

Annex I: Development of Indicators

A report by Evidence

Reviewing variants to the indicators

1. As the REF methodology develops, it is likely that some changes to the initial set of preferred indicators will be appropriate. We can anticipate some of the issues that are likely to arise and provide background on these.

- With many subject categories and a wide diversity of institutions it is almost certain that some of the submissions made under the REF will be relatively small. But at what point does 'small' become 'too small to provide sound indicators'?
- The UK's assessment cycle has historically been exclusive, so that the census periods for successive assessments did not overlap. The length of the assessment cycle needs to take into account the time window required to build sufficient citation data for a timely but reliable estimate of impact. How short and recent can the citation window be?
- Conventional indicators have referred the performance of a paper, or the average for a group of papers, to a multiple of world average. An alternative is to look at the location of papers within the spread of a distribution, using the profile of centiles, deciles and quartiles. What are the characteristics of such indicators?

Threshold sample size

2. How small a sample of papers is acceptable for the evaluation of research performance via bibliometric indicators?

3. While the citation count for individual papers can be indexed against appropriate standards for year and field, it is generally agreed that the normalised impact measures thus generated should not be used in isolation. An individual paper may have a spuriously low or high citation count for reasons that cannot be automatically detected and thus have a misleading citation impact. However, there is evidently a general correlation between the average citation impact of large groups of papers and judgements arrived at through e.g. peer review. Larger samples are therefore likely to be more informative, by absorbing undue influence from exceptional values.

4. The challenge is to determine the threshold sample size below which the average impact for a collection of papers (i.e., for the REF, the total set of papers submitted by an HEI within a subject area) becomes susceptible to the undue effects of a single (or small number of) outlier(s).

5. For collections that fall below the threshold, it would be necessary to apply more careful or intensive peer review or other manual evaluation than for larger collections.

Australia and the ERA

6. Threshold sample size has been a critical consideration for Australia in the implementation of the Excellence for Research in Australia (ERA) methodology.

7. It is apparent that the ERA system is likely to be faced with a high frequency of relatively small collections of publications in a given subject category. Australia has a

small number of research intensive institutions (the Group of Eight), a number of other institutions with strong research performance in specialist areas and a range of other institutions. It also has an evaluation system based on 157 research categories determined by a four-digit hierarchical coding system: the Australia-New Zealand Standard Research Classification (ANZSRC).

8. The ERA Indicator Development Group (IDG) reviewed the problem and discussed possible routes to modelling a solution. While the expert view was that samples as large as 100 papers would be required to provide consistently reliable results, an administrative view was that samples as small as 20 papers might be frequently encountered in evaluation. If all such small samples were beyond the reach of indicator evaluation then this would challenge the utility of the ERA approach.

9. No clear conclusion was reached in the IDG's report. However, the first ERA pilot project, which focuses on physical and earth sciences, has shown that a minimum sample size of 50 outputs in total for a research group over the census period can provide a sufficient volume of material to generate indices that expert panels agree are a reasonable representation of relative research performance.

Modelling

10. The Australian IDG recognised that modelling an outcome for small samples was problematic.

11. A simple approach could be to take a very large sample that reflected a real population of publication-active researchers (such as a large School of Biological Sciences). The citation impact for the total population would be analysed. Random samples of progressively smaller numbers of papers could then be taken and the degree of variance of the sample citation impact from the total population citation impact could be evaluated.

12. Real publication submissions for evaluation are not random samples. First, not all researchers will necessarily be evaluated as the employing institution may make a staff selection. Second, the staff complement is in any case likely to be structured with more and less experienced researchers, not a uniform staff group. Third, not all of the publications by selected staff will be evaluated as each researcher may be able to make a selection from among their 'best' outputs. The conclusion was that a random sample is frequently likely to depart from similarity to a real-world structured sample.

13. We have explored possible alternative routes to creating a sensible non-random sample but such alternatives are beset by the problem of applying an appropriate structure to the data. A significant level of additional information about the authors is in fact required.

14. Our view is that random sampling is the only available option but that such random sampling will produce outcomes that are innately likely to have greater variance than would be found in true sub-sets of the parent population. Thus, if random samples as small as 50 were found to give reasonably low variance from parent population then this would veer on the side of safety in setting this as a threshold minimum.

