Arch. Immunol. Ther. Exp., 2009, **57**, 19–32
PL ISSN 0004-069X
DOI 10.1007/s00005-009-0003-3

The use of bibliometrics to measure research quality in UK higher education institutions

Jonathan Adams

Evidence Ltd, Leeds, UK

Received: 2008.12.10, **Accepted:** 2008.12.17, **Published online:** 2009.02.14

Abstract

Research assessment in the UK has evolved over a quarter of a century from a loosely structured, peer-review based process to one with a well understood data portfolio and assessment methodology. After 2008, the assessment process will shift again, to the use of indicators based largely on publication and citation data. These indicators will in part follow the format introduced in 2008, with a profiling of assessment outcomes at national and international levels. However, the shift from peer assessment to a quantitative methodology raises critical issues about which metrics are appropriate and informative and how such metrics should be managed to produce weighting factors for funding formulae. The link between publication metrics and other perceptions of research quality needs to be thoroughly tested and reviewed, and may be variable between disciplines. Many of the indicators that drop out of publication data are poorly linked to quality and should not be used at all. There are also issues about which publications are the correct base for assessment, which staff should be included in a review, how subjects should be structured and how the citation data should be normalised to account for discipline-dependent variables. Finally, it is vital to consider the effect that any assessment process will have on the behaviour of those to be assessed.

Key words: assessment, indicators, profiling, peer review.

Corresponding author: Jonathan Adams, Evidence Ltd, 103 Clarendon Road, Leeds LS2 9DF, UK,
e-mail: jonathan.adams@evidence.co.uk

ORIGINS

Bibliometrics represent the latest stage in the development of mechanisms and processes to direct research funds selectively in the UK research base, with the policy intention of ensuring that these funds are used as efficiently and effectively as possible in support of "the best" science. We start by reviewing the origins of the present system.

In 1889, HM Treasury established an ad hoc Committee on Grants to distribute £15,000 it had set aside for 11 university colleges. At the end of 1916 the government created the Department for Scientific and Industrial Research (DSIR) to support civil science and to coordinate and commission its own research (Varcoe 1974). While the University Grants Committee (UGC) block grant paid for salaries and preliminary investigations, the DSIR gave university scientists funds to carry out specific research. Selectivity was not an important factor in the small, embryonic, dual-support system of the early 1920s. For example, there were only 24 DSIR-funded university postgraduate research studentships in 1917 and this number had grown to no more than 81 by 1938. The dual-support system did, however, provide performance indicators against which a selective funding system could gear.

After 1945, selectivity became more apparent. The UGC grant was moved to the Board of Education and its new terms of reference required the UGC to take a more active stance on university policy rather than provide only passive advice on grant allocations. The numbers of trained research professionals in universities doubled during the 1950s, science research spending in 1962 was at least tenfold that in 1945 (Wilkie 1991), and the spending on conventional science had at least doubled in real terms. Although this growth through both the UGC and DSIR routes was good news, the UGC seems always to have operated some level of selective funding after 1945. This was partly through a system of subject-based subcommittees that kept areas under review, sometimes very actively, and represented a store of intellectual capital on which the UGC could draw.

Within the 1947–1967 quinquennial cycles, the UGC accepted that there was a need for national initiatives in particular disciplines, in response to needs identified by

the government (pace the origins of Foresight and its priorities). The device of earmarked grants was adopted and this amounted to about 30% of all recurrent grants during each Quinquennium (Shattock 1994).

Throughout the growth period during the 1964 Wilson government, there was clear evidence of overt and very marked selectivity. For example, there was the so-called take-over exercise: "the continued financing of research projects, hitherto funded by the Research Councils, which it was agreed should be continued as part of the normal activities of the universities" (UGC 1966). For takeovers, the UGC made earmarked increases to the block grants of selected institutions, identified as about £1.8M per year in a total grant for 1966–1967 of about £134M (UGC 1967). The "take-over exercise" continued until the UGC Annual Survey of 1970–1971.

The UGC also asked higher education institutions for returns that divided expenditure into UGT, PGT, and R, and this came in for a good deal of criticism. The Committee argued that it was right to seek a firmer basis for apportioning expenditure, but exactly how the data it collected were used was never entirely clear. J. C. Walne presciently commented that if the UGC were to make public its methods, then universities might respond by arranging their affairs to increase their entitlement or, at least, to perpetuate existing patterns (Walne 1973). Success in attracting research grants was evidently significant (UGC 1984) and the nature of the UGC's research algorithm caused, then and later, much comment. Lively correspondence in the Journal of the Royal Statistical Society discusses special factors for research: a concentration effect – "an increasing payment per student in respect of non-teaching functions as universities get larger" and speculation about "an unusually high level of support for certain privileged areas" (Cook 1976; Cook 1977; Dainton 1977).

The UGC Annual Survey 1976–1977 saw the breakdown of Quinquennial grants, but introduced planning figures based on four-year projections. The UGC noted that the dual-support system was under great strain, and cuts needed to fall on the small area where economy was possible. While the well-found department (the first use of this term) could no longer be guaranteed, the UGC would, after discussion with the Research Councils, make a selective allocation of some £500k to enable some 11 institutions specific participation in identified areas of high priority.

The Annual Survey for 1979–1980 (UGC 1985) announced that the Committee had, with the Advisory Board for the Research Councils (ABRC), decided to set up a review of arrangements for dual support; this became the Merrison review. In the meantime, "the current distribution of equipment grant (£72M for 1980/81 cf. a recurrent grant of £987M) takes into account each university's past record of attracting outside research grants and thus provides a slightly better equipment base for those with a proven research capability".

The Merrison Report coincided with severe cuts in public funding. Merrison commented on both internal and external selectivity in this regard. The Report concluded that "we are convinced that whatever research is done should be of high quality and properly supported, and this means that universities will need to concentrate research funds into selected areas" (ABRC/UGC 1982). Universities should set up internal research committees to make selective allocation not on a project basis, but to create a strong infrastructure, and this might mean that some university departments would not be able to sustain a viable research base. For the system as a whole, it was not desirable to identify the total of research (R) spending, as this might send the wrong message about the integral nature of teaching (T) and R activity in universities. Moreover, while the overall grant could be determined in proportion to research grants, this would lay the wrong emphasis on the nature of dual support, to the detriment of seed-corn activity and innovation; this might, however, be a factor to take into account.

The Morris Report also expressed support for selectivity among research committees. "Each university, through its research committee, will have to choose which of its staff to support and which not to … some departments may well develop into departments dominated by their teaching activity" (ABRC 1983). The Joint Report of the Chairmen of the ABRC and Advisory Committee on Applied Research and Development (ACARD) proposed that there should be national and overt policy of selectivity among research objectives.

