June 2009

Interim report of the REF bibliometrics pilot exercise

Report by HEFCE

Contents

Executive summary

Purpose

1. This interim report discusses the main potential models for producing and using bibliometric indicators in the Research Excellence Framework (REF) which have been tested through the pilot process. A summary of the Expert Advisory Groups' discussion of the outcomes and their advice on the use of bibliometrics in the REF is included, as is initial feedback from pilot institutions.

Potential models

2. Through the bibliometrics pilot we have identified three main approaches or models for producing bibliometric indicators in the REF:

a. Model 1, based on institutional address. In this model the papers associated with each higher education institution (HEI) are taken directly from either the Web of Science (WoS) or Scopus, based on address data within the databases. Papers are assigned to HEIs based on the addresses of their authors, and they are assigned to a subject category (or multiple subject categories) depending on the journal in which they are published.

b. Model 2, based on authors, all papers. In this model, an attempt is made to identify all the papers published by specified groups of staff within each unit of assessment (UOA). For the purposes of the pilot process, we included all staff that were selected for the 2008 Research Assessment Exercise (RAE), in relevant UOAs (although variations to this were also tested).

c. Model 3, based on authors, selected papers. In this model, only the most highly cited papers by staff that were selected for the 2008 RAE are considered.

3. This interim report presents outcomes of the pilot exercise for each of these three models, for a selection of UOAs. This is to illustrate variations between these three models. A fuller analysis, including outcomes for all UOAs covered by the pilot, will be published in autumn 2009.

Key points

Expert Advisory Groups' discussions

4. The REF Expert Advisory Groups reviewed the outcomes from the pilot, and advised us on the robustness of the data and potential use in the REF. There was a strong consensus that bibliometrics are not sufficiently mature to be used formulaically or to replace expert review, but there is considerable scope for citation indicators to inform expert review in the REF.

5. There was widespread agreement that the most appropriate approach is to focus on citation indicators for selected papers by the staff in each submission, rather than attempt to capture all papers.

6. There are a number of ways in which bibliometrics can be used to inform expert review to enhance the reliability of the process and in some cases reduce assessment panel workloads; the particular ways in which the data are used could vary between panels.

Feedback from pilot institutions

7. The pilot HEIs have provided initial feedback. There was consensus that an author-based selective model maps most strongly onto institutions' perceptions of research excellence and onto existing institutional management systems.

8. There were concerns about institutions' ability to identify all papers published by authors at their institution, felt to be necessary for both an authors' 'all papers' model and a model where papers are identified by institutional address. Furthermore there were concerns about the robustness of mapping papers to UOAs in the institutional address model.

Introduction

9. Following the government's announcement about the reform of the research assessment and funding framework in 2006, HEFCE was asked to develop a framework based on metrics (bibliometrics, research income and research student data) for the science-based disciplines and on light-touch expert review for the other disciplines.

10. During 2007 we commissioned a scoping study to investigate the potential for using bibliometrics in the new framework – the Research Excellence Framework (REF)[1]. In November 2007 we consulted on a set of proposals on this basis, and committed to substantial further work to develop and test bibliometric indicators, through a pilot exercise.

11. Following the consultation, in spring 2008, we announced some key changes to the REF, to develop it as a unified framework using a combination of expert review and metrics (including bibliometrics) as appropriate to each subject.

12. We then initiated a pilot exercise with 22 higher education institutions (HEIs), with the following aims:

- to explore which subjects should use bibliometric indicators under the new framework
- to assess which categories of staff and publications should be included in future bibliometric exercises
- to test the main sources of citation data (the Web of Science and Scopus)
- to develop the process for collecting and managing bibliographic data
- to develop and test methods for analysing citations and benchmarking against international norms
- to identify our preferred means of constructing the indicator in the form of a citation profile
- to develop proposals for how citation indicators should be used to assess research quality within the REF
- to explore what supplementary information the process can usefully generate.

13. This report discusses the main potential models for producing and using bibliometric indicators in the REF which have been tested through the pilot process. The scope, processes and methodologies used in the pilot process are described, and some initial outcomes are presented. An earlier version of this report was presented to the REF Expert Advisory Groups'; a summary of their discussion of the outcomes and their advice on the use of bibliometrics in the REF is included here.

[1] Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education institutions. Report to HEFCE by the Centre for Science and Technology Studies, Leiden University November 2007
http://www.hefce.ac.uk/pubs/rdreports/2007/rd18_07/

14. Further reports on the pilot process will be published in the autumn, including:
    a. A fuller analysis of the outcomes.
    b. A full report of the pilot data collection process (by Evidence Ltd).
    c. A report of the pilot institutions' feedback on the pilot process (by Technopolis).

Potential models

15. Through the bibliometrics pilot we have identified and tested three main approaches or models for producing bibliometric indicators in the REF (within these are a number of sub-variants, discussed further below).

Model 1 – Based on institutional address

16. In this model we took the papers associated with each HEI directly from either the Web of Science (WoS) or Scopus, based on address data within the databases. Papers are assigned to HEIs based on the addresses of its authors, and they are assigned to a subject category (or multiple subject categories) depending on the journal in which they are published.[2]

17. This model is potentially a low cost approach, as – in principle – citation indicators for each discipline at each HEI could be produced without input from institutions. In principle, it provides a comprehensive picture of all an institution's outputs (within the WoS or Scopus). However, papers are not linked to specific members of staff, and they are assigned to units of assessment (UOAs) on the basis of the journals that they are published in.

Model 2 – Based on authors; all papers

18. In this model, we attempted to identify all the papers published by specified groups of staff within each UOA. For the purposes of the pilot process, we included all staff that were selected for the 2008 Research Assessment Exercise (RAE), in relevant UOAs (although variations to this were also tested).

19. This model provides a better link between papers, staff and UOAs. However, considerable effort was required to collect the data.

Model 3 – Based on authors; selected papers

20. In this model, we looked only at the most highly cited papers by staff that were selected for the 2008 RAE.

[2] Note that these subject categories are essentially clusters of journals, as defined within each of the two citation databases.

Pilot scope

Selection of institutions

21. The pilot process was being conducted with 22 HEIs. These HEIs were selected from a pool of 44 volunteers who responded to a survey sent out to the 63 HEIs who originally expressed an interest. HEIs were selected to provide coverage across a spectrum of types of HEI, research management modes, geographical spread and discipline area.

Institutions participating in the pilot exercise

Bangor University	London School of Hygiene and Tropical Medicine
University of Bath	University of Nottingham
University of Birmingham	University of Plymouth
Bournemouth University	University of Portsmouth
University of Cambridge	Queens University, Belfast
University of Durham	Robert Gordon University
University of East Anglia	Royal Veterinary College
University of Glasgow	University of Southampton
Imperial College London	University of Stirling
Institute of Cancer Research	University of Sussex
University of Leeds	University College London

Selection of UOAs

22. The pilot process was designed to test the applicability of bibliometrics across a range of subjects and to determine in what areas bibliometrics provides useable information for the purposes of the REF. To achieve this we selected those UOAs in RAE 2008 for which there was moderate coverage of research outputs (of 40% or greater) in either WoS or Scopus. A list of these UOAs can be found on the web-site: http://www.hefce.ac.uk/research/ref/Biblio/projects/datacoll/UOA.pdf

Units of Assessment included in the REF bibliometrics pilot process

UOA	UOA name
1	Cardiovascular Medicine
2	Cancer Studies
3	Infection and Immunology
4	Other Hospital Based Clinical Subjects

5	Other Laboratory Based Clinical Subjects
6	Epidemiology and Public Health
7	Health Services Research
8	Primary Care and Other Community Based Clinical Subjects
9	Psychiatry, Neuroscience and Clinical Psychology
10	Dentistry
11	Nursing and Midwifery
12	Allied Health Professions and Studies
13	Pharmacy
14	Biological Sciences
15	Pre-clinical and Human Biological Sciences
16	Agriculture, Veterinary and Food Sciences
17	Earth Systems and Environmental Sciences
18	Chemistry
19	Physics
20	Pure Mathematics
21	Applied Mathematics
22	Statistics and Operational Research
23	Computer Science and Informatics
24	Electrical and Electronic Engineering
25	General Engineering and Mineral & Mining Engineering
26	Chemical Engineering
27	Civil Engineering
28	Mechanical, Aeronautical and Manufacturing Engineering
29	Metallurgy and Materials
32	Geography and Environmental Studies
34	Economics and Econometrics
40	Social Work and Social Policy & Administration
43	Development Studies
44	Psychology
46	Sports-Related Studies

Staff and outputs

23. We wished to test a range of different criteria and models in the pilot process and therefore we asked for a comparatively wide range of staff and outputs from pilot HEIs. We appointed Evidence Ltd to run aspects of the pilot process, and jointly with them developed a specification for the data required from institutions; Evidence Ltd then managed the process of collecting the data from institutions and reconciling it with the WoS data.

24. Our minimum data requirement was all staff submitted to the 2008 RAE, and all available outputs by these staff published between 2001 and 2007. We asked pilot HEIs to go beyond this where possible and return details of additional staff and outputs.

25. The citation analysis in the bibliometrics pilot considers only journal articles and review papers. However, we requested data on all types of research outputs from HEIs in order to assess the levels of coverage of the citation databases. Conference proceedings are not included in the pilot process but are increasingly covered within the databases and we will conduct further analysis including conference proceedings in due course.

26. We found that HEIs took different approaches to extending the data beyond staff selected for the 2008 RAE.

27. Although we collected outputs published between 2001 and 2007, for the purposes of analysis have only included papers from 2001 to 2006. This was to enable time for new outputs to become cited.

28. A full specification for pilot data collection is available on the HEFCE web-site: http://www.hefce.ac.uk/research/ref/resources/Bulletin.pdf

29. The pilot process sought to use readily available data where possible, and in particular to build on HEIs' existing RAE databases. It therefore used similar definitions and timeframes. We have analysed each of the three main models using proxies based on such definitions. These are described below.

Citation databases

30. We are using data from both the WoS and Scopus in the bibliometrics pilot. Both databases are under continued expansion and development and the data that we have taken from them represents a snapshot at the date at which we took it. Evidence Ltd are contracted to work with WoS data, HEFCE are working with Scopus data.

Pilot methodology

31. The pilot process involved the following stages:

a. Data collection: Data about institutional papers were harvested directly from WoS/Scopus for the address model, and data about staff and papers were collected from the pilot HEIs and matched to WoS/Scopus for the author-based models.

b. Citation counts and normalisation: The number of citations to each paper was counted, and this count was normalised by field, year and document type (in the same way for all papers in the author and address models).

c. Analysis: So far, we have produced citation indicators for the papers associated with each UOA at each pilot HEI, using each model (and their sub-variants). Further analysis is in progress.

Data collection

Institutional address model

32. For the institutional address model Evidence Ltd and HEFCE used data from WoS and Scopus that could be linked to a pilot HEI by its address. For Scopus, HEFCE used the 'affiliation ID' field to extract records associated with each pilot HEI. Where there was ambiguity about whether activity at associated medical schools etc. should be included in the extract, we took an inclusive approach.

33. For the WoS, Evidence Ltd used an address-mapping between UK addresses in the WoS data and identifiable organisations. This mapping has been developed over several years and is most detailed for UK HEIs. Where there is doubt, the reconciliation is verified by checking addresses via researcher web-pages. Evidence Ltd used this mapping to extract records associated with each pilot HEI. Where there was ambiguity about whether activity at medical schools etc. associated with HEIs should be included in the extract, Evidence Ltd took an inclusive approach.

