

Other respondents examined in greater detail the claim that quantitative methods were useful because they injected an element of objectivity into peer review. There was some sympathy for this claim, but it was not unlimited. It was pointed out that subjectivity entered into the selection of which quantitative measures to use, into the aggregation of the messages coming from different indicators and, particularly, into the way these messages were fed into the judgement-forming process. The claim for objectivity is central to the

attractiveness of performance indicators; respondents generally thought that claim should be handled with some care.

The same goes for the claim that quantitative methods can help to democratize peer review by making it more open. If peer review ultimately depended on the judgement of informed peers—and respondents would not have it any other way—then there were inherent limits on how open it could be. Indeed, it could be argued that quantitative methods are more useful in defending judgements than in influencing them.

The experience of the Earth sciences exercise was that bibliometrics clearly identified the strongest, and the weakest, 15% or so of departments—which were already well known anyway. Bibliometrics proved much less incisive in ordering the middle 70% of departments, which is where the difficult and controversial decisions have to be made.

Time lags

One characteristic limiting the usefulness of all quantitative methods to a greater or lesser extent is their lack of timeliness. Quantitative methods of research assessment tend to measure things that have already happened. There can be delays of many months between completion of a piece of research and its appearance in print; further delays (notwithstanding oral dissemination, conferences and preprints) before results are incorporated in other research and publicly cited; and yet further delays before measurable esteem accrues to the original researcher. These time lags mean that quantitative methods may be measuring the performance of a group of people very different from those now in the department. Many respondents argued that the problem of time lags severely diminished the usefulness of quantitative methods. Some discussed ways of reducing the time lags, for example by using journal impact factors instead of counting actual citations, but this did not greatly alleviate the problem.

Opportunity cost

A more practical, and to respondents equally visible, problem with quantitative methods was the time taken to produce them—the opportunity cost in terms of time taken from research. A key feature of the quantitative approach is that, by involving the academics concerned in generating the data, it enables them to have an input to the assessment process. Many respondents, however, were clearly worried about the demands this would impose on their time. Some institutions, on the other hand, already routinely collected publication and esteem (though not citation) data, for example for use in annual reports, so this problem may be more perceived than real.

Non-verbal outputs

Several respondents, often in engineering but also in other disciplines, stressed the need to take account of non-verbal outputs of research. This was important for the departments themselves; it was also important for actual or potential customers for the departments' work. Exclusive emphasis on verbal output would be detrimental to departments that saw their research mission as focused on the needs of, for example, industrialists wanting a functioning model or audiences wanting a live performance. Even where the initial output was in a verbal form, account should be taken of any subsequent non-verbal outputs.

Annual and periodic use

Usefulness, as mentioned earlier, depends on context. For periodic assessments (such as the UFC selectivity exercises, carried out at intervals of several years), relatively subtle and complex methods may be used. The infrequency of these exercises, coupled with the fact that they lead directly to strategic decisions of long-term importance, justifies a relatively high investment of time and effort; and because periodic assessments are generally carried out by groups of peers with relevant first-hand experience, data requiring expert interpretation may safely be used. Annually published data, on the other hand, must be reasonably cheap to collect, must be fairly stable from year to year and, above all, must be interpretable by outsiders. Respondents were greatly concerned that annually published data could be used by those without a proper understanding of the complexities of the situation to substantiate simplistic—and by implication detrimental—judgements. They were not unconcerned about the possible misuse of periodic data, but felt that it might be easier to control.

Publications

The Research Indicators Subcommittee has been conducting trials across a variety of disciplines to assess the feasibility, and usefulness, of collecting annual publications data. The responses to the present consultation exercise suggest that such an exercise might be useful, provided individual disciplines could determine which sorts of publications were included and provided any published outcome included, loud and clear, all necessary caveats about interpretation. Once the series was established, the use of three-year rolling averages would help to smooth out misleading year-to-year fluctuations. Cumulative totals would provide a useful input to periodic assessments.

Citation analysis

There was virtually no support for the inclusion of citations in annual data series and, in view of the earlier comments about validity, little support for citation analysis as a useful input to periodic assessments. Respondents sent a firm message that, at departmental level, citation analysis added nothing to peer review. The only way to assess the quality of published departmental output was to invite the department to select representative items and then for the reviewing panel to read them.

Esteem indicators

The disparate nature of esteem indicators militated against the usefulness of their inclusion in annual data series. However, they were regarded as a useful input to periodic assessments. They should therefore be collected systematically (i.e. annually) so as to be readily available when needed.

(iv) Conclusion

Usefulness

One outcome of this consultation, then, is to dispel any notion that quantitative methods provide a quick fix to problems of assessment of departmental research. Under certain circumstances some quantitative methods can be useful; but they are not particularly quick and they do not provide an easy fix. They are most useful, and most acceptable to academic researchers, when employed as an input to periodic peer review.

Cross-disciplinary comparisons

Nor can quantitative methods be used directly to compare different disciplines. Many respondents, and not only in the humanities, were worried that they would be faced with a set of indicators inappropriate to their own disciplines. It is evident that no set of indicators of research performance can be validly applied across all disciplines. Sensitivity to the specific circumstances of individual disciplines is more important than superficial uniformity.