The need for accountability and selectivity across the system was now firmly on the table. "A Strategy for Higher Education into the 1990s" suggested (at par. 5.14) a "more selective allocation of research support among universities" in order to ensure that resources for research were used to the best advantage (UGC 1985). The UGC confirmed that it had in the past "taken account of research achievement, but … not previously implemented the principle of interaction or responsiveness" identified in its recommendation. Subsequent paragraphs developed the rationale for a selective approach that would be related to, but not passively dependent on, the existing dual-support system. The text confirms that the information supplied to the UGC by the Research Councils was a major driver in the research grant algorithm.

The 1985 UGC circular letter to universities said that the distribution of research funds would take account of work of special strength and promise, so as to maintain the quality of research in UK universities. It also said that there was no intention to identify particular areas for preferential funding.

In 1986, the UGC operated its first Research Selectivity Exercise. This asked universities to complete a four-part questionnaire covering various aspects of income and expenditure, planning, priorities, and output. This was used by the UGC subcommittees to establish evaluative ratings, on a four-point scale, in consultation with the Research Councils. The ratings were

then used for selective allocation of a part (Judgment--related Research component, JR) of the research resource. The ratings were announced in a descriptive form (below average, etc.), but the absence of an absolute standard caused much confusion.

There was extensive criticism of the 1986 exercise. Many of these early criticisms were made against the subsequent Research Assessment Exercises of 1989, 1992, and 1996 and some challenges (relevance in the arts, the evaluation of interdisciplinary research, and assessment standardization) may be problematic in any system of this kind. In particular, in regard to concentration, it was suggested that there appeared to have been a bias in favor of larger departments. It was further suggested that the assessment of research laid undue emphasis on just one part of the higher education mission and would both cause staff to neglect their teaching and skew the public impression of universities.

In 1987, the ABRC's "Strategy for the Science Base" (ABRC 1987) recognized that "the changes that are taking place in the approach to organisation and management of research … the development of selectivity and more directive management … can be seen as the inevitable response to the challenge of managing science within finite resources". It expressed support (par. 1.21) for the Oxburgh review on the Earth Sciences, which concluded that resources for that subject were over-dispersed. The ABRC also suggested (par. 1.25) that the allocation of Research Council grants to scientists in below-average departments was "not conducive to the concentration of effort that we believe generally to be in the national interest" (ABRC 1987).

These considerations led the ABRC to conclude that polices then in place would not lead quickly enough to the degree of concentration required to maintain the international competitiveness of university research (then, among 60 institutions). The proposal that emerged was for an R-T-X system of institutions (R – research, T – teaching, X – mixed economy) differentiated according to the pervasiveness and breadth of their research strength, with some 15 institutions in the top R category of substantial international research across most fields. This represented a key shift in the overt rationale for selectivity because it introduced the idea of selectivity among institutions with the express purpose of concentrating funding to create centers of excellence. This rationale depended on the underlying, and unproven, concept that excellent physics research was more likely to emerge if it were done alongside excellent chemistry research, and so on.

The 1988 "Guidance" from the Secretary of State for Education to the new Universities Funding Council (UFC) emphasized that the enhancement of the strength and quality of the science base required greater concentration and selectivity of research work. The UFC's first annual report (1987) had focused on arrangements made by universities to bring about a selective distribution of resources and the second report (1988) looked at selectivity processes and the way in which these supported groups or departments rated above-average by the 1986 selectivity exercise. The DES sought to maintain the momentum created by this, but the UFC's third report reflected growing discomfort with mechanical selectivity that depended on "yesterday's successes" and which might stifle creativity. It did note (par. 10) the much greater planned concentration on areas of demonstrated research strength and (par. 28–32) the growth of selectivity within departments.

The UFC was applying a research funding formula which took account of both research income, i.e. DR (Dual Research component) from the Research Council and CR (Contractual Research component) from other sources, and peer judgment of comparative research strengths, i.e. JR. There was a progressive shift from the old volume measure reflected in a fourth component, SR (Staff-related Research component), towards JR. The relative growth of JR should have led to a shift of funds to the strongest research institutions. Institutions argued, however, that there were problems either if this did happen or because it did not. On the one hand there was concern that concentration was suppressing the emergence of novel research, but on the other hand there was concern that those who had headroom gained resources at the expense of those who were already excellent.

In the early 1990s, the shift in the dual-support boundary transferred funds from the UFC/HEFCs (Higher Education Funding Counsils, HEFCs) to the Research Councils to cover the direct and some indirect costs of research projects. This created an expectation of some concomitant increase in selectivity driven by project-based peer review.

The new HEFCs replaced elements of the old block grant research funding algorithm by QR, a new Quality Research factor relying wholly on research grades. This should have markedly concentrated funds, but was partly mitigated by DevR (Development Research), a factor introduced to support research innovation in the former polytechnics. Thus, even after a clear policy favoring the concentration of funding had been introduced, other confounding factors continued to play a role.

INTRODUCTION TO THE PRESENT

In December 2006, the UK government announced that a new framework for assessing and funding university research would be introduced following the completion of the next research assessment exercise in 2008 (Treasury 2006). It is the government's intention that the current method for determining the quality of university research, the UK Research Assessment Exercise (RAE), should be replaced after the next cycle is completed in 2008. Metrics, rather than peer-review, will be the focus of the new system and it is expected that bibliometrics (using counts of journal articles and their citations) will be a central quality index in this system.

The higher education sector welcomed the key features of the announcement, which includes the creation

of a new UK-wide indicator of research quality. The intention was that the new framework should produce an overall "rating" or "profile" of research quality for broad subject groups at each higher education institution.

The objective of any change in the assessment method should be to sustain recent improvements in UK research performance. To do this, the metrics system will need not only to be technically correct, but also acceptable to and inspire confidence among the researchers whose performance is assessed.

Bibliometrics is probably the most useful of a number of variables that could feasibly be used to create a metric of some aspect of research performance. Thomson Scientific maintains sound international databases of journals and their citations with good time, subject, and institutional coverage. These data have characteristics (particularly in terms of the publication and citation cultures of different fields), which means that they must be interpreted and analyzed with caution. They need to be normalized to account for year and discipline and the fact that their distribution is skewed. These factors will affect analyses and must be addressed with care.

There is evidence that bibliometric indices do correlate with other, quasi-independent measures of research quality, such as RAE grades, across a range of fields in science and engineering, but such correlations leave a substantial residual variance and average citations per paper would be a poor predictor of grade. Furthermore, there may be fundamental differences between informed researcher perceptions and simple metrics of research quality.

There is a range of bibliometric variables as candidate quality indicators. There are strong arguments against the use of (i) output volume, (ii) citation volume, (iii) journal impact, and (iv) frequency of uncited papers. A number of new methods have attracted interest, but they are either superficial (for example, the h-index) or remain unproven for the present (for example, web-ometrics). Output diversity is a potentially valuable attribute, but challenging to index.