Conclusion

15. HEFCE will require further modelling to satisfy itself as to the preferred threshold sample size, which may vary from one broad discipline area to another. However, such modelling is problematic and the outcomes will need to be reviewed by an expert group rather than treated as a technical definition. The outcomes of the Australian ERA pilot suggest that sample sizes of 50 are likely generally to be adequate, but this estimate may need to be refined in the light of further trials and in the UK environment.

Time windows

16. A publication accumulates citation counts when it is referred to by more recent publications and citations accumulate at different rates in different subjects. The timing of citation accumulation is discussed fully in the earlier section on normalisation. In the present section we explore the advantages and disadvantages of different lengths of time window for citation analysis.

17. Early citations have been found to be a general indicator of subsequent citations but there are limitations to using early citation data. One drawback is that the effect of whether a paper is published early or late in the calendar year is undiluted. If we consider a two-year window (citations to date for papers published in the last two years), a paper published in January has had twice as long to get cited as a paper published in December of the same year. Under current methodology, both papers' citation counts would be normalised against the same annual baseline. Another factor to consider is that citation patterns are subject to annual fluctuations, particularly in the early years.

18. To overcome these potential limitations, it is common practice to smooth out citation counts by considering time windows of typically five years or more. It is preferable to use citation counts beyond at least the first year or two after publication, particularly if citation data are used to inform evaluation.

19. There is a trade-off between timeliness and picking an appropriately long window: if the evaluation is to be used to drive funding, we would wish to minimise the unavoidable time gap between the evaluated research(ers) and the funded research(ers), so we would like the time window to be as short as possible; but if the time window is too short, the reliability of the bibliometric indicators will decrease.

20. In the remainder of this section we assess whether a four-year time window is suitable for the purposes of REF pilot bibliometric analyses.

Methodology for comparing time windows

21. Our methodology was based on comparisons between two-year, four-year and six-year windows. We assumed that a six-year window gives sufficient citation data for evaluation purposes, then we tested whether a four-year window would be long enough to give significantly similar results. We asked two questions:

- Does a four-year window provide data that are sufficiently close to the six-year window to mean that extension to six-years would be unnecessary?
- In terms of similarity to a six-year window does a two-year window provide results that are significantly different from the four-year window?

22. For this test, we used papers from the address-based model, grouped by UOA and HEI. For the papers from each HEI-UOA combination, we calculated the average rebased impact (RBI) for citations to end of 2007 to papers published in each of three time windows:

Two years	2005-2006
Four years	2003-2006
Six years	2001-2006.

23. We focused on four UOAs:

- 02 Cancer studies
- 14 Biological sciences
- 19 Physics
- 28 Mechanical, aeronautical and manufacturing engineering

24. First, we compared the four-year RBI with the six-year RBI, for each HEI within each UOA.

25. Second, for each HEI we calculated the absolute difference between:

- the two-year RBI and the six-year RBI
- the four-year RBI and the six-year RBI

and made a pair-wise comparison between these values within the UOA.

26. What we are interested in is the proximity of the two- and four-year values to the six-year value, not whether they are above or below it. For that reason, we consider the absolute difference and ignore the direction.

Results of comparing time windows

27. The average RBI values are illustrated graphically in Figure 11.a-d. There is one graph per UOA. The vertical left-hand axis displays RBI 0-3 times world average. The HEI names have been anonymised and are represented by the upper case letters along the horizontal axis. Within each UOA, the HEIs have been arranged in descending order of the average RBI values from the six-year window so that the line slopes from top left to bottom right in each chart. There are three lines: solid, dashed and dotted which represent the six-year, four-year and two-year average RBI values respectively.

Figure I1.a. Citation impact by pilot HEI for three citation-accumulation time-windows for data in Cancer studies (UOA 02)

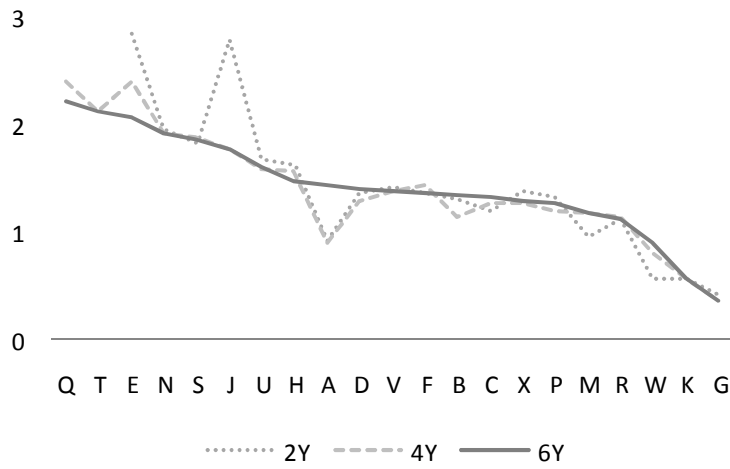


Figure I1.b. Citation impact by pilot HEI for three citation-accumulation time-windows for data in Biological sciences (UOA 14)

