Research report

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

Universities UK

Research Reports

This series of Research Reports published by Universities UK will present the results of research we have commissioned in support of our policy development function. The series aims to disseminate project results in an accessible form and there will normally be a discussion of policy options arising from the work.

This report has been prepared for Universities UK by Evidence Ltd.

Evidence Ltd (http://www.evidence.co.uk) specialises in research performance analysis and interpretation. It has extensive experience with and databases on research inputs, activity and outputs relating to research both globally and particularly for the UK research base. It has also developed innovative analytical approaches for benchmarking international, national and institutional research impact. The company has worked closely with national research funding agencies and individual staff have experience at national policy level and as senior institutional research managers.

Evidence holds or has access to a range of publication and citation databases derived primarily from the databases of Thomson Scientific in Philadelphia, US. The core data are the expanded Citation Indexes from which Thomson ISI Web of Science® is derived. Thomson ISI Web of Science® currently covers publications from approximately 8,700 of the most prestigious, high impact research journals in the world. These are augmented by additional information on publication usage in universities derived from research and consultancy work by the company and its predecessors.

Contents

Foreword

In December 2006 the 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. The sector has welcomed the key features of the announcement, which includes the creation of a new UK-wide indicator of research quality. The intention is that the new framework should produce an overall 'rating' or 'profile' of research quality for broad subject groups at each higher education institution.

It is widely expected that the ratings will initially be derived from bibliometric-based indicators rather than peer review. These indicators will need to be linked to other metrics on research funding and on research postgraduate training. In a final stage the various indices will need to be integrated into an algorithm that drives the allocation of funds to institutions. The quality indicators would ideally be capable of not only informing funding but also providing benchmarking information for higher education institutions and stakeholders. They are also expected to be cost-effective to produce and should reduce the current assessment burden on institutions. The Higher Education Funding Council for England is currently working on the development of the new arrangements. Its work includes an assessment of how far bibliometric techniques can be used to produce appropriate indicators of research quality and to evaluate options for a methodology to produce and use these indicators.

In preparation for the funding council consultation that will take place later in the year, Universities UK commissioned Evidence Ltd to explore some of the detailed and technical issues that arise when using metrics in the research assessment process. Its report does not suggest a preferred approach but identifies a number of issues for consideration. We are publishing this report with the aim of briefing the higher education sector on the kind of issues that it will need to consider when responding to the forthcoming consultation. The report will also help Universities UK to formulate its position on the development of the new framework in the context of its support for replacing the research assessment exercise after 2008. It is essential that the sector fully engages with this important consultation with the aim of ensuring that the new arrangements maintain the excellence of the UK research base and have the full confidence of the academic community.

Professor Eric Thomas

Chair

Research Policy Committee

Universities UK

1 Summary

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 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 to be acceptable to and inspire confidence among the researchers whose performance is assessed.

Bibliometrics are 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 holds sound international databases of journals and their citations with good time, subject and institutional coverage. These data have characteristics (particularly in terms of publication and citation cultures of different fields) which mean that they must be interpreted and analysed with caution. They need to be normalised 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 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 recognised 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 then that 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 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 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, 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 sensibly to follow traditional faculty structures, this is no longer the unique structure for research activity. The most important aspect of the subject grouping, however, is the strategy that is used subsequently to normalise 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 normalisation 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 normalisation 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 licence 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 normalised 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 normalising the raw citation data prior to analysis will be central and critical.

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

2 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 (RAE2008). 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 postgraduate training. And the various indices will need to be integrated into an algorithm that drives the allocation of funds to institutions.

These quality indicators would ideally be capable of not only informing funding but also providing benchmarking information for institutions and stakeholders. They are also expected to be cost-effective to produce and to reduce the current assessment burden on institutions.

This document reviews and comments on background issues related to HEFCE's stated aims:

- to assess how far bibliometric techniques can be used to produce appropriate indicators of research quality; and
- to evaluate options for a methodology to produce and use these indicators.

This focus is specifically on published journal articles and not on published patents.

3 Bibliometrics as indicators of quality

3.1 HEFCE's tender specification

What is addressed?

Early in 2007, HEFCE invited contractors to determine whether bibliometric techniques could provide indicators that are:

- acceptable and valid measures of research quality;
- comprehensive across science, engineering, technology and mathematics (STEM)[2] disciplines and all UK higher education institutions;
- robust and reliable when applied at the level of broad subject groups;
- capable (at a broad level of aggregation) of identifying high quality research and of discriminating between varying degrees of excellence.

What is not addressed?

It will be noted that HEFCE's specification does not incorporate any explicit reference to academic confidence in the outcome. However, it does make a reference to the question of 'acceptability' although this begs the question of 'to whom?'

HEFCE will also need to consider the implications of using bibliometric-based indicators of research quality within a new funding and assessment framework. Some of these implications will not become clear until the system is implemented although, at the outset, cautionary statements might be made regarding the likelihood that:

- once any social or economic indicator or other surrogate measure is made a target for the purpose of conducting policy, it will lose the information content that would qualify it to play such a role[3];
- there are potential behavioural effects of using bibliometrics as researchers respond to indicators instead of reality;
- there will be scope for any indicator system to be manipulated over time, especially by 50,000 intelligent academics; and
- substantive criticisms from key stakeholders will arise because metrics focus on select aspects of research, particularly the outputs of fundamental research, rather than on the process as a whole.

3.2 Why bibliometrics?

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 then 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.

Indicators

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.

Income

One indicator of competence is the ability to acquire a high level of scarce income for research support. Such an indicator would be made sharper if we restrict the analysis only to income from peer-reviewed sources such as the research councils.

Income is a problematic indicator, however, and economists might challenge the use of 'input' as an indicator of quality under any circumstances. In this instance, there is a cap to the total available income: a limitation determined by policy as much as the scarcity or abundance of quality recipients or the size of the field as a whole. Furthermore, cost varies between theoretical and practical projects within a field. So, for these and other reasons, inputs are usually taken as only a partial measure, even if they are limited to a 'peer reviewed source'.

Outcome

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 satisfactorily to index the value of activity.

Outputs

Outputs overcome some of these problems. Furthermore, citation of outputs provides an apparent quality measure. For these reasons, bibliometrics provide 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 Scientific® Inc holds 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.

Because the Thomson databases provide the most effective and comprehensive 'currency' for indexing research performance they have become the de facto standard for many research evaluations in the natural sciences.

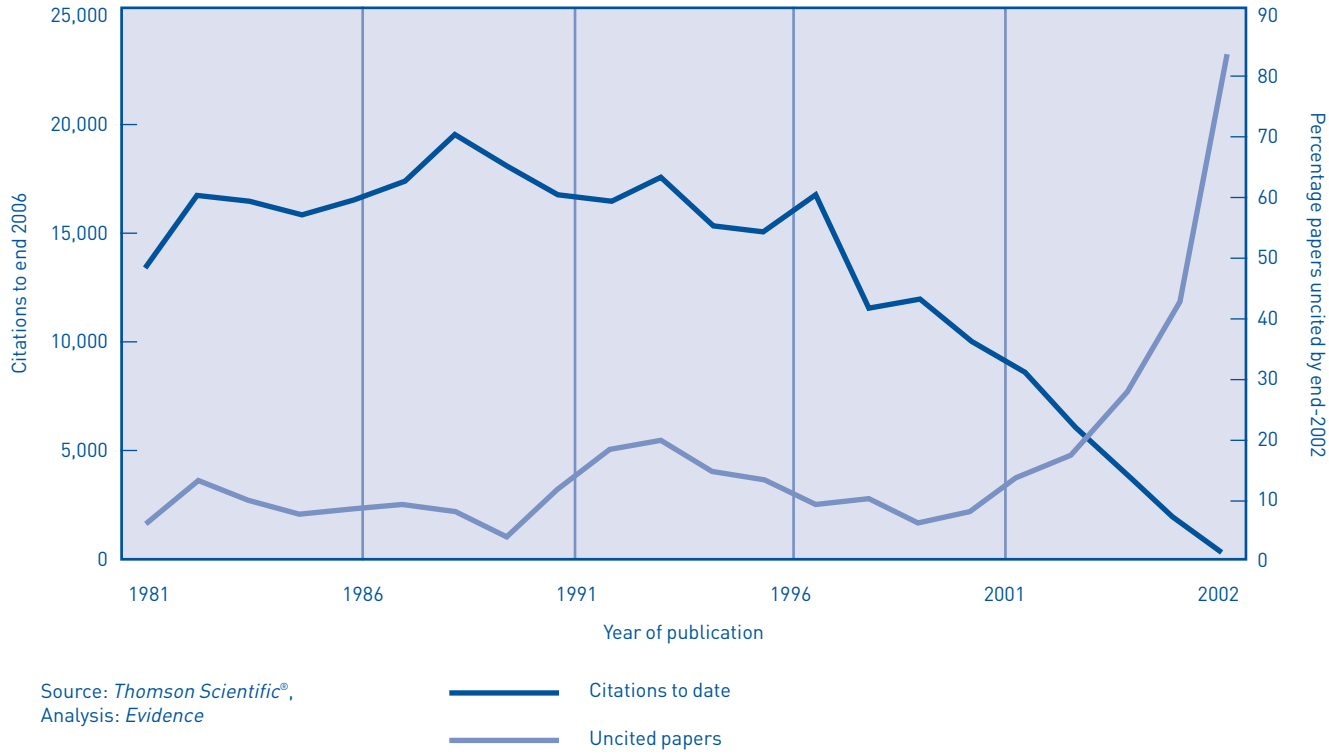
3.3 The nature of bibliometric data

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.

Citation accumulation

Once a paper is published it starts to attract interest from other researchers who may then use it as a reference point in their own work, adding it as a cited reference in subsequent publications. Thus, citations accumulate over time and uncited papers for any one year gradually fall in number.

Figure 1:
Citation accumulation for papers in Geological Sciences.



Source: *Thomson Scientific*®,
Analysis: *Evidence*

In this example, citations in geological sciences rise and then plateau after eight to ten years, but this would happen earlier in some other fields. The numbers of uncited papers fall over five years but some are never cited.

It is necessary to take these dynamics into account in any bibliometric analysis. 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. It is therefore necessary to adjust data by year or 'normalise' to some common standard. Normalisation strategies are discussed in more detail in a later section.

Disciplines differ

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.

In crude terms, biomedical researchers tend to produce more, shorter papers where methodology and prior knowledge are extensively codified in a dense array of citations. Physical scientists and engineers produce less frequent but longer papers, with more detailed content and fewer cross-references. These characteristics, not relative quality, affect typical citation rates.

Table 1:
UK publication and citation totals by broad subject group, 2002-2006

Subject group	Average cites per paper	Citations to date	Papers in Thomson journals
Molecular biology and genetics	16.15	205,597	12,733
Whole-organism biology	4.82	95,387	19,804
Physics	5.32	177,398	33,352
Engineering	1.96	50,696	25,886

These are stereotypes of the differences between fields, but they point to the challenge of comparability. It should also be noted that the stereotype is a partial truth only, because there is great variation between fields within these broad subject groups. Molecular biologists do not use the literature in the same way as organismal biologists and bibliometrics cannot compare the two directly. This further increases the complexity of satisfactory quantitative evaluation.

