

Annex C: Field Categories and normalisation

A report by Evidence Ltd

Managing the data

1. Several pieces of work are required to make effective use of the bibliometric database. Here, we review some prior work which has shaped what we have done and some new pieces of work developed for the REF pilot project.

- How should the data be categorised? On the one hand, we recognise that fields of science correspond to more or less closely integrated domains with similar cultures in terms of objectives, methodology, publication patterns and citation behaviour. We recognise these boundaries but we also want to group the information in a way that is useful for evaluation, to allow suitable comparisons between sub-fields and links across common interdisciplinary boundaries.
- Given the differences in publication and citation behaviour, we seek to identify the level of categorisation at which actual citation counts should be indexed (normalised) so that we can make sensible comparisons across years and subject boundaries. We need those comparisons because what is published by, for example, physicists is not just physics research but also chemistry, biology, mathematics and engineering.
- Actual citation counts have conventionally been normalised against a world average but citation data are highly skewed. The world average is far from the median (the middle of the citation distribution). We therefore consider whether there may be any better alternative to indexing against averages.

Categorisation

2. It makes little sense, for management or evaluation, to treat all research disciplines as one. It is better to establish an informed categorical system to sub-divide and group research activity. But such a system will be arbitrary. There is no unique, preferred or optimal way of categorising research except insofar as it serves a particular and limited purpose. There are many conventional categorical systems, used by national agencies (e.g. UOAs in the UK RAE) and international agencies (e.g. Fields of Science in the OECD system). There are a many ways to customise these for a particular purpose (e.g. by grouping OECD minor Fields of Science to national programmes).

3. This section is about grouping data for data management, not for reporting. The decision about what categorical system should replace the UOAs of the former RAE is one for HEFCE.

4. The REF introduces a quantitative element. Once data, analysis and indicators are invoked then decisions must be made about how the data will be grouped before they are processed and analysed. The grouping for analysis is quite independent of the grouping for reports.

- For data analysis and quantitative comparisons, we need a category system that brings cognate activity together, so the data from one unit can be reasonably compared with like data from another unit. This might be a finely-grained level

(for example, 'entomology') or it could be a broad comparison (for example, 'biological sciences').

- The category for reporting could be broader so, for example, entomology analysis nests alongside biochemistry analysis in biological sciences, or it could be finer.

5. Categorisation for analysis must be appropriate for purpose but it should also make sense to the parties being evaluated, corresponding to their perception of the way their research world is structured. In this section we discuss some of the factors that should be taken into account in arriving at a decision about how finely to categorise the data. In the next section, on normalisation, we discuss how the citation data should then be treated.

Functional mis-match

6. At a gross level, it is a challenge to develop a categorical system for knowledge that enables a mapping correspondence to be made between narrowly defined areas of science (categorising articles and doctoral programs), narrowly defined areas of technology (categorising by patent classes) and narrowly defined areas of industry (categorising by governmental or financial market definitions of industry).

7. We try to map fields of research onto economic sectors in order, for example, to link between OECD's data about funding and national data about research training. The OECD system has a major category that is 'Natural Science' but this subsumes biology, physics and chemistry while its major category of 'Agriculture' splits out an area that is much less central as a research focus. However, among OECD minor categories there are multiple versions of 'Biotechnology' aimed at functional utilisation in agriculture, health, and so on.

8. Universities daily find a challenge in mediating the relationship between the organisation that supports the teaching function and is grouped around degree schemes and conventional departmental structures, and the research function that is based around multiple, overlapping and increasingly trans-disciplinary research groups. Some institutions actually split staff contracts between a teaching affiliation and a research affiliation.

What is Physics?

9. If we take a familiar and conventional discipline, such as Physics, then what we find differs between one institution and another. One department is strongly theoretical; another has extensive laboratories and major facilities; while a third is primarily focused on space science. But they are all recognisably Physics and would submit to the same panel in an assessment exercise.

10. We can take three further, different perspectives on 'Physics'. A Research Council committee oversees a Physics programme, but the grants awarded go to a wider range of units than just Physics departments. The Physics departments themselves receive funding from Chemistry, Engineering and medical sources. The physicists publish in a wide range of journals, while the journals associated with physics have articles authored by researchers from a diversity of disciplines. Indeed, we could ask the question, if we looked at any one university: 'what is physics research?' Is it the research funded for

physics objectives, the research done by the physicists or the research published in physics journals? How different would our three answers be?

11. The physics problem is repeated in all conventional disciplinary areas. Among researchers there are many different views on what might be an optimal solution.

Identity

12. If we try to define a subject, do we also seek to define its distinctiveness from other subjects? When we seek to categorise research activity as 'Psychology' are we also seeking to label it as 'not Education' and that this activity is 'Biochemistry' not 'Ecology'?

13. For a qualitative evaluation, if we want to benchmark the citation counts for a physics paper then we probably want to make a comparison just with other physics papers. Our indexing should be exclusive. But we might look at a group of physicists and note that their research portfolio includes work relevant to chemistry and to mathematics. We can take apart their activity and benchmark each component but we need at some point to reintegrate the different fields of research and make a judgement as to whether the whole thing works well or not.

14. For normal purposes we need to recognise that the overlap and boundaries of a subject are not absolute and that we need to make a sensible judgement about what works.

15. Researchers themselves have a fairly fine-grained view of the world. Responses to HEFCE's recent consultations on the possible routes to development of the REF have tended to point towards more categories and a greater level of disaggregation than towards a system with few, generic categories. Clinical researchers are not 'medics' but 'cardiac' specialists or 'rheumatologists'.

16. The problem with finer-grained categories is that they require greater definition to distinguish them from other equally fine-grained categories. That does not work well with an increasingly inter-disciplinary research base where the underlying philosophy of any traditional categorical structure is contested. The more boundaries created, the more activity is found at those boundaries and must be managed. Alternatively, we must avoid any residual multidisciplinary categories and disaggregate all material, including multidisciplinary journals, to our specific categories.