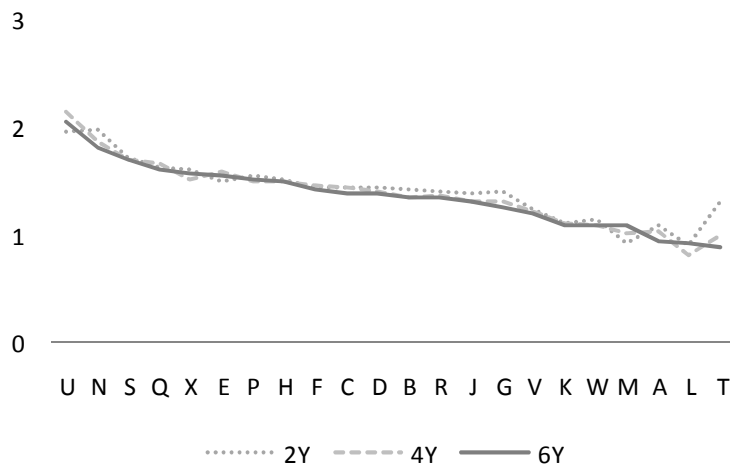


Figure I1.c Citation impact by pilot HEI for three citation-accumulation time-windows for data in Physics (UOA 19)

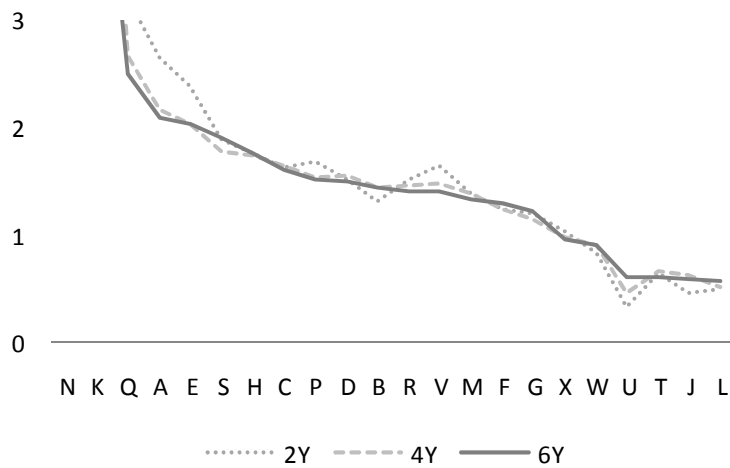
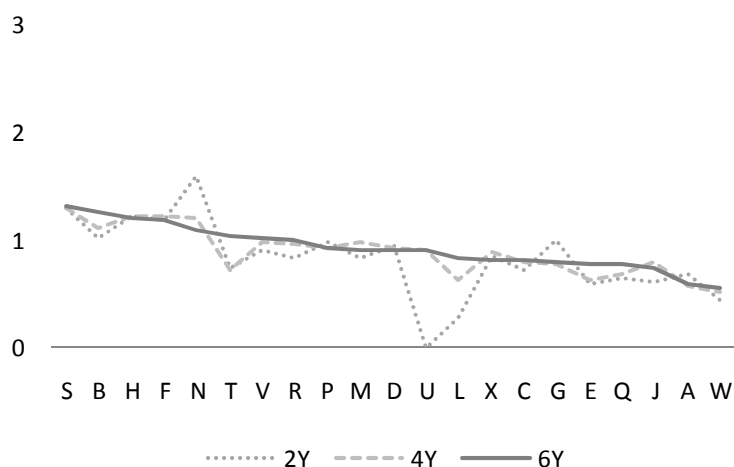


Figure I1.d Citation impact by pilot HEI for three citation-accumulation time-windows for data in Mechanical engineering (UOA 28)



28. To allow comparison between UOAs, all four graphs have been reproduced to the same RBI scale. This means that for UOA 19, the outlying high RBI values have been omitted from the presentation. However, they were included in the analysis.

29. We can use the graphs and statistical analysis to address the key questions:

- Does a four-year window provide data that are close to the six-year window?
- Does a two-year window provide results that are significantly different from the four-year window?