Impact on academic behaviour

Finally, many respondents voiced anxieties that the introduction of performance indicators would distort academic behaviour, for example by encouraging excessive publication at the expense of quality or short-term fashion at the expense of long-term importance. A key aim of quality control mechanisms is, of course, precisely that of influencing behaviour, by clarifying what is most likely to reap rewards. In introducing new, additional methods of assessing departmental research it is therefore important to consider carefully whether the behaviour likely to be encouraged by those methods is consistent with the academic mission of advancing and disseminating knowledge.

ANNEX A

MEMBERSHIP OF THE CVCP/UFC SUBCOMMITTEE ON RESEARCH INDICATORS

Chairman: Professor M.B. Harris, Vice-Chancellor, University of Essex

Members: Professor E.W. Abel, Professor of Inorganic Chemistry, University of Exeter
Mr S.R. Bosworth, Registrar, University of Salford
Dr P.M.D. Collins, Director, Science & Engineering Policy Studies Unit
Mr B.D. Cullen, DES
Mr K.S. Davies, Principal Assistant Secretary, CVCP
Professor M. Hart, FRS, Professor of Physics, University of Manchester
Mr M. Markus, DES
Ms A. Frost, Advisory Board for the Research Councils
Mr J. Irvine, Science Policy Research Unit, University of Sussex
Ms J. King, Cancer Research Campaign
Professor J.R. Quayle, FRS, Vice-Chancellor, University of Bath
Professor J. Sizer, CBE, Professor of Financial Management, University of Loughborough
Mr D. Tupman, Senior Administrative Officer, CVCP

Secretary: Mr C.R. Doherty, Principal, UFC

ANNEX B

TEXT OF THE CONSULTATIVE PAPER

I INTRODUCTION

The current emphasis on greater selectivity and greater accountability in universities has given rise to considerable interest in performance indicators - quantitative or semi-quantitative data that can be put alongside qualitative assessments in forming judgements about performance. The CVCP and UFC have established a joint committee to investigate performance indicators. A separate subcommittee is examining research. This paper seeks views on how best to develop certain potential indicators of research performance.

Assessment of individual performance (or potential) is a central feature of academic life—giving places to students, awarding degrees, appointing or promoting staff, accepting papers for publication and awarding research grants are the most evident examples. Performance indicators have a slightly different purpose, in that relatively objective data can be used both to assist the forming of judgements and to strengthen their public acceptability. Moreover, these indicators tend to operate best at the institutional or departmental rather than the individual level.

This paper is concerned with indicators of the outputs and outcomes of research. An overall assessment of research performance must, of course, also pay attention to the inputs (money, people, equipment etc) and to the processes by which inputs generate outputs. One needs, at the very least, to express data in terms of output per unit input, and this is one of the aims of the CVCP/UFC subcommittee. It is however useful to separate the various stages when considering how best to tackle them.

There are several different types of output from research: new knowledge, new linkages with outside groups, newly trained people, etc. This paper is concerned solely with ways of quantifying the generation of new knowledge.

The CVCP/UFC subcommittee is charged in the first instance with establishing the groundrules for the regular compilation of statistics on the outputs and outcomes of university research. The aim is to build up a reasonably comprehensive database which, over time, will give a reliable picture of the way university research is evolving. It is hoped that the database will provide a firm factual basis for discussion of both the volume and, to some extent, the quality of university research.

II VOLUME OF RESEARCH OUTPUT

New knowledge and new ideas generated within the academic sector normally reach published form at some stage. Commercial and other confidential material is not published, but to a first approximation one may assume that the output of new knowledge by the academic sector can be monitored by careful analysis of publications. It is therefore of interest to establish agreed procedures for measuring the volume of published output, and to determine the extent to which the results fairly represent the volume of research output.

The CVCP/UFC subcommittee has already undertaken a trial exercise to count the amount of published material, with the main object of encouraging universities to set up appropriate databases for future use. In autumn 1988, all university departments of chemistry, economics, history and physics were asked to provide comprehensive lists of their published output for the calendar year 1987. Respondents provided extensive data; many also commented on such issues as the types of publications to be included, the inclusion or exclusion of particular categories of staff (for determining what should count towards total departmental output), multiple authorship, the period over which data should be collected and

differences between science and humanities disciplines. The experience thus gained is now being developed in a second trial exercise focused on the disciplines of accountancy, chemical engineering, dentistry and French.

III QUALITY OF RESEARCH OUTPUT

Achieving a degree of consensus on how to measure the volume of research output, complex as that may be, is not enough to produce an adequate statistical description of academic research. One should also tackle the more difficult issue of quality. There is at the present time no objective, quantitative measure of quality applicable to academic research. One is therefore faced with devising surrogate measures and defining the limits within which they are useful.

Some flexibility will be necessary here. In the humanities, for example, the judgements applied to the various elements of a publications count are likely to differ markedly from those used for science disciplines. The same applies to measures of quality: a surrogate measure may prove acceptable in one field but not in another.

The UFC already has some relevant experience in this field. In connection with implementation of the Oxburgh report on university earth science, all relevant departments were asked to provide comprehensive lists of published output for the previous ten years, and citation data for the previous five years. In addition to its value for the Oxburgh implementation, this exercise provided unique experience in bibliometric analysis based on comprehensive data provided by the departments themselves.

