On REF and citations – how much can we read about sub-areas of computing?

Alan Dix

http://alandix.com/ref2014/

This version (v2): 9th May 2015
First draft (v1): 24th March 2015

Abstract

This documents describes citation-based analysis of the computing outputs of REF2014, the latest round of research assessment in the UK. The main output of REF is an institution-by-institution research quality profile within different disciplines, However, The Computer Science and Informatics sub-panel also revealed profiles for sub-areas of computing, which are, or have the potential, to be influence decision making in computing departments. The citation-based analysis suggests that, despite every effort to ensure fairness, substantial latent bias has emerged between these sub-area. Not only does this mean that the REF computing sub-area profiles are misleading, but this may also have affected and textual feedback to departments, funding to institutions and possibly gender-neutrality.

Keywords: REF, research assessment, bibliometrics, research funding

Important Note: The author was a member of REF2014 Computer Science and Informatics sub-panel 11. However, the analysis in this paper is based *solely on public domain data*. Sources are listed at end and the website includes links to these and with some of the intermediate derived data used in this analysis. The author would be very pleased for others to replicate and extend the analysis.

1. Introduction

Are there weaker and stronger areas within UK computing research?

Does REF give a reliable measure of the relative strength?

The REF process was designed for comparative evaluation of departments; so it is not obvious that it should be reliable for finer grain distinctions. The analysis of citations data in this document suggests it is not.

Apparent differences between sub-areas of computing that emerge from the REF process are reduced, disappear, or are reversed when alternative, and commonly assumed to be reliable, metrics are used.

That is, the REF process may be (relatively[1]) robust for the purpose that HEFCE and the other funding agents established it, inter-institutional money allocation; however, the evidence suggests it is not reliable for intra-disciplinary comparisons within computing.

2. REF and its impact

REF2014 was the latest instalment of an approximately six yearly review of research across all academic areas within the UK. This is ostensibly about funding determining how HEFCE and the other national bodies distribute money for the next five years. However, just as important is that the grades allocated are used in web sites and literature, especially important when marketing courses to increasingly lucrative international postgraduate students. Indeed, many institutions chose to restrict the staff they entered in order to maximise these scores rather than direct income.

Because of this, the REF process affects each university and department's research strategy, as Vice Chancellors, Deans and Department Heads pore over the results to see how they can do better next time.

3. Computing and the ACM topic classification

While the over REF process is similar for every discipline, there are variations. In particular computing had additional information available as submitting institutions were asked to give each submitted output (paper, book, software, etc.) a topic code from the ACM standard list. These was requested initially to automate the process of assigning around 7000 outputs to panellists to read (computer scientists automate everything!). However, it was also used for two other purposes.