30. Visual inspection of the graphs in Figure L1 suggests that the answer to both questions is 'yes' for each of the UOAs.

31. We tested whether the differences and similarities between the windows are significant by using the Wilcoxon paired rank test. The Wilcoxon test for paired data ranks the absolute values of the differences between the paired data in sample 1 and sample 2 and calculates a statistic on the number of negative and positive differences.

32. Table I1 summarises the probabilities within each UOA of a significant difference in (a) the impact for a four-year (4Y) and a six-year (6Y) window and (b) the disparity compared to a six-year window between a two-year (2Y) and a four-year window. Values less than 0.05 indicate a statistically significant difference whereas those closer to 1.0 indicate very little difference.

Table I1. Summary statistical outcomes for comparisons of citation impact for three different time-windows for citation accumulation

UOA	Is 4Y significantly different to 6Y?	Is 4Y significantly closer than 2Y to 6Y?
02 Cancer studies	0.63	0.06
14 Biological sciences	0.12	0.01
19 Physics	0.22	0.01
28 Mechanical engineering	0.26	0.00

33. Despite the apparently erratic outcomes generated in Figure I1.a, due to small samples for some pilot HEIs, the results for the four-year window are not statistically significantly different from those for the six-year window. The probability that these differences could be generated by chance (Table I1) are large.

34. By contrast, the improvement in similarity to a six-year window by moving from a two to a four-year window (i.e. how much the shift from two years to four years reduces the difference compared with six years) is significant for all the selected UOAs. The probability that these differences could be generated by chance are low. The significance varies across subjects: it is highly significant for Mechanical engineering (less than 0.001, or one in a thousand) and least (indeed, barely) significant for Cancer studies, with Biological sciences and Physics in between.

Conclusion

35. A four-year publication window for citation analysis provides a measurable benefit over a two-year window in producing results that are not significantly different from those that would be obtained by using a six-year window.

36. The rate of citation accumulation is most rapid in bio-medical subjects so we would expect there to be less benefit in such subjects in moving to a four- rather than two-year window. By contrast, we know that citation counts accumulate more slowly and plateau at lower numbers in engineering so the first two years are less informative while the next two years bring a significant improvement.

37. These expectations are borne out in the analyses here. The average impact for a four-year window is more similar than a two-year window to that of a six-year window, but the improvement is barely significant in Cancer studies whereas it is very significant in Mechanical engineering. There is essentially no difference in outcome for a four- or six year window in Cancer studies, whereas the difference is more evident, albeit not statistically significant, in other disciplines. This information about the relationship between the subjects and the outcomes is important because it reinforces the coherence of the individual analyses.

Centiles as indicators

38. The REF pilot bibliometric database held citation counts and rebased (normalised) impact values for each paper. These paper-level values were filtered and aggregated in a variety of ways to produce bibliometric indicators at the level of pilot HEI and UOA. This section deals with alternative methods of indexing the volume of frequently cited papers.

39. By definition, a paper with an RBI of less than one ($RBI < 1.0$) has a citation count lower than world average whereas a paper with $RBI > 1.0$ must have been cited more times than world average for a paper in its field and year. One way of identifying more frequently cited papers is to apply a threshold value to the RBI and then to express the number of papers which exceed that threshold as a proportion of the set of papers being analysed. Typically, these thresholds can be set at rising multiples (e.g. twice, four and eight times) of world average. Doubling is used to maintain constant relativity across these data.

40. When aggregating the outcomes of the REF pilot model, HEFCE used the 'percentage of papers above 2x world average citation rate' and the 'percentage of papers above 4x world average citation rate' as indicators of the volume of frequently and very frequently cited papers. These indicators, for each UOA and pilot HEI, were presented to participating institutions and to the Expert Advisory Groups.

41. There is a drawback to using such thresholds. There are few or no papers above the threshold for some subject areas. This occurs where the citation counts for papers within a field are clustered close to the global average. This varies between disciplines but it means that a particular threshold may have more relevance in one discipline than in another.

42. Another way to identify frequently cited papers is to rank all the papers in a field, by citation count, and index the proportion of a sample that falls in, say, the top 10%, of world output. The advantage of centiles over absolute thresholds is that they do not rely on the field distribution.

43. Papers in the top 10% of a distribution are referred to as being above the 90th percentile, or in the top decile. Papers in the top 25% are referred to as being above the 75th percentile, or in the top quartile.

