

Annex H: Availability of citation data by subject

A report by Evidence Ltd

What subjects and outputs can be indexed?

1. It may be inappropriate to apply an evaluation methodology, however sound, to all subjects. The differences in publication behaviour and culture may mean that data in some disciplines are simply not sufficiently informative to produce indicators that would be useful to an evaluation committee. Here we look at three issues related to subjects and outputs.

- First, we looked at the concept of 'citation density'. In some disciplines, typical publication and citation rates mean that there is little difference in the citation counts for papers judged to be relatively important or only of marginal interest. In other fields the abundance of citations means that some papers are quite obviously relatively highly cited. Can we assess which fields are likely to be a problem in this regard?
- The coverage of the commercial databases is not universal. In some fields only some journals are included while in other fields many cited items are not journals at all. Either way, current research refers to and therefore presumably relies on material which is not in the scope of our evaluation methodology. We should be aware of which fields are affected.
- Some disciplines make particular use of conference proceedings. What is the prospect for using conference proceedings for improved research evaluation in the future, and is this possibility likely to influence researcher behaviour?

Citation density

2. HEFCE determined that some subjects should be included in the REF pilot exercise while others would not be included at this stage based on estimates of coverage within the citation database

3. The threshold criterion used by HEFCE for the pilot was one where the output records submitted within a subject (specifically the Unit of Assessment) were made up of at least 40% journal articles covered by widely available commercial databases.

4. What other factors would affect the relative value of bibliometric analysis as a tool for evaluating research performance?

- Do researchers tend to see journal articles as reasonably representative of their work?
- How many articles that can be subsequently identified in commercial citation databases does a typical researcher publish per year?

This is about the portfolio from which work might be drawn. In an earlier part of the exercise we collected data to put in context the balance of material of

different output types (see Table F1 in Annex F of 'Pilot study of bibliometric indicators of research quality: Development of a bibliographic database'¹).

It is possible that a researcher might be able to identify four journal articles over a five-year evaluation cycle but actually publish no more than one article per year. The total volume of material for the UK community in a subject would be fairly sparse. In another subject, the norm might be as many as five or even more articles per researcher per year.

Clearly the potential richness of data for subjects with very different publication rates would be different and this might then affect the weight given to any evaluation based on such data.

- How many references are there in each article?

This is about reference rates, not citation rates. It is about the numbers of citations there are to be spread about rather than the average number of citations acquired by an article. Obviously the two are related. In social science journals the reference lists tend to be short, while in molecular biology they are long.

Relative publication rates

5. We assembled relevant data from the Evidence UK Higher Education Yearbook. This aggregates data at a relatively coarse level but the outcome is then a fair representation of broad areas of research and the differences between them. There are certainly additional sub-field issues and there is variation within some of the fields, e.g. within Biological Sciences we might expect some contrast between organismal and molecular areas, but the broad trends remain a good indicator.

6. The absolute size of a field is important as well. In a small field there is an appreciable cap to the volume of available citations, even if output/FTE and references/output are high. So, the data on average citations per paper gives us a measure of average availability per paper but the variation is potentially greater if there are a lot of people producing a lot of papers.

7. For 2002-2006, we can look at the annual average of Thomson Reuters indexed papers per staff. Our relative output calculations are based solely on the numbers of academic staff since these are the principal investigators. We have also displayed HESA data on the numbers of additional staff on research grades, who will likely contribute to the overall publication impetus (Table H1).