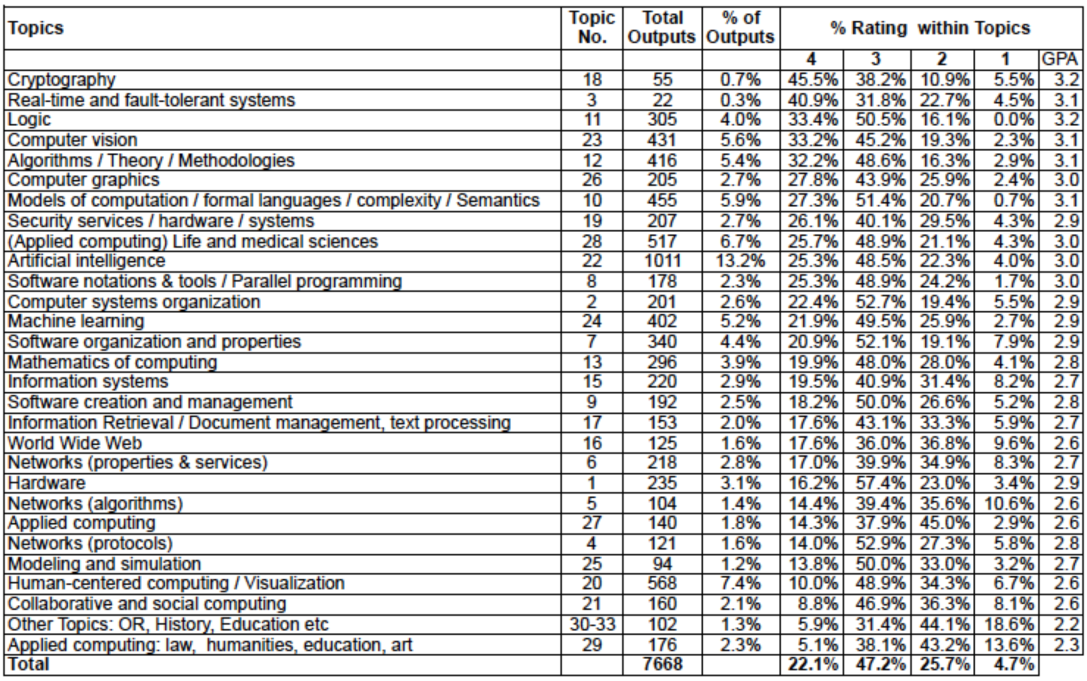
It was used to structure textual feedback to departments/schools on the relative strength of areas.

It was also used to create summary profiles of each area presented in Morris Sloman's slides [Sl15], which have been widely distributed, and also in the sub-panel 11 overview report [REFa, p.49-51]

The crucial, and to some extent controversial slide is below, with areas ranked by how well they fared in REF, from best to worst. At top of the table, with more than 45% of outputs ranked in the top 4* category is Cryptography, whereas towards the tail with only around 10% of outputs raked 4* are a number of areas, including 'Human-centered computing' and and 'Collaborative and social computing', the HCI/people-focused areas.

[1] Although probably not perfect, see section 13 for inter-institutional effects.

In general if you look through the list, the more theoretical/mathematical areas (e.g., logic, formalism) are higher on the list, whereas more applied areas (e.g. the web, more practical areas of software engineering, HCI) are lower.



Topics	Topic No.	Total Outputs	% of Outputs	% Rating within Topics				
				4	3	2	1	GPA
Cryptography	18	55	0.7%	45.5%	38.2%	10.9%	5.5%	3.2
Real-time and fault-tolerant systems	3	22	0.3%	40.9%	31.8%	22.7%	4.5%	3.1
Logic	11	305	4.0%	33.4%	50.5%	16.1%	0.0%	3.2
Computer vision	23	431	5.6%	33.2%	45.2%	19.3%	2.3%	3.1
Algorithms / Theory / Methodologies	12	416	5.4%	32.2%	48.6%	16.3%	2.9%	3.1
Computer graphics	26	205	2.7%	27.8%	43.9%	25.9%	2.4%	3.0
Models of computation / formal languages / complexity / Semantics	10	455	5.9%	27.3%	51.4%	20.7%	0.7%	3.1
Security services / hardware / systems	19	207	2.7%	26.1%	40.1%	29.5%	4.3%	2.9
(Applied computing) Life and medical sciences	28	517	6.7%	25.7%	48.9%	21.1%	4.3%	3.0
Artificial intelligence	22	1011	13.2%	25.3%	48.5%	22.3%	4.0%	3.0
Software notations & tools / Parallel programming	8	178	2.3%	25.3%	48.9%	24.2%	1.7%	3.0
Computer systems organization	2	201	2.6%	22.4%	52.7%	19.4%	5.5%	2.9
Machine learning	24	402	5.2%	21.9%	49.5%	25.9%	2.7%	2.9
Software organization and properties	7	340	4.4%	20.9%	52.1%	19.1%	7.9%	2.9
Mathematics of computing	13	296	3.9%	19.9%	48.0%	28.0%	4.1%	2.8
Information systems	15	220	2.9%	19.5%	40.9%	31.4%	8.2%	2.7
Software creation and management	9	192	2.5%	18.2%	50.0%	26.6%	5.2%	2.8
Information Retrieval / Document management, text processing	17	153	2.0%	17.6%	43.1%	33.3%	5.9%	2.7
World Wide Web	16	125	1.6%	17.6%	36.0%	36.8%	9.6%	2.6
Networks (properties & services)	6	218	2.8%	17.0%	39.9%	34.9%	8.3%	2.7
Hardware	1	235	3.1%	16.2%	57.4%	23.0%	3.4%	2.9
Networks (algorithms)	5	104	1.4%	14.4%	39.4%	35.6%	10.6%	2.6
Applied computing	27	140	1.8%	14.3%	37.9%	45.0%	2.9%	2.6
Networks (protocols)	4	121	1.6%	14.0%	52.9%	27.3%	5.8%	2.8
Modeling and simulation	25	94	1.2%	13.8%	50.0%	33.0%	3.2%	2.7
Human-centered computing / Visualization	20	568	7.4%	10.0%	48.9%	34.3%	6.7%	2.6
Collaborative and social computing	21	160	2.1%	8.8%	46.9%	36.3%	8.1%	2.6
Other Topics: OR, History, Education etc	30-33	102	1.3%	5.9%	31.4%	44.1%	18.6%	2.2
Applied computing: law, humanities, education, art	29	176	2.3%	5.1%	38.1%	43.2%	13.6%	2.3
Total		**7668**		**22.1%**	**47.2%**	**25.7%**	**4.7%**	