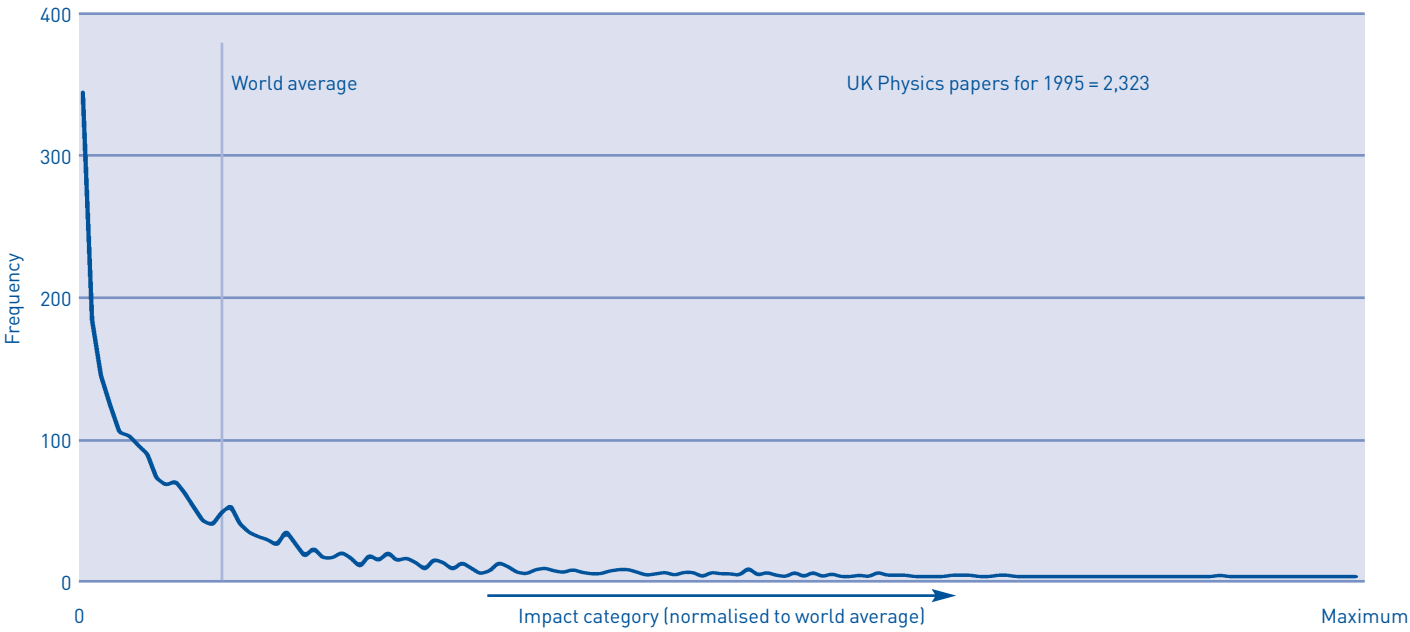
Rebased impact

Eugene Garfield, the founder of ISI, drew attention in the 1960s to processes for normalising impact by field and year. The process of normalisation to enable comparison across years and disciplines is also referred to as 'rebasing' the citation counts to a common standard. For this reason, normalised 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).

Figure 2:
A typical skewed distribution for citation data.



Source: *Thomson Scientific*®,
Analysis: *Evidence*

This is UK physics after ten year's citation accumulation. Most papers are cited less often than world average although the UK average is above the median and the world average.

Skewed data are difficult to compare visually and to interpret. The average is nowhere near the centre 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[4].

3.4 Can bibliometrics produce appropriate indicators of research quality?

Evidence has argued that bibliometric techniques can create indicators of research quality that are congruent with researcher perception.

Is the data source valid?

Are the target data – i.e. the Thomson journal databases – likely in principle to produce a valid outcome?

Thomson indexes a huge and diverse range of journals (serials) from countries around the world. It examines the citation relationships between articles in these journals and selects about 8,000 to 9,000 titles for inclusion in its leading bibliographic products, such as *Web of Science*, *Science Citation Index* and so on.

To be included in the database it is an absolute requirement that the serial should appear regularly and that it should have a well documented and implemented editorial and refereeing policy.

The selection of journals was in the past informed by panels of leading researchers but is increasingly influenced by relative citation levels. Journals are dropped when they become cited less often compared to similar journals and added as their citation profile rises.

Geographical and disciplinary coverage is also a factor. In the past it has been clear that the database had a biomedical, Anglophone and American-centric bias. That is still a partial problem but our analyses for the European Commission have shown that a wider range of languages and countries is now represented and that the relative coverage of social sciences and humanities is improving.

Thomson relies on commercial credibility. It is in the company's interest to ensure that what it covers is what researchers (for whom national agencies are usually the proxy customers) would agree is a sound representation of the best current research.

Overall, therefore, the database is a reasonable representation of higher quality research publications. Analytical outcomes of these data should lead to a valid indicator.

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 measure 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.

Our experience is that bibliometrics can be of sufficient utility to prove amenable and commercially valuable to research managers. We have extensive experience in identifying and addressing their limitations. Over fifteen years we have built up a comprehensive understanding of the procedures by which publication and citation data are collected and processed at Thomson Scientific. In-house, we constantly seek to improve the ways in which we manage, analyse and interpret Thomson data.

Nonetheless, because many limitations remain, we believe that:

- bibliometric indicators should not be used in isolation if they are presented as single citation averages;
- citation data should only be presented in the context of funding and activity data. This informs the design of our core products; and
- where bibliometric data must stand-alone, they should be treated as distributions and not as averages. *Our Impact Profile*™ product moves away from a citation average to profiling across the quality spectrum.

Do bibliometrics parallel other quality measures?

In a series of studies for HEFCE, Universities UK, the former Office of Science and Innovation and other UK agencies we have analysed 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 (a counter argument put by Professor J E Midwinter, University College London, is that researchers choose to submit a sample of 'typical' work [personal communication]).

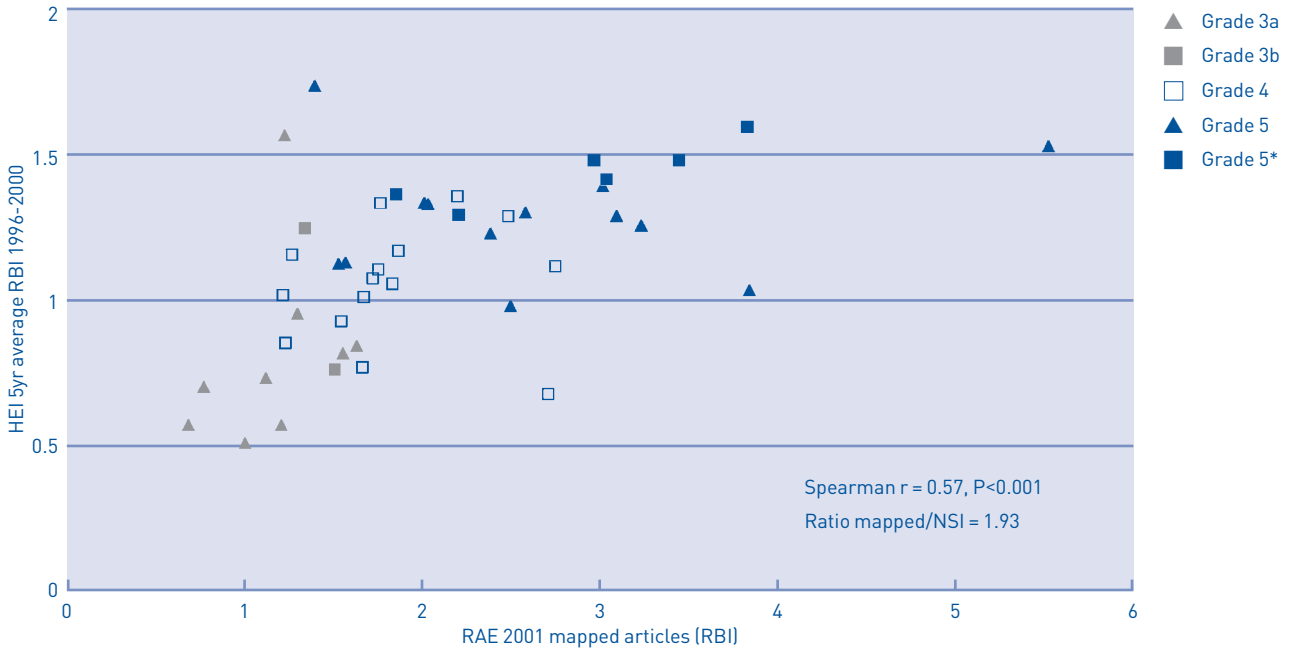
Second, a high proportion of the submitted articles are in journals catalogued by Thomson. This is particularly so for journals that are present at relatively high frequency in data on research outputs submitted in form RA2 for the RAE.

Third, within a unit of assessment, 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 earlier assumption about 'best work'.

Fourth, as is also evident from the following diagrams, the average citation impact tends to increase with the grade awarded by the peer review panel. The 'goodness of fit' between impact and RAE grade can be looked at via a more direct plot.(Figure 4)

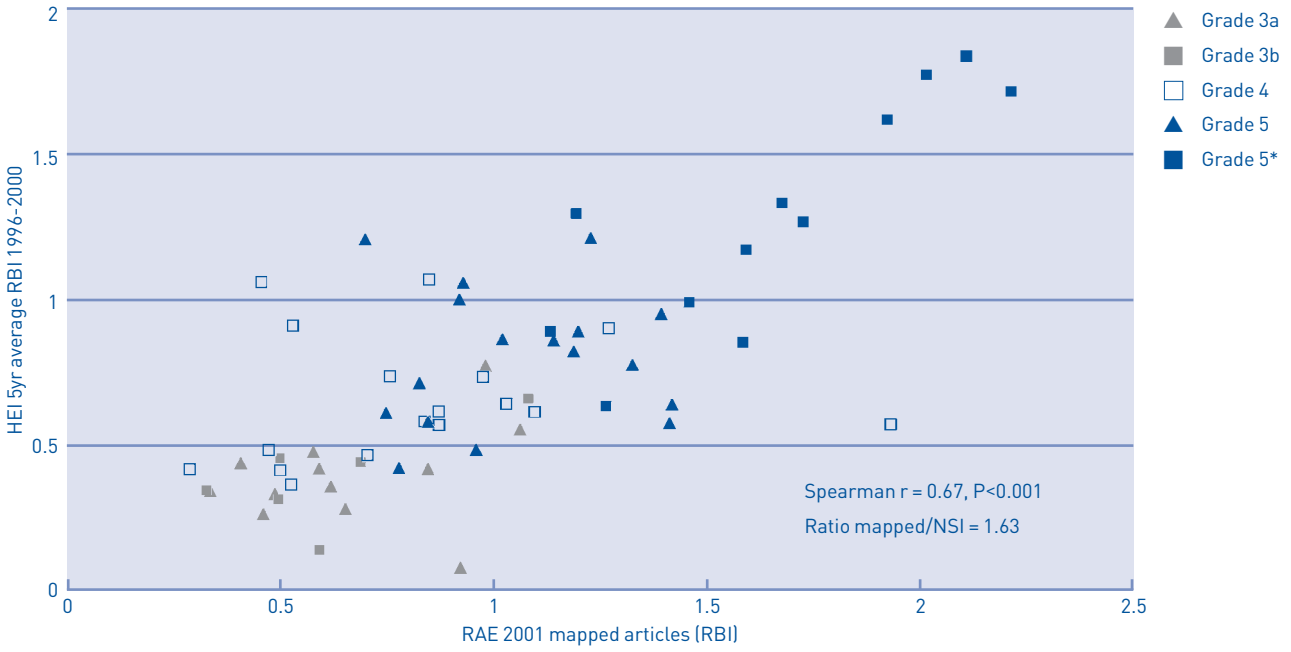
Figure 3a-b:
The correlation between average impact of publications submitted to the RAE by a unit and the average impact of all publications in that discipline by the same higher education institution over the same period.