17. The individual researcher may indeed argue that their work can only be properly evaluated by a relatively small group of peers who are part of their field, a field they define. But an evaluation system has to work both within and across fields. This work is 'good' for its field, but is the field 'good' in comparison with other related fields?

Practical approaches

18. The key consideration in the present context is how we can best aggregate data about journal articles for the purpose of research evaluation in the REF. We need to consider the categories at which we should analyse the data (discussed in the following section on 'normalisation' of citation counts) and report the data.

19. In 1997, Evidence carried out work for HEFCE to establish the feasibility of developing international benchmarks for research evaluation as reflected in the RAE. To do this, a map was created between two categorical systems: the Units of Assessment

(UOAs) in the RAE and the journal categories in Web of Science. The map was fitted by using the information about articles submitted by staff within each UOA.

20. Within each UOA, all outputs that were labelled as journal articles (or variant descriptions equivalent to this) were selected, cleaned and matched where possible to items on the Web of Science. This cleaned and validated list provided information about the frequency with which any journal was represented. The journal lists were then ranked by descending frequency and matched to the lists for the journal categories on Web of Science. Pair-wise matching was made with all possible categories.

21. The best matching category inevitably also included journals not represented in the UOA but used by researchers in that field elsewhere in the world. This therefore provided the 'best available' match. For many UOAs this 'best available' category covered many RAE journals and did not link to an undue proportion of non-RAE journals.

22. Additional Web of Science categories were associated with each UOA, increasing the coverage of RAE-submitted journals but inevitably adding increasing proportions of 'non-RAE' journals which nonetheless provided the necessary global contextualisation. At the same time, note is taken of the extent to which each additional category also overlaps with other UOAs. A decision point is reached when the next category would add relatively few RAE journals, would add many non-RAE journals or would create an undue overlap with another UOA.

23. The outcome of this exercise has since been widely used to provide a snapshot of bibliometrics by UOA. It is an address-based mapping, not an author-based mapping. The outcome for each institution is of 'research in subject journals' not 'research by subject researchers'. However, it provides what many research managers who have used Evidence analyses agree is an informative and robust system.

24. Problems arise because not all UOAs readily attract a well-defined set of journals or journal categories. For some UOAs, there is no journal set that does not map better to another UOA. For example, UOA25 General engineering is used by institutions for the assignment of collections of work not assigned to specific engineering categories such as civil, mechanical and electrical engineering. It therefore includes an exceptionally diverse range of journals but includes none of these at a high frequency. Other categories that are problematic are those that have a strong methodological element, such as statistics. Generally, categories which have a professional association are also problematic to map.

Australia and the ERA

25. The Australian Research Council (ARC) is currently implementing the Excellence in Research for Australia (ERA) evaluation system. ERA makes use of the Australia and New Zealand Standard Research Classification (ANZSRC) which was developed by the Australian Bureau of Statistics, Statistics New Zealand and the New Zealand Ministry of Research, Science and Technology. ANZSRC represents three related classifications created for the measurement and analysis of research and development in Australia and New Zealand (http://www.arc.gov.au/media/releases/media_31march08.htm).

26. The ANZSRC is a hierarchical system with progressively finer two-, four- and six-digit codes. The 22 divisions at the two-digit level are intended to align with the OECD Fields of Science classification which, as noted above, is primarily intended for economic

rather than research purposes. Thus Division 01 covers just Mathematics while Division 11 covers all of Medical and Health Sciences. The hierarchy is intended to allow ready drill down from the two-digit to the 157 Groups at the four-digit level, and equally direct aggregation upwards into clearly defined categories. Division 06, for example, is Biological sciences and within this are nested other Groups including Group 0601 Biochemistry and cell biology, and 0602 Ecology. However, 0602 overlaps with Division 05 Environmental sciences while other parts of Division 06 overlap with Division 07 Agricultural and veterinary sciences.

27. For the purposes of ERA it was necessary to establish journal mapping for each of the two- and four-digit ANZSRC codes. To do this, ARC set up a number of expert panels through the four Australian learned academies which assigned all the journals frequently used in Australia to a four-digit Group. This domestic list was then augmented with a further round of assignments of the non-Australian journals so as to extend the contextualisation to a global set of 19,500 journals. The assignment was primarily exclusive (a journal was intended to be assigned to a single four-digit Group) but in fact such exclusivity proved infeasible even at this level.

28. A further problem is with journals and research which fits uneasily with a single four-digit code. Within each Division there was a residual category for 'other' research; for example, Division 01 Mathematics includes Group 0199 Other mathematical sciences, and so on. This leaves a significant body of research that can be classified at a two-digit level but not more finely. In fact, even this is unclear and there are some journals which remain lodged at the two-digit Division level and are not even assigned to 'other'.

29. In practice, more problems emerged once the system was implemented. Nanotechnology was recognised as a significant area for economic development and hence of critical interest within the research base. In running the first pilot cluster for ERA, which focused on physical sciences, it was found that no category adequately captured nanotechnology in any coherent fashion. It largely spread across several Groups within Division 02 Physical sciences but, because no single 02xx Group proved ideal, linked also to 03 Chemical sciences and 09 Engineering. Had the aggregation been made at the Division level, then most nanotechnology research could have been reasonably assigned and assessed in one tranche.

Clustering

30. A further route to aggregation is by clustering items with similar attributes. This is essentially what Evidence did to create the map between UOAs and Web of Science categories. However, the same logic can be applied to the UOAs themselves or to journals.

31. We can take the journal frequency lists for each UOA and then calculate a similarity coefficient between one UOA and all other UOAs based on the similarity of journal usage. We can coalesce the most similar pair, recalculate and then progressively add most-similar pairs together until we have a single branched 'tree' linking all our UOAs. For the RAE data, this shows that 'materials science' is more closely associated, in terms of journal usage, with physics and chemistry than with any engineering

discipline. The hierarchical assembly creates clusters at multiple levels, akin to the panels and super-panels of RAE2008.