Figure 1. Computing Topic Analysis
Extract from "Sub-panel 11: Computer-Science and Informatics. REF Analysis" [Sl15]. Also in "Research Excellence Framework 2014: Overview report by Main Panel B and Sub-panels 7 to 15" [REFa, p.49-51]

4. Implications for HCI and other applied areas

As might be imagined, colleagues in the UK HCI (Human Computer Interaction) community discussed this and wondered whether those on the panel who are connected with HCI (including the author), were perhaps hard judges of their fellows. This is not an unreasonable suggestion, as it has been noted as a problem in grant reviewing for EPSRC and their bodies. However, HCI was not the only apparently weak area, and also the nature of the outputs allocation process within computing meant that a paper in any area would be judged by just one expert and two more general computer scientists.

This said, the author felt a degree of responsibility to both interpret and understand these results, so far as is possible given confidentiality. This is not least because across the UK, university PVCs and department heads will be looking at the above table each time a new post becomes available or there is proposed departmental expansion and think "is it worth investing in this area?".

As well as affecting individuals' careers, hiring and investment decisions in applied areas will disproportionately affect women and non-academic 'impact', the latter of particular interest to funding agencies and government.

5. Data and meaning

One of the first things is the table above is data not information. While it shows that applied areas had relatively fewer outputs ranked 4*, it does not say why.

There are three reasons any or all of which could cause this:

1) The best work in HCI and other applied areas in the UK really is weaker.

2) There is a 'long tail' whereby weaker researchers are less likely to choose theoretical areas. That is the best work is not weaker, just that there is more weaker work leading to a lower percentage at the top.

3) There is something in the processes by which the REF computing sub-panel evaluated outputs that favoured some topic areas relative to others.

During discussions it became clear that it would be possible to disentangle these partially using public-domain data about the REF submissions, which includes, amongst other things, Scopus citation data.

6. Of citations and metrics

There is considerable distrust of the use of citations or other metric-based approaches to research assessment. After RAE2008, there were proposals that the next assessment (that is REF2014, which has just happened) should be based on metrics, largely to reduce the substantial costs and effort that goes into the process. As well as the considerable time the panels spent evaluating the submissions, there was substantial central investment in admin and IT support, and even greater time across UK universities preparing the submissions.