Author models

34. For these models, the pilot HEIs provided Evidence and HEFCE with information on their staff and publications within the census period of the pilot process. These were supplemented by additional papers found by Evidence and verified by the pilot HEIs. Evidence and HEFCE then matched these data to WoS and Scopus respectively. These data were aggregated into UOAs by linking each paper to members of staff, and hence UOA.

35. Further discussion of data collection is at **Annex A**.

Citation counts and normalisation

36. Each database contains keys indicating the citation links that exist between items in the database. We use these to generate a citation count of the number of times that each item has been cited by other items in the database. We are dependent on the accuracy of the database providers capturing of these links; there is some evidence that they are not 100% accurate. We anticipate doing further work to examine the accuracy of the links in each database in due course.

37. Each database is a snapshot of papers captured up to the end of 2007. In order to allow a reasonable amount of time for each paper to accumulate citations, we only include papers published between 2001 and 2006 in our models. This gives them a clear year to be cited, before being included in the analysis.

38. Although only articles and reviews are included in the bibliometric analysis, we count citations to these items from **all** items in the database. This includes conference proceedings (in Scopus), letters, notes etc. We believe that that Scopus' broader coverage of conference proceedings may be having a significant effect on the citation counts in UOAs where these are a common mode of publication.

39. The number of times that papers are cited is not in itself an informative indicator; citation counts need to be benchmarked or normalised against similar research. In particular, citations accumulate over time so the year of publication needs to be taken into account; citation patterns differ greatly in different disciplines and so the field of research needs to be taken into account; and citations to review papers tend to be higher than for articles (note that for the pilot analysis only documents classified as 'article' or 'review' by the appropriate bibliometric database are included in the analysis) and this also needs to be taken into account.

40. For each paper in the pilot exercise (whether in the institutional address model, author-based models or both) we calculate:

a. The total number of times the paper has been cited up to the end of 2007.

b. The 'normalisation factor' – this is the average number of citations to all papers (worldwide) within the WoS or Scopus database of the same type (i.e. either articles or review papers), published in the same year, and published in the same 'field' (i.e. subject category).

c. The normalised citation score for each paper = a/b.

41. In terms of 'field' or subject category, for the WoS analysis we used WoS journal categories (there are 247 of these); the Scopus analysis used four figure 'asjc'[3] codes (there are 334 of these).

[3] All science journal classification.

42. Papers in journals that are assigned to more than one subject category are normalised against the mean normalisation factor for all the subject categories to which the journal is assigned.

43. Articles in multidisciplinary journals that are not assigned to a single subject category have been linked to subjects where possible by analysing the most frequent categories among the journals in the material they cite.

44. Normalisation is discussed further in **Annex B**. Within the pilot process, Evidence undertook some analysis to test normalisation at broader subject categories than WoS journal categories; however there were no apparent advantages to doing this. On the whole we expect that researchers would want their work normalised against relatively specific fields of research rather than broader groupings.

45. At a later stage we anticipate further work to refine approaches to normalisation, as the journal categories within the WoS and Scopus may not be the most suitable basis in all disciplines.

Analysis

Model 1 – Based on institutional address

46. To allow comparison with the author-based models, we produced indicators for each UOA. Papers linked to the pilot HEIs could not, through this model, be linked to their staff, so we assigned papers to UOAs by reference to the journals they are published in as follows:

a. We used the existing categories of journals within the WoS and Scopus databases. There are 247 WoS subject categories and 334 Scopus categories. The journals assigned to each WoS category can be found at http://scientific.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D .

b. We have a mapping of WoS subject categories to RAE 2008 UOAs, and a mapping of Scopus subject classifications to RAE 2008 UOAs. These were constructed by finding a best fit for each subject category or classification with reference to the journals in which items submitted to each of the RAE UOAs appeared. This allows us to associate papers to UOAs.

c. Where a journal (and hence the items published in it) is assigned to several subject categories, the work is included in the address model for all UOAs it is mapped to. Each paper is counted once per UOA.

d. Each database includes a 'multidisciplinary' category. Journals mapped to this include journals such as Nature, Science and Proceedings of the National Academy of Sciences of the USA (PNAS). Where possible, we assign each item in journals in this category to a more appropriate one by looking at the subject categories of the items that it cites.

47. The mappings for Scopus and WoS were developed independently by HEFCE and Evidence respectively.

48. These mapping processes map each subject category to a UOA. However, this also allows some UOAs to have no journal categories and hence no outputs associated with them. Because each journal can be assigned to several subject categories, some journals may be assigned to several UOAs.

49. This process provided us with a set of papers associated with each UOA at each pilot HEI, and indictors were produced as described below. Note that some papers, although produced by, for example, physicists, may not have been published in physics journals. As such, some work may be mapped to UOAs outside the scope of the pilot process and/or the UOAs that particular pilot HEIs did not provide data for.

Model 2 – Based on authors; all papers

50. In the author-based models we are able to associate each paper to one or more members of staff at the pilot HEI. From these links we are able to associate papers to UOAs, based on the staff who wrote them.

51. When looking at the 'all papers' model we include all papers linked to an RAE submitted member of staff associated with each UOA. We only include papers published between 2001 and 2006. An output will not appear more than once per submission, regardless of the number of authors from that submission who are linked to it. Papers can be attributed to several HEIs and/or UOAs, if linked to staff from them.

Model 3 – Based on authors; selected papers

52. To select papers for this model, we create a list of papers associated with each RAE submitted author, ranked by normalised citation score. We keep only the six papers with the highest normalised citation score for each author.

53. If a paper has been co-authored between two members of staff associated with a UOA, and it is in both of their 'top six' papers, the paper will only appear once in the submission. We are aware that this means one of the author's top six papers has been 'wasted' and that author may have fewer than six papers in the model. As such, we present this model as a proxy for a selective model rather than a model that could be implemented in practice. Further, for the purpose of the pilot process, we have selected papers algorithmically based purely on citations; in practice institutions could select their best papers.

54. Through the pilot process we have tested these three broad models, although for the two author-based models we have used proxies for the selection of authors and papers, rather than asking the HEIs to select staff and papers specifically for the pilot process. We have also tested some variations within these models.

Model variants

55. We looked at a range of model variants for the author-based models. These included limiting the papers included to those that were written while at least one of the authors was at the HEI and looking at papers associated with a wider spectrum of researchers than those they submitted to the RAE. The outcomes for these models tend to be quite similar to the 'Author-based; all papers' model discussed above. Unfortunately the data quality for these additional models is variable between HEIs, making comparisons between them difficult. We anticipate performing some further analysis on these models, where possible. We also plan to look at the effect of considering a shorter publication time window on the outcomes.

Overview of outcomes

56. For an initial overview of outcomes, we produced indicators for each of the three main models, using WoS and Scopus data, for each UOA included in the pilot process.

57. The indicators we use for this initial overview are:

> **% above 2x world average**: percentage of the HEI's papers that are above twice the world average (normalised) citation score
>
> **% above 4x world average**: percentage of the HEI's papers that are above four times the world average (normalised) citation score

58. Graphs showing these initial outcomes for a selection of UOAs are shown below. These are a representative sample of UOAs from across RAE main panels A to H, J and K. The graphs are presented in decreasing order of coverage within the citation databases.

59. The graphs are presented without interpretation at this stage. We have taken and relied on advice from the Expert Advisory Groups and interim feedback from the pilot institutions. These are summarised in the remainder of the report (paragraphs 65 onwards)

60. The pilot HEIs are in the process of reviewing the data in detail; until we have had further feedback from them about the data, we present these outcomes anonymously. Given that we include data from the 2008 RAE, we have also anonymised the UOAs for the purposes of this interim report. The final report due to be published in autumn 2009 will include data for all UOAs included in the pilot process and will not be anonymised. At this stage, the graphs are intended to highlight differences between the three models overall, rather than between individual UOAs or institutions.

61. There are two charts for each UOA; one showing each model using the lower threshold (2x world average) and one showing each model using the higher threshold (4x world average).

62. To aid interpretation of these data, we include indicators from the 2008 RAE output sub-profile (this sub-profile is more relevant than the overall quality profiles which include environment and esteem factors). This is in order to provide an initial 'sense check' of the outcomes of the pilot process against the only other available quality indicators (the RAE). This is not intended as a direct comparison between the RAE and the bibliometrics pilot outcomes, as there are a number of differences between the scope, coverage, assessment criteria and methods used for each exercise. For these reasons, the preferred bibliometric model should not necessarily be the one which provides the closest fit with RAE outcomes.

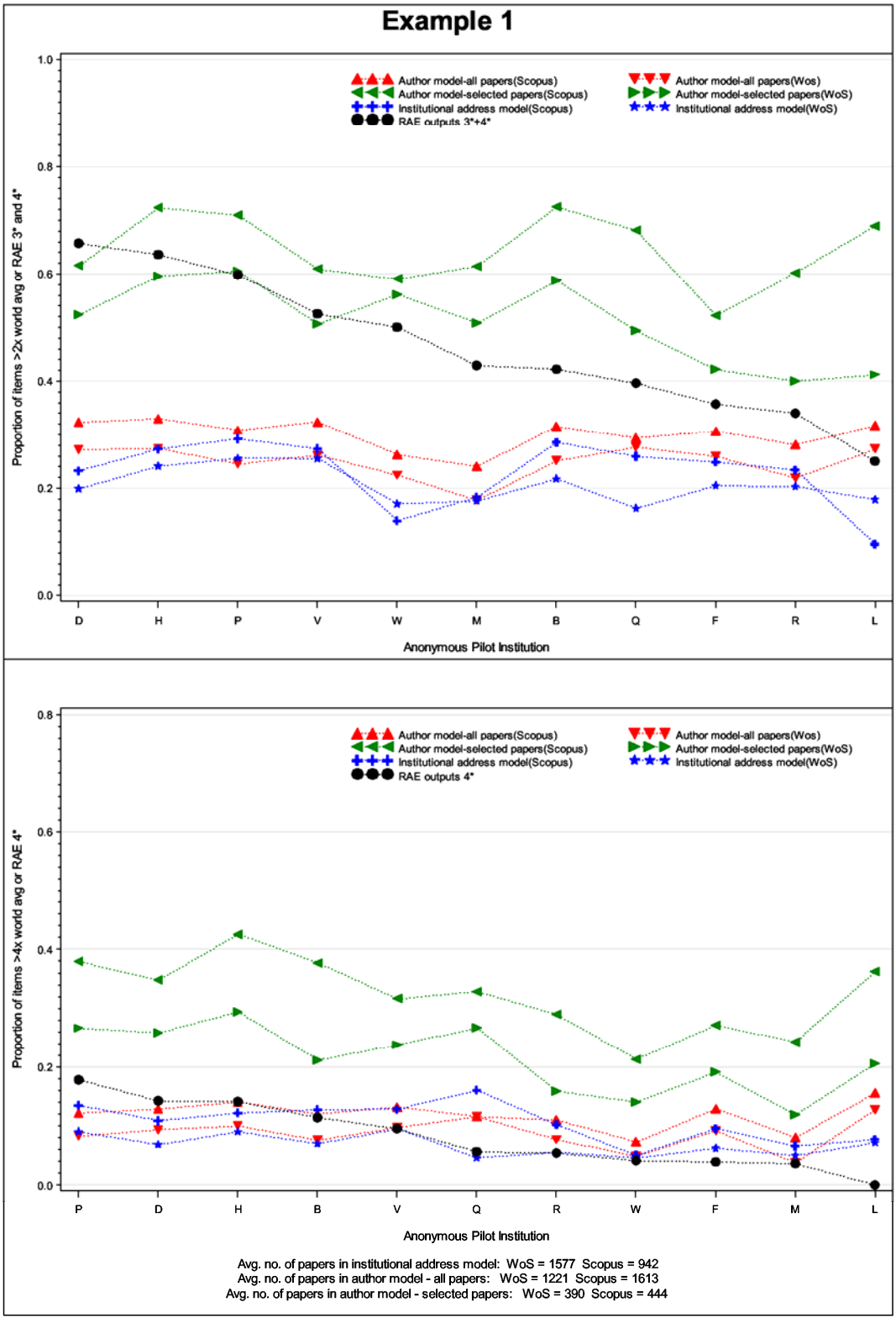
63. For the purposes of this 'sense check' against the pilot outcomes, we present the % above 2x world average bibliometric indicator alongside the proportion of outputs rated