32. The clustering of UOAs makes the assumption that the UOAs themselves have some prior validity. For the REF there is no intention that such a priority would apply. Instead, the possibility of a novel discipline structure can be invoked. How should clustering be developed for this?

33. As a single article has a set of citation relationships to other articles, which it cites or by which it is cited, so journals can be related to other journals. The set of articles in the journal will cite not only other articles in the same journal but articles in other journals. Some of these citing relationships will be stronger (more frequent) than others. We can create a matrix of pair-wise citation links and look for clusters of relatively high association. Such a cluster would identify a cognate group of journals: journals that draw on one another's knowledge content more than they do on the rest of the literature.

34. The advantage of such a mapping is that it draws on the underlying intellectual relatedness of the literature rather than on value judgements made by an expert group. Decisions still have to be made about where a cluster might be deemed to be sufficiently distinct from other clusters but the process is entirely transparent.

35. The disadvantage of such clustering is that it is explicitly exclusive. Once a journal is allocated to a cluster it cannot then be associated with a second cluster at the same level, only to a parent cluster at a more aggregated level. This therefore creates a problem for the increasingly multidisciplinary part of the literature, which can only be avoided by agreeing that journals showing split affiliations are treated in a different way. The objectivity of the process is then compromised.

Conclusion

36. There is no technically ideal approach to categorising research by discipline.

37. A finer-grained approach is, on the whole, preferred by researchers because it captures the scope of the research community with which they identify. However, very specific categories may make accurate assignment and field comparison problematic. A coarser-grained approach more appropriately encompasses multidisciplinary and is more likely to capture appropriate sample sizes.

38. Setting a boundary to categories needs to take into account both identification and overlap. The community identifies disciplines in an organic fashion that tends to collapse into traditional structures because current research fronts are dynamic and mobile. However, such traditional categories allow a degree of overlap. More objective identification of cognate areas tends to create relatively exclusive categories which would also need regular review.

39. A clustering approach, using citation linkages between journals, would be feasible but onerous if wholly new categories were sought. An analysis of commonality in journal usage between RAE UOAs would be relatively straightforward if some aggregation of UOAs into broader categories were sought.

40. The next section discusses the process used to index actual citation counts, which vary by year and subject, to a common benchmark so that they can be managed to produce performance indicators within each subject category.

Normalisation

41. In the previous section we discussed the factors that might influence decisions about the finer or broader subject categories used to group bibliometric data. In this section we discuss what should be done with citation counts once outputs are grouped by category, so that useful comparisons of citation impact can be made across years and between subjects.

42. If a paper has been cited 20 times, is that a good citation count? We cannot know without referring the actual count to an expected value. The answer depends on when the item was published, the subject to which it relates and the type of publication. Bringing the disparate citation counts – the raw impact – for a portfolio of material to a common basis is variously called ‘normalisation’ or ‘rebasng the count’. The reference benchmark most often used for normalisation is the relevant world average.

43. The three key attributes of any piece of research activity data are the time, subject and location associated with the activity. Each of these attributes is a variable that affects the bibliometric data (publication and citation counts) linked to outputs. Because of these influences, a direct measure of citations per paper may be misleading as an index of relative citation ‘impact’.

44. In what way do these factors affect bibliometric data?

a. **Document type** – the nature of the document affects its utility. Review papers, for example, are frequently cited not because of their originality but because they collate a background literature. One general citation to a review substitutes for a plethora of specific references. For this reason, the citation rates for reviews might reasonably be treated differently from those for ‘standard’ articles, insofar as review papers can be separately identified. Letters to the editor, (as distinct from Letters in Nature) are, by contrast, rarely cited and are usually excluded from assessment.

b. **Time** – citations accumulate over time. A paper published in a recent year has had less time to be cited than a paper published at an earlier date. We therefore generally need to take the year of publication into account, but for very recent papers there will also be some effect within-year between a paper published in December and one published in January, 11 months earlier.

c. **Subject** – disciplines have their own publication and citation culture. At a general level, bio-medical sciences tend to produce shorter and more frequent papers where much of the standard background and methodology is reduced to a shorthand summary by referring to other papers. As a consequence, there are many more papers produced in these fields, and each paper carries more citations, so there are relatively high citation rates compared to e.g. physical sciences. This is a ‘natural’ outcome of the field rather than a reflection of differences in impact.

d. **Location** – there is some evidence of differences in citation culture between countries. When comparisons are made within a country this is not a problem, nor is it a problem with large samples and multi-national analyses. Some consideration may be appropriate, however, for smaller samples and comparisons between just two countries.

Articles and reviews

45. Reviews are widely considered to be different to ‘normal’ journal articles. Reviews are, by their nature, about original knowledge developed in other publications. It is not true that reviews cannot in themselves contribute original intellectual property. The reviewer may create an effective new synthesis by bringing together many disparate elements of work in their field. They may thereby identify new problems as well as signposting potential pathways to solutions.