There is considerable evidence that in many subjects, in particular STEM, relatively simple metrics are a good proxy for manually intensive assessment [SM02, Op95, Op98, HO01, HC03], although in some subjects including music [OS08] and economics [LT10, CP11] their coverage or accuracy is arguably less effective.

Figure 2 shows this is also roughly true for REF2014. The graph shows the REF 'Power' vs a similar metric calculated using citation counts. The power is the size of submission (FTE of staff entered as researchers for the exercise) multiplied by the GPA, a weighted sum of the scores. As is evident the correlation between the results obtained through roughly 10 person-years or so of panellists time (for computing alone), and that by simple metrics is strong.

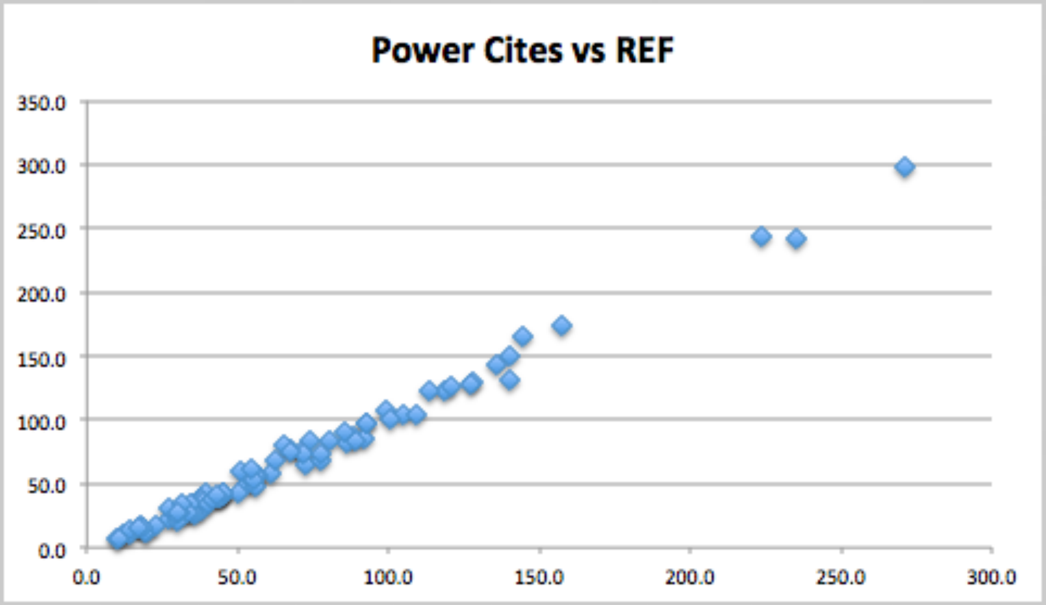


Figure 2. Comparison of REF scores and simple citation based metrics (Scopus, all years)

There are caveats, not least if you ignore the small number of absolute high-flying institutions, things do become a bit more muddy if you zoom into the lower left corner, but the difference *in this measure* is relatively insubstantial.

One of the most persuasive arguments against adopting metrics is not that they fail to measure past performance, but rather the way they would modify future behaviour. We see this in the health service and in schools. The various targets and metrics are established measure effective health and education (not to mention financial 'efficiency'!), but as soon as there are targets and metrics, management focuses on maximising the metrics, not the thing they are intending to measure.

Many breakthroughs in academia have been wrought after significant periods of apparent unproductivity, for example, when Andrew Wiles proved Fermat's Last Theorem after six years. If it were known that the university were to be funded based on metrics, there would be great temptation to only hire those who produced the right kind of work.

In addition, there are kinds of outputs of research that are important, but by their nature attract no cites at all, for example, most patents, software, and, of course, very recent publications.