Methodology and data sources

44. We calculated indicators based on centiles for the REF pilot bibliometric data set and compared the percentile indicators with the '% above x times world average' method of indexing frequently cited papers.

45. The top decile for our global data set contains the most frequently cited 10% of the world's papers. A unit with more than 10% of its papers in that decile is producing more 'frequently cited' papers than is typical of the world as a whole.

46. The data set for this analysis is the set of papers included in the bibliometric analyses of REF pilot data (address-based model). Where a journal is assigned to more than one category, more than one percentile value is calculated. For the purpose of this analysis we have used the average percentile value for multi-category articles.

47. For a particular unit (e.g. research group within pilot HEI) we expressed the number of papers above the threshold percentile (i.e. the number with rebased impact above the rebased impact value of the paper at the 75th or 90th global centile) as a percentage of all the papers produced by that unit. This indicates whether the group is producing more or fewer frequently cited papers than world average.

48. For example, let the RBI at the 90th centile be 2.41. A research group has published 126 papers in our data set. Of those, 15 have an RBI > 2.41. Thus $15/126 = 11.9\%$ of their papers are in the top decile. Thus, this group has more relatively highly cited papers than is typical for the world as a whole.

Percentile analysis

49. In the total pilot HEI REF bibliometric data set, there are 22 pilot HEIs which each cover a diversity of but not all UOAs. The analysis generated indicators for 650 UOA-HEI combinations. Some of these were very sparsely populated, with only a handful of papers, so centile analysis was uninformative since it was affected by small shifts. For the purposes of comparing indicators, we applied a volume threshold to exclude units with fewer than 100 papers.

50. To illustrate the outcome of these analyses graphically, we focused on the five well-populated UOAs which have been used as exemplars in other analyses. These are 02 Cancer studies; 14 Biological sciences; 17 Earth and environmental sciences; 19 Physics; and 28 Mechanical, aeronautical and manufacturing engineering.

51. There are four variables in each chart (Figure I2.a-e), which are the percentage of each HEI's papers that were in each of four categories of frequently cited papers:

- Top quartile.
- Above 2x world average citation impact.
- Top decile.
- Above 4x world average citation impact.

52. The horizontal grid lines on the charts are at 10% of papers and 25% of papers for comparison with top decile and top quartile curves respectively. The institutions are sorted in order of increasing percentage of papers in the upper quartile by citation impact. Each UOA had data for between nine and 20 HEIs, and there were 72 data points across the five UOAs.

53. Where the curve for the institutional percent in the top quartile exceeds the line for the global quartile marker, then the HEI has more than 25% of its papers in the global top quartile. Similarly for the curve for the top decile and the line for the global decile marker.

Figure I2.a Comparison of citation impact indicators using centiles and multiples of world average, for Cancer studies (UOA 02)

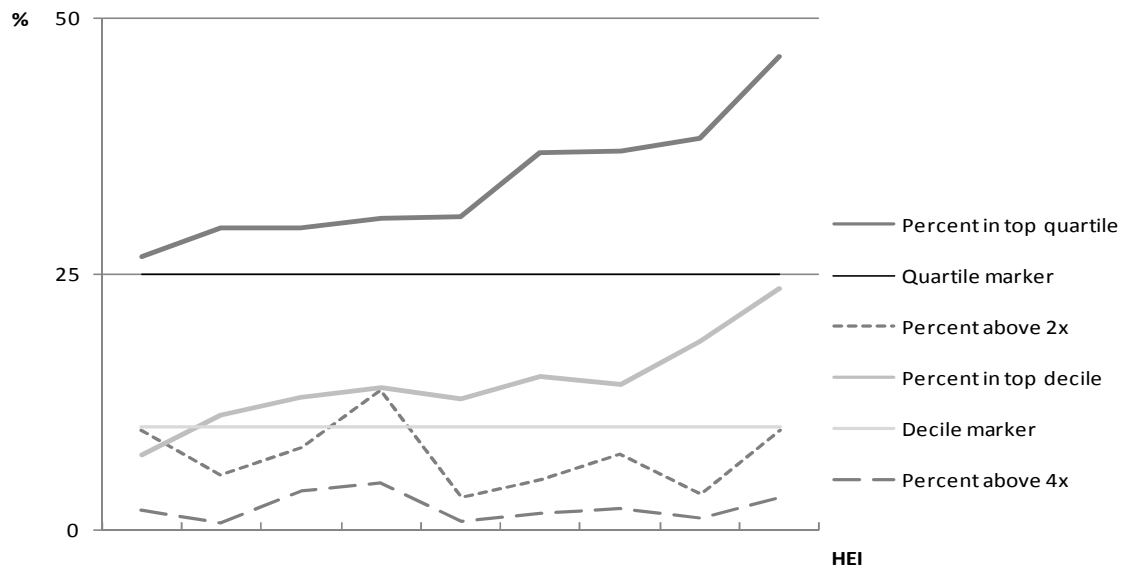


Figure I2.b Comparison of citation impact indicators using centiles and multiples of world average, for Biological sciences (UOA 14)