IV CITATION ANALYSIS

The best known measure used as a surrogate for 'quality' is citation analysis. One would not equate citations with quality in any simple, linear fashion: there are several reasons why the most creative or far-reaching papers are not necessarily the ones that are cited most frequently. But it can reasonably be argued that highly cited papers have a significant impact on the development of their fields. There is a good deal of evidence to support the thesis that, if one is dealing with large enough groups—say those producing more than 50 papers per year (the threshold number depends on the discipline)—or with a stable group over a number of years, there is a correlation between measures such as average citations per paper and the reputation of the group among its peers.

There are several different forms of citation analysis. The one most appropriate to the present context involves producing a comprehensive list of a department's output and then counting (by reference to the Science Citation Index, the Social Sciences Citation Index, the Arts & Humanities Citation Index or the CompuMath Citation Index) how often each item is cited during, say, the first three years after publication. Due account must be taken of the proportion of items in the list that is other than papers in journals scanned by the various citation indices.

The advantages of citation analysis along these lines include:

- that it gives reproducible results about a factor other than the sheer volume of research output. It thereby adds an important additional perspective to the picture produced by the counting of publications. The bibliometric study of university earth science, for example, found no significant correlation between the total departmental output of publications, or the average output per member of staff, on the one hand, and the average number of citations received by each publication on the other. An assessment based solely on publication counts may therefore give a misleading impression;
- once the analysis has been going long enough to allow calculation of rolling averages, it gives trend data;

- if desired, self-citation or own-group citation can be eliminated;
- in addition to the basic calculation of average citations per publication, complementary indicators can be derived, such as the proportion of publications that are highly, or never, cited.

Disadvantages include:

- that it is labour-intensive, even if some aspects are computerized;
- that one has to wait at least three years after publication before assessing the citation performance of a publication, which means that the results may reflect projects initiated five or more years previously, sometimes by people no longer in the department;
- that some departments may be too small to generate a statistically usable number of publications and citations;
- that publications in sources not scanned by the Institute for Scientific Information (which produces the citation indexes) are *ipso facto* considerably more likely to have their citation records underestimated by the indexes. This is not a problem just for the humanities: 44% of publications by university earth sciences departments are not scanned by ISI;
- that citation practice (as well as publication practice) may vary so much between fields, and especially between science and the humanities, that the results have to be interpreted in different ways.

Citation analysis would thus appear to have some value in the present context. It is a *partial* indicator of the impact of the published output of research, and is capable of generating a variety of useful measures such as citations per paper, numbers of highly cited papers, numbers of uncited papers. It is most reliable when applied to the larger departments or to departments that are stable over time. The more obvious abuses (eg excessive self-citation) can be circumvented. On the other hand, one has to wait several years before the citation record of a paper becomes apparent. Moreover, citation analysis may be less relevant to those disciplines where the typical form of published output is not a paper in a referred journal. It is therefore open to discussion whether the insights generated by citation analyses are worth the effort and cost of collecting and analysing the data.

ESTEEM INDICATORS

In attempting to quantify the quality of a department's research, it is important where possible to build up a portfolio of surrogate measures. Where the various measures prove to be mutually consistent, one can build up a composite picture with a reasonable degree of reliability. A complementary approach to citation analysis is to examine variables that can be interpreted as measures of the esteem in which the department as a whole or its individual members are held. Possibilities include:

- membership of learned societies at a level where admission is by competitive election only (eg FBA, FRS, FEng);
- major prizes;
- numbers of visiting scholars supported by formal schemes;
- journal editorships.

VI CONCLUSION

In the light of the foregoing discussion, universities and their departments or subject groups are invited to comment on what measures can most usefully be adopted to complement measures of the volume of departmental research output provided by publication counts.

ANNEX C

ANALYSIS OF RESPONDENTS TO THE CONSULTATIVE PAPER

The consultative paper was sent to the Vice-Chancellors or Principals of all 45 universities in the UK; replies were received from 41. Of the two federal universities, Wales provided replies from each of its six constituent colleges and London provided replies from its central Planning and Development Division, from nine constituent colleges or schools and from six medical schools.

In nearly every case it was clear that there had been extensive consultation within the institution. Vice-Chancellors and Principals sought the views of heads of faculties, schools and departments as well as academic advisory boards, boards of studies, research committees and special advisory committees. Many of the heads in turn consulted their staff, either in writing or during regular or specially convened meetings. The typical response from an institution therefore comprised an overall view—either the central office's own view or its summary of the inputs it had received—accompanied by a selection of the various consultations conducted within the institution. The latter included summaries of departmental etc meetings and written submissions from individual academics. The exercise thus came as close as could practicably be expected to a comprehensive consultation of the academic community in UK universities. Comments were also received from the Committee of Heads of University Law Schools, the History at the Universities Defence Group, the Universities Council for the Education of Teachers and the European Association of Science Editors.

If one may record the fact without being accused of literally weighing the evidence, the response received by the CVCP/UFC Subcommittee totalled well over 500 pages.

8. Proportion of publications receiving at least three citations in any one year, by year of publication, 1981–1985
9. Proportion of publications never cited, 1981–1985

For reasons of confidentiality, all departments are described by a randomly allocated number between 1 and 22.