"Citations per paper" is a widely accepted index in international evaluation. Highly cited papers are recognized as identifying exceptional research activity. These are not usually applicable to individual researchers, but if incorporated in an approach to profiling the overall output of research units, they could prove of value. If such profiling were associated with an analysis of performance trends, it could lead to an acceptable analysis if other concerns can be satisfied.

Citation counts, their accuracy and appropriateness, are a critical factor. There are no simple or unique answers. It is acknowledged that the Thomson databases necessarily represent only a proportion of the global literature. This means that they account for only part of the citations to and from the catalogued research articles and that the coverage is better in science than in engineering. The problems of obtaining accurate citation counts may be increasing as Internet publication diversifies. There are also technical issues concerning fractional citation assignment to multiple authors, the relative value of citations from different sources, and the significance of self-citation. The time frame for assessment and for citation counting relative to the assessment will also affect the outcomes and may need to be adjusted for different subject groups.

The population to be assessed needs to be defined, in principle and operationally. In particular, is the assessment to be of individuals and their research activity or is it of units and of the research activity of individuals working in them? How will this affect data gathering, and how will that be influenced by the census dates for more frequent assessment? There are equal-opportunity issues to be considered. It is unlikely that bibliometrics will exacerbate existing deficiencies in this regard, except insofar as research managers perceive a sharper degree of differentiation, but metrics have an inability to respond to contextual information about individuals.

The definition of the broad subject groups and the assignment of staff and activity to them will need careful consideration. While the RAE subject groups might appear to follow traditional faculty structures sensibly, this is no longer the unique structure for research activity. The most important aspect of the subject grouping, however, is the strategy that is subsequently used to normalize and aggregate the data for finer-grained subjects within each group. This is likely to be complex and to vary by group, but the precise level of normalization of data will have a profound effect on outcomes. It is noted that similar considerations will apply to any other data on funding or training.

Differences between subjects (at a broad and fine level) mean that no uniform approach to data management is likely to prove acceptable if all subjects are to be treated equitably. There will need to be sensitive and fine scale adjustments of normalization and weighting factors and of weighting between bibliometrics and other indicators. There is also a challenge to be addressed in the management of interdisciplinary research where, again, the insensitivity of metric algorithms will miss the benefits of peer responsiveness.

The management of the bibliometric data will need to be addressed. The license cost will be significant and there will be a substantial volume of initial work to set up an effective database for this purpose. In the longer term, this development may produce a net return to institutions by providing additional local management information. Internal research management will be unchanged and much the same information will ultimately be required. In this context, the role of peer oversight needs to be clarified.

Profiling methodologies, based on normalized citation counts, appear to be the most likely route to developing comprehensive and acceptable metrics. They should also prove useful in differentiating excellence for benchmarking, but the strategy for normalizing the raw citation data prior to analysis will be central and critical.

A number of potentially emergent behavioral effects will need to be addressed, although experience suggests both that many behavioral responses cannot be anticipated and that some of these responses could jeopardize the validity of the metrics themselves in the medium term.

BACKGROUND

In December 2006, the UK Government announced that a new framework for higher education research assessment and funding would be introduced following the next national research assessment exercise (RAE 2008). The Higher Education Funding Council for England (HEFCE), in collaboration with other national higher education funding bodies, is developing this framework. An early priority is to establish a UK-wide indicator of research quality (for science-based subjects in the first instance). The intention is that the framework should produce an overall "rating" or "profile" of research quality for broad (faculty-based) subject groups at each higher education institution.

It is widely expected that the index will initially be derived (in part) from bibliometric-based indicators, but expert subject panels would be involved in producing the overall ratings[1]. The bibliometric indicators will also need to be linked to other metrics on research funding and on research in postgraduate training. The various indices will also need to be integrated into an algorithm that drives the allocation of funds to institutions.

The quality indicators would ideally be capable of doing more than informing funding. They also need to satisfy an additional key role provided by current research assessment: they need to provide benchmarking information for institutions and stakeholders. At the same time, they must also be cost effective to produce and must reduce the assessment burden on institutions.

BIBLIOMETRICS AS INDICATORS OF QUALITY

The research process can be simplified as:

Inputs – activity – outputs – outcomes

What we are really interested in is the quality of the research activity. If it is high, we might reasonably expect that the output will be good and that will lead to beneficial outcomes. However, we cannot measure the quality of research activity directly. Although peer experts can usually establish fairly quickly whether a laboratory or group in their field is any good or not, that perception does not translate into an objective measure.

To overcome our limitations we use "indicators" – and that is all they are: they indicate what we want to know, but do not measure it directly. They are proxies.

One indicator of competence is the ability to acquire a high level of scarce income for research support. Income is a problematic indicator, however, and economists might challenge the use of "input" as an indicator of quality under any circumstances. A cap to the total available income is determined by policy as much as the quality of recipients or the size of the field. Furthermore, cost varies between theoretical and practical projects within a field. Thus, for these and other reasons, inputs are usually taken as only a partial measure, even if they are limited to a "peer-reviewed source".

Outcomes from basic research, which comprises much of the public-sector research base activity, can be disconnected from the original research. First, the outcome may not be clear for many years. Second, the outcome may be affected by many original discoveries and one discovery may likewise have many influences on outcomes. In the absence of a one-to-one relationship it becomes challenging to index the value of activity satisfactorily.

Outputs overcome some of these problems and the citation of outputs provides an apparent quality measure. For these reasons, bibliometrics provides an attractive source of research performance data. Further benefits of using such data are that they cover many fields in a similar way and therefore enable some measure of comparability. They also cover many countries in the same way and provide further value in comparisons. And Thomson Reuters® Inc. maintains a database initiated by the Institute of Scientific Information (ISI) back in the 1960s, so there is a well-developed data structure and a powerful back-resource on which to draw.

Citations between papers are signals of intellectual relationships. They are a natural, indeed essential, part of the development of the knowledge corpus. They are therefore valuable as an external index about research because they are produced naturally as part of "what researchers do" and because they are related naturally to "impact" and "significance". Not all indicators have such attributes.

Citations accumulate over time and uncited papers for any one year gradually fall in number. Older papers are likely to have had more time to increase their citation count. There are likely to be fewer uncited papers in samples from more distant years. Time is not the only factor causing systematic differences in samples of publication and citation data. Different disciplines have innate, cultural differences in the way in which they use the literature in terms of article length, frequency, and citation structures. This further increases the complexity of satisfactory quantitative evaluation.

[1] Bibliometrics are indicators of research performance based on data associated with journal articles. Research publications normally refer to (or cite) prior work which serves as an authority for established knowledge, methodologies, and so on. Publications may then be cited by later outputs. Citations therefore provide a network of association between items within the accepted corpus of knowledge. Researchers generally agree that more frequently cited items have a greater significance than those that remain uncited. Eugene Garfield, the founder of the Institute of Scientific Information, proposed that citation counts should be seen as an index of "impact" within the relevant field of research. That "impact" has later been interpreted as related to quality, with highly cited papers having greater impact and therefore being of higher quality than less frequently cited items.