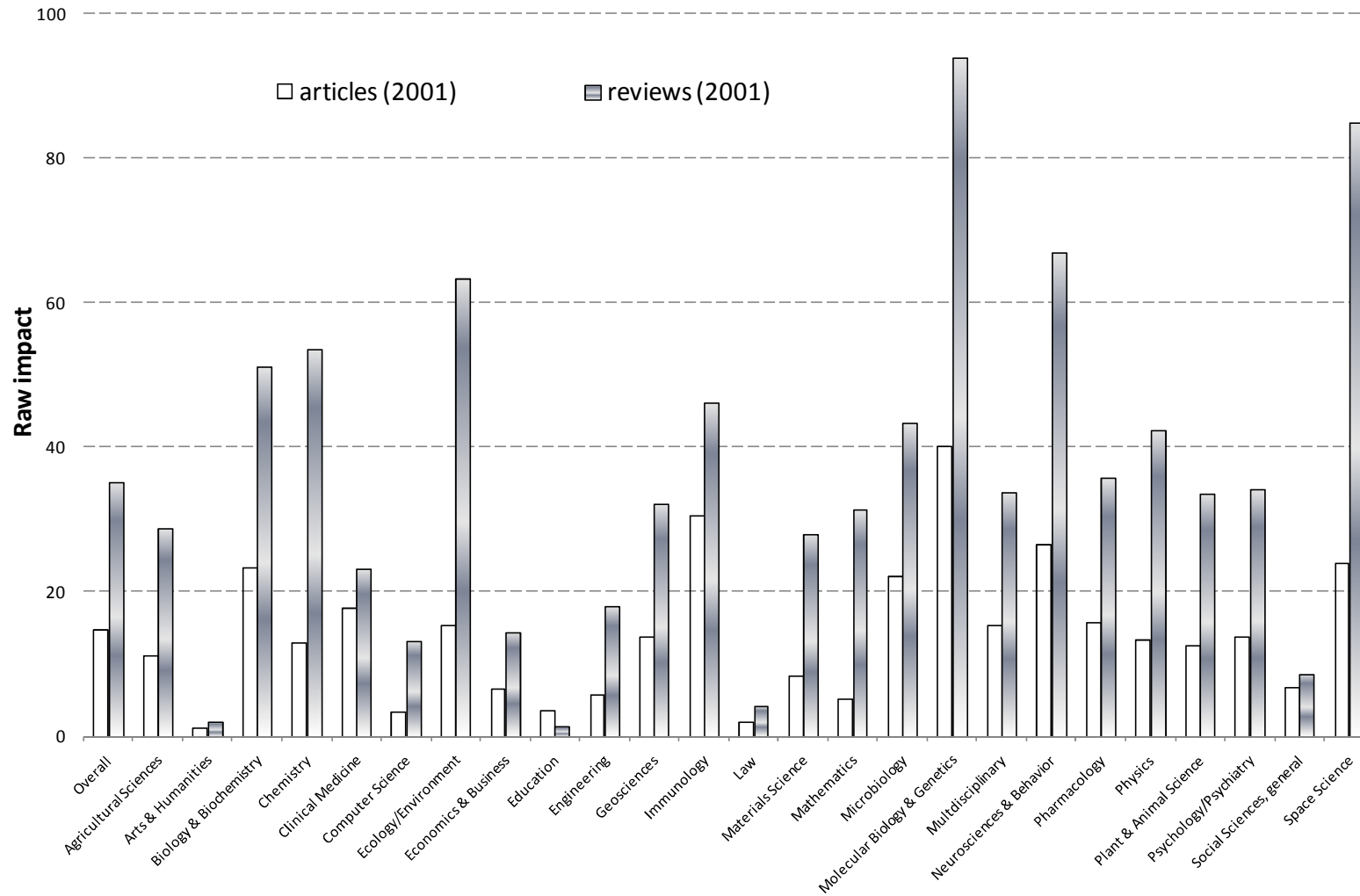
46. Reviews do also, however, provide a short-hand reference to multiple other pieces of work, because they summarise a body of literature, and they could thereby attract high citation rates. It is therefore argued that they should be treated separately from articles. Observation confirms that reviews do indeed accumulate more citations on average than articles (Table C1).

Table C1. Average citation counts to articles and reviews published between 2001 and 2007 and recorded in Thomson Reuters’ UK National Citation Report

Year	2001	2002	2003	2004	2005	2006	2007
Articles	14.75	13.60	10.78	8.44	5.55	2.55	0.48
Reviews	35.17	31.25	26.08	19.25	12.29	5.46	0.96

47. This can be broken down by field. In Figure C1, which illustrates data for publications in 2001, it can be seen that whereas by 2007 there was little difference in citation accumulation (raw impact) for articles and reviews in clinical medicine, there was a two-fold difference in molecular biology and a more than three-fold difference in space science.

Figure C1. Comparative citation counts for UK authored articles and reviews, analysed by journal category for items published in 2001



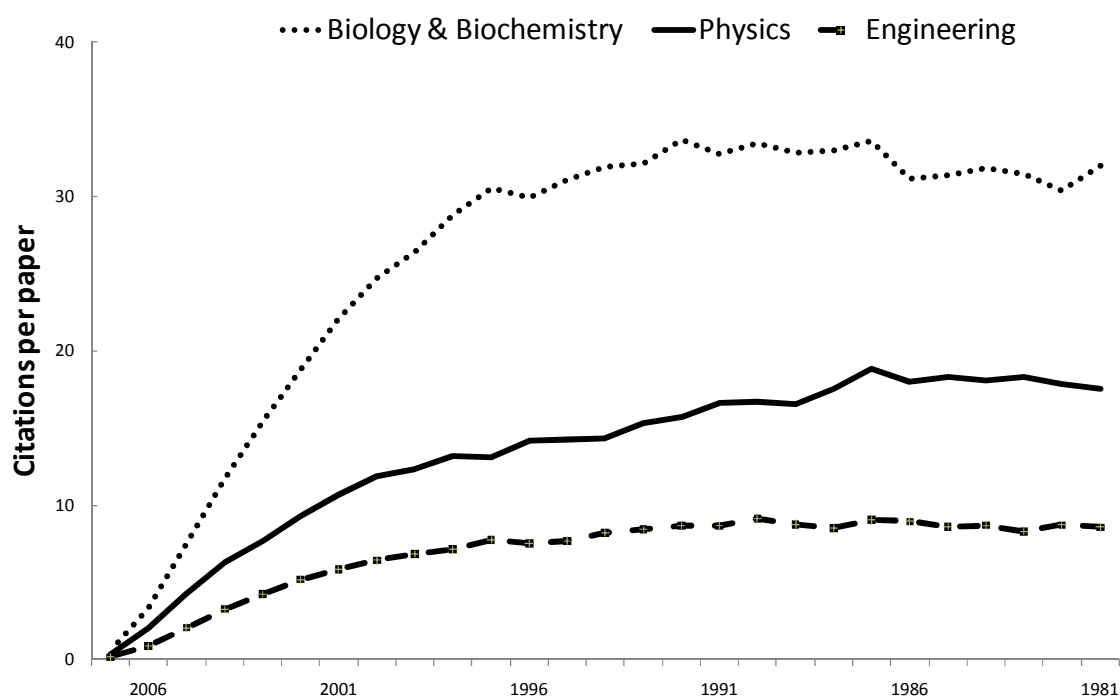
48. This is not the only complicating factor. Reviews are also a more common type of publication in some fields than others. Thus the balance of articles and reviews varies by field, as does the average citation ratio. And, of course, the balance of reviews among outputs for a specific institution is even more variable. On balance, therefore, there is a sound justification for treating articles and reviews separately. The actual citation counts for an item should be benchmarked against items of a similar kind.