2. Scale of output

The number of staff in the 22 departments is shown in table 1. The average number of staff per department, 16.5, was very close to the figure found for the 36 department set (16.6), so in this respect the smaller set was a representative sample. Department 15 fell from 42 total staff in 1982 to 38 in 1986, and in the latter year dropped below department 1. The third largest department, 6, was significantly smaller than departments 15 and 1 but significantly larger than the fourth placed department. It gained three staff during the period. The three smallest departments were 14, 21 and 13, averaging 6-8 staff each. The three largest departments (14% of the total) had, on average, 29% of total staff, while the three smallest had 6%.

Departments 1 and 15 produced substantially the largest number of publications during 1978–1986, as shown in table 2. Towards the end of the period, however, department 15 was dropping towards the third placed department 6. Output of publications was a little more concentrated than staff numbers: the three most productive departments accounted for 33% of publications, and the three least productive for 5%.

A different picture emerges when publication output is compared with staff numbers in each department (table 3). The highest ranked departments by this indicator, averaged over 1982–1986, are 11, 13, 8 and 1 (in that order). However, department 8 dropped from first place in 1982 (with 4.94 publications per staff) to seventh place in 1986 (with 2.61 publications per staff). The most consistent improvement was by department 19, bottom in 1982 (1.00 publications per staff) up to fourth in 1986 (3.45 publications per staff).

There were some dramatic changes from year to year, particularly among the small departments: for example, department 14 fell from second in 1983 (at 3.58 publications per staff) to second last in 1984 (1.75 publications per staff), and department 21 dropped from 2.67 publications per staff in 1985 to only 0.47 publications per staff in 1986.

The Oxburgh report recommended that major research departments should have at least 25 members of staff. In 1986, only five departments out of the full set of 36 met this criterion; of these five, three were in the top seven in terms of publications per member of staff, and the other two were twenty-sixth and twenty-seventh. The smallest department—13—was one of the most productive per head, and several other small departments also did well. The six most productive departments averaged two or three times as many publications per staff member as the six least productive departments. They also tended to be about one third bigger. For the full set of 36 departments, the correlation coefficient between publications per staff and total staff numbers per department was 0.31; for the 22 department set it was 0.07. The evidence from tables 1–3 is that the case for large departments is to be argued not on grounds of productivity but on grounds of, for example, the range of special skills needed to teach and research across all main areas of earth science and to exploit major items of equipment.

3. Type of output

Tables 4, 5 and 6 show how published output is divided between journals scanned by the SCI, non-SCI journals and non-journal material. The proportion of journal papers published in non-SCI journals is shown in table 4. For the 22 department set, as for the

36 department set, this proportion averaged 20% over 1978–1986. Department 11 had the highest proportion of non-SCI papers over the period as a whole, though it dropped sharply in 1986 (from 54% to 21%). Departments 1, 14 and 22 had consistently the smallest proportion of their journal papers published in non-SCI journals.

Department 14 also had the smallest proportion of publications that were other than journal papers (table 5). However, there are substantial year-to-year variations on this indicator, and few clear patterns emerge.

Tables 4 and 5 are combined in table 6 to show the proportion of each department's total publications that was other than papers in SCI journals. Department 11 had consistently, and by some way, the highest proportion of non-SCI publications, and department 14 the lowest proportion. For both the 22 department set and the 36 department set, the proportion of total publications that was other than papers in SCI journals consistently averaged 44%.

4. Citation data

Departments were asked to show how often each publication from 1978 onwards was cited in each year 1982–1986. One of the characteristics that one looks for in a performance indicator is that it should respond reasonably quickly to changes in performance. The total number of citations received in 1986 by an individual department could be significantly affected by a few highly cited papers published six or eight years earlier whose authors might no longer be members of that department: in other words, the total would not, in itself, distinguish a department that had recently produced highly cited work from one living on past glories. On the other hand, papers typically take several years to become highly cited. Earth science papers on average receive between a third and a half of all their citations within four years of publication.

We therefore calculated an indicator of citations per publication, defined as the number of citations received in a given year to publications produced that year and the preceding three years, divided by the number of publications produced that year and the preceding three years. The results are given in table 7.

Department 6 was consistently, and by some way, the best performer in the 22 department set in terms of citations per publication. The next five departments averaged over 1982-1986 (10, 22, 14, 1 and 5 in that order) form a set that is generally distinguishable from the remaining departments. The lowest departments, on average, were 19, 21 and 13. There was no significant correlation between citations per publication and the total number of publications produced by a department, nor between citations per publication and publications per staff member.

A second citation indicator is the number of highly cited publications, defined here as publications receiving three or more citations in any single calendar year. We calculated this indicator for each year of publication from 1981 to 1985, though it should be noted that the more recent publications had less time to become highly cited (1986 publications were omitted from table 8 for this reason, and the inclusion of 1985 publications is marginal). Department 6 was the highest ranked within the 22 department set by this indicator, and five of the top six departments in terms of citations per publication were in the top six in terms of proportion of publications that were highly cited. As might be expected, there is a strong correlation (correlation coefficient of 0.90) between citations per publication and proportion of publications that are highly cited; but there are no significant correlations between proportion of highly cited publications and total numbers of publications or numbers of publications per staff.