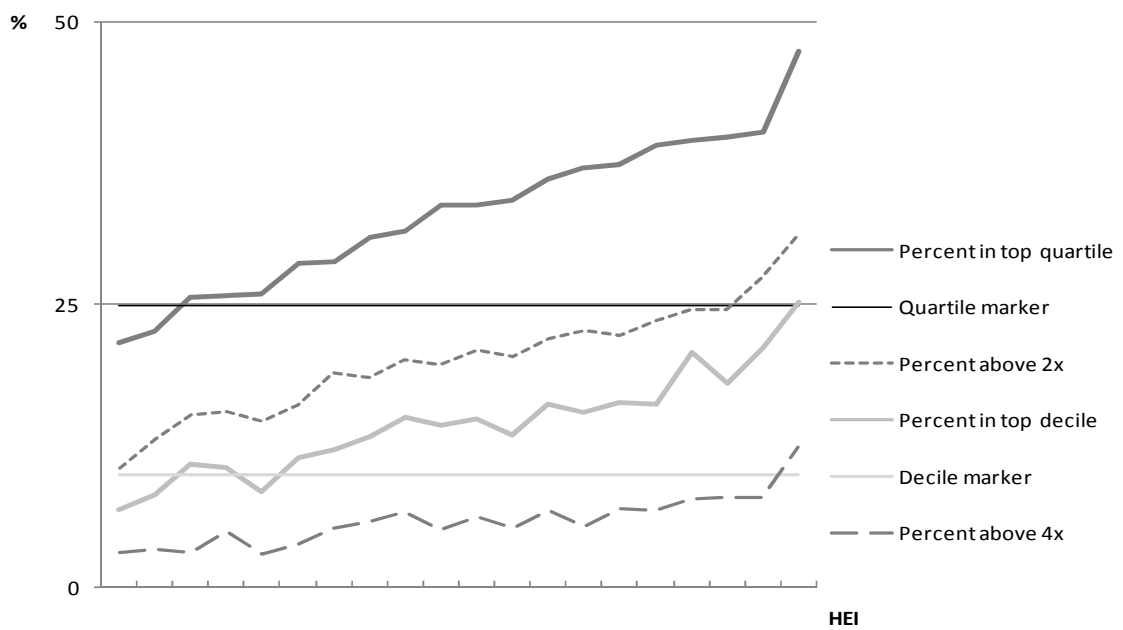


Figure I2.c Comparison of citation impact indicators using centiles and multiples of world average, for Earth and environmental sciences (UOA 17)