As well as mitigating against grand opus work, different areas of study have different levels of citation. Notably, the engineering panels for REF decided not to use citations at all in their deliberations, mostly because applied work tends to accumulate less citations, as one of them said, "it doesn't get cited, but it is used to build a bridge". More sophisticated metrics-based approaches can correct for these differences, but these adjustments make them even harder to apply to smaller units.

It was the problem of assessing small units that was also at the heart of Higher Education Policy Institute 's opposition to REF2014 being citation based, "citation analysis does not provide stable and valid analysis at lower levels of aggregation" [Li07].

For the computing panel, we had Scopus citations available to use as part of our professional judgement, for example, to help decide the 'significance' of work, but not as a major part of assessment. The intention was to have Google scholar citations available, as these are believed to be a better metric for computing research, but HEFCE was not able to source these reliably.

7. HEFCE and citations

While each panel was free to choose whether or not they used citation data, HEFCE itself uses citations as a way of validating the fairness of the REF process. While on an output-by-output basis citations are seen as unreliable for the reasons above and others, en mass, in subject areas where it makes sense (in particular STEM), and when adjusted appropriately for different citations patterns, they do appear to have reliability.

That is, they may be a poor way to measure research, but they are a good way to validate the measurement of research. For example, if there were differences between, say, Physics, Computing and Mathematics in terms of the percentage of 4*s awarded, is this defensible?

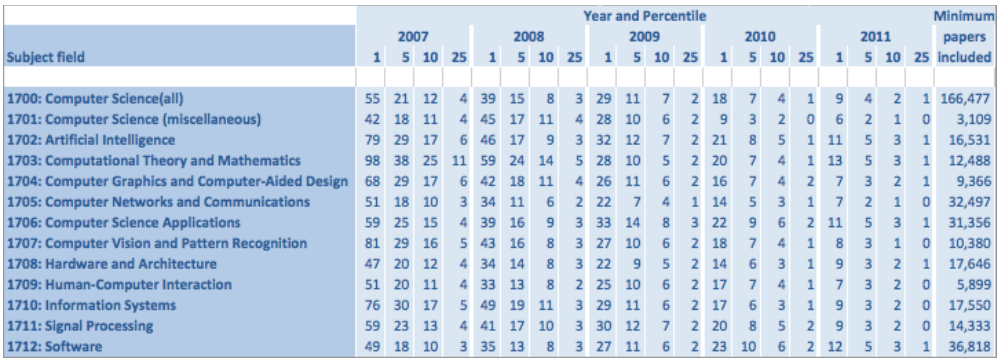
8. Applying metrics to sub-areas of computing

The scores of individual outputs and documentation on each panel's decisions are confidential and due to be destroyed. However, the submissions themselves are a matter of public record and easily accessible for the REF website's download pages [REFd]. Notably these include for each output both the ACM classification used in Sloman's analysis (performed before the data was destroyed), and the Scopus citation count (at a census point late in 2013).

Given older outputs naturally gather more citations, outputs were divided into groups depending on year of publication, sorted within year and then allocated to quartiles. This effectively gives each paper a four level score, lower (lowest 25%), lower-mid (25-50%), upper-mid (50-75%) and upper (75-100%) quartiles. These were then merged and divided by ACM code in order to give a profile for each ACM topic area. The same process gives a profile for each institution used to obtain the graph presented earlier.

The process was repeated, but taking into account only2008-2011 as citation counting might be more reliable during these years.

A slightly different analysis was performed using 'contextual data' distributed by the REF team to universities before the REF submission in 2013 [REFb]. This data (sourced from Scopus) gives for each subject area the number of citations that would be expected for the top 1%, 5%, 10% and 25% of articles within the area. Notably, this includes a breakdown within computing into 12 sub-areas, which can, more or less, match ACM topics.