Finally, we looked at the proportion of publications that were never cited (table 9). As with highly cited publications, we calculated this indicator for the years 1981– 1985. For

publications produced at least three years before the end of the period for which citation data were collected (i.e. produced in 1981, 1982 and 1983), over one third had not been cited by 1986. The six departments averaging the smallest proportion of uncited publications were 14, 15, 6, 1, 17 and 10; the departments averaging the highest proportion were 19, 18, 4 and 3. Department 4 produced 30 publications in 1984: not one had been cited even once by the end of 1986 (though it should be noted that half these 30 publications were other than articles in SCI journals). There is some negative correlation (- 0.71) between proportion of uncited publications and proportion of highly cited publications, as one might expect.

5. Conclusions about performance

What conclusions can be drawn from the previous sections about the relative performance of individual departments in the 22 department set?

Two notes of caution should be sounded before an answer is attempted. This exercise has generated a number of indicators, each of which may be regarded as a partial indicator of performance and the aggregate of which, it is hoped, will give a reasonably clear message about overall performance. But there is no simple way to aggregate the partial indicators into a whole. So an element of subjective judgement enters at this stage.

The second note of caution concerns statistical significance. If, for the sake of argument, it is accepted that bibliometric analysis can say something about research performance, there is nevertheless a threshold of activity below which bibliometric analysis cannot be expected to say that something consistently or reliably. The threshold may be expressed in terms of the numbers of papers published annually by a department. The value of the threshold will vary with discipline, and with the extent of disciplinary spread within a given department. Reliable estimates are lacking, but the threshold for statistical significance is unlikely to lie much below 50 papers per department per year. In 1986, only 5 departments in the 22 department set (and 8 in the full 36 department set) met this criterion. One must therefore deal with more than a single year's output. This averaging procedure has the incidental advantage of damping down the sometimes substantial year-to-year oscillations noted earlier.

It should also be emphasised that the following discussion concerns the 22 department set. The 'best' (or 'worst') department in the 22 department set may not be the best (or worst) department in the UK.

It should further be stressed that the bibliometric profile of a department will be affected by the specializations of its staff. This is particularly true of the smaller departments in which only a few specializations may be represented: a poor bibliometric performance could reflect not poor work but rather work in an area that typically produces few publications and/or few citations. We have not attempted to make allowances for this, and it is not clear how it could be done. Moreover, the departments included in the 22 department set cover not only geology but also, in some cases, meteorology, geophysics and environmental sciences.

Table 10 shows, for each of nine indicators, the departmental rankings averaged over the most recent three- or five-year period. This suggests a simple (if crude) way of aggregating the individual indicators; it is included here to stimulate discussion of the results rather than to provide detailed verdicts on the relative performance of all 22 departments.

The largest departments in terms of staff numbers or publication output were 1, 15, 6 and 5; the smallest were 20, 13, 22, 14 and 21. The lack of correlation between

publications per staff and staff numbers has been noted already. If one attempted to aggregate the volume and productivity indicators (staff, publications and publications per staff), department 1 emerged as the strongest, followed by department 15 and then by a group comprising, in no particular order, departments 5, 6, 8, 9, 11 and 17. The weakest department was, by some way, department 21, with department 22 second from last.

There is greater consistency among the three citation indicators (citations per publication, proportion of highly cited publications, proportion of uncited publications), which makes aggregation of those three indicators less problematic. Departments 6 and 14 were the strongest performers, followed by 1 and 10 and then, at a bit of a distance, by 17 and 2. Departments 13 and 19 were the weakest, followed by 3 and 20.

The combination of scale, productivity and citation indicators points to departments 1 and 6 as the strongest, followed (in no particular order) by 5, 10, 15 and 17. Department 21 emerged as the weakest, followed (in no particular order) by 3, 13, 19 and 20.

As discussed in more detail in section IV.2 below, departments publishing in journals not scanned by SCI, or producing non-journal publications, are likely to be penalized in citation analysis. Citations per publication scores do, indeed, show some negative correlation with proportion of non-SCI journal papers (correlation coefficient of -0.63), with proportion of non-journal publications (-0.62) and with proportion of publications other than papers in SCI journals (-0.66). But poor citation performance cannot be ascribed solely to a mismatch between publication strategy and the selection policy of SCI. Among the departments with poor citation performance, this mismatch plays a significantly greater role in accounting for the performance of departments 3, 13 and 20 than of department 19.

IV METHODOLOGICAL ISSUES

1. Departmental bibliographies

There are essentially two possible approaches to producing departmental bibliographies: asking each department to provide a list of its publications, or interrogating a computerized database. This report deals with a survey that used the first approach. The problems encountered were described in section II: they are all surmountable if care is taken in specifying exactly what is wanted and if departments follow the instructions, especially as regards presentation, multiple authorship and inclusion only of appropriate publications. The burden placed on departments by requests for such information will diminish as departments increasingly collect publication data routinely for their own management purposes and as a standardized format is adopted. Provision of such bibliographies in computerized form will in due course facilitate analysis.