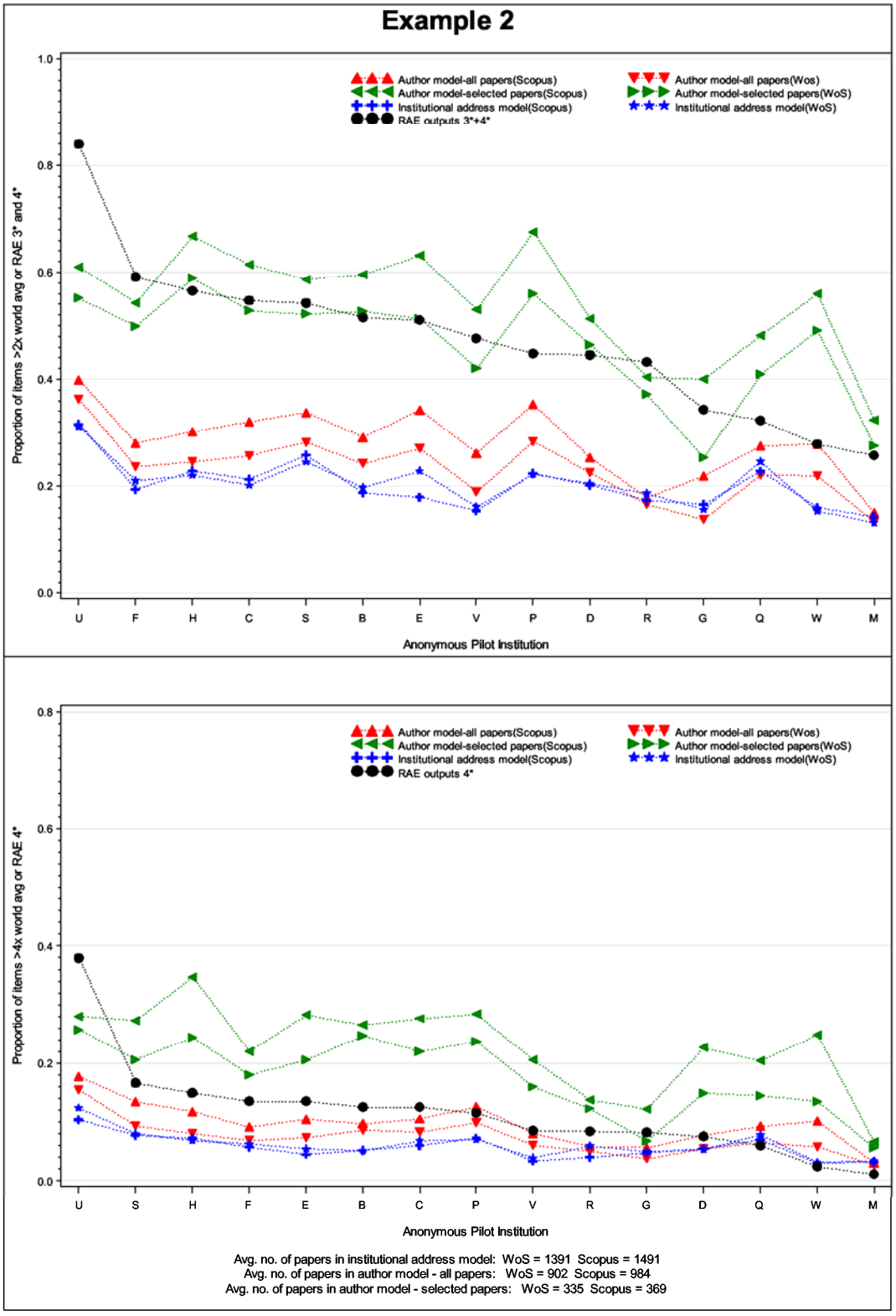
3* or above in the RAE. We present the % above 4x world average bibliometric indicator alongside the proportion of outputs rated 4* in the RAE. This is for illustrative purposes and does not mean that we regard 2x world average as equivalent to 3* or 4x world average as equivalent to 4*.

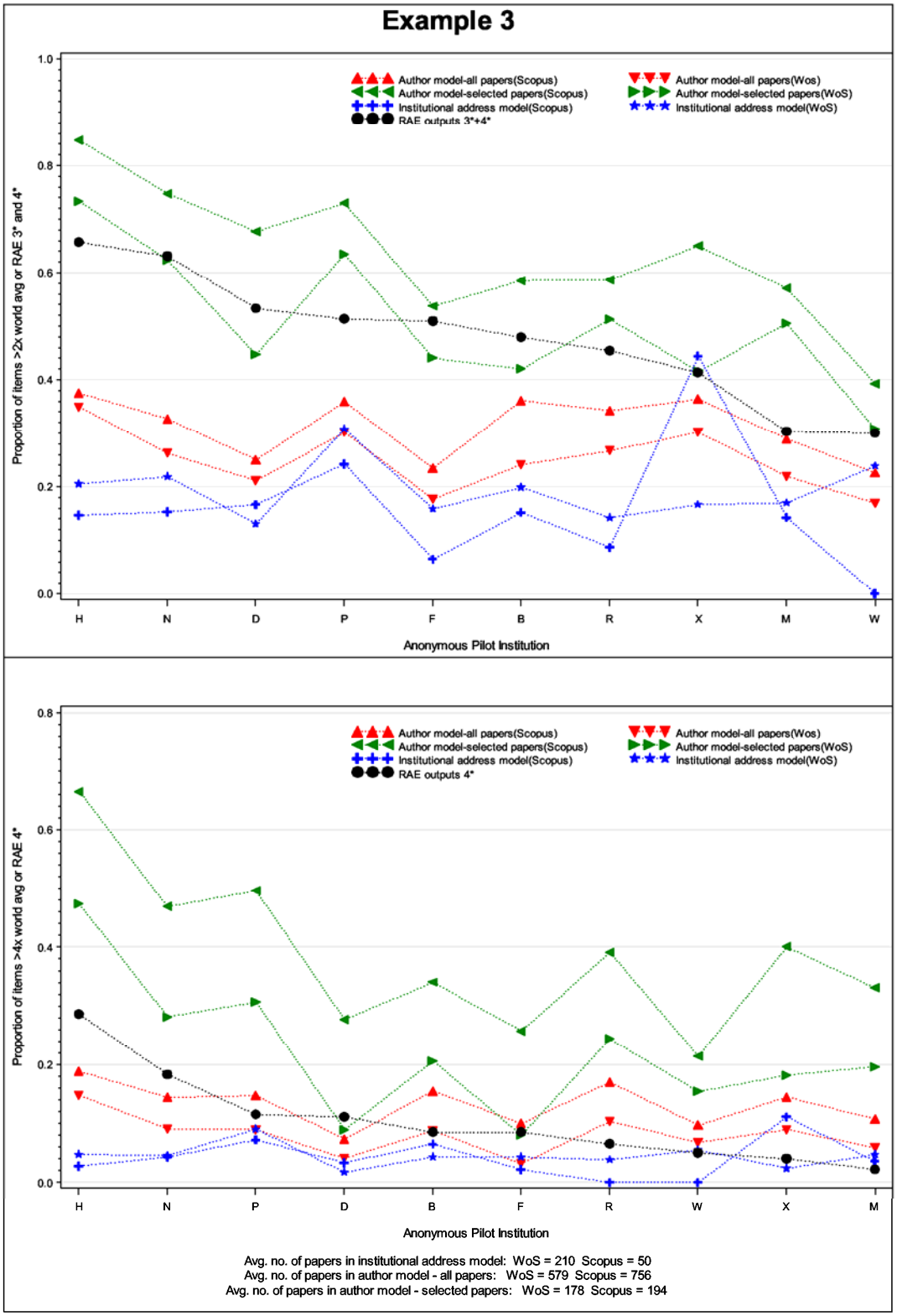
64. In reporting the bibliometric outcomes, we have applied a size threshold to the summary graphs where submissions with fewer than 50 papers have not been included.

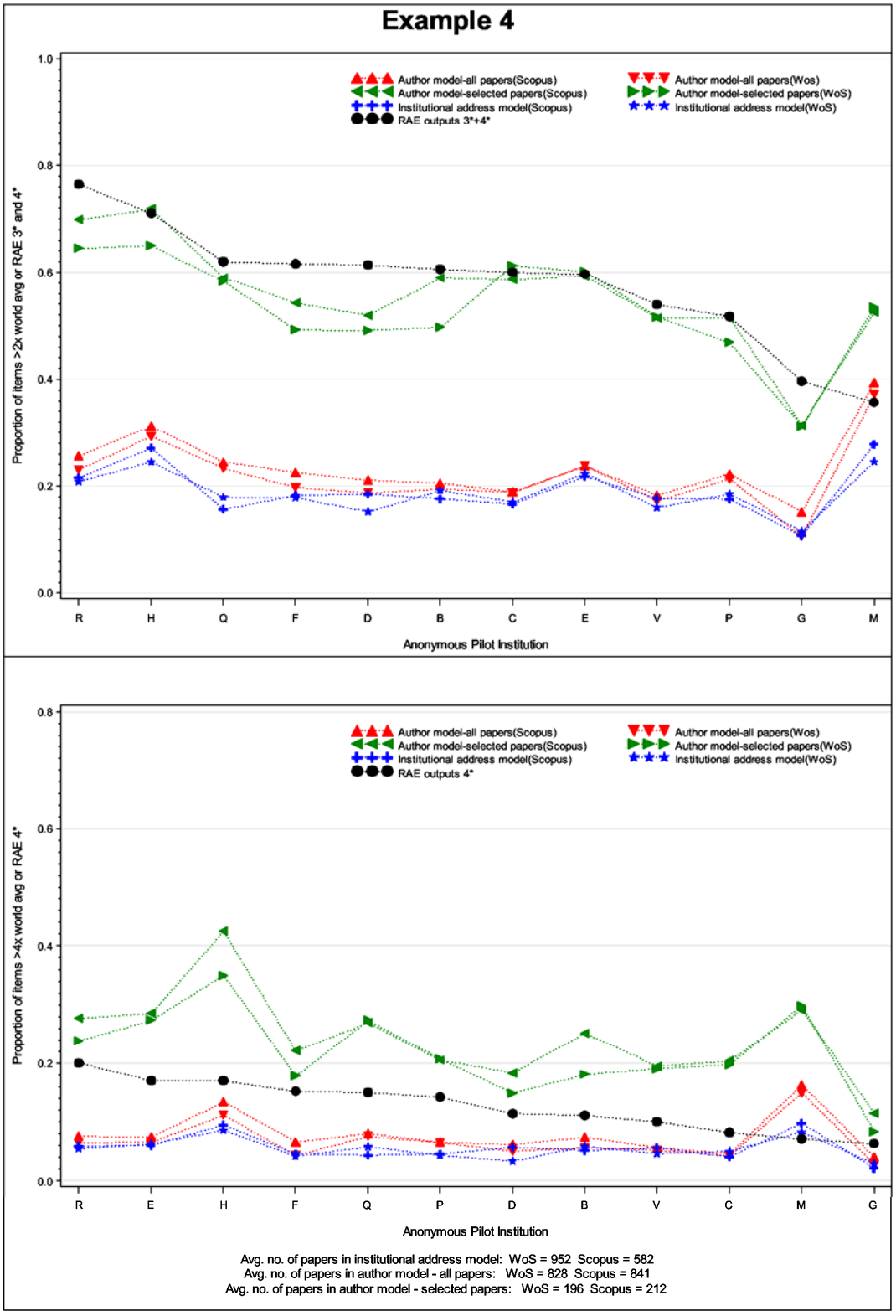
Graphs of the initial outcomes of the citation analysis for an anonymous sample of UOAs in descending order of coverage within the citation databases.