	Year and Percentile																				Minimum
	2007				2008				2009				2010				2011				papers
Subject field	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	included
1700: Computer Science(all)	55	21	12	4	39	15	8	3	29	11	7	2	18	7	4	1	9	4	2	1	166,477
1701: Computer Science (miscellaneous)	42	18	11	4	45	17	11	4	28	10	6	2	9	3	2	0	6	2	1	0	3,109
1702: Artificial Intelligence	79	29	17	6	46	17	9	3	32	12	7	2	21	8	5	1	11	5	3	1	16,531
1703: Computational Theory and Mathematics	98	38	25	11	59	24	14	5	28	10	5	2	20	7	4	1	13	5	3	1	12,488
1704: Computer Graphics and Computer-Aided Design	68	29	17	6	42	18	11	4	26	11	6	2	16	7	4	2	7	3	2	1	9,366
1705: Computer Networks and Communications	51	18	10	3	34	11	6	2	22	7	4	1	14	5	3	1	7	2	1	0	32,497
1706: Computer Science Applications	59	25	15	4	39	16	9	3	33	14	8	3	22	9	6	2	11	5	3	1	31,356
1707: Computer Vision and Pattern Recognition	81	29	16	5	43	16	8	3	27	10	6	2	18	7	4	1	8	3	1	0	10,380
1708: Hardware and Architecture	47	20	12	4	34	14	8	3	22	9	5	2	14	6	3	1	9	3	2	1	17,646
1709: Human-Computer Interaction	51	20	11	4	33	13	8	2	25	10	6	2	17	7	4	1	7	3	2	0	5,899
1710: Information Systems	76	30	17	5	49	19	11	3	29	11	6	2	17	6	3	1	9	3	2	0	17,550
1711: Signal Processing	59	23	13	4	41	17	10	3	30	12	7	2	20	8	5	2	9	3	2	0	14,333
1712: Software	49	18	10	3	35	13	8	3	27	11	6	2	23	10	6	2	12	5	3	1	36,818

Figure 3. REF contextual data [REFb]

Using this table, each output can be classified as within the top 1%, 5%, 10% or 25% of papers within its sub-area worldwide.

Finally, Google Scholar citations (at a census point late in 2014) were obtained for each output [can I mention source?] and compared with Scopus citation data in Sloman's analysis [Sl15], as these are often deemed to be more reliable in computing. The same Google Scholar data[2] was used in this analysis to create quartile-based calculations similar to those for the Scopus citations. There are two variants, one where missing values (outputs which are not found in Google Scholar) are deemed 'missing' (so not contributing to profiles), and another where they are included as zero values.