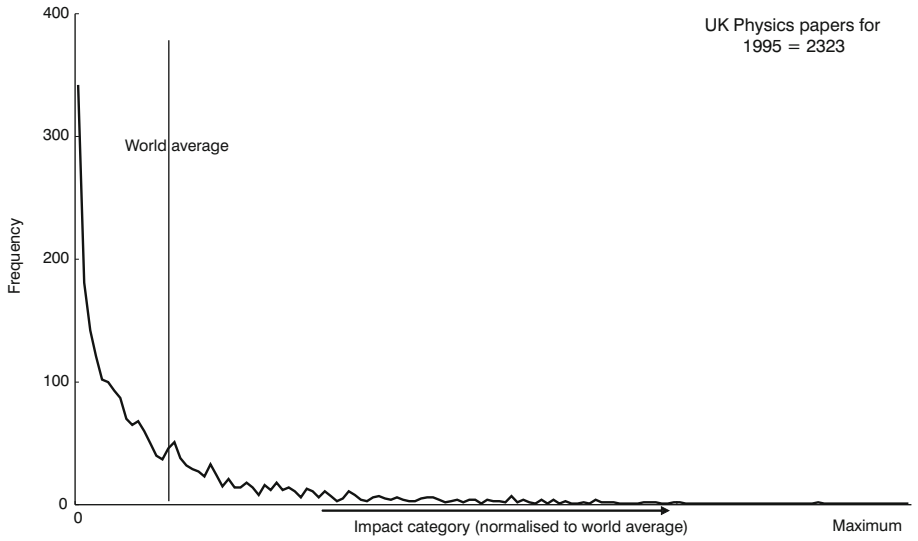

Fig. 1. A typical skewed distribution for citation data. This is UK physics after ten years of citation accumulation: most papers are cited less often than the world average although the UK average is above the median and the world average. Source: Thomson Reuters, Analysis: Evidence.

Eugene Garfield, the founder of ISI, drew attention in the 1960s to processes for normalizing impact by field and year. The process of normalization to enable comparison across years and disciplines is also referred to as "rebasing" the citation counts to a common standard. For this reason, normalized impact indices are referred to as rebased impact or RBI (RBI appears in various figures in this document).

SKEWED DISTRIBUTIONS

The distribution of almost all research data is skewed: there are many low-index data points and a few very high-index points. This applies to funding per person or per unit and it applies to papers per person and to citations per paper. These positively skewed distributions typically have a mean (average) that is much greater than their median (central point) (Fig. 1).

Skewed data are difficult to compare visually and to interpret. The average is nowhere near the center of the distribution and is no guide to the median value. Because they follow a negative binomial distribution, they cannot be handled using parametric statistical analyses and it is therefore necessary to transform them in some way in order to arrive at a more intuitive presentation and manageable analysis (see Adams et al. 2008; Leydesdorff and Bensman 2006).

ARE BIBLIOMETRIC OUTCOMES LINKED TO RESEARCH QUALITY?

There are very few reports that comprehensively establish a relationship between bibliometric impact and any other, independent, evaluation. That is not to say that the efficacy of bibliometrics should necessarily depend on establishing any correlation. It may be that bibliometrics measures one dimension while another metric approaches a different dimension. The cartography created by a plurality of partial indicators may then reconcile to a third, subjective perception.

In practice, the presence of an article in a journal with good editorial practice suggests it has at least some intrinsic merit established by the peer review of the editor and referees. If that article is then widely cited, that adds a second level of peer recognition (and if the citations endorse the work, approval). It would be surprising, therefore, if there were no match between bibliometric indicators and peer perceptions.

In a series of studies for HEFCE, Universities UK, the former Office of Science and Innovation, and other UK agencies, we have analyzed the relationship between variables associated with research activity and the categorical grading assigned by the RAE. First, RAE data confirm that journal articles are the preferred mode of output submitted for research assessment in the STEM areas for which HEFCE seeks to apply a metrics-based system. We assume that the items that are submitted normally represent material that indicates the highest available level of achievement for the individual. Second, a high proportion of the submitted articles are in journals catalogued by Thomson. This is particularly so for journals that are present at relatively high frequency in data on research outputs submitted for the RAE. Third, within a subject area, the average impact of the submitted articles for an institution is correlated with the impact of the total output for the institution but is somewhat higher, confirming our assumption about "best work" (Fig. 2).

Fourth, the average citation impact tends to increase with the grade awarded by the peer review panel. This "goodness of fit" between impact and RAE grade can be looked at via a more direct plot (Fig. 3).

This shows several things. At a gross level, the average rebased impact at each grade (i.e. the average impact taken across all the units that were awarded that grade) progresses upwards steadily with grade. The

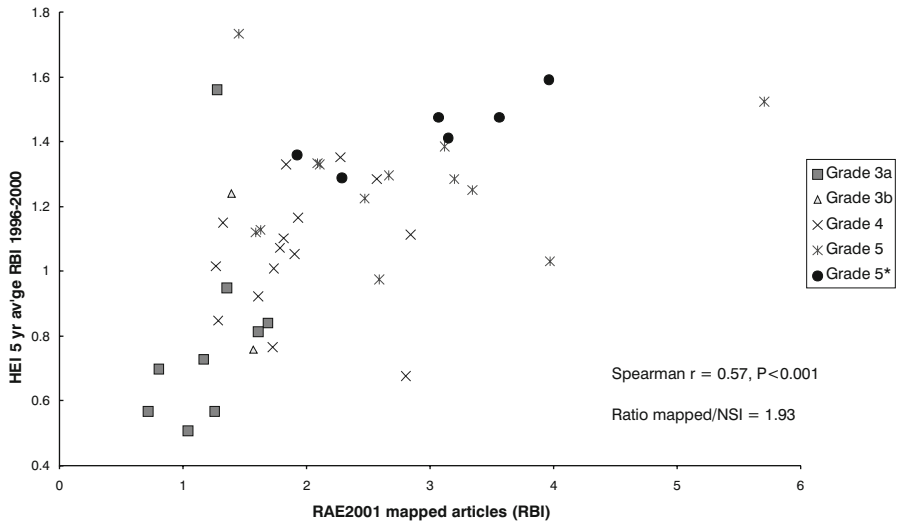

Fig. 2. The correlation between average impact of publications submitted to the RAE by a unit and the average impact of all publications in that discipline by the same higher education institution over the same period: Chemistry (unit of assessment 18). Source: Thomson Reuters, Analysis: Evidence