Figure 3a
Chemistry (unit of assessment 18)



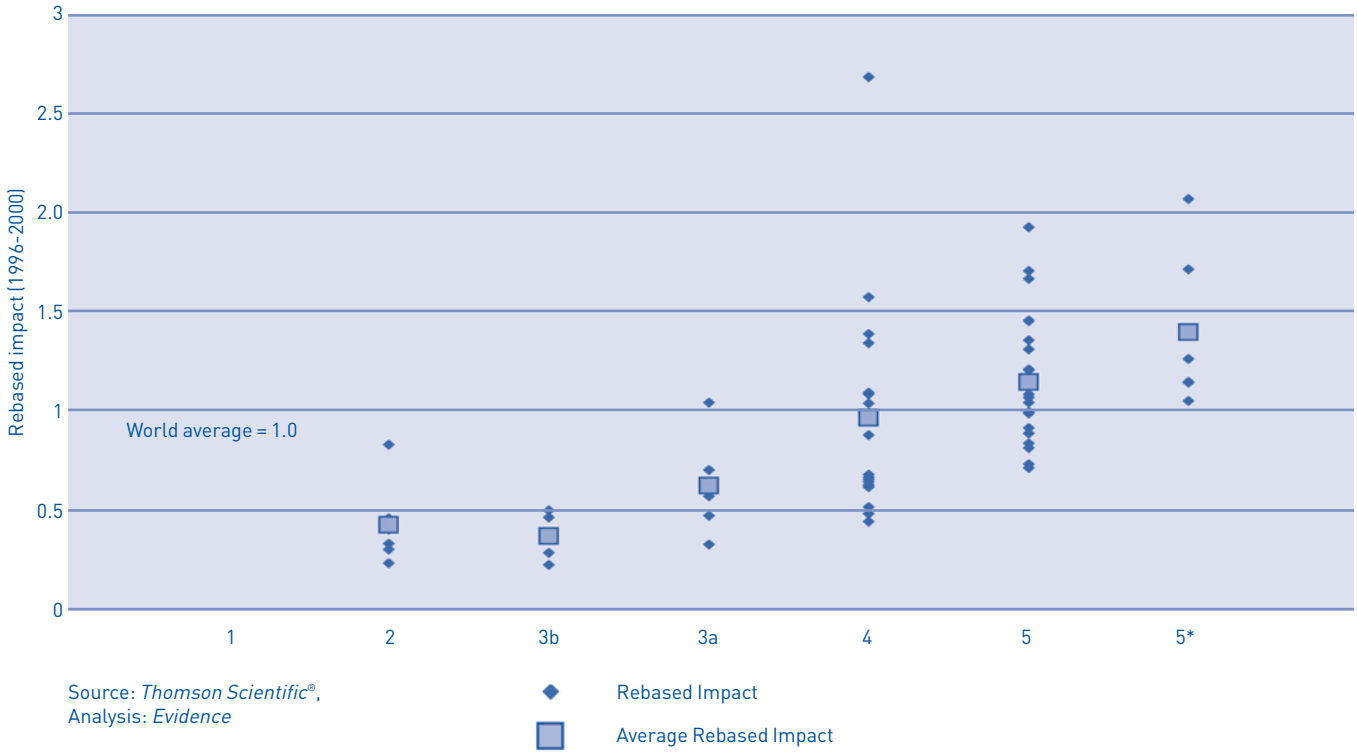
Source: *Thomson Scientific®*,
Analysis: *Evidence*

Figure 3b:
Psychology (unit of assessment 13)



Source: *Thomson Scientific®*,
Analysis: *Evidence*

Figure 4:
Average citations per paper - data from RAE2001 for biology (unit of assessment 14).



Source: *Thomson Scientific*®, Analysis: *Evidence*

Each blue data point is the average impact of the papers in an institutional submission and each grey square is the integrated average at a stated grade.

This shows several things. At a gross level, the average rebased impact at each grade (that is, the average impact taken across all the units that were awarded that grade) progresses upwards steadily with grade. The average value for grade 4 units is around 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.

Residual variance

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 world average would not enable one to tell whether that unit was 4 or 5 graded, or whether it might even be a very good 3a or a bibliometrically weak 5*.

Conclusions on the validity of bibliometric indicators

At a grand level, there is sufficient evidence available from experience and analysis to justify the general use of bibliometrics as an index of research performance. This has been found useful as part of management information in universities. But the application of this as a determining factor at a more narrow level is not justified by these analyses and there are many factors (considered below) that would affect specific outcomes.

3.5 Can bibliometric techniques provide indicators that are transparent and comprehensible?

As well as the technical limitations of bibliometric techniques there are perceived limitations and potential criticisms from stakeholders. A past lack of transparency has been one of the causes of a wide range of misconceptions that have developed. The transparency of the indicators is an issue of interpretation:

- how do most researchers and other stakeholders currently interpret bibliometric information?
- how can bibliometric indicators be better presented so as to improve their transparency?

The literature is insufficiently informative because most analysis and interpretation is done by experts (scientometricians, bibliometricians and other analysts), or at least those fairly familiar with the data and methods. As a result, some of the literature raises issues which are certainly worthy of consideration and help us to understand specific outcomes but which do not necessarily lead to pragmatic solutions that would support management applications. For example, few scientometric groups have experience in research management or in the differentiation between formal and workable outcomes.

Our experience with staff in institutions whom we have met during surveys, consultancy and at conferences is that a wide range of misconceptions about bibliometrics are in circulation, and have been for many years despite regular rebuttal[5].

There will need to be substantial work to dispel concerns. HEFCE will, of course, have the responses from the earlier DfES/HEFCE consultation on the original proposals for a shift to metrics. This is probably the most current database of informed opinion and might allow it to anticipate and prepare for many of the concerns that should be addressed.

It would be reasonable to expect that the next stage of consultation would make explicit reference to that. The outcomes could be compiled into a brief document for the relevant website, as part of the frequently asked questions that are likely to be required.

3.6 Bibliometric indicators of research quality in non-STEM disciplines

It should be noted that HEFCE's immediate priority is to develop appropriate metrics for STEM subjects with research in non-STEM areas still being assessed by a peer process of some kind.

Despite this, it is worth considering the limitations on the use of bibliometrics for non-STEM subjects. They fall into two categories. First, there are data-limitations where researchers' outputs (of various types) are not comprehensively catalogued in bibliometric databases. Second, there are 'behavioural' limitations because of the ways in which researchers in these disciplines cite, or do not cite, previous work.

Mathematics

For mathematics and statistics, journal articles formed over 90 per cent of items submitted on form RA2 for the RAE in 2001. Fewer than 70 per cent of these articles were in Thomson-indexed journals, however, which compares to 75 per cent on average for STEM subjects. This suggests that there are data limitations, which are compounded by behavioural factors that produce an unusually high rate of 'uncited' papers.

Social sciences

In 2005/06, *Evidence* carried out two studies for the Economic and Social Research Council (ESRC) on the validity of bibliometrics and the use of research performance indicators in UK social sciences[6].

Our study involved a review of available data sources, a bibliometric analysis of RAE2001 RA2 submissions for each unit of assessment and a consultation with researchers and learned societies. Bibliometric measures were reasonably robust for some subjects (for example: psychology and economics) because they were already in current use and accepted by stakeholders. In other subjects, particularly in applied and policy-related areas where key outputs were less likely to be in the form of journal articles, bibliometric indicators were of less significance, technically difficult to produce and less likely to be acceptable to researchers as a measure of quality.

In a second study we found that many performance indicators used in science were also employed by social scientists, but in an individualistic and expert way that did not translate into a systemic algorithm. There was a credibility gap between what researchers would themselves use within their community and what they would accept in application to their community.

Arts and humanities

Publication mode is different again in the arts and humanities, with a strong preference for books and monographs. Such outputs are not generated rapidly, nor are citations widely distributed and catalogued. There are fewer resources to pay for on-line databases, and none to pay for staff to index and cross-reference the citations in the publications.

An evaluation by *Evidence* confirmed that the Web of Science® publication and citation data for Arts and humanities disciplines were rather limited. It was found to be strongly Anglophone, limited in scope and with a marked American focus. It could be used to make some assessment of leading research units in the UK in some areas where journals were more widely used, but other areas were hardly touched and international comparisons were invalid.

There is unlikely to be an independent range of solutions for each of the potentially problematic issues which will arise. The outcome has to be methodologically workable, so we see the need for optimisation within a pragmatic model. This implies that there may be some policy compromises, i.e. a workable, cost-effective outcome that satisfies a high proportion of factors may still leave some issues unresolved. Ideal solutions for each issue could, by contrast, make an overall solution unworkable or prohibitively expensive.

In this context, we note that there is presently a focus on bibliometrics in isolation, whereas the ultimate implementation must set this indicator alongside others on funding and training (for example). Conclusions reached at this stage may therefore differ from those that would be reached where the research process is considered as a whole.

4.1 What variables should be included?

For HEFCE, this should be the core of its work to develop a new assessment system, assuming that stakeholder reservations on principles and methods can be met.

The bibliometric data can provide a number of component variables. This can be developed 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; and
- 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.

The fractional assignment of output to authors and institutions has also been mooted. This would potentially have a severe impact on collaboration, since it would reduce the net value of an output to an author who 'shared'. In discussing the assignment of citation counts (below) we note that fractional assignment of outputs favours those who collaborate least, such as the United States which compares better with the European Union in analyses using this approach.

There are some circumstances where volume might be useful for management information. For example, trends in output volume might indicate some aspect of activity while the relative size of two units might be an indicator of their research capacity. But these are issues for further analysis, not for an indicator process.

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. So although this is an informative indicator it is likely to favour larger institutions 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; and
- if more papers are published then the likelihood is that 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. (See also Table 1).

- 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:

- the numbers of uncited papers in any 'cohort' or sample falls over time, so account needs to be taken of time since publication;
- 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 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.

New methods

There are alternative methodologies of which the community will be aware and may wish to make comparison. For example:

- Hirsch (h) index – this is unlikely to prove effective for HEFCE's purpose because it works better for high output-high impact researchers and it produces only a single metric with low information content[7]. It is not applicable to the general body of researchers.
- Citing rank index – this is akin to Google page rank indexing, in that the number of links is moderated by the quality of those links. A citation from a high impact journal is worth more than citation from a low impact journal (we discuss this further below). This might 'control' for spurious self-cites from multiple low-value sources but it increases data requirements and management costs and the outcomes are, as yet, unproven.

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[8].

This index of research quality is widely-used by the scientometrics community. It has been employed extensively for many years by Thomson Scientific® 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 (which endorses it under the label of a 'crown indicator') and in our own PSA target indicators for the Office of Science and Innovation[9].

The characteristics of citation accumulation mean that impact must always be contextualised. That is to say, account must be taken of both the year and field of publication so as to normalise 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. Normalisation 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.

Here the naturally skewed distribution of citation data has been made visually more acceptable by sorting the data into 'bins' relative to the world average[10].

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.

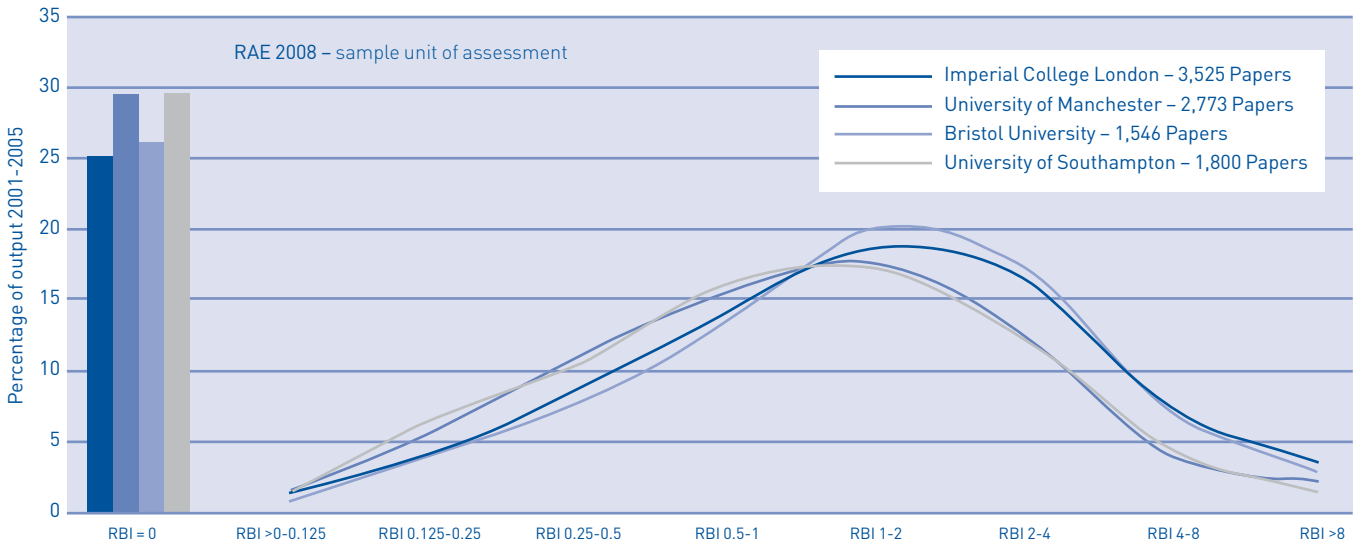
There is no doubt that highly-cited papers are associated with exceptional research, but the metric is a poor index of more general research activity. The threshold is so high that for many fields there would be few UK institutions that had more than a handful of papers in the index.