Unless one is prepared to use a different specialized database for each discipline, and on each occasion to invest the effort needed to convert it to a suitable form, the only option for online generation of bibliographies is to use the main Institute for Scientific Information (ISI) database or a database derived from it. It is possible, though difficult, to extract from the ISI database lists of earth science publications produced in UK universities. Such lists would necessarily be confined to papers published in the 3000 or so journals scanned by SCI and, as table 6 shows, would omit nearly half (for some departments, considerably more than half) the total published output. What would be omitted is the 20% of papers that are published in non-SCI journals and the 30% of publications that are not journal papers at all. One might choose to argue that the 20% should be disregarded as being of lower quality than the articles published in SCI journals, though there is certainly room for more than one view about SCI's journal selection policy. It would be difficult to argue that the 30% of publications that are not papers but

books, chapters in books, maps, project reports, etc should likewise be disregarded. And since these proportions vary considerably between departments, ignoring non-SCI material severely distorts the picture of the relative performance of the departments.

It could be argued that the SCI will pick up material inadvertently omitted from a departmentally produced bibliography, but any such gain is unlikely to outweigh the nearly 50% loss of material just mentioned. Moreover, as departments get into the habit of producing bibliographies routinely, the amount of material inadvertently omitted is likely to be insignificant.

The SCI therefore cannot safely be used to produce a profile of universities' overall output of Earth science publications. [It is possible, however, that the SCI provides more comprehensive coverage in other fields of research, if they typically produce fewer non-journal publications and if the SCI journal selection policy more closely mirrors UK publishing habits.] There is no reliable alternative to asking individual departments to supply bibliographies.

2. Citation analysis

The Science Citation Index (SCI) is the only systematic source of data on citations and is thus the fundamental tool for citation analysis. It is compiled by examining everything published in a selection of over 3000 journals and recording everything cited in these journals. Most journals contain more citations to themselves than to any other single journal. A paper published in a journal not scanned by SCI may be cited by papers in SCI journals and therefore appear in the SCI, but a substantial proportion of the citations it receives (e.g. those from other papers in the samc journal) will not be recorded. Non-journal publications (books, reports, etc) similarly appear in the SCI if cited in SCI journals, but, again, the citations they receive in other non-journal publications will not be recorded in the SCI. A department publishing exclusively in SCI journals will, *ipso facto*, produce a better citation score (i.e. have more citations recorded in the SCI) than an equally competent department publishing in non-SCI journals or producing non-journal publications. In the context of a bibliometric description of university departments, it is therefore necessary to measure the extent to which the SCI selection of journals adequately represents each department's output.

The Institute for Scientific Information claims that SCI journals cover 75% of the world's significant science research literature. One might be tempted to argue that good researchers, by definition, publish in SCI journals and that, in counting citations, one should not attempt to take account of the extent to which citations are lost because of the SCI's selection policy. This argument may have some validity in the context of comparing relative trends in the research performance of the major scientific nations, but it cannot be assumed to hold good at departmental level.

Quite how one should take account of the mismatch between departmental publishing strategies and SCI selection policy is another matter. Personal judgement is likely to play a part. One relevant consideration will be the extent to which the journals that are included in SCI cover the particular research interests of a given department.

3. Conclusions

Bibliometric analysis is of interest in three distinct policy contexts:

(i) review of the provision for individual disciplines;
(ii) periodic review of departmental performance;
(iii) production of regular (annual) performance indicators.

The present exercise was undertaken initially in the context of (i). In the event, decisions about the reorganization of earth science provision within the university system were taken before the results of the bibliometric analysis were available, though had the original schedule been followed it is likely that the bibliometric results would, at least, have been of interest. As it is, the results record the performance of the system during the years prior to reorganization; a similar exercise at some point in the future should reveal the effects of the reorganization.

The present methodology can be used as part of the process of producing periodic rankings of individual departments. The first such ranking exercise was carried out in 1986, and a second is envisaged for 1989. Because of the analytical labour involved and the need to build up and maintain teams with the relevant expertise, it would be preferable to tackle a proportion of the disciplines amenable to bibliometric analysis every year rather than to tackle all disciplines at one go every few years.

For regular performance indicators, it might be preferable to stick to total publications and publications per staff member. Care would be needed to ensure uniformity of what was included under 'publications' and what constituted a staff member. It might be necessary to devise some weighting system for different types of publication. If citations per publication were included, it would be important to ensure that the indicator was so defined as to take proper account of time lags, and to find a way of giving due credit to publications not scanned by the *Science Citation Index*.

Finally, it would be instructive to take advantage of a future discipline review to test the feasibility of the methodological recommendations made in section II above and to examine whether the findings about the relevance of the SCI to the published output of earth science departments in UK universities apply also to other disciplines.