49. Finally, it is necessary to record that not all reviews are actually labelled formally as such. While some serials are explicitly titled 'annual review of ...' and so forth, some journals may publish occasional review articles alongside normal journal articles. Often the editor will flag these as reviews. However, this is not universally the case and some items which are de facto reviews may not be designated as anything other than standard articles within a database.

Time

50. Citations accumulate over time, but at different rates in different subjects. Recently published papers will tend to have fewer citations than those published in the past. Papers in biochemistry will tend to have more citations than papers published in the same year in engineering (figure C2). Papers in physics will go on accumulating citations for over ten years whereas papers in biochemistry plateau in about eight years (figure C2).

Figure C2. Citation accumulation in three fields of science. Average citations per paper is shown for UK-authored papers published in the years 1981 through to 2007



51. These are generalisations but it is clear that the year of publication is something that must be taken into account in establishing the relative citation impact of a paper. Is

16 citations a good tally? It is for a paper in engineering, and it is not bad for physics, but it is good for a paper in biochemistry only if it was published in the last four years.

52. The year of publication might appear to be unproblematic as a reference. Unfortunately, we need to recognise that the available databases include two different date fields: publication year and database year.

53. The publication year is that assigned by the publisher to the journal volume. The problem with publication date is that it only roughly follows the actual appearance of an item. Two items with similar cover dates could actually be published several months apart.

54. For the Journal of Animal Ecology, Wiley InterScience identifies the six bi-monthly issues in 2007 as Volume 76, and the issues in 2008 are Volume 77. Volume 77, part 6, is dated November 2008. A sister-journal, the Journal of Applied Ecology, has Volume 45 in 2008 and is also bimonthly. In fact, Volume 45, part 6, dated December 2008, was already available on-line in November 2008.

55. Note that the rate at which the successive issues of each volume appear varies between journals and not all journals are available as promptly as these examples. Some journals experience a lag between actual publication and the nominal, cover date. The end-year issue, nominally of November or December, may not be available until early the following year. It is worth noting that timeliness of publication is a factor taken into account by commercial database compilers in deciding whether a journal should be included in their products.

56. The database year is that set by the compilers of the publication and citation database, which are Thomson Reuters for Web of Science and Elsevier for Scopus. Like the journal cover date, database year is only partly linked to the underlying calendar year.

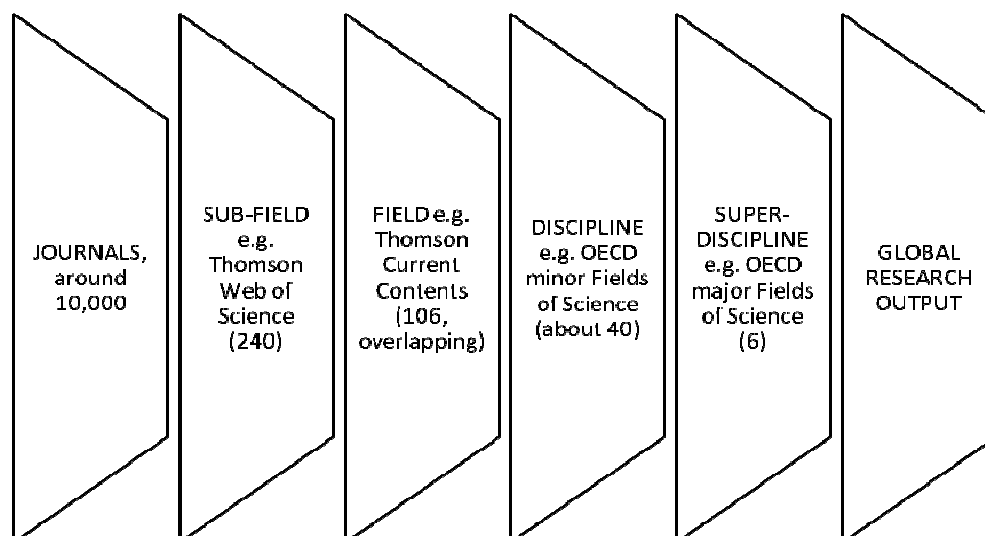
57. Typically, the cut-off date for an annual database compilation will be some time after 1 January, in order to fit in with other work schedules. This can vary: it might be in the first week after New Year or not until a week or so later. Each additional week would add an additional 2% volume to the closing database and reduce the volume of the following year.

58. It is infeasible to carry out an analysis that adheres strictly to the calendar year. The choice is between publisher year and database year. In practice, almost all previous analysts have relied on database year since the alternative is to create a reference benchmark from scratch using raw data.

Subject

59. We can imagine an aggregation spectrum of subject categories from as fine a grain as the journal volume in which an article is published to as coarse a grain as the total global publication output for a given period (Figure C3).