¹ 'Pilot study of bibliometric indicators of research quality: Development of a bibliographic database. Report to UK HE funding bodies by Evidence', http://www.hefce.ac.uk/pubs/RDreports/2009/rd14_09/

Table H1. An estimate of annual indexed journal articles (Thomson Reuters' data) relative to numbers of academic staff (HESA data) for main subject areas in UK HEIs

Subject area	Academic staff	Additional research staff	Average annual output
Clinical	7,109	7,841	4.63
Biological	5,026	5,727	2.80
Physical	7,372	5,298	3.99
Engineering & technology	12,463	5,889	1.21
Visual & performing arts	3,979	338	0.08
Humanities & languages	8,223	908	0.23
Social sciences & business	22,208	3,007	0.35
Health	12,383	3,128	0.33

8. There are clearly three broad areas, the core sciences, where the output per person per year provides a significant flow of information about research activity as reflected in publications. In engineering the flow is much reduced, perhaps because many outputs are in conference proceedings which will become increasingly well documented in the near future.

9. In the social sciences, humanities and arts the typical output is so low as to make bibliometrics only a marginal guide to research activity and hence only a possible very partial indicator of performance.

10. 'Health sciences' is a mixed area. Many staff are professionally engaged and used practitioner journals not well covered by commercial data. Others are in areas that provide excellent bibliometrics. This is therefore an area in which bibliometrics may be seen by some commentators to provide useful information but which in practice cannot be assumed to provide sufficient information without support from peer review.

Relative reference rates

11. Counting the number of references from journal articles is not as informative as might be thought at first glance. The problem is that not all references are to other journal articles, let alone to other journal articles in the underlying database. Also referenced are book, chapters in books, conference proceedings, web-sites, and a variety of reports. None of these will add to the citation currency available to create citation counts to articles.

12. We have therefore fallen back on the use of average citation rates as the most informative index for purpose. We are therefore asking a question about the total numbers of citations to articles over a period.

13. It is likely that the REF would take as its census period a four- or five-year window. How many citations would on average be made to an article in any field over such a period?

14. We looked at Thomson Reuters ESI data for the five-year period 2003-2007. We aggregated data by the major categories which approximate to the five subject areas on which we focused in prior reports (Cancer, Biology, Physics, Earth Sciences, Mechanical Engineering). We also included a generic analysis of social science to explore the extent to which citation rates differ (Table H2).

15. The average citations per paper shown in Table K2 is reflected in the normalisation factors used to change actual citation counts to citation impact.

Table H2. The spread of citations per paper averaged over the period 2003-2007 for publications in a spread of research fields

UOA in pilot study	Relevant ESI categories	World papers	Average citations per paper
Cancer	Clinical Medicine	946,944	5.64
	Immunology	58,158	9.88
Biological sciences	Molecular Biology & Genetics	129,274	11.51
	Biology & Biochemistry	269,094	7.40
	Plant & Animal Science	256,695	3.07
Earth sciences	Geosciences	128,203	3.79
Physics	Physics	443,052	3.98
	Space Science	57,807	7.15
Engineering	Engineering	370,916	1.86
Social science	Social Sciences general	177,249	1.88

16. There is an immediate contrast between the science areas and engineering. Although the citation rates vary significantly between subject areas within both biology and physics, all of these have markedly higher rates than does engineering which has citation rates more akin to those of social sciences.

17. There are some consistent patterns in the two analyses. Science researchers publish more and they cite more. We can take some headline data from Tables H1 and H2 and from this create a simplistic but perhaps useful index (Table H3) which suggests the sorts of numbers of citations to other journal articles generated by each researcher during a typical year. This expresses the availability of citations to other papers. Where availability or density is relatively high then the spread of possible citation frequency outcomes is greater and therefore better at distinguishing between the average, the good

and the excellent in terms of the number of times a paper might be used and referenced by later work.

Table H3. An index of relative citation density based on estimates of typical output per academic staff (Table H1) and citations per paper (Table H2) for different fields of research

	Papers per person	Citations per paper	'Citation density'
Clinical	4.63	5.64	26
Biology	2.80	7.40	20
Physics	3.99	3.98	16
Engineering	1.21	1.86	2
Social science	0.35	1.88	0.66

18. Recall that this analysis does not take account of citations to articles from non-journal material nor the citations from articles to non-journal material (the latter is discussed in the next section).