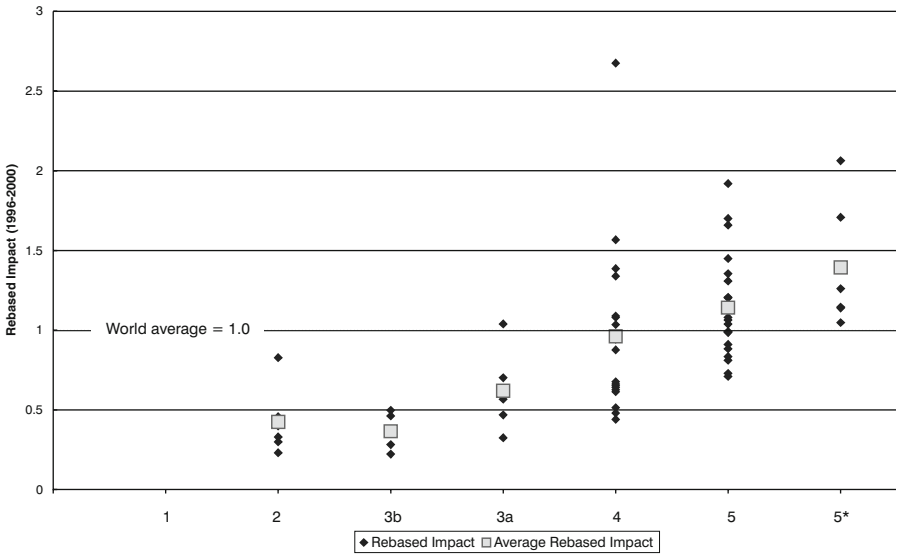

Fig. 3. Average citations per paper, data from RAE2001 for Biology (unit of assessment 14). Each red data point is the average impact of the papers in an institutional submission and each gray square is the integrated average at a stated grade. Source: Thomson Reuters®, Analysis: Evidence.

average value for grade 4 units is around the world average, which also makes sense in terms of the RAE criteria. So not only is there a similar progression, but the relationship is coherent.

There is rather less good news when one considers the variation within any grade. It then becomes evident that the average impact for any stated unit within the grade band can be very variable. There is, in other words, a great deal of residual variance whatever the value of the correlation. To put this another way, in a metrics-based system the information that a unit had an average impact close to the world average would not enable one to tell whether that unit was graded as a 4 or 5 by peer review or whether it might even be a very good 3a or a bibliometrically weak 5.

WHAT VARIABLES SHOULD BE INCLUDED?

Bibliometric data can provide a number of component variables. These can be integrated as a well-structured analysis, but the elements could equally be seen as different indicators. Components that could be addressed by an appropriate model would include the following.

Output volume

The number of papers produced by a unit, department, or institution should not be included as a bibliometric indicator for the following reasons:

- the quantity of outputs has no direct bearing on quality
- the UK's output has in the last few years reduced as share of world total without any detriment to quality. In fact, the UK is producing fewer uncited papers.

Diversity of outputs, by journal and subject

The subject diversity of papers produced by a unit, department or institution is informative, but should probably not be included as a bibliometric indicator.

A key indicator that we have developed for the former UK Office of Science and Innovation is based on research diversity. The concept is that a more diverse research base is also more agile and responsive. This is therefore a desirable attribute and might reasonably be one to encourage.

Diversity is not necessarily scale independent, however, because greater capacity gives greater room for sustainable diversity. Hence, large units are more likely to carry diversity than smaller units. Therefore, although this is an informative indicator it is likely to favor larger institution and departments irrespective of their quality. We would not recommend using it as a metric for research assessment without further investigation.

Citation volume

The number of citations acquired by a unit, department, or institution should not be included as a bibliometric indicator for the following reasons:

- this is an indicator of market share, not of performance
- if more papers are published, then the likelihood is that the citation count will increase because there will be some cross-reference and there are more "targets" to be cited.

Journal impact

Journal impact factor for assessed publications should not be included as a bibliometric indicator because:

- typical citation rates vary between broad subjects: biology papers are on average cited more frequently than physics papers
- citation rates also vary within broad subject groups and thus affect individual journal citation rates
- the variance is due to characteristics such as field size, publication frequency, and citation culture, not to any innate difference in quality.

The impact factor of a journal is an issue of significant commercial interest. There is no doubt that publishers seek to increase their average citation rates and believe that by doing so they will increase the number of subscribers and perhaps affect the quality of papers submitted for inclusion.

It is not true that papers published in lower impact journals are innately of lesser quality than other outputs. On the one hand, the process of getting a paper accepted for publication in Nature (for example) is highly competitive. To pass the editorial and refereeing process is an indication of significant interest and likely value. On the other hand, many papers submitted to relatively low impact journals are targeted at specific channels that increase the likelihood they will be read by either a particular group of researchers or a particular practitioner or user group.

UK soil science is an example of an area with low impact journals but where outputs are deliberately targeted at users. Our work for the Department of Environment, Food, and Rural Affairs (Defra) has shown that UK soil science is of high relative international impact and its utility within the UK and elsewhere is unchallenged. It would be extremely unfortunate if such research were coerced into high impact journals not read by the relevant users.

Uncited material

The number of uncited papers produced by a unit, department, or institution should not be included as a bibliometric indicator for the following reasons:

- we do not know why any specific papers may fail to be cited: it may be poor quality, but it may contain important but negative results;
- the numbers of uncited papers in any "cohort" or sample falls over time, so account needs to be taken of the time since publication.

The frequency of uncited papers provides useful management information, but it is not necessarily useful for a metrics algorithm since it requires a reasonable level of informed interpretation to make sensible use of the information.

There has been little work on the nature of uncited papers or on methodologies for accounting for the important work that identifies less fruitful areas of investigation, but which itself remains uncited (if indeed it does remain uncited). It is argued that publication of negative results is a desirable component of a cutting-edge research base. It is certainly not to be discouraged because it increases efficiency by avoiding repeated errors in choosing paths to explore.

Average citations per publication

The average citation count of papers produced by a unit, department, or institution could be included as a bibliometric indicator.

This is often referred to as a measure of "impact", from the original recognition by ISI's founder, Eugene Garfield, that papers cited more frequently than average within their field have a greater "impact" on the work of others (Garfield 1955). This index of research quality is widely used by the scientometrics community. It has been employed extensively for many years by Thomson Reuters® and by ISI, its predecessor. More recently it has been used in, for example, the European science and technology indicators, by the CWTS bibliometrics research group at the University of Leiden (who endorse it under the label of a "crown indicator"), and in our own PSA target indicators for the Office of Science and Innovation (Evidence/OSI 2007).

The characteristics of citation accumulation mean that impact must always be contextualized, which is to say that account must be taken of both the year and field of publication so as to normalize or "rebase" a specific citation count against a relevant average and thereby enable comparison between years and, if necessary, across fields. The problem with using an average citation count is that the average in a research performance distribution has little to do with the median because the data are highly positively skewed (see above). Thus the average tells us little on its own about the balance of work between poor and high quality.

It is critical that the data should be appropriately treated before being aggregated. Normalization strategies are a critical part of any metrics-based methodology and will be discussed in more detail below.