Figure C3. A scale spectrum for possible levels of categorisation of bibliometric data



60. If we look at a single article in the context of its journal then we may observe that it has more or fewer citations than we would expect if we took an average across the whole of the volume in which it is published. The ratio between observed/expected (O/E) is a useful indicator: is this an article cited more frequently than is typical for that journal volume?

61. If we move to any higher level of aggregation then we introduce some arbitrariness into our categorisation. The definition of any 'sub-field' may be highly individual: a piece of work might be seen as ecological, behavioural or evolutionary biology according to the career of the author as much as the content. For systematic purpose, however, we have to work at the level of journals so that, with the exception of a defined set of multidisciplinary journals of which Nature and Science are obvious examples, the whole of a journal – all its articles – are allocated to a single category. We have to decide how that category is defined and how comprehensive it should be.

62. The journal categories used by commercial databases are broadly influenced by the citation links between journals. The effect of this is that material that cross-refers at a high frequency tends to end up in the same category. This represents a relatively natural grouping of published material. Small, fine-grained, highly connected categories can be nested within or progressively aggregated to form larger and coarser categories.

63. Once we have decided what journals are included in our category then we can look at the total number of articles in those journals for a given period, collate the total number of citations to those articles and thus create a grand average citations/article. This is now a reference benchmark against which any individual article can be compared.

64. The ratio of the observed number of citations for an article to the category average produces our normalised, or rebased, impact. Our level of categorisation for normalisation can differ from that for analysis and reporting, which allows us to take important variations in citation rates into account.

Location

65. Is 'good' research 'good' in world terms or is it good in terms of the subject within the UK? For the purposes of the present exercise, the appropriate benchmark is global activity. 'Good' needs to be defined in terms of relative international excellence. We are not interested in whether an article is frequently cited only in a UK context. The journal categories used by commercial database compilers include an international selection of journals and thus contain all the articles published in those journals irrespective of the location of the authors. They are, therefore, an appropriate international reference set.

Levels of aggregation

66. What effect does the level of aggregation have on outcomes, such as normalised impact? What is the 'correct' categorisation to use?

67. Categorisation has to be fit for 'purpose'. The categorical structure used for international comparisons across the breadth of the research base would usually be much coarser than that for an institutional management study. Surveys of researchers suggest that they tend to see themselves as part of a relatively small, well-defined sub-field – towards the left-hand end of this spread. This is the comfort zone in which they make 'peer' references to their own and others' activity.

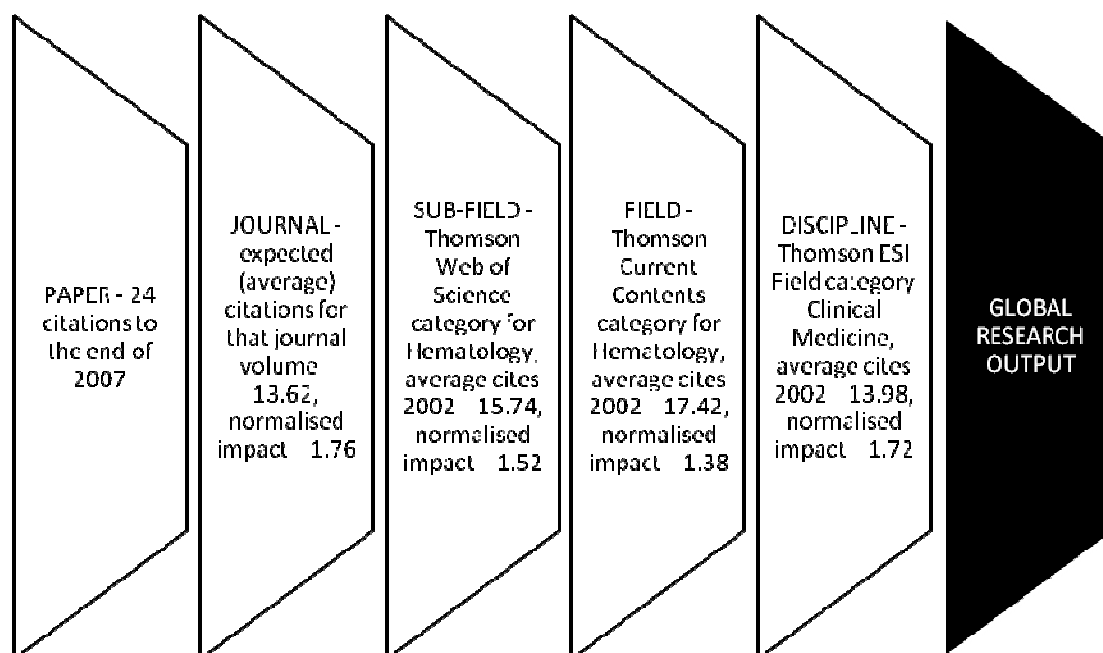
68. This spread has also been seen in 'vertical' terms by Michel Zitt (University of Nantes) who refers to the optical 'zoom' from a close and detailed focus on a piece of research, pulling back to gradually reveal the same research in a progressively broader and more diverse context.

69. However, some journals are perceived by the research community to be more prestigious than others. There is competition to get published in a journal which is not only an ideal outlet for a given piece of work (ideal because of its editorial focus, content balance and readership) but which is also relatively highly rated by the community. So an article that has a good O/E for its journal may nonetheless be seen to be of less than average quality in its sub-field because the journal is perceived to be of minor significance and not rated highly by peers.

70. Moving up or across the scale, some sub-fields are of greater or lesser significance within their field, and some fields of greater or lesser significance within their discipline. A given piece of work may be of notable recent impact in its field, but that field might currently be seen as 'mature' within its broader discipline (offering less scope for innovation and originality). Cutting-edge research has moved on, into other fields and sub-fields and this work of local impact has, in fact, little significance for innovation and development of the broader subject.