19. This is not a definitive analysis and we accept that it is entirely open to criticism. A complete analysis of citation density would need to take into account the numbers of researchers in a well-defined community, their typical publication rates, the detailed and disaggregated patterns of references in their papers and – perhaps also – the way in which they cite themselves, major authorities and other parts of the literature.

20. While this analysis is only a rough guide, to the extent that it does throw light on the utility and application of bibliometric analyses, it suggests that bibliometric evaluation should work well for most areas of science, but that it can only be used with caution in engineering and the social sciences. This outcome would broadly accord with many community surveys.

Coverage of cited items

21. Apart from considerations of publication and citation rates there is also a need to consider what part of the literature used by the research community is actually covered by the commercial databases.

22. Thomson Reuters' policy on journal coverage is clearly stated as one which seeks to include the most highly cited part of the serial literature. To this extent it is an approach focused firmly on research excellence, because that part of the literature used most by researchers (hence most frequently cited in the past) is likely also to be that most sought by other researchers in the future.

23. Not all fields of research rely solely on journal articles to disseminate key research outcomes. Our question is therefore, what proportion of the material that is referenced by articles in the Thomson Reuters database are also items in that database? How much of what is cited is an indexed item (i.e. one to which the citation can be recorded)? If many

cited items are not in the data then we are only getting a partial view of what is important to the researchers.

24. Thomson Reuters' databases hold limited information about citations to non-indexed items (i.e. items in non-indexed journals or non-journal outputs). To process the data, because there is no reference database to non-indexed items, it is necessary to analyse the cited items by relatively intensive processes compared with the management of journal citations. It is therefore feasible to do this for specific data samples rather than the entire UK database.

25. We analysed all the indexed articles for a leading UK research-based university with a diverse portfolio including a strong social science base. The university was among those which submitted to a very large number of UOAs in RAE2001 and RAE2008. However, the data may not necessarily be representative in the range of subject areas, output modes or citing behaviour. Less research intensive institutions are probably less, rather than more likely to use journals indexed by Thomson Reuters.

26. We reviewed all the information in the database about citations from these papers to non-indexed items. The cited items in this analysis are classified by UOA according to the subject matter of the citing article. We reported to HEFCE in November 2007 on the spread of subject categories of cited items².

27. It is important to note that, although potentially useful, the data on citations to non-indexed items are complex, uncategorised and non-standardised. Thus the data are of sufficient quality to count non-indexed cited items but, for example, it is not then possible readily to analyse these cited items by subject category.

28. Table H4 shows that there is a substantial variation between UOAs in the percentage of cited items that are non-indexed items. The overall average percentage of indexed items across all subjects was about 60%. Under this gross average, it is important to note the contrast between the natural science and the social science subjects. It is apparent that whereas at least 75% of the material cited in the science UOAs is also Thomson Reuters-indexed journal article material, this is true for only about 25% of the social science articles.

² 'Report to the Higher Education Funding Council for England: Bibliometric analysis of interdisciplinary research', http://www.hefce.ac.uk/pubs/rdreports/2007/rd19_07/

Table H4. Percentage of items cited by articles indexed by Thomson Reuters that are not also in the indexed data (data sample taken from a large research intensive UK university)

Subject area	Number of source articles	Number of cited items	Cited items not indexed by Thomson Reuters
All subject areas	62,965	890,876	38%
Clinical laboratory sciences	13,125	219,613	20%
Biological sciences	9,501	202,200	24%
Chemistry	5,149	96,768	27%
Psychology	4,620	93,720	27%
Computer sciences	589	9,955	69%
Mechanical engineering	1,466	24,368	41%
Geography	1,589	20,644	54%
Sociology	629	11,408	75%

29. The analysis in Table H4 suggests that the natural science literature, including Psychology, makes extensive use of literature that is indexed by the commercial databases. A high proportion of cited items are also indexed items so that more than three-quarters of citations are captured within the Thomson Reuters database. This is less true of engineering, but the majority of citations are still captured within the database. Within social sciences this seems no longer to be true. Indeed, for Education (small data sample not tabulated) some 85% of citations are outside the database.