Highly cited papers

The number of highly cited papers produced by a unit, department, or institution could be included as a bibliometric indicator. Thomson has established a criterion for "highly cited" which captures the most frequently cited one per cent of outputs after taking into account the field and year of publication. The UK has about 13.3 per cent of world papers that meet this criterion, which is even better than its share of total papers. Rather than looking at the total output, some evaluations focus on these publications on the assumption that if they are unusually highly cited, they are likely to have made the greatest contribution within their field or to innovative products and processes.

There is no doubt that highly cited papers are associated with exceptional research, but the metric is a poor index of more general research activity. The threshold is so high that for many fields there would be few UK institutions that had more than a handful of papers in the index.

Profiles

This is the most informative approach to bibliometric assessments of research performance. We noted above that bibliometric data should be considered in terms of distributions rather than averages where they must be used in isolation. This helps to overcome the extent to which the "average" disguises the natural skew in the data. For management information, a profile of "impact" is a helpful illustration that shows how performance is distributed and how it compares with a reference profile. For HEFCE's purpose, the issue would be how to extract key variables from such a profile in order to capture the essential characteristics algorithmically.

The naturally skewed distribution of citation data can be made visually more acceptable by sorting the data into "bins" relative to the world average (see Adams et al. 2008). What advantage does this offer compared with average normalized impact? The company Evidence Ltd. recently completed analyses for a UK research council which showed the value of looking at citation profiles as well as averages. Due to extremely highly cited reviews in Nature from an international project, one group had a much higher average rebased impact, but the citation profiles (below) were almost indistinguishable (Fig. 4).

What values might be abstracted from such a profile?

- uncited papers as a proportion of the total
- the proportion of papers cited less often than the benchmark (national average, world average)
- the proportion of papers above the benchmark
- the proportion of papers cited at exceptionally high levels for field and year (where "exceptional" requires a threshold, such as ≥4 times the world average, to be defined as highly cited).

All of these values could be used as a metric on their own, but drawing them in a structured way from a pre-

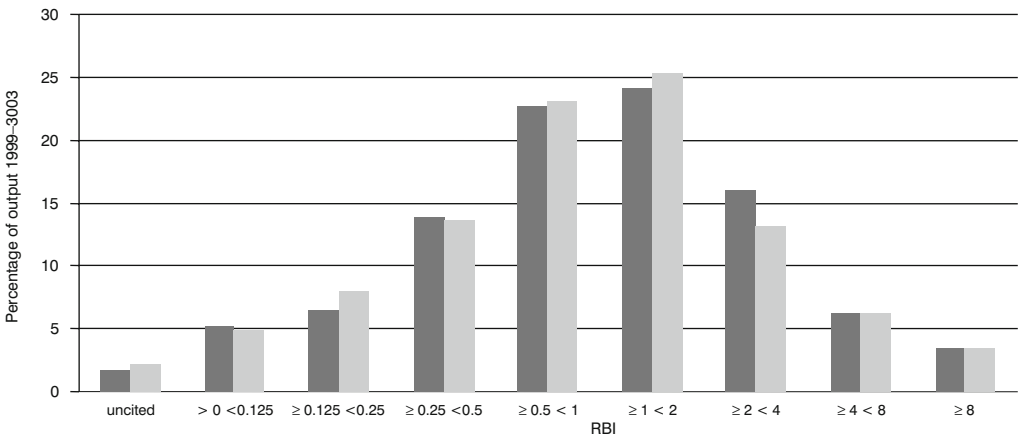

Fig. 4. Comparative impact profiles for bibliometric data from two sets of researchers working in the same field in research council units and in higher education institutions. The average normalized citation impact of the two groups differs markedly (2.39 vs. 1.86) because of exceptionally high outliers in one group. Source: unnamed RC, Thomson Reuters, Analysis: Evidence

scribed format that can itself be observed by both the assessed and the assessors may help to make the process more transparent and acceptable. Not least, it shows how the components link to the underlying data and how they are derived, and it tests whether they make sense.

WHAT IS THE POPULATION TO BE ASSESSED?

"What is being assessed" is an important issue. Research outputs are associated both with people (researchers as authors) and institutions (where they work). When people move between institutions, should the "credit" associated with their metrics move with them or should it retain its association with the institution where the activity occurred? Should all their outputs be included or only a selection of the best?

Metrics could cover population in terms of:

- discipline = "all the chemists"
- management unit = "all the research staff in chemistry"
- staff grade = "all the permanent academic staff in chemistry"
- journal category = "all the publications that are in chemistry journals".

It seems likely that there are issues to be addressed in this regard, but it is equally likely that the same issues would have arisen under the RAE peer review as they are systematic rather than specific to bibliometric or other indicators.

HOW SHOULD SUBJECT GROUPS BE DEFINED AND CONSTRUCTED?[2]

The aggregation of analysis has an effect on the assessment and on the outcome for institutions. In other words, it affects the way the data are handled and it affects the way the results are perceived.

Evidence Ltd. established a methodology for aggregating research activity into subject groups at the fine and coarse levels in work for HEFCE in 1997 (Adams 1998). That methodology has stood the test of time and has been widely employed since. It has recently been tested and validated in work for the research councils.

Customized clustering could be based on links between Thomson's output databases, RAE publication databases, and information about the funding and location of researchers. It could be developed at the outset or left until a later stage when the methodology has been agreed upon after consultation.

[2] HEFCE has indicated *a priori* that it anticipates somewhere in the region of 5–8 groups to cover the sciences, engineering, technology, and medicine.

What is physics?

The approach Evidence took (Adams 1998) was to look at the use that different subject groups made of the literature. We can see that Physics (Unit of Assessment, UoA19) submits a given range of journals for RAE2001 assessment, whereas Chemistry (UoA18) submits a different but overlapping range. Both are similar to Materials (UoA32) but quite different from Biology (UoA14).

We can therefore cluster physics (as seen by institutions for RAE purposes) with chemistry and materials and draw a distinction with a separate biology cluster.

Data from the Engineering and Physical Sciences Research Council, drawn from publications associated with researchers funded by different physics-related programs (i.e. as seen by the research council for research purposes), map well onto this RAE analysis.

Therefore whether we look at university units or at research communities, we find a coherent and common association through the literature. Physics can be robustly defined through a set of journals and this journal set identifies links to cognate disciplines and delimits its boundaries with different subject areas.

Are we assessing the subject or staff within the subject?

A fundamental question is whether the evaluation of "research quality" is about a group of staff (via their publications) or about a discipline (via the publications of the staff in that subject). This is a practical issue as well as conceptual. Material needs to be assigned to the subject groups for assessment, and that can be done either by assigning evidence of "research activity" or by assigning evidence linked to staff.

Select publications to match staff

For bibliometrics, selection of staff would be meaningless unless publications are also selected to match staff. The more direct or explicit the link between individual staff and specific publications, the more costly the methodology for the funding agency and the institutions.