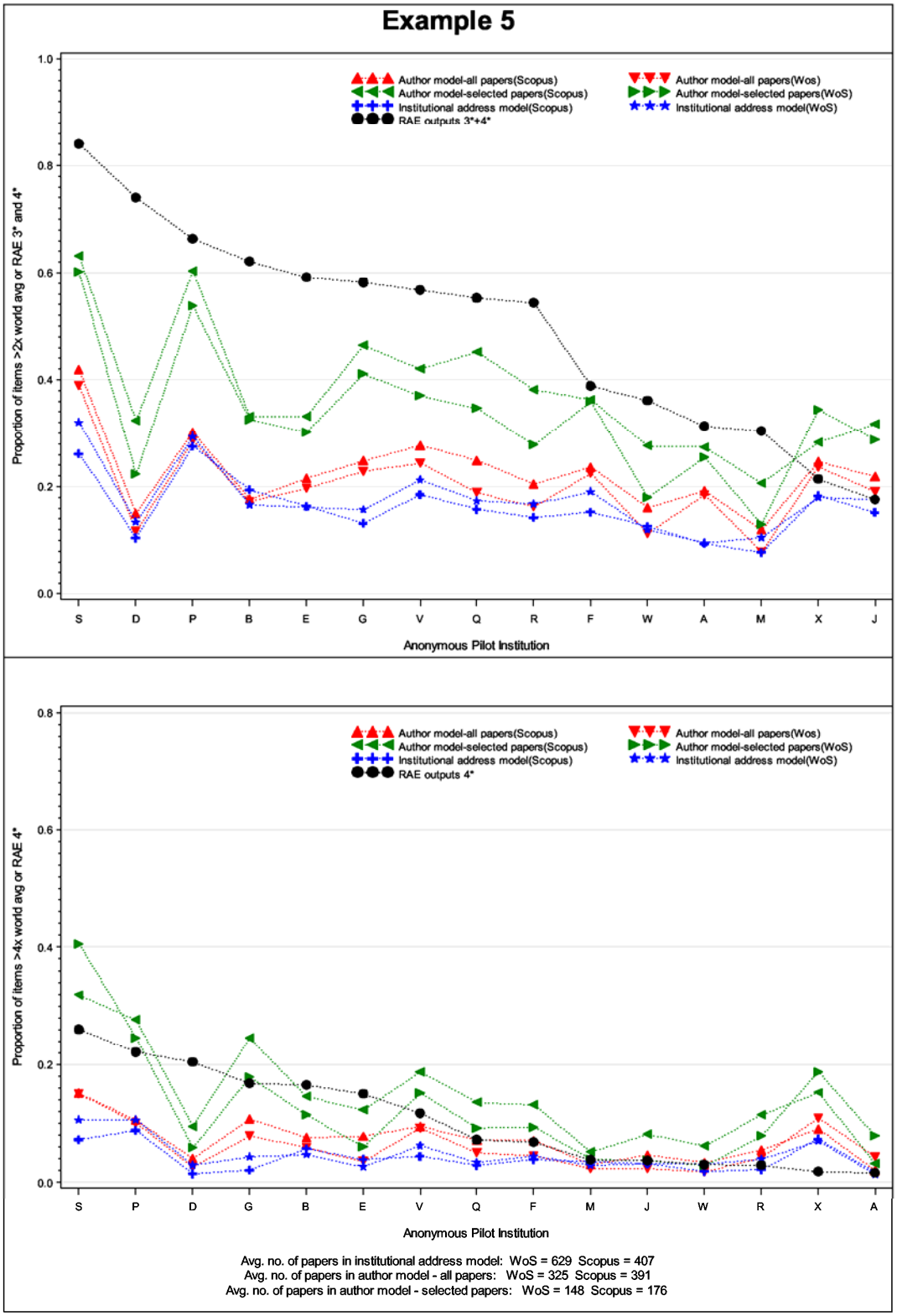
The top graphs show the proportion of outputs that were 2x world average in the analysis and the proportion of outputs rated 3* and 4* in the RAE outputs sub profile. The bottom graphs show the proportion of outputs that were 4x world average in the analysis and the proportion of outputs rated 4* in the RAE outputs sub profile (see paragraphs 57 to 61 for a full explanation).

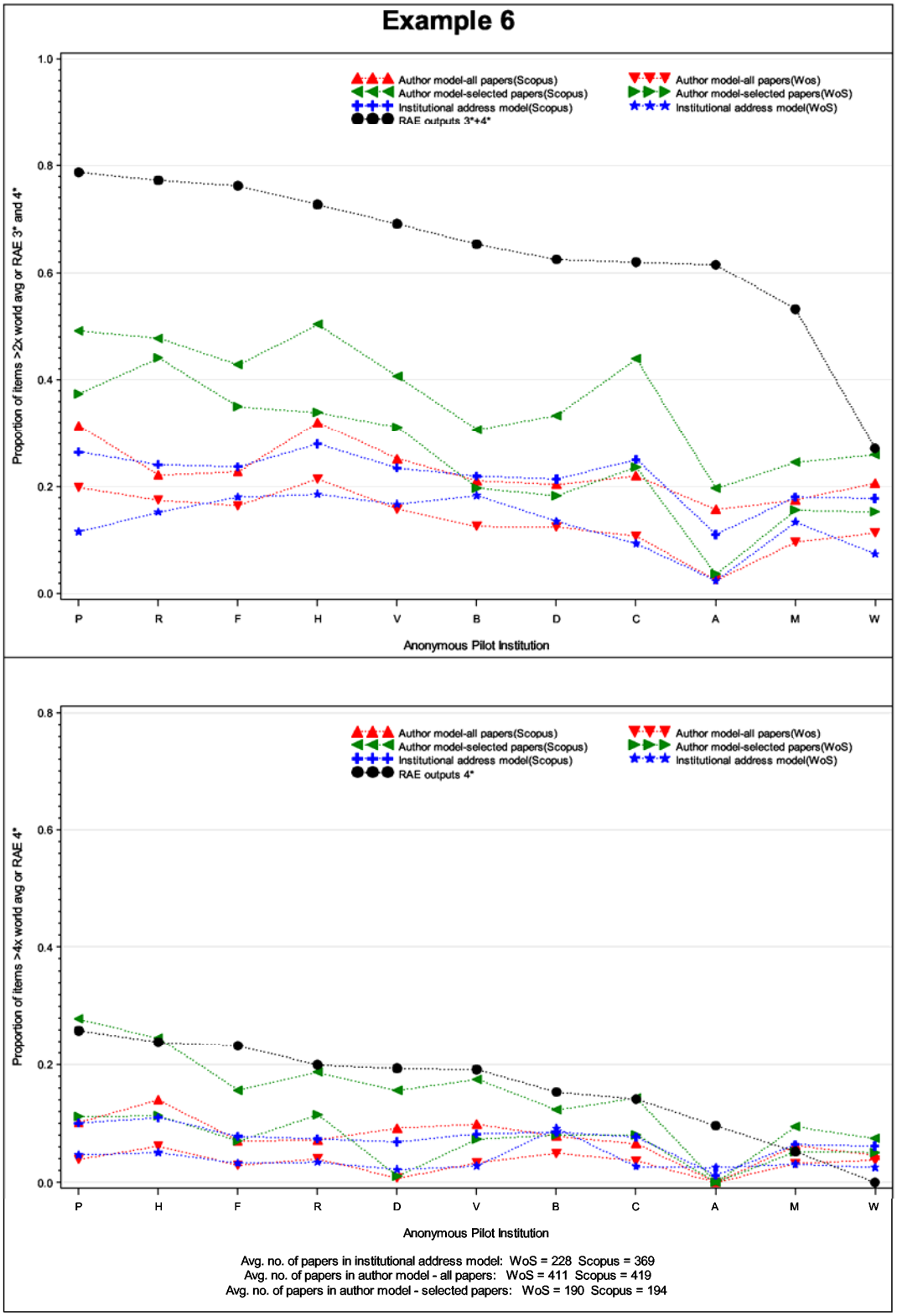

Example 1
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
D H P V W M B Q F R L
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
P D H B V Q R W F M L
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 1577 Scopus = 942
Avg. no. of papers in author model - all papers: WoS = 1221 Scopus = 1613
Avg. no. of papers in author model - selected papers: WoS = 390 Scopus = 444

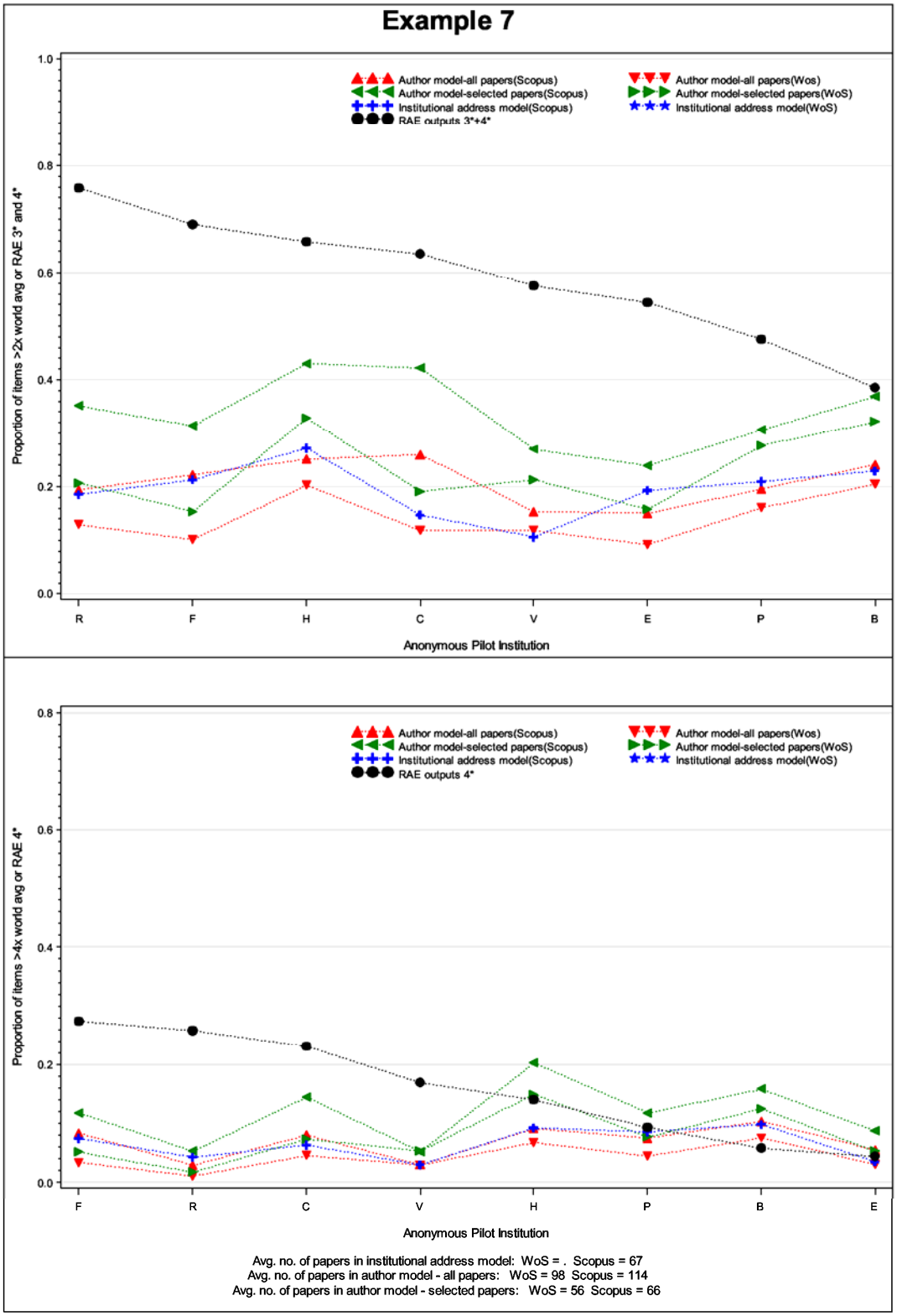

Example 2
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
U F H C S B E V P D R G Q W M
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
U S H F E B C P V R G D Q W M
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 1391 Scopus = 1491
Avg. no. of papers in author model - all papers: WoS = 902 Scopus = 984
Avg. no. of papers in author model - selected papers: WoS = 335 Scopus = 369