30. Overall, our conclusion is that coverage for natural sciences and – with some caution – engineering seems to conform with the expectations of the methodology for evaluation, that a high proportion what is cited by the target literature should contribute to the impact measures for other parts of the evaluated literature. This is not true, however, for social sciences.

Conference proceedings

31. Bibliometric analysis has hitherto been concerned primarily with journal articles, while the range of output modes used by researchers is actually much more diverse. This diversity is particularly true outside the core clinical, biological and physical sciences.

32. In some disciplines, such as social sciences and humanities, monographs have played a major role as the key mode for dissemination of the most significant research outcomes. There is a suggestion that this pattern is changing in the social sciences as they adopt a more 'US' and less 'European' research style which favours the wider use of journals, but it is generally agreed that the humanities are likely to continue to use predominantly non-journal outputs.

33. In technology based areas, including much of engineering and computer sciences, there has always been a substantive use of conference proceedings as a key dissemination mode. One reason for this is that this is a particularly effective route for communication with key user groups in more applied research areas. Research users attend conferences in order to find out about the most current research developments, and then make use of the proceedings, whereas they are less likely to read widely among academic journals. As a consequence the major professional institutes and associations support and manage many prestigious and widely recognised conference series.

34. Conference proceedings may, of course, also be important in other disciplines. While there are fewer conference series in some natural science areas than is true of engineering, there are often conferences on special topics where the content is certainly of similar status to key review publications.

35. Evidence from successive RAEs, summarised in Table 5, suggests that there may be some shift towards journal articles, at least in terms of what is submitted for evaluation among researchers' 'four best publications'.

Table H5. The balance of different output types by main subject areas in successive rounds of the UK Research Assessment Exercise

RAE1996	Science		Engineering		Social sciences		Humanities and arts	
	Outputs	%	Outputs	%	Outputs	%	Outputs	%
Books and chapters	5,013	5.8	2,405	8.1	16,185	35.1	22,635	44.4
Conference proceedings	2,657	3.1	9,117	30.8	3,202	6.9	2,133	4.2
Journal articles	77,037	89.8	16,951	57.3	22,575	49.0	15,135	29.7
Other	1,104	1.3	1,122	3.8	4,154	9.0	11,128	21.8

RAE2001	Science		Engineering		Social sciences		Humanities and arts	
	Outputs	%	Outputs	%	Outputs	%	Outputs	%
Books and chapters	1,953	2.5	1,438	5.4	12,972	28.6	25,217	46.5
Conference proceedings	751	0.9	3,944	14.9	857	1.9	1,619	3.0

Journal articles	76,182	95.8	20,657	78.1	29,449	65.0	17,074	31.5
Other	618	0.8	408	1.5	2,008	4.4	10,345	19.1

RAE2008	Science		Engineering		Social sciences		Humanities and arts	
	Outputs	%	Outputs	%	Outputs	%	Outputs	%
Books and chapters	1,048	1.2	216	1.2	12,632	19.0	21,579	47.6
Conference proceedings	2,164	2.5	326	1.8	614	0.9	897	2.0
Journal articles	80,203	93.8	17,451	95.4	50,163	75.5	14,543	32.1
Other	2,125	2.5	301	1.6	3,018	4.5	8,287	18.3

36. There is evidently a marked shift from conference proceedings to journal articles for engineering, and from books to journal articles for social sciences, in the successive assessments of 1996, 2001 and 2008. This is partly a genuine reflection of changing cultures and partly a reflection of expectations about what kinds of output are the most effective evidence of research performance. While conference proceedings remained of significance for engineering research evaluation in 2001, by 2008 they represented less than 2% of what was submitted.