71. We can take a real example that shows this changing contextualisation. The paper of interest is 'Sex ratios and the risks of haematological malignancies' published in 2002 in the British Journal of Haematology, Vol 118, pp. 1071-1077. This 2002 paper has 24 citations – an above average count for its journal in that year. Figure C4 reveals that it may have a normalised impact between 1.38 (at field level) and 1.72 (in its broader discipline).

Figure C4. The effect of varying category scope on the normalised citation impact of a specific paper



72. Common sense suggests that ‘the right level for normalisation’ should not be too fine a level, which becomes unduly self-referential. Nor should it be too coarse a level, which loses any sense of disciplinary and cultural context. But in-between there are important nuances about the relative significance of fields and sub-fields.

73. This is obviously not simply of academic interest. It will affect the weight, the relative value, which that paper gives to any sample in which it is included to create an index of research performance. Nor is the answer solely a technical one, if it is technical at all. It is also, perhaps largely, political. Decisions about the level of normalisation will be value judgements.

74. For example, consider fields A and B where A has a lower average citation rate than B and both are set within some parent discipline X (Figure C5).

Figure C5. Effects on indexed impact for two papers in different fields within the same discipline

Discipline X – average of 15 citations per paper for year y	
Field A, within X Average of 12 citations per paper for year y	Field B, within X Average of 20 citations per paper for year y
An article with 20 citations and linked to A has a normalised impact of $20/12 = 1.67$ at field level	An article with 20 citations and linked to B has a normalised impact of $20/20 = 1.0$ at field level
An article with 20 citations from A or B has a normalised impact of $20/15 = 1.33$ at discipline level	

75. This is likely to influence researchers' views on 'correct' solutions. The papers in field B have a lower normalised impact if that normalisation is done at field level, because the average field citation rate is higher. If the normalisation is at discipline level then, because their field citation rate is higher than the discipline average, their normalised impact improves.

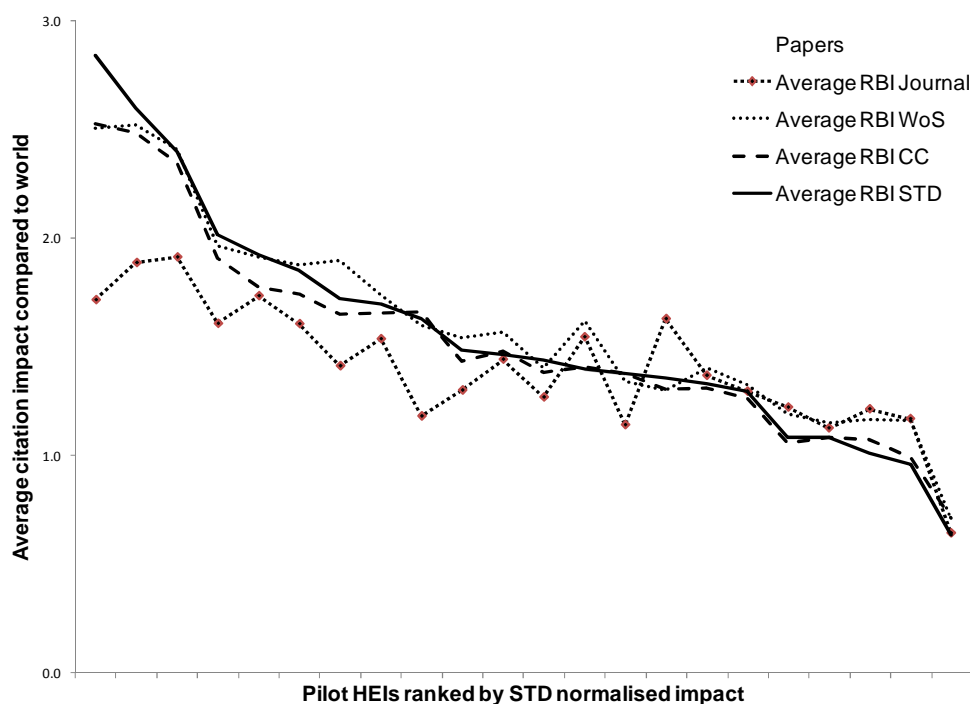
Practical outcomes

76. We analysed the effect on the apparent relative research performance of institutions by normalising the citation counts for papers provided by the 22 REF pilot HEIs at three different levels. We used: the 240 fine-grained categories of the Web of Science (WoS); the 106 categories of Thomson Reuters Current Contents (CC); and the 24 categories of Thomson Reuters Essential Science Indicators (ESI). Additionally, for reference, we also normalised citation counts against the expected citation count for the journal volume in which each paper appeared (Journal).

77. For each paper submitted by the pilot HEIs and matched to Thomson Reuters, we took the actual citation count to end-2007 and then calculated the four relevant benchmarks created for each paper by analysing the complete journal and article set globally for the three different levels and all articles in the relevant journal volume.

78. Note again that the reporting level was the same for all these four different normalisation levels. Irrespective of the level of normalisation, we can aggregate the papers by UOA, or by any other category. In practice, we departed from the UOAs and instead aggregated data at the ESI level. The four values for each pilot HEI, ranked by the average impact at the most aggregated ESI normalisation level, are shown in Figure C6.

Figure C6 The outcome at HEI pilot level, measured as citation impact, of applying different levels of normalisation to publications in biology/biochemistry



79. It is apparent that the three categorical levels of normalisation are closely correlated. The outcomes normalised at journal level depart from these three. This pattern was seen in all other analyses as well.

80. Normalisation at the journal level tends to pull down the average impact of the high-performing units because they publish in relatively highly cited journals so their high average citation counts are indexed against other high impact papers. Conversely this lifts the lower-performing units which are frequently publishing in less well cited journals. This would obscure real differences in relative citation rates and therefore this is an inappropriate level of normalisation to use for this evaluation.

81. Since there was no great variation between the three other categories, there is no clearly 'better' option among these. However, in several analyses it appeared that normalisation at the level of the finer-grained WoS categories tended to result in higher average impact. This finer level of normalisation also corresponds to the level most frequently identified by respondents to HEFCE consultation as the level at which they identify their 'field'.