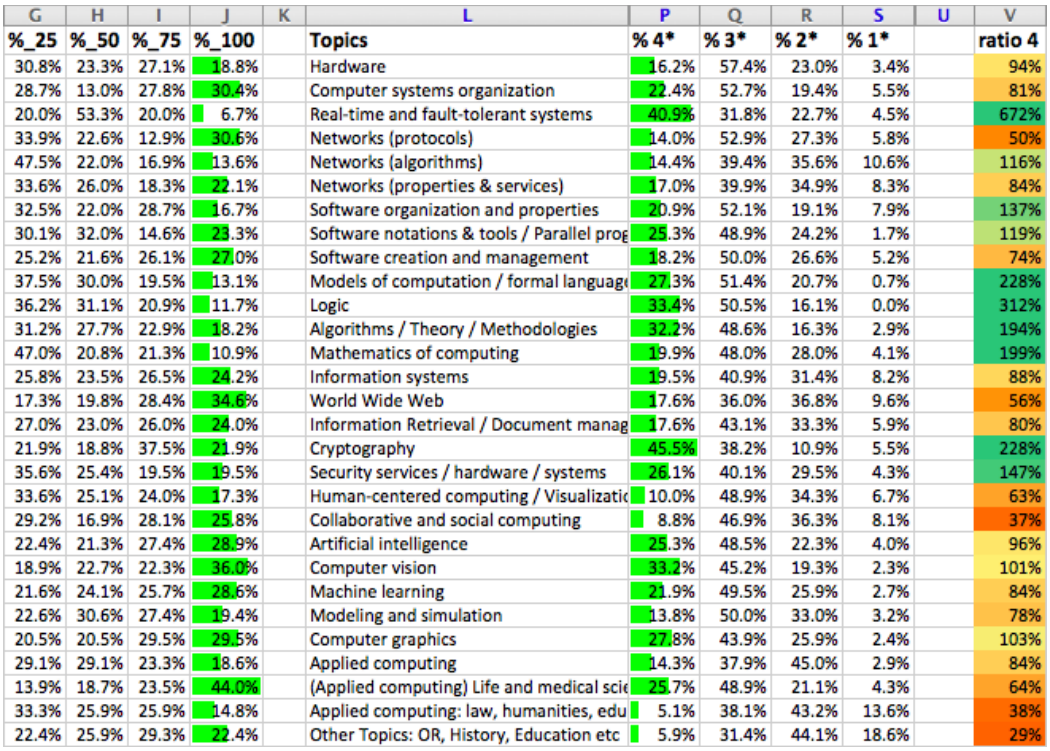
This results in seven variations giving profiles for each area. These are all included as separate worksheets in [Dix15]. The breakdown within computing of 4*, 3*, 2* and 1* is 22.1%, 47.2%, 25.70%, 4.70%. That is 4* should correspond roughly to the top quartile, 3* to the mid two quartiles, and 2* and 1* to the lower quartiles. The match to the context-corrected top world-wide 1%, 5%, 10%, 25% is more complex, but the spreadsheet includes a column where the 'expected' proportion of 4* within the topic area (as predicted by citations) is compared with the actual proportion.

9. What do the metrics they say?

It was expected that there would be a slight disfavouring of HCI and other applied areas, as many in computing do not appreciate softer methods, more discursive papers, etc. Similarly one might expect a slight favouring of more mathematical work, as, unless one checks the proofs line-by-line and finds an error, they are more obviously rigorous. The computing panel made every attempt to be as even handed as possible, but in the end is a human process, so some variation is to be expected, maybe 20-30%. However, the figures that

[2] Some additional hand correction was applied to this as the automatic matching had mismatched a small number of outputs.

emerged from the analysis were so shocking that the author asked a colleague, Andrew Howes to verify the results independently. He replicated the analysis using R and obtained the same results.