TABLE 6
continued

PUBLICATIONS OTHER THAN PAPERS IN SCI JOURNALS
AS A PROPORTION OF ALL PUBLICATIONS, 1978–86

Average 1978-86	Dept	Average 1978-80	Dept	Average 1981-83	Dept	Average 1984-86	Dept
0.69	11	0.72	11	0.70	11	0.66	11
0.59	3	0.66	21	0.66	13	0.63	20
0.58	12	0.61	9	0.64	3	0.61	12
0.55	13	0.58	12	0.56	12	0.58	4
0.54	21	0.56	3	0.54	9	0.57	3
0.53	4	0.52	4	0.53	17	0.56	13
0.53	20	0.52	2	0.51	20	0.55	15
0.52	9	0.49	8	0.50	19	0.54	17
0.50	17	0.49	19	0.49	4	0.52	21
0.49	8	0.45	6	0.48	15	0.52	8
0.49	15	0.45	15	0.47	8	0.50	5
0.47	2	0.44	20	0.46	22	0.49	2
0.46	19	0.43	13	0.44	21	0.45	22
0.43	5	0.42	17	0.41	2	0.42	9
0.38	6	0.41	16	0.39	5	0.42	10
0.38	16	0.41	5	0.39	16	0.40	19
0.38	7	0.40	7	0.38	7	0.38	6
0.37	22	0.37	18	0.35	18	0.35	16
0.35	10	0.32	10	0.33	6	0.35	7
0.31	18	0.31	1	0.31	10	0.33	1
0.31	1	0.19	22	0.29	1	0.21	18
0.16	14	0.15	14	0.20	14	0.13	14
AVERAGE:							
ALL DEPTS 0.44		0.44		0.43		0.46	
TOP 6 DEPTS 0.64		0.63		0.65		0.65	
BOTTOM 6 DEPTS 0.29		0.29		0.28		0.31	

Table 7

CITATIONS PER PUBLICATION*, 1982-1986

	1982	Dept	1983	Dept	1984	Dept	1985	Dept	1986	Dept	Average 1982-86	Dept
	1.75	22	1.66	6	1.65	6	1.65	6	1.59	6	1.65	6
	1.72	6	1.60	22	1.59	10	1.20	14	1.23	22	1.38	10
	1.62	10	1.51	10	1.36	14	1.14	22	1.20	2	1.32	22
	1.48	14	1.27	14	1.10	1	1.12	2	1.18	10	1.26	14
	1.36	8	1.20	1	0.98	5	1.09	1	1.17	1	1.14	1
	1.14	1	1.08	5	0.97	2	1.03	10	1.01	5	1.04	5
	1.10	5	0.79	8	0.90	22	1.00	5	0.99	15	0.87	2
	0.89	16	0.74	18	0.75	8	0.80	16	0.97	14	0.86	8
	0.74	17	0.72	11	0.69	16	0.72	15	0.91	17	0.70	16
	0.66	18	0.68	3	0.60	15	0.70	18	0.78	20	0.70	17
	0.60	19	0.65	16	0.60	18	0.68	8	0.72	8	0.69	15
	0.60	2	0.60	17	0.58	17	0.66	17	0.58	18	0.66	18
	0.59	11	0.57	15	0.55	4	0.54	7	0.56	7	0.54	11
	0.58	15	0.48	12	0.52	7	0.53	11	0.56	9	0.50	7
	0.53	4	0.48	21	0.51	12	0.50	12	0.50	4	0.49	3
	0.49	7	0.44	2	0.49	9	0.47	3	0.49	16	0.46	12
	0.48	3	0.39	7	0.41	11	0.41	9	0.46	12	0.44	4
	0.47	20	0.34	13	0.38	3	0.38	19	0.46	11	0.41	9
	0.35	12	0.25	9	0.31	20	0.37	4	0.44	3	0.41	20
	0.34	13	0.25	4	0.30	13	0.36	21	0.42	21	0.37	19
	0.32	9	0.22	19	0.25	19	0.31	20	0.38	19	0.35	21
	0.31	21	0.15	20	0.20	21	0.15	13	0.29	13	0.28	13
AVERAGE:												
ALL DEPTS	0.91		0.84		0.82		0.80		0.87		0.85	
TOP 6 DEPTS	1.41		1.35		1.26		1.21		1.23		1.29	
BOTTOM 6 DEPTS	0.37		0.26		0.34		0.34		0.42		0.35	

* i.e. the number of citations in year 'n' to publications published in year 'n' and the preceding three years, divided by the number of publications published in year 'n' and the three preceding years

TABLE 8 PROPORTION OF PUBLICATIONS CITED AT LEAST THREE TIMES IN ANY ONE YEAR, BY YEAR OF PUBLICATION, 1981–1985

1981	Dept	1982	Dept	1983	Dept	1984	Dept	1985	Dept	Average 1981-85	Dept
0.60	14	0.58	14	0.54	14	0.40	2	0.31	22	0.43	6
0.60	6	0.58	6	0.41	1	0.39	6	0.20	1	0.41	14
0.43	1	0.48	10	0.40	6	0.26	17	0.17	6	0.34	1
0.40	22	0.40	17	0.37	2	0.25	1	0.16	14	0.30	10
0.38	10	0.38	1	0.34	10	0.24	10	0.16	2	0.28	22
0.33	7	0.32	22	0.28	17	0.20	5	0.10	21	0.28	2
0.30	8	0.31	5	0.24	5	0.20	22	0.07	5	0.25	17
0.29	4	0.26	16	0.22	7	0.18	15	0.06	10	0.20	5
0.27	21	0.25	18	0.22	4	0.18	16	0.06	15	0.19	7
0.25	16	0.24	8	0.19	18	0.17	20	0.06	8	0.19	8
0.25	2	0.23	2	0.19	11	0.17	8	0.06	17	0.17	4
0.24	17	0.22	19	0.18	8	0.17	13	0.05	11	0.16	16
0.24	11	0.21	15	0.17	12	0.16	7	0.04	3	0.14	15
0.22	3	0.20	7	0.16	9	0.16	9	0.04	4	0.13	11
0.19	9	0.19	4	0.15	22	0.14	14	0.04	18	0.13	9
0.19	5	0.18	12	0.13	3	0.11	12	0.03	7	0.13	18
0.17	18	0.17	21	0.11	15	0.10	4	0.03	16	0.12	21
0.16	15	0.14	9	0.09	20	0.07	19	0.03	12	0.11	12
0.13	13	0.13	11	0.08	16	0.07	11	0.02	9	0.10	3
0.12	20	0.11	3	0.00	19	0.06	21	0.00	13	0.09	20
0.09	12	0.07	20	0.00	21	0.00	18	0.00	19	0.07	13
0.00	19	0.07	13	0.00	13	0.00	3	0.00	20	0.06	19
AVERAGE:											
ALL DEPTS 0.30		0.28		0.24		0.19		0.08		0.22	
TOP 6 DEPTS 0.46		0.45		0.39		0.26		0.18		0.35	
BOTTOM 6 DEPTS 0.13		0.15		0.08		0.05		0.02		0.08	