Profiles

This is the most informative approach to bibliometric assessments of research performance.

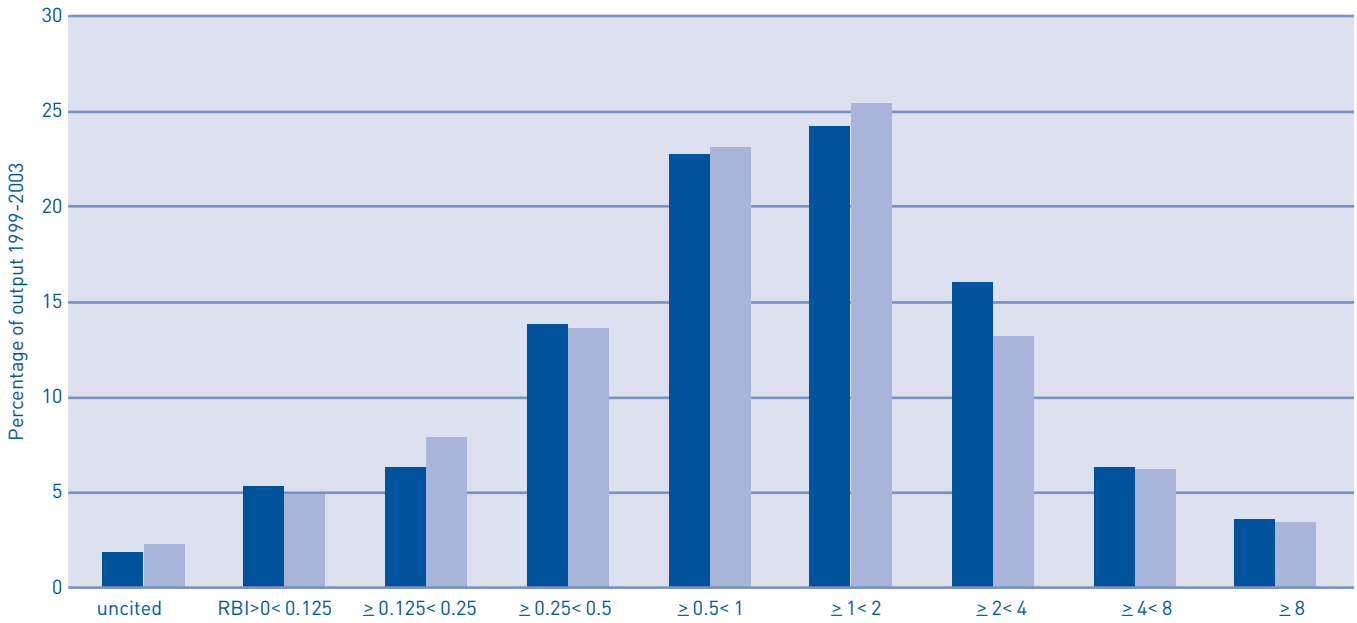
We noted above that bibliometric data should be considered in terms of distributions rather than averages where they must be used in isolation. This helps to overcome the extent to which the 'average' disguises the natural skew in the data. For management information, a profile of 'impact' is a helpful 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.

Figure 5:
What an *Impact Profile*™ would look like



Source: *Thomson Scientific*®,
Analysis: *Evidence*

Figure 6:
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



Source: unnamed RC, *Thomson Scientific*®, Analysis: *Evidence*

What advantage does this offer compared to average normalised impact? *Evidence* recently completed some analyses for a 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 (above) were almost indistinguishable.

The average normalised citation impact of the two groups differs markedly (2.39 versus 1.86) because of exceptionally high outliers in one group.

What values might be abstracted from such a profile?

- uncited papers as a proportion of total;
- proportion of papers cited less often than benchmark (UK average, world average);
- proportion of papers above benchmark; and
- proportion of papers cited at exceptionally high levels for field and year (where 'exceptional' requires a threshold [such as > 4 times world average] to be defined as for highly-cited).

All of these values could be used as a metric on their own, but drawing on them in a structured way from a prescribed distribution 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 metric components link to the underlying data and how they are derived, and it tests whether they make sense.

Trends in output and quality

It is arguable that any point-metrics, even those developed from a profile, are problematic because they only capture performance at an instant. An informed observer, by contrast, should be able to take account not only of position but also of trajectory.

It may therefore be desirable to include a factor that indexes current (or recent) performance against past performance and gives a greater weighting to improvement and to a sustained profile than one that is declining. There is a two-fold gain:

- less weight is placed on historical or reputational aspects and there is likely to be a greater reward for rising stars than old troupers;
- there is a reward for the management ability to sustain performance.

This also emphasises the integrated performance of a unit rather than the peak performance of individuals.

Conclusions on variables to use as metrics

None of these metrics is uncontroversial. Different methodologies would more or less readily produce such a range of variables. The h-index, for example, would produce only a single variable. *Impact Profiles* would produce data for all the above variables.

Not output volume. It should reflect productivity, but if used as a performance indicator it could spur unnecessary levels of trivial output. Fractional assignment by author and institution would be even more problematic. The required balancing factors make the system unduly complex.

Not uncited publications. They include both important work, which remains uncited because it valuably refers to blind alleys (innovative researchers are more likely to detect these than followers) and inconsequential work, which remains uncited for its lack of value. Indexing 'uncited' volume may be open to manipulation.

Not journal impact factors. They are not an index of quality and their use in metrics would perturb behaviour. Variation in journal quality is linked to other factors that need to be considered in data normalisation strategies (see below).

Overall, we anticipate that a methodology emphasising the distribution of cited papers is most likely to be acceptable to most stakeholders, but they will need convincing of fairness, balance and sensitivity to disciplinary culture and nuance.

The adoption of some measure that takes trajectory as well as snapshot performance into account would be desirable, to reward improving and sustained overall performance as well as individual peaks.

4.2 Combining indicators to produce overall ratings or profiles

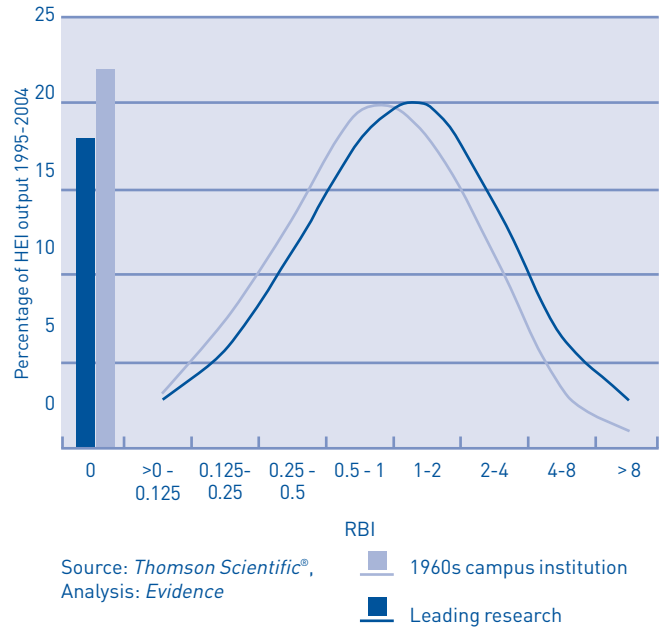
We can consider a series of steps to combining indicator components:

- for each adopted component it would be necessary to produce a single number or index value;
- indicators are then combined by scaling the numbers to bring them within the same range (as one would do with head-counts of PhDs and £millions of income);
- suitable weightings need to be determined for the contribution to quality and value judgments of other (funding and training) data; and
- the outcomes need to be integrated.

It is too early to propose a 'best' approach to combining indicators but it is recognised that Universities UK will want to have some idea of possible approaches and the following is as an example using the notion of a profile of increasing impact (normalised for year and by a subject related factor)[11].

The curve in the figure shows the proportion of papers in any category, not the numbers. Table 2 (over page) shows an example of sampling and weighting to produce a single metric for these profiled papers.

Figure 7:
Impact profiles for two research-based institutions.



Source: *Thomson Scientific*®, Analysis: *Evidence*

Table 2:
The weighted score from the metrics process is compared here with an 'impact power' measure derived from RAE data.

Category	Uncited	Cited < world average	Cited > 1, < 4 world average	Cited > 4 world average	Weighted score	Impact power
Weighting	0	1	2	4		
University A	640	993	1,053	303	4,311	5,703
University B	159	246	202	25	750	882

Source: *Thomson Scientific*®,
Analysis: *Evidence*

These are real institutions. The profile shows the distribution of quality, while table 2 above translates this into the categorised numbers of papers. These categories can be changed according to the significance attached to different parts of the profile.

Once collated into the designated categories, the numbers of papers are weighted and the weighted counts summed to produce a score: the 'metric' of research quality. The weightings can also be adjusted to attach greater value to some areas.

The methodology as implemented reduces the degree of selectivity between A and B, but a change to the weighting (for example changing the cited > 4 weighting from 4 to 8) would restore that differential.

The weighted score, once the categories and weightings are agreed, would then feed into a further stage in the development of HEFCE's metrics. At that point the bibliometric index would need to be scaled against the funding and studentship components.

It will be seen that this approach has some similarities to the revised 1* – 4**** grade system proposed by Sir Gareth Roberts and adopted for RAE2008. This principle would affirm for stakeholders a link in the progression from RAE2001, through RAE2008 and into the metrics system.

4.3 What is the correct citation count?

It is essential that there should be confidence that if citation impact is to be used to index quality then the citation counts are accurate and complete. It is also desirable that there should be agreement on how the accumulated citations are then categorised.

Database accuracy and completeness

Thomson uses well-established algorithms developed by ISI to link new publications through its reference lists to older publications and hence to create citation counts for those older outputs.

This is not a simple process. While the researcher could take a reference list and immediately interpret where to go in a library to find a specific item this is not always the case for a machine. Authors use diverse abbreviations for journals, provide inaccurate year and volume numbers, incorrect pagination and imaginative variations of article titles.

Thomson algorithms use field combinations to make a match, and the accuracy of citation counts is not usually seen as a serious issue. Even so, not all citations are collated, which may be an issue of concern to those in fields with typically low citation rates where the difference between three and four cites for an article may be significant. Opportunities to validate data may be desirable, to create a higher level of confidence in the underlying data.

More importantly, it should be understood that Thomson data provide an internal linkage. What is being counted in the citation count for a stated article is the sum of citations from other articles in the journals catalogued by Thomson. Citations from other (non-source) items including books and conference proceedings are not counted.

Table 3:
Single higher education institution sample: material cited by Thomson articles and not within the database

Unit of Assessment		Number of source articles	Number of cited items	Cited items that are not Thomson articles (%)
All subject areas		62,965	890,876	38
1	Clinical laboratory sciences	13,125	219,613	20
14	Biological sciences	9,501	202,200	24
18	Chemistry	5,149	96,768	27
30	Mechanical engineering	1,466	24,368	41

Source: *Thomson Scientific®*,
Analysis: *Evidence*

Does this matter? The table above shows the numbers for a large sample of articles for a leading research university. As much as a quarter of cited material in science is 'outside Thomson' and much more in engineering. We do not know the number of 'outside' items that cite into Thomson sources but we might assume that it is proportionate across fields.