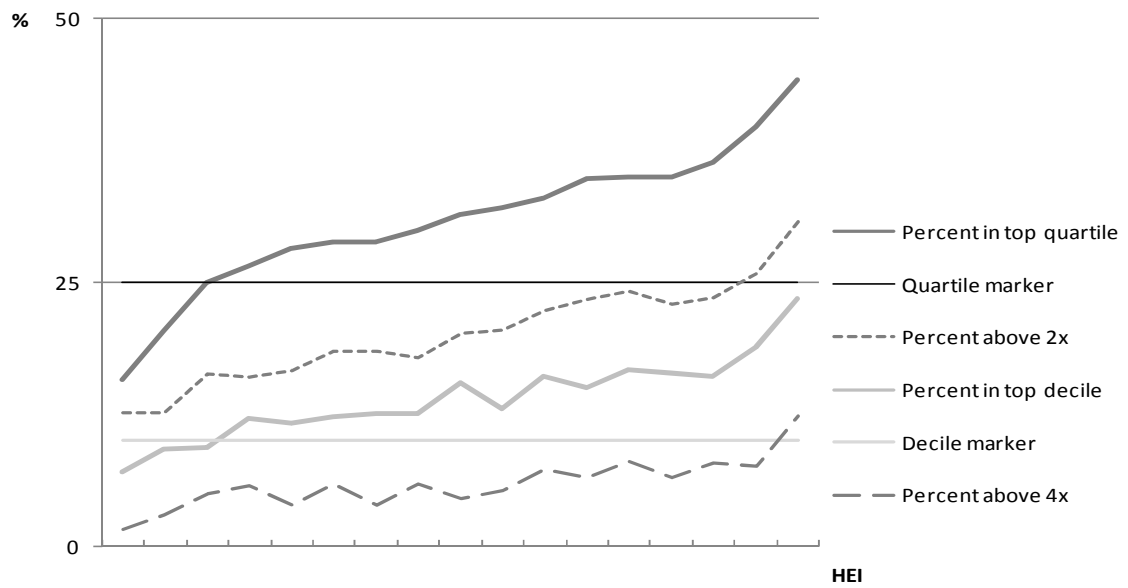


Figure I2.d Comparison of citation impact indicators using centiles and multiples of world average, for Physics (UOA 19)

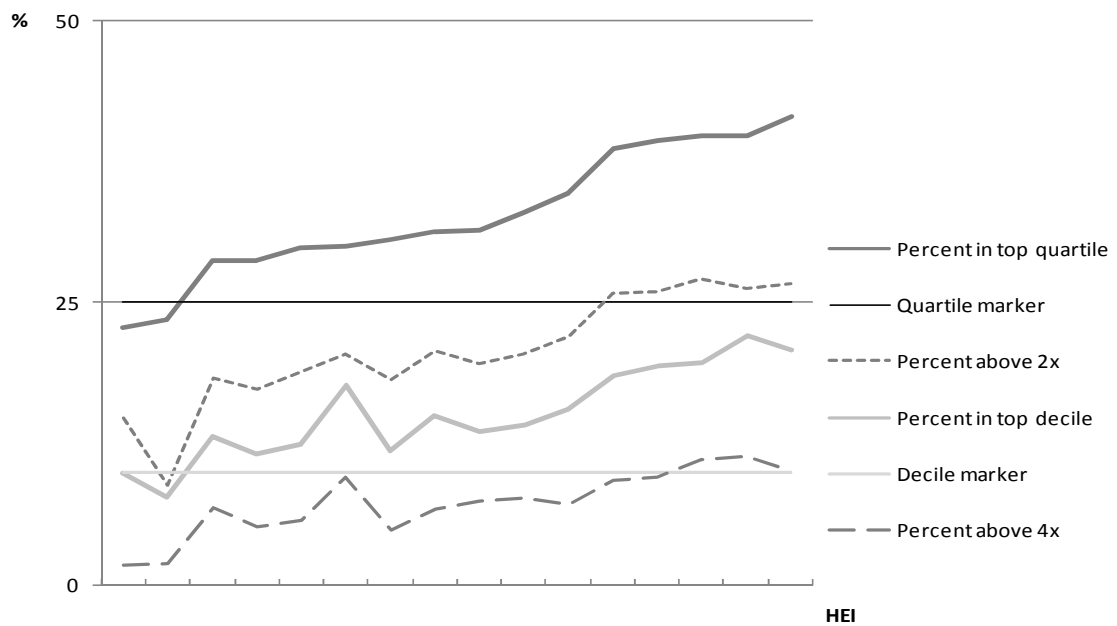
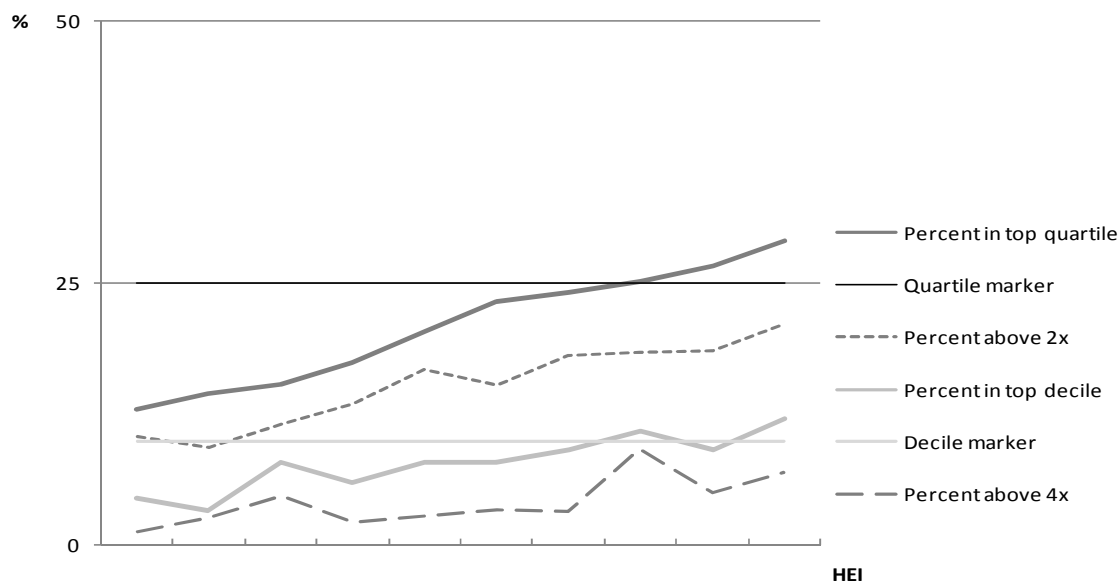