Staff who move

A twist to the question of what is assessed comes at this point. When staff move between institutions, are their publications reassigned with them or do they constitute part of the legacy activity of the institution they are leaving?

Level of analysis

This assignment of evidence may seem slightly arcane, but it is both fundamental and practical. The methodology will require some time to explore in proper detail. The outcome will affect, first, researcher confidence that what is assessed by the metrics is a true rep-

resentation of their work and, second, the costs to the institutions of working with the system. Any proposals therefore need to be scrutinized with extreme caution and in fine detail.

For example, the pathway to identifying and analyzing the impact of publications at higher education institution X associated with chemistry research (a set of journals associated with chemistry as a discipline) is different from that required to identify and analyze the publications of the staff employed by X within its school of chemistry.

By carrying out the analysis at the level of five to eight STEM subject categories, HEFCE will have reduced, but not removed, the difference between the subject and people analyses compared with an analysis at, for example, the unit of assessment level. It will not have addressed staff mobility.

If there is a clear argument suggesting that bibliometric analyses at aggregate subject level are indistinguishable from analyses at staff level, this would significantly reduce the costs of any subsequent part of this development. But this is unlikely to prove satisfactory from a researcher perspectve.

NORMALIZATION STRATEGIES AND AGGREGATION

Creating a basis for data comparability within five to eight broad subject areas will, we believe, be problematic. We noted above that the availability of funding can vary substantially between sub-fields, as does publication culture. There will have to be correcting (normalization) factors to enable data to be brought together for comparison because there are differences between subject categories in rates of citation accumulation and in typical citation plateaus. For this reason, both time since publication and journal category are taken into account when normalizing or "rebasing" citation counts to enable indexing and comparison.

It would be inappropriate to aggregate data at the level of HEFCE's broad subject groups unless they are first made comparable by a satisfactory normalization at some finer level. Over-normalization will be as problematic as under-normalization because it will remove the subtle differences that the exercise seeks to identify. It is therefore critical to determine the appropriate level for normalization.

Citation rates vary between field (and sub-fields) and citation counts accumulate over time. At which level should bibliometric data be normalized? It could be the broad subject field, fields below this (for example, units of assessment), the Thomson or Scopus journal categories, or at the level of journals themselves.

We have evaluated the effects of these on quality rankings in psychology, biology, and physics. We calculated the normalized citation performance of UK research units for each of three levels of article aggregation (journal, journal category, and unit of assessment, where several categories map to each unit of assessment). We compared this with the grade awarded to that unit in RAE 2001. We found that the correlation between average normalized citation impact and peer-reviewed grade does indeed vary according to the selected level of granularity. There is little difference between grade-related impact when citation counts are normalized at the journal level. However, more highly graded units had a statistically significantly higher impact when the normalization was relative to the Thomson journal category or to the journal sets mapping to the unit of assessment (Adams et al. 2008) (Fig. 5).

The implication is that the material submitted by grade 4 units is actually sourced from journals of lower

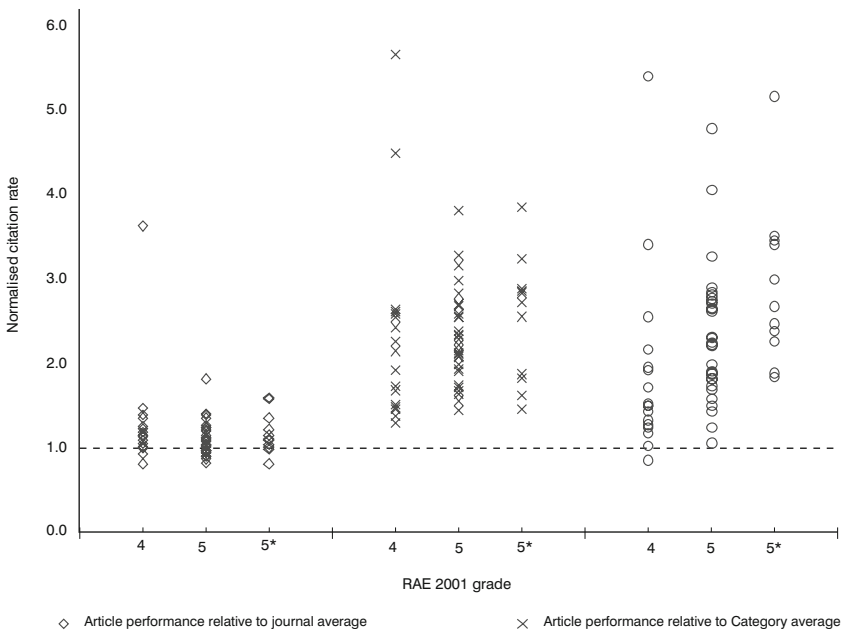

Fig. 5. Variation in citation impact for Biological Sciences (unit of assessment 14) using RAE2001 submitted articles and grades awarded. Each point is the average citation count to the end of 2005 for the set of journal articles submitted by a stated institution within this unit of assessment, with the impact for each article normalized against a world average. The data are grouped according to the RAE grade awarded to the institution.

average impact than the material submitted by the grade 5 units. Thus when the level of analysis is relative to journal, these items appear to be of similar impact relative to the medium in which they are published. When the viewpoint is zoomed out to the broader category level, then the higher absolute citation count for the articles produced by the more highly graded units becomes apparent, and even more apparent at the unit of assessment level.

While the pattern varies between broad fields, an upper and lower boundary to the granularity of sensible normalization is apparent. Above the unit of assessment level, the differentiation between fields is lost. Below the Thomson journal category, the differentiation between peer estimates of quality is lost. It will be vital that data are thoroughly reviewed and the right level of normalization is set for each broad field in any metrics system.

There is a further note of caution. This analysis applies only to bibliometric data. A parallel analysis would be required for funding data and for training data if these are used elsewhere in the metrics system.

AGGREGATION

Pulling material together could follow a diversity of routes, but we suggest that a sound method would seek to follow the natural hierarchy of similarity within the source data. This is best reflected in the similarity of journal usage between cognate research areas.

Fields that have similar journal usage will be most amenable to similar treatment in bibliometric indexing (for example, reducing the need to use many different normalization factors) and will have natural affiliations. We already know that chemistry, physics, and materials science show strong affiliation as a natural "physical science" group that also shows clear separation from a "biological science" group (Adams et al. 1998).

DATA ACQUISITION, COLLECTION, AND PREPARATION FOR ANALYSIS

Preparation for analysis and associated quality assurance will clearly be a key part of the development of the relevant methodology. It may be appropriate to consider the implementation of an assurance methodology once the basic process is agreed, but the need to have a process for which such assurance is feasible is an absolute requirement.

Conflicts between HEFCE's central data and the data provided or validated by institutions will also have to be addressed and accommodated. Thomson catalogues about 100,000 articles every year that have one or more UK-based authors. Because of new address variants, Evidence spends significant staff time every year analyzing address variations and determining the actual institution with which the author is associated. For example, by processing 25 years of legacy data we have increased the linkage of articles to the University of Oxford by 40 per cent compared with raw Thomson data assignment.