82. The only caveat against using fine categories is that the critical comparison is made within these small fields: is this good research in this category? We are not comparing the relative performance of these different fields: is the research in this category good? While we do not seek to make comparisons between, for example, biology and engineering there is an argument that the evaluation might reasonably make comparisons between fields within biology. In that case, there is an argument for using somewhat broader categories than the 240 WoS fields, such as the 106 Current Contents fields.

Conclusion

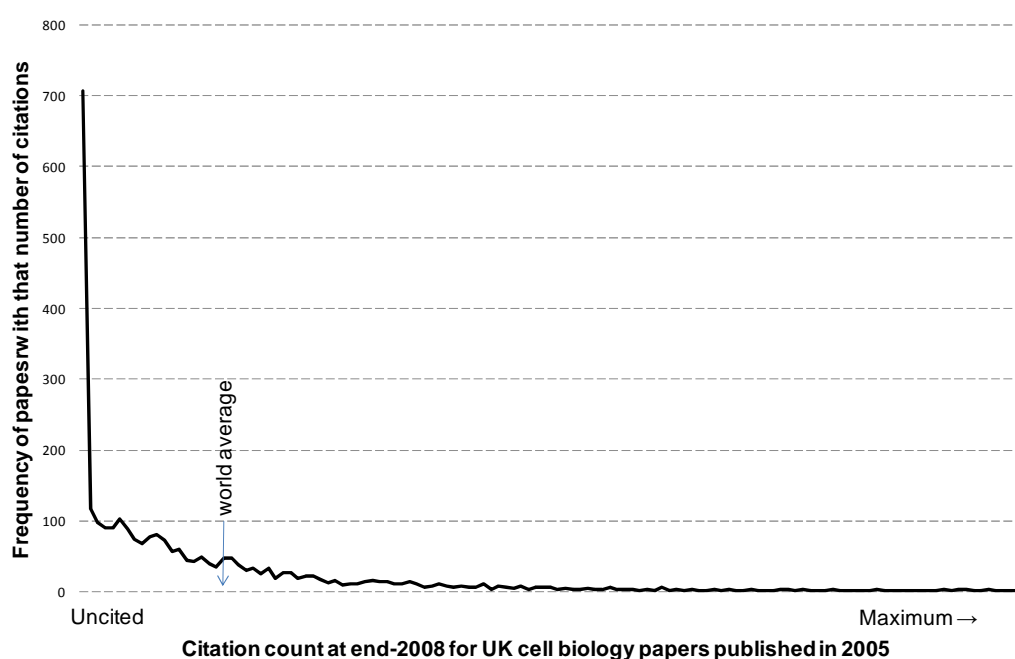
83. For the avoidance of doubt, as we noted earlier, the levels of aggregation used to normalise the data and then to report results are completely independent. They can be the same, but do not have to be.

84. We have reviewed a range of issues which affect normalisation. In practice, a relatively fine-grained approach appears to be workable, but this should be at a level well above the 'journal' comparison of actual and expected citation rates which tends to obscure differences between high and low performing units.

Median and average

85. Research activity distributions are almost invariably skewed with a high frequency of low value events and relatively few incidences of exceptionally high value events. This is true of income per researcher or per unit, numbers of postgraduate trainees, publications per person and citations per publication.

Figure C7. Citation counts to end-2008 for UK cell biology papers published in 2005. Citation and impact data are skewed, with many papers either uncited or cited only infrequently while a few papers receive exceptionally high citation counts



86. When a distribution is skewed, the average value for the data (citations per paper) is no longer a good guide to the centre of the distribution. The average is in fact 'dragged upwards' by the small number of exceptionally high values. Most papers will have citation counts that are much less than the average: most UK cell biology papers have fewer citations than the world average and, indeed, many are uncited three years after publication (Figure C7). However, the UK average citation impact is actually well above world average (it has in fact been close to 1.2 times world average since 1981).

87. The centre of the distribution is marked by the median or central value. This is an informative indicator, which has been made available to the REF Expert Advisory Groups. It gives guidance as to where the overall distribution lies in terms of world average as well as the UK average.

88. Conventionally, citation counts are converted into citation impact measures by looking at the ratio between the actual citation count for a paper and the relevant world average for the year and field of publication (see paragraphs 41 to 84). The average departs far from the centre of the distribution. So, we should consider whether there may be some benefit in departing from convention and calculating the citation impact as a ratio of [actual]/[world median] instead of [actual]/[world average]?

89. We sought to evaluate the benefit of using the median to calculate impact. It immediately became apparent that in at least some fields the outcome would be of no value.

90. The median is the central value among a distribution of citation counts. It is by definition a whole number of citations. The problem is that if we look at recent years then the median value used for the world reference baseline is frequently '0' citations per paper. Even where it is '1' we still end up with superficially distorted outcomes because the few papers that actually have a more substantial immediate tally appear as outstanding. The difference between papers published early and late in the year appears even more disproportionate than when using the average.

91. Irrespective of the more marginal considerations, the relative frequency of occurrence of world medians that are 'zero citations' removes any possibility of using this as a normalisation factor except for older sets of papers. It is almost never the case for any field that the median could be used for the most recent year. The REF may well not use the most recent year, but base its evaluation on papers that are at least 12 months old. However, even at two years there are some fields where the median is still zero and many where it is still only one.

92. Our conclusion is that, while the use of the average means that the actual citation counts are indexed against a value skewed towards the upper end of the citation distribution, the median is not a valid alternative index because in recent years it would be an unusable value.

93. There is another option, which is to avoid using thresholds set against the world average and to look at the centiles (and deciles and quartiles) of each distribution separately. This would mean that items at the high end of a clumped distribution, close to world average, would still be picked out as lying at a high centile for their discipline. We discuss this in a later section.