Figure I2.e Comparison of citation impact indicators using centiles and multiples of world average, for Mechanical engineering (UOA 28)



54. Each chart has the four lines stacked one above the other in the same order. The number of papers in the upper quartile is always greater than the number above 2x world average impact. The number of papers in the top decile is always less than the number above 2x world impact but greater than the number above 4x. This is the case for each HEI in each of the sample UOAs.

55. The condition that 'percentage in the top decile' \geq 'percentage of papers above 4x world average' also applies across the other pilot UOAs where the number of papers in the sample exceeds 100, with the exception of one institution in one UOA where there are six papers above 4x world average RBI but only four in the top decile.

Ranking

56. We looked at the effect of using 'percentage of papers in upper quartile' compared with using 'percentage of papers above 2x world average RBI' in order to rank the institutions in each of the five samples. Across the five selected UOAs, 28 (39%) of the 72 rank positions remained the same for the two indicators, with a further 32 (44%) altering only by one place up or down. Ten moved by two places and two moved by three places. The table for UOA 02 is reproduced as Table L2. Our conclusion is that using the upper quartile as a threshold indicator would produce broadly similar results to using twice world average.

Table I2. Comparison between rank on % in top quartile and % above twice world average for Cancer studies (UOA 02)

Pilot HEI	Rank on % in top quartile	Rank on % above 2x world average	Change in rank
M	1	1	0
R	2	3	1
F	3	2	-1
D	4	5	1
S	5	4	-1
B	6	7	1
H	7	8	1
P	8	6	-2
U	9	9	0

57. The percentile calculations can be used to compare the citation impact of the papers produced by these institutions against the world average for each percentile range. These decile and quartile benchmarks provide us with indicators analogous to the ‘x times world average’ thresholds for rebased impact. The charts show how closely they are related.

58. We can also use the decile and quartile markers to make comparisons between subjects. Thus, if we compare across the selected UOAs, we see that for 14 Biology, 17 Earth Systems and 19 Physics, the volumes of frequently cited and very frequently cited papers exceed world average for all but two or three of the institutions. By contrast, for 02 Cancer Studies and 28 Mechanical Engineering, the converse is true – only two or three institutions are producing frequently and very frequently cited papers in volumes that exceed world average. There are quite different densities and distributions of highly cited papers in different subject areas.

Conclusion

59. By comparing the two types of indicators – rebased impact and centiles – for frequently cited papers, we have shown that centiles provide similar information to rebased impact thresholds, but have the advantage that they reduce the risk of sparse data for subject areas where citation counts are clustered around the world average.

60. The centile indicators also provide information about whether an institution is performing above or below world average at particular levels of impact. We conclude that percentile analyses provide a viable alternative to rebased impact thresholds.

61. Centile indicators would be useful where there are few or no papers in relatively high impact categories. They can also be used to compare volume of frequently and very frequently cited papers against world average volumes.

62. A further refinement might be to draw a percentile-profile – similar to an Thomson Reuters Impact Profile but one that shows the distribution of a sample against a standard world distribution. The shape of the world distribution is determined by the number of uncited papers. If half the world papers are uncited, then the remaining half is split equally between the two upper quartiles.

63. It should be borne in mind that both centile and threshold indicators work best when the number of papers is large enough to cancel out the potentially distorting effects of one or two outlying papers. These aspects are addressed elsewhere in this report. For a discussion of thresholds see other sections for a discussion of the levels of granularity appropriate to different types of bibliometric analyses.