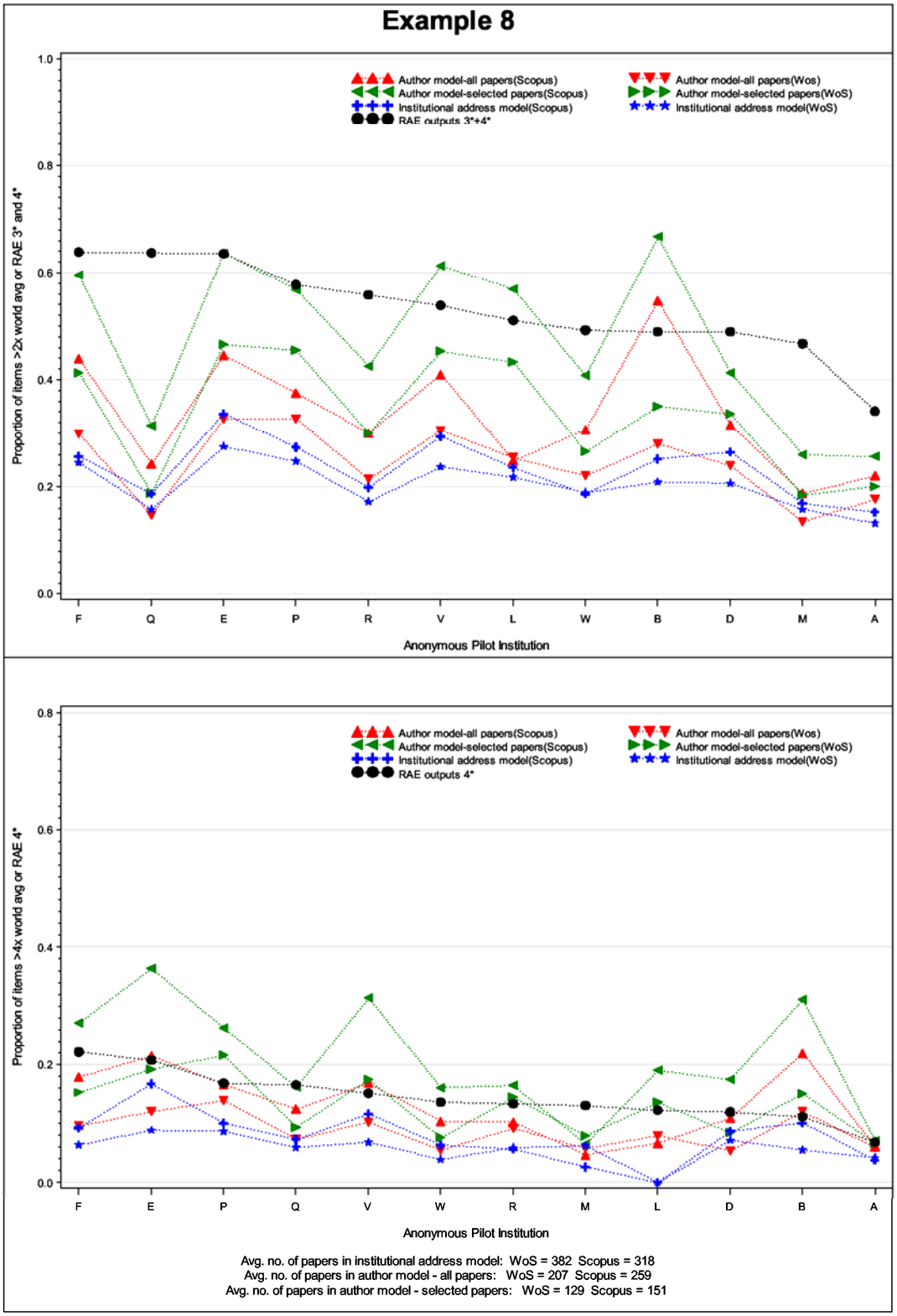

Example 3
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
H N D P F B R X M W
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
H N P D B F R W X M
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 210 Scopus = 50
Avg. no. of papers in author model - all papers: WoS = 579 Scopus = 756
Avg. no. of papers in author model - selected papers: WoS = 178 Scopus = 194

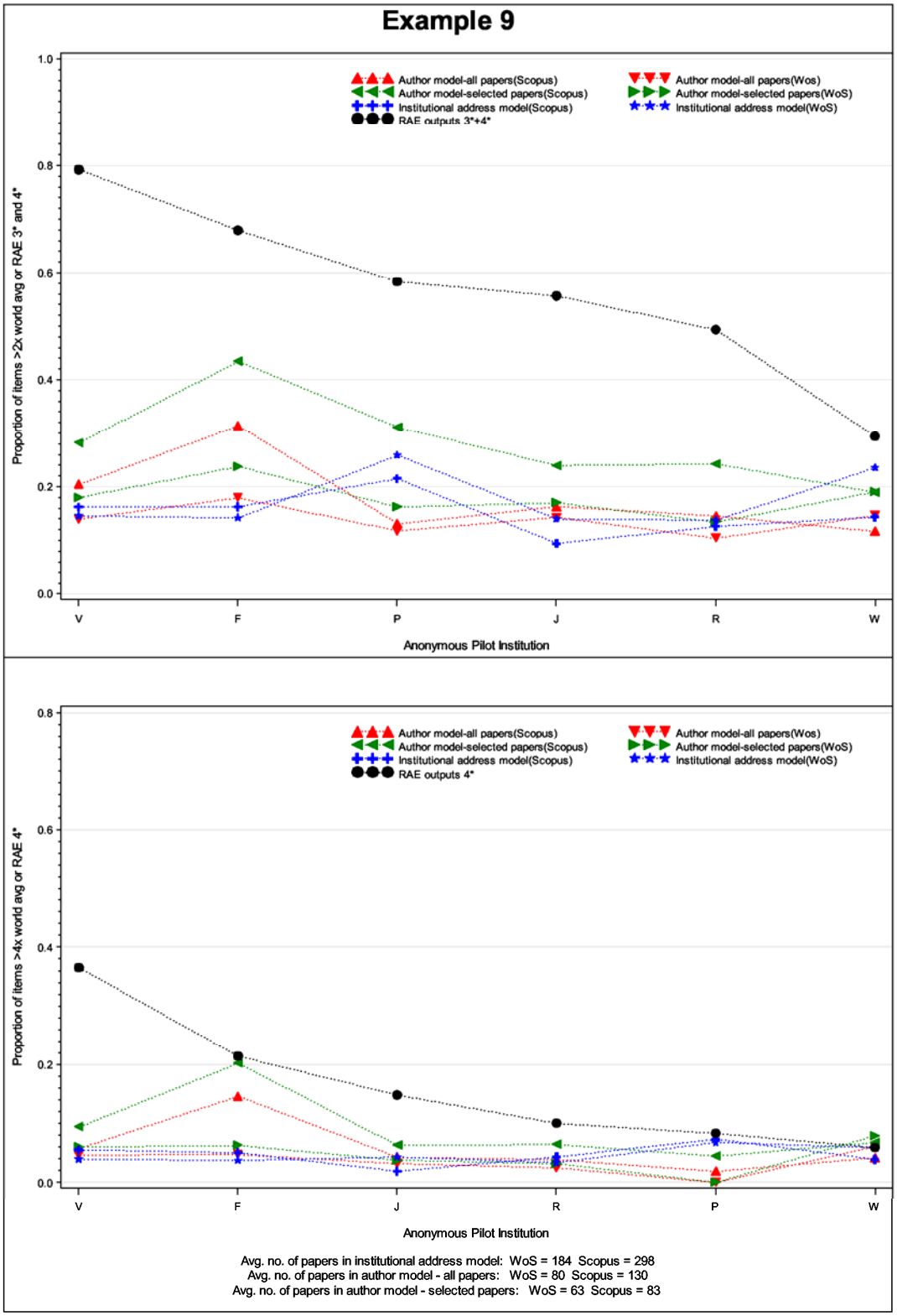

Example 4
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
R H Q F D B C E V P G M
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
R E H F Q P D B V C M G
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 952 Scopus = 582
Avg. no. of papers in author model - all papers: WoS = 828 Scopus = 841
Avg. no. of papers in author model - selected papers: WoS = 196 Scopus = 212