37. It is also true that much of what appears in regular conference series is subsequently published in journals. Indeed, as HEFCE's own analysis shows (Annex L) some of what is published in journals is in fact a direct transmission of conference proceedings. The conference paper represents an early announcement of key outcomes, disseminated in order to make the findings available as quickly as possible. Although the subsequent journal article may be regarded as the 'item of record' it may well be the first publication that attracts attention and subsequent citations.

38. Overall it is the case that conference proceedings represent a significant body of research reports, and that these proceedings are likely to receive an assessable balance of citations from later research publications. It is therefore likely that the incorporation of conference proceedings into bibliographic databases and their subsequent indexing for citation analysis will strengthen the information content of any publication-based research evaluation.

Database content

39. It has been widely noted that the Scopus database, contains many conference series. It is also the case that the Web of Science database, now managed by Thomson Reuters, has always included a wide range of conference proceedings. In both cases that range of conference coverage is now being enhanced, in the case of the Web of Science by combining the existing data with a pre-existing conference database that had been run in parallel.

40. Both databases are likely to extend and refine the range of conference proceedings that they index. The Web of Science has well advertised criteria that determine the range of journals that are included, which are primarily the most frequently cited international journals plus leading regional and national journals in select subjects.

41. It is unclear how conference series would compare with journals on these criteria and therefore the effect of the additional material will need to be evaluated. Similarly, little is known of the effect in Scopus of changing the balance between journal and conference elements.

Effect on impact

42. While much scientometric research has explored the citation characteristics of journal articles, less attention has been paid to the relationships between conference proceedings and the material to which they refer.

43. Little is known about researcher culture in the use of conference proceedings. Do researchers tend to see a proceedings publication as an essentially ephemeral item, which may contain some key references but will not take a full view of the literature, or do they write up the item and incorporate references in the same way as they would a typical journal article?

44. The balance between types of reference may also be different. For example, a proceedings item necessarily refers to particularly topical outcomes and may therefore contain an exceptional proportion of references to other recent work rather than the longer established corpus. Or, because it is innovative, it may anchor itself more formally in major authorities and refer disproportionately to well-recognised older work, in order by association to bolster the authority of the new. Or proceedings may refer less to methodology, because of time and space constraints.

45. Equally little is known about the way in which conference proceedings are cited. Do other researchers seek to make particular use of proceedings because they represent the first announcement of new discoveries or do they prefer to refer to the article of record?

46. Other possible variant examples will occur to researchers in different disciplines from their own experience and awareness of the various disciplinary cultures in this regard. The overall conclusion is that proceedings remain an unknown quantity in terms of citation data.

47. The potential here is that the additional citations from proceedings may change the balance of citations to different articles and thereby alter our perception, from indicator data, of what is relatively good. This may seem unlikely but it remains at present an unknown.

48. It is certainly desirable that conference proceedings should be treated separately from other types of publication for the purpose of citation indicators. This would follow the practice of separating articles and reviews, which evidently have different citation characteristics.

Conclusion

49. Data on the publication and citation frequency of conference proceedings extends the information available for evaluation. But it is not just more, it is also different information. We must treat it with caution until we fully understand its properties and its interpretation.

50. Publishing cultures are changing, as the data from successive RAE suggests and other reports also indicate. The predominance of the journal article as the key output mode in natural sciences may become an increasingly common characteristic in other fields. For the present, however, conference proceedings remain an important output mode in engineering and technology.

51. Is it therefore not yet possible to be sure:

- How the incorporation of conference proceedings will affect the balance of citations to articles that are more or less frequently cited by other articles.
- Whether conference proceedings will tend to cite particular types of article, such as reviews or methodology, more or less.
- Whether the time spread of references in proceedings is particularly topical or more historical than articles.
- What the citation characteristics are of proceedings compared with articles.

52. A further and more extensive report on the issues raised in this section is expected from Thomson Reuters later in the year. That report will be made available to HEFCE.