The conclusion must be that a significant number of citations to published material are not indexed, that this matters because it varies between fields, and that this should be taken into account in normalising data before aggregation.

Cites to multiple versions

Because of the accelerating pace of research and because of the possibility of making material available electronically long before it is available in print, there is an emerging problem of 'version confusion'.

It has been a practice in engineering to make some papers available both as 'published proceedings' in well-established conference series and as journal articles. This is an unproblematic part of engineering culture, but it can reduce the count of citations to the journal version because researchers are already familiar with and are citing the proceedings version.

These cites to multiple version could be collated, in theory, and a single reconciliation could be produced. In practice it is infeasible mechanically to distinguish between the identical and the similar. This problem of reconciliation becomes much worse with draft versions and preprints available on institutional websites. Titles, length and content may also differ slightly but sometimes significantly between these versions.

If this tendency proliferates, and because immediacy is valued it is likely to do so, then the citation tallies for journal articles may become less valid indicators of impact. This is not likely to happen at the same pace in all disciplines. Some, such as physics and computing, already make more frequent use of pre-print and electronic posting than do others such as biology.

It is claimed[12] that search engines will be able to collate citations to multiple versions and to categorise similar citations to different versions, but the relationship of this to research quality remains unproven.

This problem may be significant for automated metric systems but is less important for peer review systems, where the peers are able to recognise a submitted article that stands as a signifier of record for what may have been several interim versions.

The consequence is that citation metrics that are valid now will need to be re-evaluated regularly and across different fields to ensure that they remain valid as publication practices change.

Fractional assignment of cites to authors

It is a moot point as to whether all citation credit should be attributed to all authors. This is likely to be an increasingly important issue in a world where a rising proportion of research papers are collaborative and where other research shows that collaboration produces higher-value papers.

For example, if an article is authored by two people rather than one should they each be allotted half of the citations for the purposes of indexing their individual research excellence? If the authors are working in separate institutions then should the institutional citation collations be reduced proportionately to take account of the multiple authorship? If there are three authors, two in one institution and one in a second then should the citations be split proportionately to people or to institutions? What if one author has recently moved institutions and was only giving the address elsewhere as a courtesy?

There are extreme examples: what should be done with astronomy papers that have over 100 authors, from up to 50 countries and a greater number of institutions? If fractional assignment is applied to the RAE data, would it affect only assignment of authorship within the UK or be applied internationally?

Should greater weighting be given to lead authors? How do we identify lead authors, when practice varies between disciplines so that the lead is sometimes the first named author and sometimes the last named?

What about fields where authorship is strictly alphabetical? In a recent study for one of the regional development agencies, *Evidence* analysed the 'lead institution' for highly-cited publications from select fields of policy interest. This appeared to show a disproportionate balance towards one institution, in environmental sciences. It emerged that, by chance, staff in that institution in that area had surnames with alphabetical precedence. The outcome, apparently indicating a particular strength at one location, was spurious.

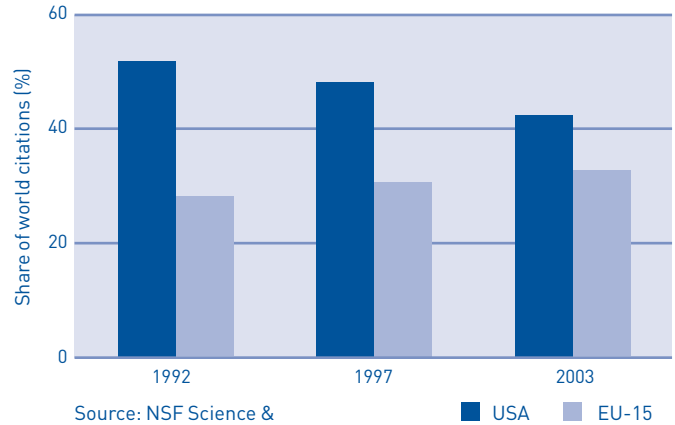
Fractional assignment and collaboration

If a procedure involving fractional assignment of citations were adopted then could this affect the level of collaboration with researchers working in and across institutions?

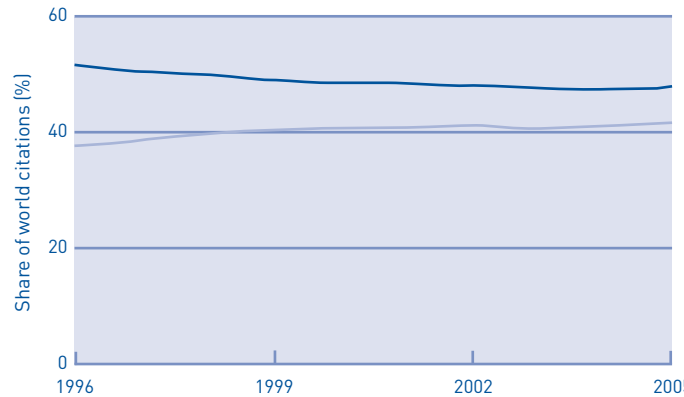
The UK has recently increased the level of its international co-authorship so that about 40 per cent of its journal articles have at least one non-UK co-author address. That is a fairly typical level of collaboration across the European Union. The United States is less collaborative (about 25 per cent of its papers have an international co-author) but is still linked to 170 of 192 countries analysed. The outcome has a marked impact on analyses of output that use fractional assignment. European Commission science and technology analyses (whole-article assignment) suggest that European Union and United States citation counts are higher than National Science Foundation analyses (fractional assignment).

The National Science Foundation analysis (top) shows that the United States and the European Union retain a smaller share of world citations whereas the Office of Science and Innovation analyses (bottom) suggest that they are now more similar.

Figure 8a-b:
Fractional assignment of citation counts to author institutions results in a lower value outcome for more collaborative entities.



Source: NSF Science & Engineering Indicators 2006



Source: OSI PSA target Indicators 2006

There are implications for the RAE metrics. Fractional assignment of papers and their citations would work against collaboration: do not share your glory. Evidence has an example of this for a collaborative programme in Austria, where the lead institutions collaborated on many outputs but reserved the highest-impact articles for single-institution authorship.

Thomson has reviewed fractional citation assignment and concluded that it does not significantly affect outcomes of large-scale analyses. If it does have an effect at a fine level, then that effect is itself open to interpretation for the reasons discussed.

For *Evidence*, our practice has been to avoid fractional assignment entirely and to allow all authors, institutions and countries the full benefit of the citations received. The arguments for treating the data otherwise are too complex to produce any wholly satisfactory general approach. Furthermore it is clear that much of the data would need to be reviewed manually rather than algorithmically and the emergent issues would then proliferate.

Cites in terms of who or what is citing

It is argued that some citations are worth more than others because of the author who is citing or because of the journal from which the citation is made. This is akin to Google's Page Rank™ algorithm (link counts are like citation counts) as we noted in a previous section.

For example, if an output from a highly-cited research group cites some prior article then the fact that they are perceived as 'high impact' themselves should make the citation more significant than a random cross-reference. Similarly, a citation from an article in *Science* might count for more than a citation from an article in *Scientometrics*.

A simple way of adjusting incoming cites for 'value' would be to use journal impact factors, but these are themselves problematic because citation rates vary between fields, as noted. Molecular biology citations would be counted as more valuable than population biology citations, which might be seen as inequitable within schools of biology.

Evidence has not sought to make use of weighted citations because we remain unconvinced that utility can be revalued in the ways indicated. If an item is found useful enough to justify a citation then that is a universal expression of value. We accept that something that contributes to cutting-edge research might be deemed particularly valuable but we would not accept that a journal impact factor is the right scale to express such relative value. We would be prepared to revise this opinion, particularly in regard to open-access and archived publication databases, if future work showed that such weighted measures made sense in terms of researcher utility.

Self-citation

Is a citation to one's own work different from a citation from someone else? Some people believe that there is a difference: that self-citation is an undesirable factor in citation analyses and that self-citation should be excluded prior to evaluation.

Citations are part of the process of building on prior knowledge. They establish a thread of development, a natural and appropriate part of which is reference back to one's earlier work. It is a cultural necessity.

A recent paper[13] has raised new arguments about the influence of self-citation: "the more one cites oneself the more one is cited by other scholars ... our models suggest that each additional self-citation increases the number of citations from others by about one after one year, and by about three after five years ... there is no significant penalty for the most frequent self-citers".

The statistics are objective but raise deeper subjective issues about why self-citation might be inappropriate, whether multiple self-citers are doing anything different from the normal sociology of research, and why self-citation might be penalised – which most researchers might find a surprising proposal!

What is self-citation?

First, what is 'self-citation'? If Adams (2005) cites Adams (1996) then the case seems simple. But what if King and Adams (2007) cite May and Adams (1998)? Does the presence of Adams taint the reference? And what if someone in May's group at Oxford cites May: is that still self-citation because it is within a small and local team or is the cross-reference now not 'self'?

If we take a large and research-active group, say a team at the Medical Research Council's Laboratory of Molecular Biology in Cambridge then we see constant cross-references within the laboratory, often with overlapping authorship. The laboratory is unquestionably cutting edge: more Nobel prizes than any other institution in the world. It is entirely logical that it makes extensive reference to its own work since that includes some of the most innovative current work.

It is infeasible to build up a significant body of self-citations without a significant body of peer-reviewed publications in the set of journals covered by Thomson. And if those self-cites have passed peer review then they must have been seen to be appropriate.

This is, we believe, why Fowler found a link between multiple self-citation and additional non-self cites[14]. These were high-profile, leading edge groups referring to their own work and being 'trailed' by many others. To 'penalise' these self-citations would be antithetical to research.

Second, how do you remove self-citation? We need to check all the papers citing Adams (1996)[15] and remove any citations from Adams over the period 1996 to present. Of course, we need to make sure these are all from that Adams (in Leeds) and not another Adams elsewhere. There is a problem, because there were actually three 'J Adams' in Leeds during the period and the target 'J Adams' was also employed at Imperial College for part of the period.

A machine check is therefore likely to be problematic. In practice, a researcher might reasonably ask for a detailed validation of all their citations to check that only the right ones were removed. It would be no good arguing that 'on average' a mechanical system would work. The individual is indifferent to the average and will rightly be perturbed at erroneous outcomes.

Third, what is the behavioural effect of removing self-citations? HEFCE would be sending a signal to the system that self-citation is wrong. Behaviour might then change as a result. People would alter the way they cited, perhaps reducing their own valid connections but seeking less valid cites from other sources. There would be a risk of a serious disruption to the underlying culture. In the short term, the reduction in self-citing would undermine the UK's international citation profile and institutions would fall in comparative league tables.

Negative citations

There is frequent concern that some papers accumulate significant citation counts 'because they are wrong'. There is little evidence of this. The bulk of ill-conceived work that does pass editorial scrutiny and reach publication probably remains uncited because it is also trivial. A well-known example of frequently cited 'wrong' work (Fleischman and Pons on cold fusion) still had a significant and non-trivial impact on work in its field. Interestingly, however, it was not published in a conventional medium because of the researchers' concern that it would not pass the initial editorial hurdles.