TABLE 9

PROPORTION OF PAPERS PUBLISHED IN A GIVEN YEAR AND NOT CITED BY 1986

1981	Dept	1982	Dept	1983	Dept	1984	Dept	1985	Dept	Average 81-85	Dept
0.00	11	0.09	6	0.15	17	0.14	14	0.17	15	0.23	14
0.00	7	0.13	14	0.22	15	0.23	15	0.35	21	0.28	15
0.03	14	0.21	16	0.28	10	0.28	1	0.43	17	0.30	6
0.18	6	0.24	5	0.33	1	0.33	2	0.44	22	0.31	1
0.20	10	0.25	1	0.33	14	0.38	17	0.46	6	0.32	17
0.22	16	0.25	17	0.33	12	0.39	6	0.48	1	0.37	10
0.23	8	0.28	15	0.33	5	0.39	16	0.51	2	0.40	16
0.24	1	0.32	10	0.37	2	0.40	8	0.53	14	0.42	7
0.27	21	0.35	8	0.38	6	0.43	18	0.58	10	0.43	8
0.29	4	0.48	2	0.42	16	0.48	20	0.63	12	0.44	2
0.33	18	0.48	7	0.44	9	0.50	13	0.67	5	0.46	5
0.35	20	0.48	9	0.44	7	0.50	10	0.67	9	0.47	21
0.36	9	0.49	12	0.44	8	0.52	7	0.68	7	0.47	11
0.38	17	0.51	11	0.45	19	0.52	11	0.71	13	0.50	9
0.44	13	0.56	20	0.48	4	0.53	21	0.72	3	0.53	20
0.45	19	0.56	19	0.48	20	0.53	22	0.74	8	0.53	22
0.46	5	0.56	4	0.56	21	0.56	9	0.76	16	0.54	12
0.47	22	0.57	13	0.56	13	0.57	5	0.77	11	0.55	13
0.52	15	0.58	22	0.56	11	0.64	19	0.78	20	0.58	19
0.52	12	0.63	3	0.63	3	0.67	3	0.79	19	0.61	18
0.54	2	0.67	21	0.65	22	0.73	12	0.81	4	0.63	4
0.56	3	0.79	18	0.69	18	1.00	4	0.81	18	0.64	3
AVERAGE:											
ALL DEPTS 0.32		0.36		0.38		0.46		0.58		0.42	
TOP 6 DEPTS 0.13		0.32		0.27		0.30		0.38		0.29	
BOTTOM 6 DEPTS 0.51		0.52		0.61		0.68		0.78		0.62	

TABLE 10

DEPARTMENTAL RANKINGS BY VARIOUS INDICATORS, AVERAGED OVER RECENT YEARS

Staff per dept	Publications per dept	Publications per staff	Citations per publication	% Pubns cited at least 3 times in a year	% Pubns not cited*	% Papers not in SCI journals	% Pubns not in journals	% Pubns not in SCI journals
1982-86	1984-86	1982-86	1982-86	1981-85	1981-85	1984-86	1984-86	1984-86
15	1	11	6	6	14	21	20	11
1	15	13	10	14	15	11	15	20
6	6	8	22	1	6	13	22	12
5	5	1	14	10	1	4	12	4
12	17	9	1	22	17	3	11	3
10	9	17	5	2	10	12	17	13
17	11	14	2	17	16	9	8	15
2	8	16	8	5	7	8	4	17
9	2	20	16	7	8	17	13	21
8	12	2	17	8	2	20	2	8
7	10	15	15	4	5	5	5	5
3	19	5	18	16	21	6	3	2
18	3	19	11	15	11	10	10	22
16	16	4	7	11	9	19	1	9
11	7	6	3	9	20	2	19	10
4	4	10	12	18	22	15	7	19
19	18	18	4	21	12	16	6	6
20	20	22	9	12	13	7	16	16
22	13	12	20	3	19	18	9	7
14	22	3	19	20	18	1	18	1
21	14	21	21	13	4	14	21	18
13	21	7	13	19	3	22	14	14

* In this column, departments are ranked from lowest scoring to highest scoring. In all other columns, the order is the opposite

6 Carlton House Terrace, London SW1Y 5AG