As noted earlier, the article records will need to be linked to staff so that they can be linked to subject groups. Two issues arise:

- author names and addresses are not linked, but grouped in separate fields. The linkage has to be made manually
- author synonyms (two name variants, one person) and homonyms (two people, same name and initials) are incredibly common (for example, there are at least three unique Dr. F. Guillemots in UK data). These can only be distinguished manually.

The UK publishes about 100,000 articles per year. For the metrics system to work, these all need to be linked accurately to named individuals. In 2004, those articles had a total of 473,046 authors, not all of whom were in the UK. The numbers and diversity of coauthors is increasing.

EMERGING BEHAVIORAL EFFECTS

In this essay, we have considered many aspects of the ways in which data can be assembled and analyzed to create bibliometric indicators of research performance. In doing this we seek to quantify an aspect of behavior. What effect does our intervention have on the behavior we seek to evaluate?

There is a risk that any metrics exercise may be intrinsically self-defeating because it depends on indicators as proxies for the activity of interest (Goodhart 1975). Once an indicator is made a target for policy, it starts to lose the information content that originally qualified it to play such a role. There is room for manipulation, there may be emergent behavioral effects, and the metrics only captures part of the research process and its benefits.

It is facile to pretend that all behavioral effects can be anticipated and modeled. The metrics system will be assaulted from the day it is promulgated by 50,000 intelligent and motivated individuals deeply suspicious of its outcomes. There will be consequences.

If citations per paper are used, this will potentially affect citation behavior across the system. The Netherlands started to use bibliometric indicators much earlier than most other European Union nations, and this has helped to support the academic development of scientometrics in that country. However, it had a wider effect on the publication and citation behavior of the Dutch as well. Output relative to the rest of the world has gone up by a factor unmatched by any other European Union country. Citation share has increased as well, partly due to output growth and partly to awareness of citation metrics as an evaluation criterion.

If there are emergent effects, some can undoubtedly

be addressed by adding modifications to the metrics, but this risks the development of an increasingly complex system that loses not only simplicity (hence ease of operation) but also transparency and so leads to a loss of institutional and researcher confidence.

Possible drivers of behavioral change

Research volume is a poor indicator because quality, not quantity, is the objective, and once volume is used as an indicator (as in RAE 1986) it begets spurious publications. The "Dutch effect" is partial evidence of this at a national level.

Citation volume is equally a poor indicator because it is linked to output volume and does not in itself prove anything. Since citation rates vary between sub-fields, there will always be differences in citation accumulation. If volume were an indicator, this would encourage lower-citing fields meaninglessly to emulate high-citing fields.

Journal impact factors are a poor indicator because of the variation in citation rate. If they were used there would be an erroneous competition to get any article into a high-impact journal, even if this were not the best medium for the output. Practitioner journals would certainly suffer, but there would be disruption of coherence within fields as the existing assortment of material by medium was disturbed.

Uncited papers are a poor index. The relative volume of uncited papers is interesting so long as it is seen as a partial and system-level measure. If used at a local level in a model it would simply lead to a systematic tendency to ensure that every individual and institutional output was cited at least once, whether for good reason or not.

Output diversity is a potentially useful indicator, but could be disrupted by the generation of spurious diversity in publication patterns.

Removing self-citations from analyses could be one way of moderating the "Dutch effect", but there are sound reasons not to attempt this. Self-citation is a normal part of research culture. If self-citation were actively penalized by HEFCE metrics, this could lead to a change in citation behavior, with transitional drops in citation rates, a failure to track links in research programs, and a loss of international prestige.

Partitioning credit for collaborative papers would also be ill-advised. Collaboration is an increasingly important part of research and provides signal benefits. A significant part of the UK's best research outputs are internationally collaborative. To send messages to the system that there is a "metrics cost" in collaboration would undermine the very things that the Office of Science and Innovation, the research councils, and a recent House of Commons report are seeking to stimulate.

In all this it should be recognized that it will not be possible to detect changes in UK behavior and outcomes for some years. By then, the UK may be set on a pathway from which it is difficult to extricate itself.

REFERENCES

ABRC (1983) The support given by Research Councils for in-house and university research (The Morris Report)

ABRC (1987) A strategy for the science base; a discussion document prepared for the Secretary of State for Education and Science by the Advisory Board for the Research Councils. HMSO

ABRC/UGC (1982) Report of a joint working party on the support of University scientific research. Cmnd 8567, HMSO (The Merrison Report)

Adams J (1998) Benchmarking international research. Nature 396:615–618

Adams J, Bailey T, Jackson L et al (1998) Benchmarking of research in England. Report to HEFCE & CPSE, University of Leeds (ISBN 1 901981 04)

Adams J, Gurney K, Marshall S (2007) Profiling citation impact: a new methodology. Scientometrics 72:325–344

Adams J, Gurney KA, Jackson L (2008) Calibrating the zoom: a test of Zitt's hypothesis, Scientometrics, 75:81–95.

Cook WR (1976) How the University Grants Committee determines allocations of recurrent grants – a curious correlation. J Royal Statistical Society (A) 139:374–384

Cook WR (1977) Curious correlations – a reply. J Royal Statistical Society (A) 140:511–513

Dainton F (1977) Comments on "How the UGC determines allocations of recurrent grants – a curious correlation". J Royal Statistical Society (A) 140:199

Evidence/OSI (2007) PSA target metrics for the UK research base. Available via http://www.evidence.co.uk/downloads/OSIPSATargetMetrics070326.pdf

Garfield E (1955) Citation indexes for science: a new dimension in documentation through association of ideas. Science 122:108–111

Goodhart C (1975) Problems of monetary management: The U.K. experience In: Papers in Monetary Economics. Volume I, Reserve Bank of Australia, 1975

Leydesdorff L, Bensman S (2006) Classification and Powerlaws: The Logarithmic Transformation. J Am Soc Inf Scientists Technologists 57:1470–1486

Shattock M (1994) The UGC and the management of British universities. Society for Research into Higher Education & Open University Press, Buckingham

Treasury HM (2006) Pre-Budget Report 2006. Chapter 3

UGC (1966) University Grants Committee, Annual Survey for the academic year 1965–1966. Cmnd 3192, HMSO

UGC (1967) University Grants Committee, Annual Survey for the academic year 1966–1967. Cmnd 3120, HMSO

UGC (1984), Annual Survey for the academic year 1983–1984. Cmnd 3245, HMSO

UGC (1985) A strategy for higher education into the 1990s; the University Grants Committee's advice. HMSO

Varcoe I (1974) Organizing for science in Britain. Oxford University Press, Oxford

Walne JC (1973) Analysis of university costs at the UGC. Higher Education 2:228–235

Wilkie T (1991) British science and politics since 1945. Blackwell, Oxford