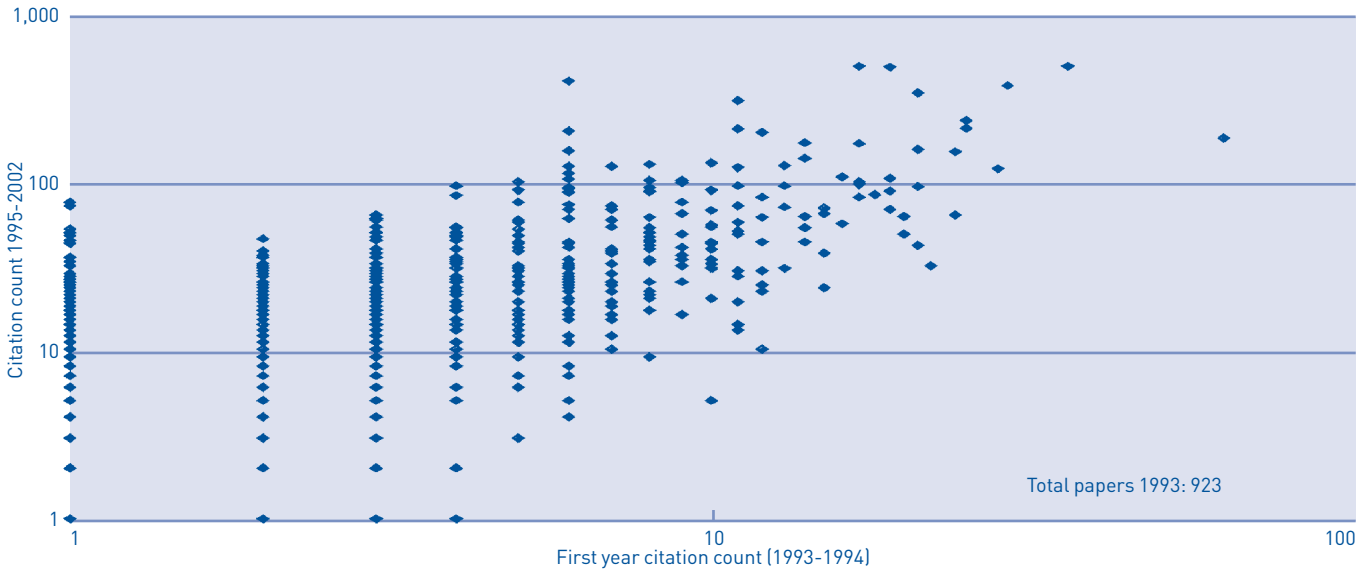
4.4 The timeframe for bibliometric analysis

Timeframe affects bibliometric data and indices. Immediacy is provided by recent data while smoothing is provided by longer time 'windows'. There are two operational issues:

- The time frame over which publications are sampled;
- The time frame over which citations to those publications are counted.

It can be argued that although papers in a short, fixed-time window will not accumulate very high citation counts, they could demonstrate sufficient differentiation to categorise quality. We have shown (figure below) that early citation rates are a good predictor of longer-term performance. But, while this may be valid for large samples, it is not readily acceptable for individuals.

Figure 9



Source: *Thomson Scientific*®,
Analysis: *Evidence*

For large samples of papers, the citations accumulated in the first two years after publication are a good guide to the likely citation count for years three to ten. For individual papers the trajectory is less certain[16].

There are different citation accumulation rates for different fields. Biochemists cite rapidly and move on while ecologists rarely expect papers to be cited within 18 months of publication. A five-year window of analysis will work well for the former group but poorly for the latter, although both would fall in a single metrics assessment area in HEFCE's structure. Disciplines within physics, such as astronomy and solid-state physics, may be affected similarly.

'Mayflies' and 'sleeping beauties'

There will be concern about whether a short assessment timeframe provides a valid picture of the impact of an individual 'outlier' publication. This may be of two kinds: the mayfly that shows exceptional early impact; and the sleeping beauty, ignored for years and then found to be of critical value.

Fleischman and Pons (1989) work on cold fusion might be taken as a mayfly, because it stimulated much attention before being rebutted. Most mayflies are of a lesser order but, like cold fusion, present stimulating ideas even if the content proves less sound. There is no convincing evidence that there are a disproportionate number of papers that gain exceptional transitory acclaim.

There are papers which suddenly attract many citations, when they had been previously unnoticed for several years after publication. These are no more frequent than would be expected by chance, however, given the underlying distribution of citation patterns[17]. It is unlikely this affects many researchers and might be little different in effect from the outcomes of peer review on unnoticed work.

4.5 The population for assessment

Population

'What is being assessed' is an issue that we refer to at several points. The metrics on research inputs and outputs are associated both with institutions and with people. 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?

The new metrics could cover population in terms of:

- discipline = 'all the chemists';
- management unit = 'all the staff in chemistry'; and
- staff grade = 'all the tenured, research-active staff in chemistry'.

Part of this rests on whether funding is there to support units within an institution, which would then put forward the staff they have to pay, or research strategy within an institution, which may then group researchers according to interdisciplinary and strategic programmes.

Research-active?

Is assessment of all the activity of an institution or only of the research-active component? This is a non-trivial question because it links into the metrics methodology in two ways:

- what volume of material needs to be made available for assessment?
- how are outputs going to be associated with people?

And there is a behavioural outcome:

- if the assessment of outputs relates to total or only to research-active, what effect does that then have on the population?

Who selects staff?

In the past, institutions have pre-selected staff for assessment. They are thought to have become more selective in doing so (volume was less in RAE2001 than in RAE1996) because outcomes affect reputation as well as funding. If institutions are to select staff then there will need, for equity, to be some consideration of whether that selection needs to be policed and validated and whether all institutions are expected (and seen) to submit broadly the same staff group or select one of their own choice.

Definitions

HEFCE must define the population in theory and in practice. The present system takes a particular census date and then institutions submit a checkable list of staff on stated grades employed at that date. This is not the total research population either at the census date or over the assessed period but it is well understood and readily verified.

For the future, on a shorter cycle, there will have to be:

- an operational definition of what is to be included, in terms of institutional and individual research metrics;
- an operational definition of who is to be included;
- a definition of when this is done in terms of cyclical census dates and staff mobility; and
- a protocol for validating the presumptive population.

4.6 Equal opportunities

One should review career trajectories in relation to publication profiles properly to assess potential implications for equal opportunities (for early career researchers and others). This would involve linking individual staff records to historical publication records, which is an intensive manual data-processing task.

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.

HEFCE analysis

HEFCE reviewed the data from RAE2001[18] and concluded that this revealed issues not about the assessment process but about the selection process in institutions prior to assessment.

- staff aged over 30 were more likely to be selected;
- men (64 per cent selected) were more likely to be selected than women (46 per cent). Like-for-like comparisons showed that men had significantly higher selection rates than women over a middle-age range between 30 and 47;
- unadjusted comparisons showed selection rates of around 58 per cent to 60 per cent for staff from most different ethnic groups;
- staff from black ethnic groups had a lower selection rate of 37 per cent. This was partly because a higher proportion of these staff were employed in departments which did not submit to RAE2001; and
- disability was not a significant factor in the propensity to be selected.

No bibliometric difference on past data

Evidence found that bibliometric impact of selected staff who submitted outputs for assessment showed no great differences:

- between men and women, if the highly-cited staff are excluded; and
- between researchers from different ethnic groups.

Peers versus metrics

Equal opportunities considerations point to a number of issues where peer review is entirely capable of adjusting perceptions for particular cases but metric algorithms can make no such adjustment.

People who have taken career breaks (typically, women researchers with a family) may have a disjointed publication portfolio. This can be flagged but a metric adjustment for this needs to be discussed.

Early-career researchers have a smaller research portfolio, are still on the learning curve in their discipline and are thus less likely to have a large or widely known body of work. They are likely to have lower citation counts relative to their field, though their trajectory would be entirely 'readable' for an appointing committee.

We have no evidence that new researchers have innately weaker outputs, and the use of fixed time-windows would put their citation accumulation on a par with established staff. As a counter-factor, their output might actually be greater and they would be citing and boosting their own work. Furthermore every profile will include some new researchers, who could be flagged appropriately by employers.

The problem of institutional responses

More systematic issues affecting gender and ethnicity are not made worse by indicators, except insofar as the metrics may accentuate differentials because the distributions are skewed. However, if there was a belief that new researchers systematically affected profiles then that would in itself be problematic.

The important consideration here will not be about the metrics but about the attitude of institutional managers to metrics in relation to different groups of staff.

4.7 How the broad subject groups should be defined and constructed[19]

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.

Bibliometrics

Evidence established a methodology for aggregating research activity into subject groups at fine and coarse level in work for HEFCE in 1997[20]. 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.

Customised 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 after consultation.

What is physics?

The approach *Evidence* took[21] was to look at the use that different subject groups made of the literature. We can see that physics (unit of assessment 19) submits a given range of journals for RAE2001 assessment whereas chemistry (unit of assessment 18) submits a different but overlapping range. Both are similar to materials (unit of assessment 32) but quite different to biology (unit of assessment 14).

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.

The EPSRC data, drawn from publications associated with researchers funded by different physics-related programmes (i.e. as seen by the research council for research purposes), map well onto this RAE analysis.

So, whether we look at university units or we look 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 boundaries with different subject areas.

Other units, other data

This journal-orientated proposal is deceptively simple. It should not be forgotten that the subject clustering has to work with a diversity of data and institutions.

What would work (has worked) readily for publication data in the older 'big civics' does not necessarily make sense for graduate schools, or for newly emergent and trans-disciplinary research areas, or for applied research in post-92 institutions.

The clustering will not only have to be mapped to data (hence to the categorisation of data held by third-party sources) for funding, training and publications but will also have to work with the appropriate reference systems for normalising all those data in a metrics' structure.

If the funding data are taken as an example, then funding availability varies as much between sub-fields in biology as does the typical citation culture.

Compensation factors

There is no simple solution to clustering. No solution is likely to be particularly satisfactory. It will be essential to appreciate the finer granularity within any clustering and then to ensure that sufficient account is taken of the fine grained differences to build in compensating factors that adjust for differences in such factors as culture, funding and researcher flow.

4.8 What is assigned to subject groupings?

Staff coverage has been referred to above in the context of 'population'. This is likely to be a contentious area and methodology must be linked to policy objectives.

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 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 HEFCE 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?

Who links staff to articles?

Options for making explicit staff-output links are either through central, procedural assignment or through assignment by the institution:

- if assignment is made by HEFCE, then institutions should require validation of the assignment, so it seems unlikely that they would not be involved in some way; and
- if the lists of publications for assessment are provided by the institution then that will require validation as well.

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

For example, the pathway to identifying and analysing 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 to that required to identify and analyse 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, 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 then this would significantly reduce the costs of any subsequent part of this development. But this is unlikely to prove satisfactory from a researcher perspective.

4.9 Normalisation strategies and aggregation

Creating a basis for data-comparability within the 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 (normalisation) factors to enable data to be brought together for comparison.

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 normalising 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 normalisation at some finer level. Over-normalisation will be as problematic as under-normalisation, because it will remove the subtle differences that the exercise seeks to identify. It is therefore critical to determine the appropriate level for normalisation.

Normalisation

Citation rates vary between field (and sub-fields) and citation counts accumulate over time. At which level should bibliometric data be normalised?