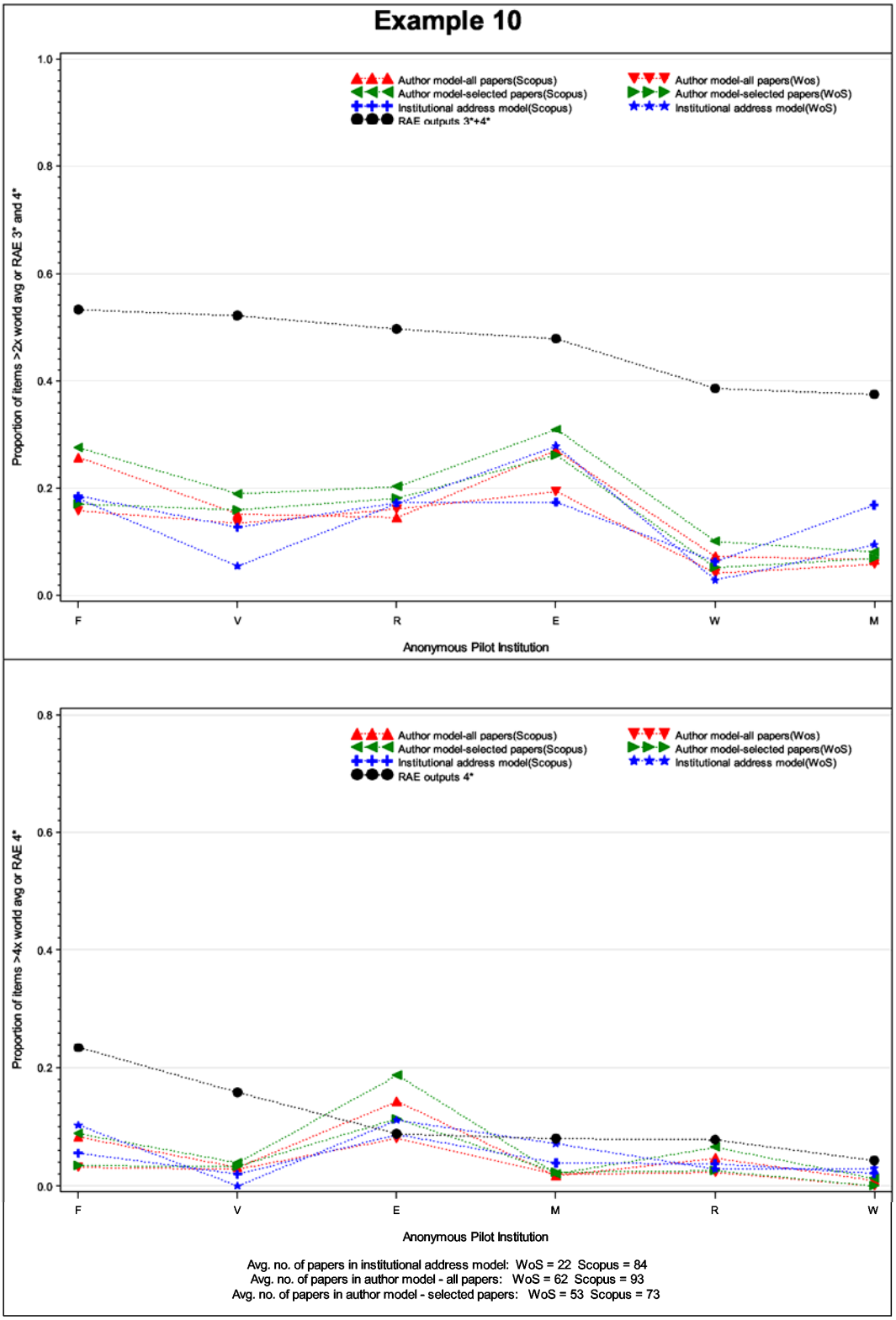

Example 5
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
S D P B E G V Q R F W A M X J
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
S P D G B E V Q F M J W R X A
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 629 Scopus = 407
Avg. no. of papers in author model - all papers: WoS = 325 Scopus = 391
Avg. no. of papers in author model - selected papers: WoS = 148 Scopus = 176


Example 6
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
P R F H V B D C A M W
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
P H F R D V B C A M W
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 228 Scopus = 369
Avg. no. of papers in author model - all papers: WoS = 411 Scopus = 419
Avg. no. of papers in author model - selected papers: WoS = 190 Scopus = 194


Example 7
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
R
F
H
C
V
E
P
B
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
F
R
C
V
H
P
B
E
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = . Scopus = 67
Avg. no. of papers in author model - all papers: WoS = 98 Scopus = 114
Avg. no. of papers in author model - selected papers: WoS = 56 Scopus = 66


Example 8
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
F Q E P R V L W B D M A
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
F E P Q V W R M L D B A
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 382 Scopus = 318
Avg. no. of papers in author model - all papers: WoS = 207 Scopus = 259
Avg. no. of papers in author model - selected papers: WoS = 129 Scopus = 151


Example 9
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
V
F
P
J
R
W
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
V
F
J
R
P
W
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 184 Scopus = 298
Avg. no. of papers in author model - all papers: WoS = 80 Scopus = 130
Avg. no. of papers in author model - selected papers: WoS = 63 Scopus = 83


Example 10
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 3*+4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >2x world avg or RAE 3* and 4*
1.0
0.8
0.6
0.4
0.2
0.0
F
V
R
E
W
M
Anonymous Pilot Institution
Author model-all papers(Scopus)
Author model-selected papers(Scopus)
Institutional address model(Scopus)
RAE outputs 4*
Author model-all papers(Wos)
Author model-selected papers(WoS)
Institutional address model(WoS)
Proportion of items >4x world avg or RAE 4*
0.8
0.6
0.4
0.2
0.0
F
V
E
M
R
W
Anonymous Pilot Institution
Avg. no. of papers in institutional address model: WoS = 22 Scopus = 84
Avg. no. of papers in author model - all papers: WoS = 62 Scopus = 93
Avg. no. of papers in author model - selected papers: WoS = 53 Scopus = 73

Discussion of outcomes by the Expert Advisory Groups

65. The Expert Advisory Groups had access to these initial pilot outcomes and additional more detailed data for their subject areas. At meetings during April and May 2009 we sought their advice on interpretation of the outcomes and the options for using bibliometrics in the REF. Below is a summary of their discussions at a series of break-out groups.

Robustness of bibliometrics pilot results

66. In many disciplines (particularly in medicine, biological and physical sciences and psychology), members reported that the 'top 6' model (which looked at the most highly cited papers only) generally produced reasonable results, but with a number of significant discrepancies. In other disciplines (particularly in the social sciences and mathematics) the results were less credible, and in some disciplines (such as health sciences, engineering and computer science) there was a more mixed picture.

67. Members generally reported that the other two models (which looked at 'all papers') did not generally produce credible results or provide sufficient differentiation.

68. A number of reasons for the variations in robustness between disciplines, and for the discrepancies in the results within a discipline, were identified:

a. Different sets of papers were looked at in the RAE and in the pilot process. Some members suggested repeating the citation analysis using only those outputs assessed in the RAE.

b. The volume of citations and the time taken to accumulate citations varies between disciplines; citation indicators are more robust in disciplines which publish and cite more rapidly.

c. The coverage of citation databases is limited in a number disciplines, particularly where non-journal outputs are common.

d. Citations measure impact on the academic community; this is only one aspect of quality, whereas the RAE results represent a rounded view of quality.

e. Citations do not provide a good measure of applied research and cannot take into account non-academic impact.

f. Other limitations of bibliometrics that could distort some of the results, such as negative citations.

g. Some of the sample sizes were small and these tend to be less stable.

h. More recent papers have had less time to accumulate citations. Even though publication year is taken into account in the analysis, the results were less robust for papers published in the more recent years.

i. Limitations with the normalisation process. In particular:

i. The categorisation of journals into fields was felt to be problematic in a number of fields (for example where diverse journals are used, such as in statistics), and for a number of journals (particularly broad journals that cover several sub-fields such as the Lancet, BMJ, Physical Review, etc).

ii. Citation rates were normalised against a worldwide 'mean' for the field; yet the distribution of citations is highly skewed.

j. Differences in the two citation databases (Web of Science and Scopus) led to some marked differences in the results. A few members noted that other databases were more widely used by their disciplines (such as arXiv and Google Scholar).

k. The way items are categorised within the databases as 'articles', 'review papers' and so on can differ from the way institutions or researchers would classify them. Some material on the databases (for example in 'trade' journals) would not be considered research.

l. The mix of sub-fields within a submission can affect citation indicators; for example a submission can be dominated by a highly cited sub-field within physics.

m. In a few cases members reported discrepancies between RAE outcomes and citation indicators, where the RAE scores appeared to reflect the prestige of the journals papers were published in, whereas the citation rates for the papers provided a different picture.

Use of bibliometrics in the REF

69. There was a strong consensus that bibliometrics would be a useful aid to expert review, but that it could not be used formulaically, due to the range of limitations and discrepancies in the data. Expert review would still be required to take these into account and to ensure the credibility of the process.

70. There was a strong consensus that bibliometrics should be applied to selected papers only. Members agreed that the 'address-based' model was undesirable for a number of reasons, not least the substantial problems in associating papers with the relevant UOA. Of the two 'author-based' models, members felt that selected papers would be more useful and informative, providing a better discrimination of quality. There was no consensus on what value information on all papers would add, and members raised concerns that assessment of all papers would disincentivise speculative research and lead to other adverse behaviours.

71. Members discussed a number of ways in which expert review panels could make use of citation data to enhance the reliability and consistency of expert review and/or to reduce panels' workloads. There was no clear consensus on a single approach a range of possible uses were identified:

a. To inform the reading of individual outputs. (Most groups supported this approach although some were concerned about using citation data in this way.)

b. As indicators for each submission as a whole, to sense check or provide a 'challenge' to the panels' scores based on reviewing the outputs.

c. To inform 'borderline' decisions.

d. To provide benchmarks against international standards and aid calibration against the quality descriptors.

e. To inform discussions about consistency between panels, or enable comparisons across disciplines.

f. To enable panels to sample and reduce the number of outputs to be reviewed in detail. Some suggested that bibliometrics could form part of a stratified random sampling procedure; however some members were sceptical about this and many generally doubted that bibliometrics would enable panels to read fewer outputs.

72. Members felt that the particular ways in which panels could make use of the data should vary as appropriate to the discipline. Also, decisions about which disciplines should use bibliometrics will need to be made depending on the coverage and robustness of the data. This varies greatly between disciplines.

73. Members discussed the type of citation information that would be useful to panels:

a. Many agreed that all panels that make use of citation data should be provided with the same types of data, but that they could use or interpret the data differently as appropriate.

b. Limitations with the normalisation method were raised. Members generally agreed that panels would want the 'raw' citation count, in addition to data that enables them to interpret this within an international context. This could be a benchmark for the field, or an indication of where the citation count falls within the worldwide distribution for the field (a centile). Panels would also be interested in a benchmark or centile for all papers submitted to the UOA.

c. There was also interest in the kinds of contextual data provided from the pilot process, relating to the sources of citation (local, national and international) and international co-authorship.

d. There was some discussion about which citation database(s) should be used and many felt that REF should not be limited to using a single database across all panels.

74. Some issues about the potential behavioural consequences of using bibliometrics were raised:

a. If bibliometrics were to be used in different ways across sub panels it could influence institutional decisions about where to submit members of staff, or their decisions about which types of outputs to submit to different panels (for example, selecting on the basis of citations for some panels, and the implication that this could favour older papers or disadvantage early career researchers).

b. Publication behaviours could be affected, for example a movement towards higher cited journals.

75. Members discussed whether the benefits of using bibliometrics would outweigh the costs. Some found this difficult to answer given limited knowledge about the costs. Nevertheless there was broad agreement that overall the benefits would outweigh the costs – assuming a selective approach. For institutions this would involve a similar level of burden to the RAE and any additional cost of using bibliometrics would be largely absorbed by internal management within institutions. For panels, some members felt that bibliometrics might involve additional work (for example in resolving differences between panel judgements and citation scores); others felt that they could be used to increase sampling and reduce panels' workloads.

Further development

76. Members suggested a number of areas for further work:

- most importantly, to develop normalisation techniques including ways of categorising journals
- assessing the accuracy of the databases
- investigating how career stage affects citations (particularly for early career researchers)
- understanding the behavioural implications of the preferred model
- investigating other citation sources (such as Google Scholar)

- understanding coverage of the databases (including, for example foreign language journals)

- extending the analysis to include conference proceedings.

Feedback from pilot institutions

77. Technopolis is currently identifying lessons learned by institutions participating in the pilot exercise. The report on the first round of consultation focussed on the data collection phase of the pilot process and has been published on HEFCE's web-site[4]. Pilot institutions found the data requirements challenging where they did not already have sophisticated research information management systems. The report states that a requirement on institutions to collect publications information for an 'authors; all papers' model would be burdensome where there were no well developed publications systems in place. Institutions also felt that even though an address-based model would be less burdensome in principle, they would still wish to verify papers associated with their institution.

78. Institutions participating in the bibliometric pilot had the opportunity to discuss their first impressions of the outcomes of the pilot process at a meeting in May 2009. Delegates attending the meeting wished to stress that this feedback is impressionistic and they have not yet had time to scrutinise the outcomes in depth or consult across their institutions. Their initial feedback was in general agreement with the Expert Advisory Groups.

79. Delegates at the meeting expressed a strong preference for an author-based, selective model for bibliometrics in the REF. This model correlates most strongly with their perception of the research excellence within their institutions and sits most comfortably with the institutions existing research information systems and reporting behaviour.

80. The 'author; all papers' model was considered to be less attractive than a selective arrangement because of the variable state of information systems across HEIs and the likelihood that for many it would be difficult to even get close to a list of all papers. Any subset would almost certainly exhibit some degree of bias and thereby produce a somewhat unpredictable assessment and ranking.

[4] Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot: Results of the Round One consultation. Report to HEFCE by Technopolis May 2009 http://www.hefce.ac.uk/pubs/rdreports/2009/rd09_09/

81. The institutional address model was the least favoured approach; delegates citied numerous problems identified through the pilot process. The most commonly reported were:

a. incorrect assignment of staff to institutions (based on addresses), thought to be widespread and rather unpredictable.

b. incorrect assignment of papers to UOAs, which can greatly affect an institution's ranking.

82. These issues were also noted by the Expert Advisory Groups and at an accountability burden workshops held for pro Vice-Chancellors of research. Technopolis is currently gathering more formal feedback from the pilot institutions, which will be reported later in 2009.

83. The full report of the pilot process and further analysis will be published in autumn 2009.

Annex A

Report on a pilot study of bibliometric indicators of research quality

Development of a bibliographic database [Draft]

The full and final report of this study by Evidence Ltd will be published in summer 2009 on the HEFCE web-site.

Executive summary

1. This report covers the work required to address the development of an initial bibliographic database to evaluate the feasibility of a Research Excellence Framework (REF) methodology. Other reports assess the workload and challenges faced by the contributing universities and colleges, but acknowledgment is made here of the extensive support and enthusiasm extended by the staff in those higher education institutions (HEIs).

2. The REF is intended to make more extensive use of quantitative research performance indicators than the Research Assessment Exercise (RAE). The metrics discussed in reference to the REF are restricted to 'bibliometrics' which are the indicators created by an analysis of research journal articles and their subsequent citations. The collation and normalisation of citation data for the bibliographic database and the evaluation of variant bibliometric analyses will be described in later reports.

3. The census period of the exercise is 2001-2007. Data were supplied by a group of 22 pilot HEIs. The pilot HEIs were selected by HEFCE to cover a wide range of research management systems and processes. Subject areas were captured within 35 Units of Assessment (UOAs) selected by HEFCE because they had 40% or greater coverage of RAE-submitted outputs in principal commercial data sources (either Thomson Reuters' Web of Science or Elsevier's Scopus). Not all pilot HEIs elected to supply data for all UOAs.

Preparation and specification

4. The project was launched in June 2008. It was expected that the development of the bibliographic database would take up to six months. This was a challenging timetable for both the pilot HEIs involved in pilot work and for the contractors. Data collection took place over the summer, when many HEI staff were on leave. It was therefore agreed that REF pilot data specification should match the RAE2008 data collection as closely as possible. This would reveal the challenge of implementing a national exercise and provide important information about the current readiness of data management systems in the higher education (HE) research base.

5. The pilot work was designed to compare two variant approaches: a low-burden *address-based* model, with data collated by address and linked to subjects via journal categories; and a more onerous *author-based* model in which outputs are linked to subjects via author-staff disambiguation. To ensure that sufficient data would be available

for each pilot HEI, the project made use of a *presumptive* dataset for each institution supplied by Evidence Ltd from prior work to collate institutional article records. The presumptive data would form the entire database required for the address-model variant. *Actual* data are those article records already collected by institutions from their staff and therefore explicitly validated as part of the publication record submitted for the REF pilot exercise.

6. An outline specification for pilot HEI data was circulated in July 2008. For staff data (table 1), the RAE specification and definitions were used as a starting point. Additional (non-RAE) data were requested to enable a determination of the effects of varying the staff selection (and hence the collated output data). For output data (table 2), the RAE specifications were again used. Some additional fields were requested to help in matching outputs to citation databases. A comprehensive list of outputs (in addition to journal articles) was sought to provide a context for benchmarking indicators and tracking publication behaviour.

7. A third, necessary and central part, of the data requirement for the REF pilot project was the association of output data with named staff for the author-based model (table 3). Institutions were asked to provide a pair-wise association between staff and publication IDs. To ensure that sufficient links would be available for each pilot HEI, the project made use of the Symplectic Publications system to enable a comprehensive search for additional links.

8. Six pilot HEIs indicated that their total REF submission would not be more than they submitted to the RAE, even if time were available. Several pilot HEIs indicated that they expected a roughly four- to five-fold additional data submission compared to RAE2008. In every case, pilot HEI estimates were less than Evidence's 'presumptive' estimate, on average by about 30%. In the outcome, most HEIs were able to extend their submission beyond solely RAE data.

Receipt and processing

9. Data development and collection was supported by regular contact between the pilot HEIs, the contractors and HEFCE. Because of the very compressed timetable set for HEFCE, the contractors agreed to accept pilot HEI data that fell outside the published specification and to clean this centrally. Additionally, some HEIs could not have submitted data to the specification required. This subsequently had serious consequences for resource capacity in later stages, but also provided valuable insight into the quality of institutional data systems.

10. Data were submitted via the HEFCE extranet in three tables in agreed formats (Excel, Access or xml). There were multiple, serial and overlapping, rather than single data submissions from some pilot HEIs. The multiple data submissions included the following total records:

- 87,641 staff and other researcher records in versions of Table 1, of which 44,136 cleaned and deduplicated records were passed to the Symplectic Publications system;

- 678,077 article and other output records in versions of Table 2, of which 328,136 cleaned and deduplicated article records were passed to the Symplectic Publications system;

- 872,132 links between staff and authors in versions of Table 3, of which 433,447 properly indexed and deduplicated links were passed to the Symplectic Publications system.

11. This compares with the roughly 50,000 staff records and 200,000 output records handled within the RAE system (based on estimates from 2001 data; the indication is that 2008 was a somewhat but not significantly larger submission).

12. It became evident that the limit to staff data that could be provided by many pilot HEIs corresponded to the staff list submitted for RAE 2008. Output data were also limited. Few institutions have in place a system for the regular submission of standard and comprehensive publication data or content by academic departments to any central database or repository.

13. Because there was a greater level of central data processing, cleaning and management than had originally been planned, the project became increasingly engaged with data management. The greatest impact on the project was in the speed of development of the core database. A second area of delay was in the processing of additional records from the presumptive data. Because of the underlying deficits on data quality, the task of linking authors with staff was also more onerous and complex than intended. The combined effect of delayed data handover between Evidence and Symplectic and the poor relative quality of the data at that point exacerbated the delay in offering enhanced data to pilot HEIs for verification.

14. Whereas it was originally intended that pilot HEIs should be offered supplementary data records in September 2008 and the opportunity to verify additional staff-author links through October and November, the outcome of data issues delayed this into a compressed period during December and January 2009. Some further data development continued into February 2009.

15. Important lessons have, hopefully, emerged. The central one is that most institutions will require a very clear and extended implementation pathway before the REF could be introduced on a national scale.

Management of staff data

16. The process of building the staff table was an iterative process of importing, reviewing, returning to pilot HEIs and amending records.

17. The quality of the data submissions was affected by two things. First, there was an enforced haste to supply information which was then submitted in a form that, had more

time been available, the pilot HEIs themselves would have corrected. For example, fields were confused or mislabelled. Second, there were fundamental deficits in pilot HEI systems that meant that data could not be readily retrieved in a particular format. Beyond data quality there is an issue of data deficits. There was also a variable outcome because of differences in the approach that pilot HEIs took to supplying data.

18. The most critical piece of data that was widely absent was any information about the prior employment record of staff currently employed at a pilot HEI. While such information is held by institutions it has not previously been a part of normal electronic database records. For the REF, the significance of this is in identifying and examining output data prior to current employment. If this is to be a standard part of analysis then there will need to be a systematic and systemic change in the way this is captured.

19. Initially, all staff data were included. It was later decided to restrict subsequent analysis to RAE-eligible staff only. After initial data review, it was determined that staff who were ineligible for the RAE appeared to have relatively few publications that were not co-authored with an RAE eligible member of staff. Evidence transferred 44,136 cleaned staff records to the Symplectic Publications system.

Management of output data

20. The integration process, to create a single bibliographic database for matching and processing, took longer than anticipated. This was because of data quality issues (both missing and erroneous data) and because more updating was required than had been expected. Twelve pilot HEIs serially submitted output data as many as six times. This was complicated by supplementary datasets, complete updates and partial replacements, not always in the same format as original data from the same institution.

21. The data request to pilot HEIs asked them to submit not just journal articles but also non-journal outputs to throw light on the broader publication context.

- About 250,000 records of 328,136 output records supplied by pilot HEIs appeared to be from research journals;
- On balance, the data suggests that articles and reviews probably account for 65-70% of significant outputs in the subject areas under examination;
- Of the residual records about 25,000 appeared to be books or chapters and 40,000 to be conference contributions.

22. Conference proceedings, which will soon be subject to much improved evaluation, account for 10-15% of significant outputs in the subject areas under examination.

23. This balance would allow the REF to explore academic impact for upwards of 80% of the potentially available material in these subject areas.

24. Cleaning regimes were applied to selected fields in the outputs database, concentrating on those essential for a satisfactory match to be made to commercial citation databases. This prioritised fields such as journal titles, volume and page numbers and unique identifiers including DOIs (digital object identifiers) and Thomson UTs (unique tags). The data were combined with the staff data and used as the subsequent output dataset for verification in the Symplectic Publications system.

Reconciliation of journal outputs to the Web of Science

25. Evidence was responsible for reconciling the article records supplied by pilot HEIs to the article records in the Thomson Reuters Web of Science commercial database. This was verified by checks through the Symplectic Publications system. Reconciliation of pilot HEI data to Elsevier's Scopus database, an alternative source of commercial supply, was carried out by HEFCE. This, and comparison between the two, is reported in the main report.

26. For apparent article and review records, DOI data were available for 49% of outputs and gave an overall matching success of 28% or 70,147 outputs. Journal and article title data were available for 85% of outputs and gave an overall matching success of 62% or 155,986 outputs. Journal title, volume and pagination data were available for 78% of outputs and gave an overall matching success of 61% or 152,440 outputs.

Issues arising from data gathering and processing

27. Many issues that arose during the REF pilot exercise are much less likely to arise during a full-scale national implementation. Nonetheless, the problems that did occur will have to be taken into account. Many of them reflect the fact that most institutions are not currently able readily to supply the data that would be required. This constraint is not limited to any particular group nor to any particular type of data.

28. Due to the constrained timetable of the pilot project, three stages were running concurrently. Running them consecutively would have increased efficiency and made it less onerous to track and trace data and to modify the design, but this would have elongated the timetable by months.

29. Pilot HEIs were permitted to make successive submissions of data. This was intended to assist pilot HEIs to keep to a tight timetable and allow us flexibility to respond to different levels of data availability between pilot HEIs. With hindsight, the cost is stark, because multiple submissions increased the central workload disproportionately. The benefits of allowing multiple submissions, however, were iterative development of data processing and cleaning techniques; flexibility to vary the requirements according to individual pilot HEIs status; and an opportunity to brief pilot HEI staff on working aspects of the relevant data.