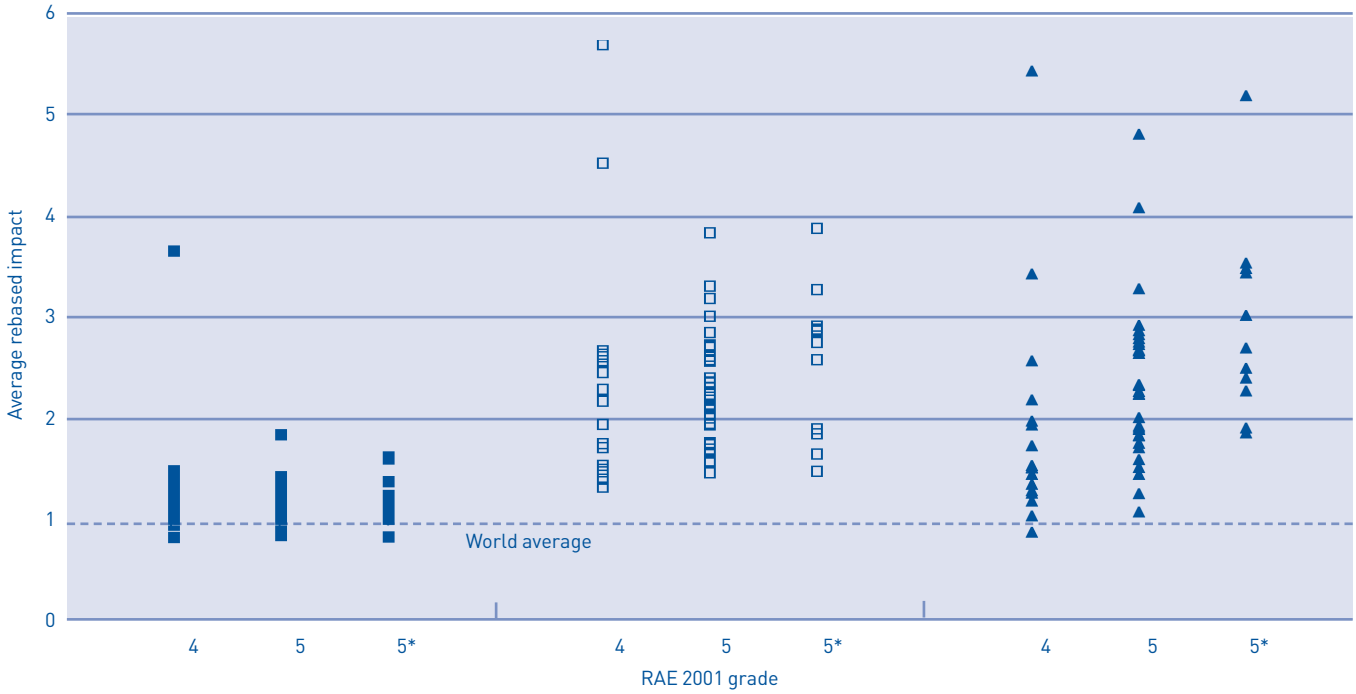
- the broad subject field;
- fields below this (for example, units of assessment);
- Thomson journal categories;
- at the level of journals themselves?

We tested some options to see the effects of these on quality rankings in psychology, biology and physics and to explore differences. We calculated the normalised citation performance of UK research units for each of three levels of article-aggregation (journal, journal category, unit of assessment – where several categories map to each unit of assessment). We then compared this with the grade awarded to that unit in RAE2001. We found that the correlation between average normalised 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 normalised at journal level. But higher graded units had a statistically significant higher impact when normalisation was relative to the Thomson journal category or to the journal sets mapping to the unit of assessment[22].

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 impact for each article normalised against a world average. The data are grouped according to the RAE grade awarded to the institution.

Figure 10:
Variation in citation impact for biological sciences (unit of assessment 14) using RAE2001 submitted articles and grades awarded.



Source: RAE Manager, *Thomson Scientific*®, Analysis: *Evidence*

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

Normalisation and applied research

Another possibility is that a finer-scale normalisation would also separate clusters of applied but lower-cited research from frequently-cited fundamental research of topical interest[23]. Thus, the relative value of applied work is lifted at the journal scale but swamped at the field level.

Conclusions on normalisation

While the pattern varies between broad fields, an upper and lower boundary to the granularity of sensible normalisation is apparent. Above unit of assessment level the differentiation between fields is lost. Below 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 normalisation 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 will 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 normalisation 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[24].

Other data

The final step in creating metrics will be the integration of bibliometric with other data. At this point there will need to be some weighting attached to the inputs from different sources: how much is the bibliometric indicator 'worth' compared to the funding and training metrics? The significance of these elements may vary between subject groups and it should not be assumed that there is one correct weighting to apply.

4.10 Accommodating research differences between subjects

Account will need to be taken of not only differences between research areas but also differences within them, particularly in the balance of mode of activity.

Engineering conference proceedings

Engineering and technology has culturally made more use of conference proceedings as a key output mode than is typical in the natural sciences. This is because conferences are a better route to communicate with the most likely user audience, in the private sector. There are a series of prestigious conferences the proceedings for which form a record of acknowledged quality.

However, engineers have made increasing use of journals over the last decade, as the data returned to RAE1996 and RAE2001 show. The differences between subjects are therefore less evident than they were historically, although computer studies is still very dependent on proceedings rather than journals.

Irrespective of this changing level of journal use, the principles of excellence should hold good within any subject areas. The problem is that bibliometrics will provide only a partial analysis in these areas whereas they will provide a more complete comprehensive analysis in other sciences.

There is therefore likely to be less confidence in the outcomes of bibliometric analysis for engineering. For this reason, it may be appropriate to give somewhat greater weight in this subject group to the metrics on funding and on training. Such considerations may also apply in other areas.

Basic and applied research

Within subjects and between institutions there is a varying balance between basic and applied research. Because the target for any applied research is partly outside the research base, the impact is only partly measured by citations coming from within the research base in later publications.

It is therefore inevitable that applied research will tend to be cited less, but not necessarily have lower impact if impact includes economic as well as scientific factors. It will be necessary to consider how this can be addressed.

This is an issue which will be of significance for some stakeholders, particularly research users, who will expect the application of research to receive some parity with blue-skies research. The implications for national policy and for government innovation strategies, are obvious. In a bibliometrics-based system there is no step at which non-academic user evaluation can be applied.

Variation between fields within subject groups

We noted that citation rates vary between areas in terms of time to first citation, citation accumulation and citation totals. Such differences can be addressed by using appropriate normalisation factors (as noted) and these can be applied between areas within macro-disciplines (such as the broad subject groups) but the differences need to be identified.

It will probably be necessary to adjust within-discipline clusters for factors which go beyond normalisation and may affect the time-frame for the analysis and other factors more readily identifiable to a subject specialist within peer review. To take the example of biology, a citation time-frame of five years would work well for molecular biologists but much less well for organismal biologists.

4.11 Interdisciplinary research

What is the argument for separate interdisciplinary assignment? The notion is that interdisciplinary work is treated and valued differently from other work. If so, it might be cited differently and its impact value (citation count) might be systematically less than core disciplinary work. This is unproven, but the notion of different treatment does require some consideration of why and how a bibliometric evaluation might be applied.

The 'interdisciplinary' nature of research has no proper (testable) definition. Unless it is possible to assign research outputs to an interdisciplinary basket then they cannot be treated separately for assessment. Assignment cannot be made at a journal level since some of the most prestigious journals (e.g. *Nature*, *Science*) are explicitly multi-disciplinary in content and clearly contain some interdisciplinary articles.

Identifying interdisciplinarity

Nature articles are assigned by Thomson to specific subject categories on the basis of their citations (i.e. an article with many references to physics journals is probably a physics article). A similar approach could be taken to a broader definition of interdisciplinarity. Those articles which cited prior knowledge across many subject categories would be objectively more interdisciplinary than those which cited only within one category.

Author definition of interdisciplinarity is probably not sensible. If a very large volume of work were to be so assigned then it would create serious methodological challenges. Less than one per cent of outputs submitted to RAE2001 was flagged for evaluation across panels but it would be reasonable to expect this to rise if it were known that a claim of interdisciplinarity led to peer review rather than metric assessment.

Innovation at the margins of core subjects

Interdisciplinary work is often at the margins or interfaces between established subject domains. It is desirable to evaluate the general association between indexed quality of outputs in core and margin of the subject clusters. Innovation also occurs at the margins. Margins can be identified on the basis of the frequency of journal usage. However, if higher education institutions determine staff selection and then assign staff to atypical subject clusters then the issues of interdisciplinarity might become marginal to the disruption to bibliometric profiling.

If the emergence of innovative research areas, which tend initially to be marginal until more widely accepted and adopted, is a 'core and margin' issue then the methodology should be adapted to avoid incentives that reduce the current dynamism of the UK research base. A panel can respond to the nuances of work recognised as innovative but low impact whereas data metrics on their own cannot.

Conclusion

A detriment to interdisciplinary work or to work at the margins of core disciplines is unproven but if it exists and is reified in metrics assessment then that would be disadvantageous to UK research innovation. It is desirable that HEFCE should confirm that no such systematic detriment exists.

4.12 Data acquisition, collection and preparation for analysis

Bibliometric database

Evidence acquires, collects and processes all the UK publication and citation data held in Thomson databases every year as part of its normal business. 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.

It will be essential to understand the nature of the data, problems with data processing and potential problems in year-to-year changes. *Evidence* has experienced many of these over the last few years. Users need to be aware that Thomson does not normally supply validating routines or documentation of changes.

Part of the process of analysis would be establishing the likely annual timetable for data collection and reconciliation. This could require the linking of staff data from the Higher Education Statistics Agency, bibliometric data from Thomson, local processing and on-line sign-off by institutions. These various elements will need to be explored and explained to institutions.

Factors that will need to be incorporated are: assurance on continuity of baseline data supply; annual cycle; time-frame for citation census and cut-offs; variation in data structure between products required for UK specification and global baseline; year-to-year variations in data compilation, structure, aggregation and content within core Thomson databases (e.g. journal additions and deletions); data conventions; article types (article, review, editorial, note etc).

Assignment to higher education institutions

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 analysing 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.

It is possible, indeed likely, that opportunities for data collection and preparation will change over the lifetime of the assessment methodology. For example, the shift to open access publishing may produce more comprehensive 'libraries' of outputs for which quality standards can be applied in a transparent and measurable way. This is not yet feasible, however, nor would many researchers accept that current methods (such as page ranking) are sufficiently well understood to serve as indicators for HEFCE's purpose.

A key issue will be whether a central publications database is held by HEFCE or whether publication data are supplied by higher education institutions. If the former, then institutions will need access to validate the correct assignment of records to be associated with their staff.

Assignment to staff

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 Guillemot's in UK data). These can only be distinguished manually.

The UK publishes about 100,000 articles a 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 co-authors are increasing.

It should be assumed that researchers will need access to the assessment database in order to validate the correct assignment of records with which their name is putatively associated.

Citations

If self-citation is identified by HEFCE as a problem, and these citations are then removed from the system, the citations to each UK paper will also need to be analysed and the self-cites removed. There were about 8.5 citations per UK paper on average over the five-year period from 2001 to 2005, or about four million citations to about 500,000 papers.

It will be argued that self-cites can be identified automatically, but each researcher in a low-citing field may want assurance about the accuracy of this procedure, if not personal validation. Furthermore, the accuracy of creating a 'self-cite-free world benchmark' for normalisation will also be an issue.

4.13 Potential costs and workload implications for HEFCE and institutions

The cost of indicators is a three-part calculation. This will involve:

- a licence from Thomson Scientific® to access the data;
- the expert cost:
  - initially, setting up the indicator system
  - periodically, running the cyclical analysis
- the institutional cost of supplying stated data to ensure matching and validation.

Only the first part of this can be readily stated at the outset, while the other two components depend largely on the specific methodology required.

Thomson data

To carry out any work to develop and test proposals for a methodology it is necessary to access appropriate data from Thomson. This involves a licence cost for the data. Once paid, that licence fee would cover any subsequent re-use of the same databases during the current calendar year (Thomson's financial year is January-December).

Some other background databases will also be required for benchmarking, including the National Science Foundation indicators which are used to establish UK and world baselines.

If a more detailed global approach is taken, to address self-cites for example, then the costs will rise proportionately.

System costs

If HEFCE decides to set up the indicator system centrally then it will need to create and develop an expert unit to administer the system.

The only function of this unit will be to run the cyclical metrics analysis, so the overhead cost of holding sufficient expertise in-house will be significant because of the peak demand and the range of issues that may then need to be addressed.

If the data are held centrally then secure institutional access for validation of address and researcher assignment will need to be developed.

Frequent updating will also be required as researchers move between institutions (assuming that the data records follow individuals rather than sticking with institutions).

Institutional costs

The costs to institutions of using bibliometric quality markers will depend on the approach taken. There are significant start-up costs but subsequent annual costs are likely to fall quite quickly once the metrics system is established.

An appropriate programme of work could put in place a bibliometrics-orientated system that would not only enable the annual cost to institutions to be made fairly small but would actually improve their local research management information systems, reducing net costs by adding substantive local value.

If data are held centrally then institutions will need access to the HEFCE database to confirm:

- the assignment of records to institutions (each record may be assigned to multiple institutions);
- the reassignment of articles to their institution when staff are recruited from elsewhere; and
- the link between records and individuals.

Institutions will need a local system to work with their staff to ensure that these assignments are correct. In effect, the institution may need to be able to demonstrate through a physical reference base that each claimed article record is valid. This will not differ from the RA2 form archives currently held by most institutions, so will not reduce the workload. The metrics' system will add the work of assignment, verification and validation.

If self-citation becomes an issue, then the institutional costs of checking will increase significantly, assuming that their researchers are not prepared to sign-off an electronic verification system.

Other data types

There will also be costs associated with the non-bibliometric data, including the funding and training data. Since this is akin to the current RA3 and RA4 sections of the RAE returns there will be little reduction in institutional costs in that regard.

The linkage of these data to individuals will be an additional cost if the funding and training activity needs to be assigned ad personam rather than as a return for a unit as a whole.

Peer involvement

It is understood that the metrics outcomes would not standalone but would continue to be subject to some degree of scrutiny by an expert panel. The range of information that such a panel would receive is unclear but it has to be assumed that it would need more than the metrics data alone if it is to have any serious role. The alternatives are either the rubber-stamping of some opaque final tables falling out of the data analysis or a complex auditing and second-guessing of the same analyses.

The obvious reference material would be statements by the institutions about their recent research performance and strategy at the level of the broad subject groups, i.e. something on the lines of the RA5/RA6.

Overall, therefore, the metrics system will draw on the same data from RA2, RA3 and RA4 as at present (but with data reconciled to individuals listed in RA1) and will then go through a final peer scrutiny drawing on something like the RA5/6. One might reflect that, despite all the superficial change, it should have a reassuringly familiar feel!

Emerging behavioural effects

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[25]. 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 behavioural effects and the metrics only capture part of the research process and its benefits.

It is facile to pretend that all behavioural effects can be anticipated and modelled. 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 then this will potentially affect citation behaviour 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. But it had a wider effect on the publication and citation behaviour 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.

The Netherlands started to use bibliometric research indicators ahead of other countries. There is evidence of emergent behavioural changes amongst its researchers, with an exceptional rise in output and citation share. By contrast, UK share of output has recently fallen slightly but at no detriment to the average research quality recorded in Office of Science and Innovation performance indicators.

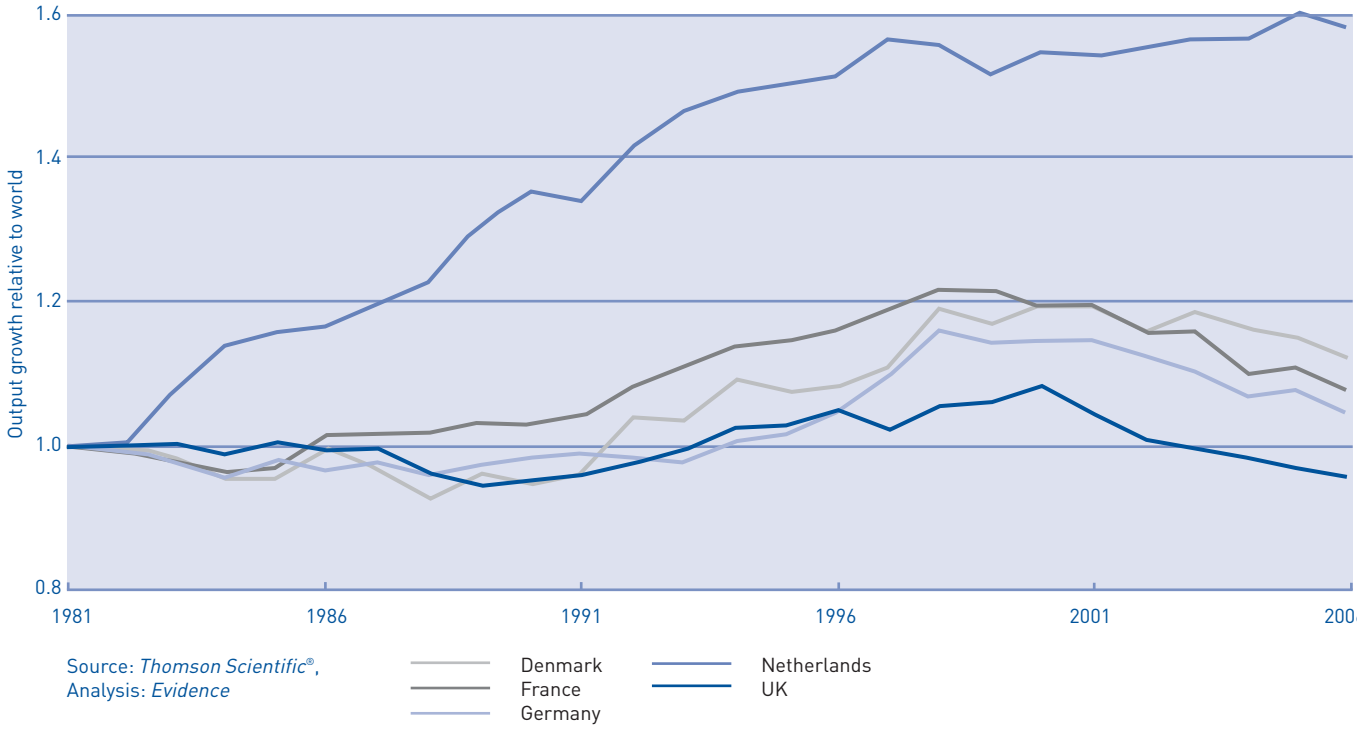
If there are emergent effects then some can undoubtedly be addressed by adding modifications to the metrics, but this risks the development of an increasingly Byzantine and qualified system that loses not only simplicity (hence, ease of operation) but also transparency and so leads to a loss of institutional and researcher confidence.

Possible behavioural changes

Research volume is a poor indicator because quality, not quantity, is the objective and once volume is used as an indicator (as in RAE1986) 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 of itself prove anything. Since citation rates vary between sub-fields there will always be differences in citation accumulation. If volume were an indicator then this would encourage lower-citing fields meaninglessly to emulate high-citing fields.

Figure 11:
Relative growth of output share in European Union nations, 1981-2006.



Source: *Thomson Scientific*®, Analysis: *Evidence*

Journal impact factors are a poor indicator because of the variation in citation rate. If they were used, then 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 was actively penalised by HEFCE metrics then this could lead to a change in citation behaviour, with transitional drops in citation rates, a failure to track links in research programmes and a loss of international prestige. Certainly, Office of Science and Innovation indicators of the UK's relative international standing will drop.

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

Differentiating excellence

How well do any of the indicators allow good to be separated from bad and good from very good?

It is broadly agreed that normalised citation impact is a reasonably good way of differentiating research quality. But it does not work well on very small samples, where individual outlier papers may distort the result (as shown in our impact profile analysis). The example given earlier in this report, for research council data, shows why profiling is likely to be a better discriminator than averages. This requires further evaluation.

The strategy for normalisation of citation counts may be the most important influence on any differentiation or submergence of relative excellence within any research area. The examples given in this report show why the choice of level for normalisation and choice of strategy for data aggregation could have profound effects on the measured relative performance of sub-fields within broad subject groups. It is vital that this is properly evaluated before any action is taken.

Benchmarking

It seems extremely unlikely that research metrics, which will tend to favour some modes of research more than others (e.g. basic over applied), will prove sufficiently comprehensive and acceptable to support quality assurance benchmarking for all institutions.

At present, almost all higher education institutions are content to let their research reputations rest – with caveats and qualifications – on RAE grades. Will they make a similar choice of RAE metrics?

If even a few institutions choose to present their research in terms other than those identified by HEFCE's metrics then this raises a challenge to credibility. Can both presentations be correct? If an institution sincerely believes that it is supporting research that is better evaluated through other measures than the funding council is using then that will presumably raise questions at researcher level; about the 'equity' of the system. It may raise question about the 'fair' distribution of resources between institutions and about the distribution of resources between researchers within institutions if metrics are seen to work more favourably for some research programmes.

Notes

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.

2 The Government has indicated that mathematics and statistics are not included for the first phase of the shift to metrics.

3 Goodhart C A E [1984] *Monetary Theory and Practice*, Macmillan: Basingstoke.

4 See Leydesdorff L and Bensman S [2006] 'Classification and Powerlaws: The Logarithmic Transformation', *Journal of the American Society of Information Scientists and Technologists*, 57, 1470-1486; Adams J, Gurney K and Marshall S [2007] 'Profiling citation impact: a new methodology', *Scientometrics*, 72, 325-344.

5 see Adams J [2007] 'Nobody expects the Spanish Inquisition', *Research Fortnight*, 25 April 2007, 18-19.

6 Evidence/ESRC [2004] *Bibliometric profiles for selected Units of Assessment* http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Bibliometric%20Profiles%20for%20RAE%20Outputs%20in%20the%20Social%20Sciences_tcm6-18357.pdf

7 see also Van Raan A J [2006] 'Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups', *Scientometrics* 67, 491-502 and Editorial [2007] 'Achievement index climbs the ranks', *Nature* 448, 737.

8 Garfield E [1955] 'Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas', *Science*, 122, 108-111.

9 Evidence/OSI [2007] *PSA target metrics for the UK research base* http://www.evidence.co.uk/downloads/OSIPSATargetMetrics070326.pdf

10 See: Adams et. al. [2007] op cit.

11 This is derived from Adams et. al. [2007] op cit.

12 notably in extensive work by Professor S Harnad, University of Southampton

13 Fowler J H and Aksnes D W [2007] 'Does self-citation pay?' *Scientometrics*, 72, 427-437

14 Fowler and Aksnes [2007] op cit.

15 As did Fowler and Aksnes [2007] op cit.

16 See Adams J [2005] 'Early citation counts correlate with accumulated impact', *Scientometrics*, 65, 567-581.

17 Van Raan A [2004] 'Sleeping Beauties in Science', *Scientometrics*, 59, 467-472

18 HEFCE August 2006/32 web only issues paper Selection of staff for inclusion in RAE 2001

19 HEFCE has indicated a priori that it anticipates somewhere in the region of 5-8 groups to cover the sciences, engineering, technology and medicine.

20 Adams J [1998] 'Benchmarking international research', *Nature*, 396, 615-618.

21 Adams [1998] op cit and in a recent EPSRC report

22 Adams J, Gurney K A and Jackson, L [2007] 'Calibrating the zoom: a test of Zitt's hypothesis', *Scientometrics*, 75 (1), in press.

23 Zitt M, Ramana-Rahary S and Bassecoulard E [2005] 'Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation', *Scientometrics*, 63 373-401.

24 Adams J, Bailey T, Jackson L, Scott P, Pendlebury D and Small H. [1998] *Benchmarking of research in England*, Report to HEFCE & CPSE, University of Leeds. ISBN 1 901981 04 5).

25 Goodhart C, [1975] 'Problems of monetary management: the U.K. experience', in: Courakis A (ed), [1981] *Inflation, Depression and Economic Policy in the West* (Totowa).

This product has been manufactured on paper from well managed forests and other controlled sources.
It is manufactured using the FSC Chain of Custody and by a company employing the ISO14001 environmental standard.

About Universities UK

This publication has been produced by Universities UK, which is the representative body for the executive heads of UK universities and is recognised as the umbrella group for the university sector. It works to advance the interests of universities and to spread good practice throughout the higher education sector.

Universities UK
Woburn House
20 Tavistock Square
London
WC1H 9HQ

telephone
+44 (0)20 7419 4111

fax
+44 (0)20 7388 8649

email
info@UniversitiesUK.ac.uk

web
www.UniversitiesUK.ac.uk

ISBN 978 1 84036 165 4
October 2007

Universities UK

evidence