Extension of output data and disambiguation of author and staff names

30. Symplectic, a subcontractor to Evidence, focussed its work around two major tasks: first, pilot HEI publication data were reconciled to Thomson Reuters Web of

Science data in an automated fashion to produce a single, inclusive pool of publications; second, records were then matched with the academic staff lists.

31. Data records were classified into three categories: output with a Thomson Reuters record alone; output with an HEI record matched to a Thomson Reuters record; output with an HEI record alone. To maximise the linkage between staff and the publications data, an automated mechanism was needed to suggest potential links. Automated methodology suffers from two major drawbacks: first, the risk of identifying false positive matches suggesting authors have written more papers than is the case; second, the risk of missing matches, thus failing to suggest the author of papers in the dataset.

32. A simple algorithm was devised to match outputs with their authors. This relied on matching institutionally supplied names and variations with 'searchable data' restricted to the portion of the article database associated with each staff member's home institution. The links supplied by each pilot HEI were then applied over these data to form a firm link of "Approved" articles.

33. Any suggested link from the automated mechanism was a "pending" link reviewed by pilot HEIs through a customised web interface or a downloadable spreadsheet. It became clear that a sampling strategy would be required where strategic approval methodology could be applied in order to ensure maintenance of the data quality and to understand weak points. Several methodologies were applied to institutional data in order to help institutions with larger amounts of "pending".

Creation of database

34. The outputs of the Symplectic Publications system were recreated forms of the key REF data tables containing deduplicated staff information (Table 1), extended, cleaned and deduplicated publication records (Table 2) and more comprehensive links between staff and authors (Table 3). The data records and links processed by Symplectic Publications and accepted by the pilot HEIs were resubmitted to the secure server.

35. The final steps in the creation of the bibliographic database required for the REF pilot project were the association with the validated publication records of their relevant citations data and the normalisation of the citations data to enable comparative analyses. A later report will describe the development of the combined publication and citation database and the decisions made regarding normalisation.

Annex B

Principles of normalisation

This is a working paper prepared by Evidence Ltd for the Expert Advisory Groups in May 2009.

Introduction

1. The three key attributes of any piece of research activity data are the time, subject and location associated with the activity.

2. 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'. It is therefore necessary, for a well-founded analysis, to standardise bibliometric data and this is normally done by reference to a common benchmark. Bringing a diverse data-set to a common framework which corrects for 'anomalies' is referred to as normalisation or 'rebasing' the citation count.

3. Examples of the sorts of factors affecting bibliometric data are as follows:

   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 have been 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 the year, such as between a paper published in December and one published in January, eleven 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.

4. Classical approaches to the normalisation of citation counts for a journal article take into account the year of publication, document type and the field to which the journal is assigned.

Reference benchmark - time

5. 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.

6. 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.

7. 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.

8. 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.

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

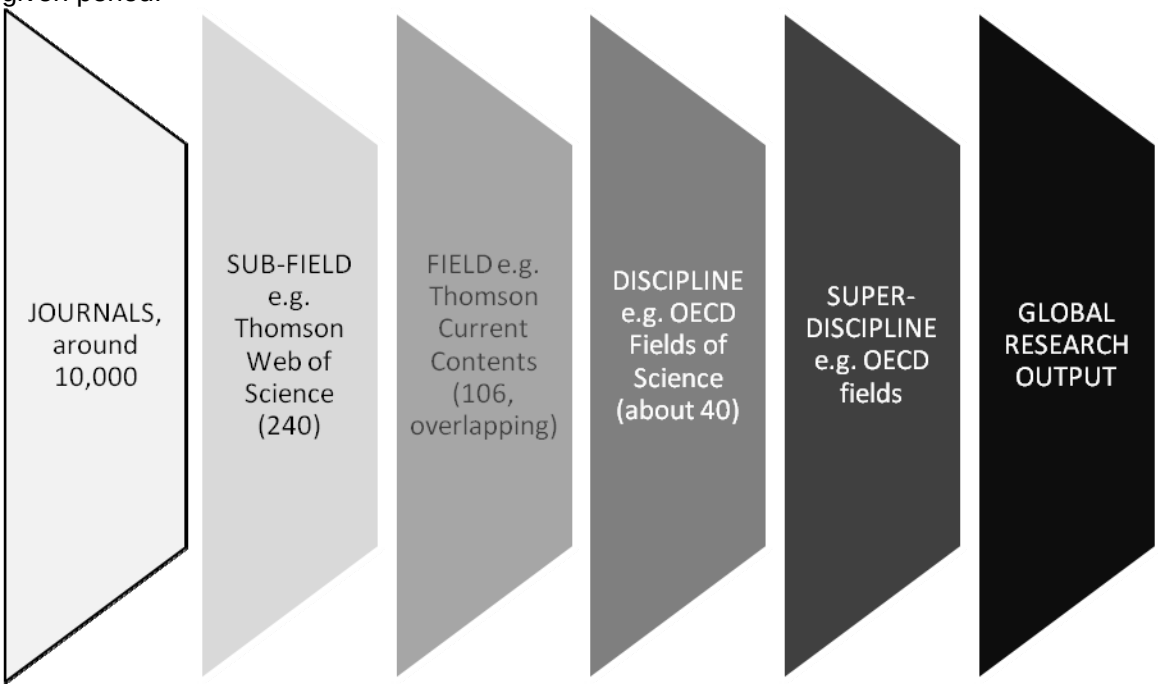
10. Typically, the cut-off date for an annual database compilation will be some time after 1 January, in order to fit in with other work schedules. This can vary: it might be in the first week after New Year or not until a week or so later. Each additional week would add an additional 2% volume to the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

11. 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.

Reference benchmark – subject

12. 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 the global publication output for a given period.



13. 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?

14. 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 systemic purpose, however, we have to work at the level of journals so that, with the exception of a defined set of multi-disciplinary 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.

15. 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.

16. 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 cites/article. This is now a reference benchmark against which any individual article can be compared.

17. 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.

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

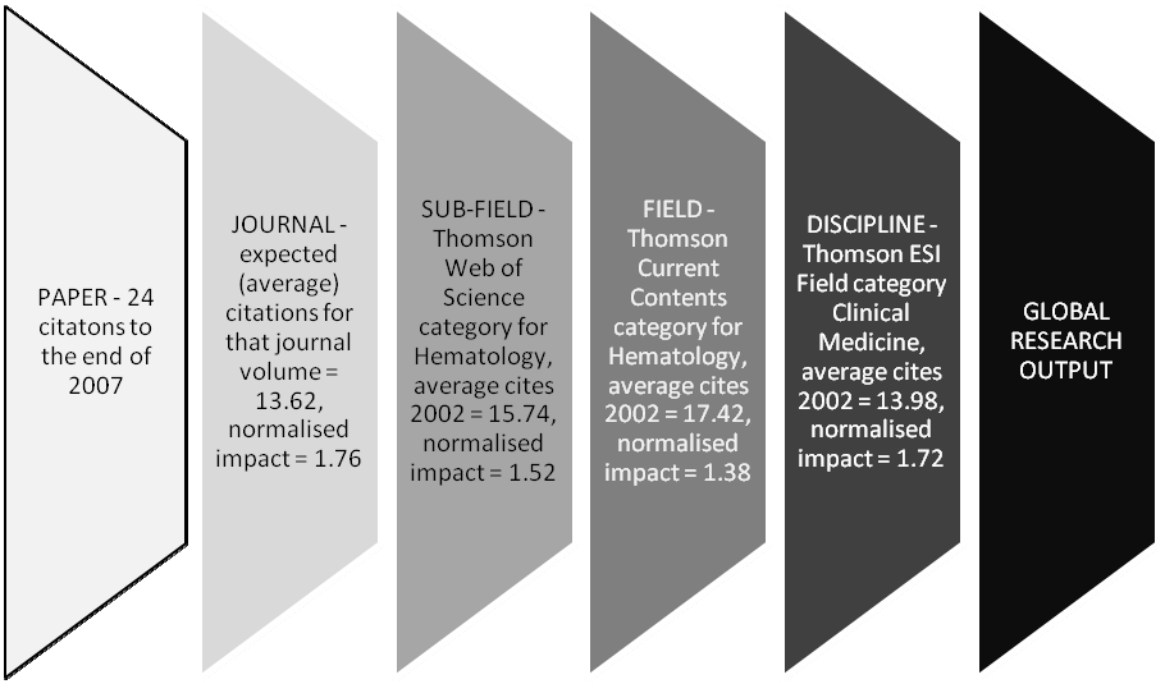
19. Categorisation has to be fit for 'purpose'. The categorical structure used for international comparisons across the breadth of the research base would usually be much coarser than that for an institutional management study. Surveys of researchers suggest that they tend to see themselves as part of 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.

20. 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.

21. Every researcher knows, however, that some journals are perceived 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.

22. 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.

23. 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.



24. So, a 2002 paper with 24 citations – an above average count for its journal in that year - may have a normalised impact between 1.38 (at field level) and 1.72 (in its broader discipline). 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.

25. There is no simple answer to the question 'what is the right level for normalisation?' Common sense suggests that it 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 between these extremes there are important nuances about the relative significance of fields and sub-fields.

26. The answer is not 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 judgments.

27. 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.

<table>
<tr><td colspan="2">Discipline X – average of 15 citations per paper for year y</td></tr>
<tr><td>Field A, within X
Average of 12 citations per paper for year y</td><td>Field B, within X
Average of 20 citations per paper for year y</td></tr>
<tr><td>An article with 20 citations and linked to A has a normalised impact of
20/12 = 1.67 at field level.</td><td>An article with 20 citations and linked to B has a normalised impact of
20/20 = 1.0 at field level.</td></tr>
<tr><td colspan="2">An article with 20 citations from A or B has a normalised impact of
20/15 = 1.33 at discipline level.</td></tr>
</table>

28. 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.

Reference benchmark – location

29. For the purposes of the present exercise, the appropriate benchmark is global activity. We are not interested in whether an article is well-cited only in a UK context but whether it is good in international terms.

30. 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.

Proposed work plan

31. The work-plan is first to normalise, clean, correct, extend and develop the databases provided by the pilot institutions. The article records on the institutional data form the core platform on which the work is based, enhanced by the additional records that Evidence can associate with the institutions and the additional author-staff links made by Symplectic.

32. Second, for each article record on the database, we will assemble the relevant citation data. We will collate citations to date by year and we will seek accurately to identify self-citations: those citations that are made from a later paper by the same author.

33. We will then normalise the observed citation counts against average 'expected' citation counts for larger sets of articles.

34. On normalisation by year, the short-run practical approach is to make use of database year, which provides a ready reference point. In the long-term, HEFCE may wish to make use of its access to global data to develop global reference sets collated by journal volume.

35. On normalisation by subject, it would be informative to create a series of analyses for each different major subject area to explore the effect of normalising article citation counts at different categorical levels within those areas. The outcomes may well differ by area.

36. We propose to compare the effects of using the journal, sub-field (WoS), field (Current Contents) and discipline (Essential Science Indicators) levels of aggregation as categories for normalisation.

37. We do not expect the journal level of normalisation to be one which researchers would wish to adopt, because of variations in journal quality, but it forms part of the spectrum of information which could inform quality assessment.

Analysis and reporting

38. 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.

39. Thus, the citation counts for individual articles might be normalised at the level of 'field' (106 categories here, e.g. Optics & Acoustics). That would be a finer-grained categorisation than, for example, the former RAE UOAs. Averages reporting, however, might be taken at the level of disciplines, broadly corresponding to UOAs (e.g. Physics), or at a coarser level, corresponding more to Schools (e.g. Physical Sciences) or Faculties.