G	H	I	J	K	L	P	Q	R	S	U	V
%_25	%_50	%_75	%_100		Topics	% 4*	% 3*	% 2*	% 1*		ratio 4
30.8%	23.3%	27.1%	18.8%		Hardware	16.2%	57.4%	23.0%	3.4%		94%
28.7%	13.0%	27.8%	30.4%		Computer systems organization	22.4%	52.7%	19.4%	5.5%		81%
20.0%	53.3%	20.0%	6.7%		Real-time and fault-tolerant systems	40.9%	31.8%	22.7%	4.5%		672%
33.9%	22.6%	12.9%	30.6%		Networks (protocols)	14.0%	52.9%	27.3%	5.8%		50%
47.5%	22.0%	16.9%	13.6%		Networks (algorithms)	14.4%	39.4%	35.6%	10.6%		116%
33.6%	26.0%	18.3%	22.1%		Networks (properties & services)	17.0%	39.9%	34.9%	8.3%		84%
32.5%	22.0%	28.7%	16.7%		Software organization and properties	20.9%	52.1%	19.1%	7.9%		137%
30.1%	32.0%	14.6%	23.3%		Software notations & tools / Parallel prog	25.3%	48.9%	24.2%	1.7%		119%
25.2%	21.6%	26.1%	27.0%		Software creation and management	18.2%	50.0%	26.6%	5.2%		74%
37.5%	30.0%	19.5%	13.1%		Models of computation / formal languag	27.3%	51.4%	20.7%	0.7%		228%
36.2%	31.1%	20.9%	11.7%		Logic	33.4%	50.5%	16.1%	0.0%		312%
31.2%	27.7%	22.9%	18.2%		Algorithms / Theory / Methodologies	32.2%	48.6%	16.3%	2.9%		194%
47.0%	20.8%	21.3%	10.9%		Mathematics of computing	19.9%	48.0%	28.0%	4.1%		199%
25.8%	23.5%	26.5%	24.2%		Information systems	19.5%	40.9%	31.4%	8.2%		88%
17.3%	19.8%	28.4%	34.6%		World Wide Web	17.6%	36.0%	36.8%	9.6%		56%
27.0%	23.0%	26.0%	24.0%		Information Retrieval / Document manag	17.6%	43.1%	33.3%	5.9%		80%
21.9%	18.8%	37.5%	21.9%		Cryptography	45.5%	38.2%	10.9%	5.5%		228%
35.6%	25.4%	19.5%	19.5%		Security services / hardware / systems	26.1%	40.1%	29.5%	4.3%		147%
33.6%	25.1%	24.0%	17.3%		Human-centered computing / Visualizatic	10.0%	48.9%	34.3%	6.7%		63%
29.2%	16.9%	28.1%	25.8%		Collaborative and social computing	8.8%	46.9%	36.3%	8.1%		37%
22.4%	21.3%	27.4%	28.9%		Artificial intelligence	25.3%	48.5%	22.3%	4.0%		96%
18.9%	22.7%	22.3%	36.0%		Computer vision	33.2%	45.2%	19.3%	2.3%		101%
21.6%	24.1%	25.7%	28.6%		Machine learning	21.9%	49.5%	25.9%	2.7%		84%
22.6%	30.6%	27.4%	19.4%		Modeling and simulation	13.8%	50.0%	33.0%	3.2%		78%
20.5%	20.5%	29.5%	29.5%		Computer graphics	27.8%	43.9%	25.9%	2.4%		103%
29.1%	29.1%	23.3%	18.6%		Applied computing	14.3%	37.9%	45.0%	2.9%		84%
13.9%	18.7%	23.5%	44.0%		(Applied computing) Life and medical scie	25.7%	48.9%	21.1%	4.3%		64%
33.3%	25.9%	25.9%	14.8%		Applied computing: law, humanities, edu	5.1%	38.1%	43.2%	13.6%		38%
22.4%	25.9%	29.3%	22.4%		Other Topics: OR, History, Education etc	5.9%	31.4%	44.1%	18.6%		29%

Figure 4. ACM Topics -- citation quartiles vs REF scores

On the left hand side of Figure 4 are the quartiles as calculated from citations, where the quartile marked 100% is the upper quartile of most highly cited papers. That is column J shows the percentage of papers in each topic that ranked in the top 25% of their year within REF. The

On the right hand side, columns P-S are the 4*, 3*, 2* and 1* percentages generated by Sloman's analysis. The little green histograms on columns J and P are show rankings within the columns.

Some topics, notably Real Time Systems, towards the top, and the two bottom rows, are based on small numbers of outputs, so cannot be sensibly compared. However, the other topics represent several hundred up to over a thousand outputs for AI.

Scanning your eye down the top quartile column (J) and the 4* column (P), it is clear that there is a significant difference. Column V shows how much more or less column (P), the REF results for 4* are compared with a predication based on the number of top quartile outputs.

Given these are reasonably large areas, one might expect a rough correlation, just as we saw for institutions in figure 2. However, here there is no such correlation. Instead we see some areas with 2 or 3 times as many 4* outputs as one might expect, and others with half or a third the number. Figure 5 plots each topic area with horizontal axis the rank order based on the % of top quartile outputs based on citations and the vertical axis the rank order 4* and the rank order based on percentage of 4* outputs. Compare this graph with figure 2, for sub-areas there is no relationship between citations and REF scores.

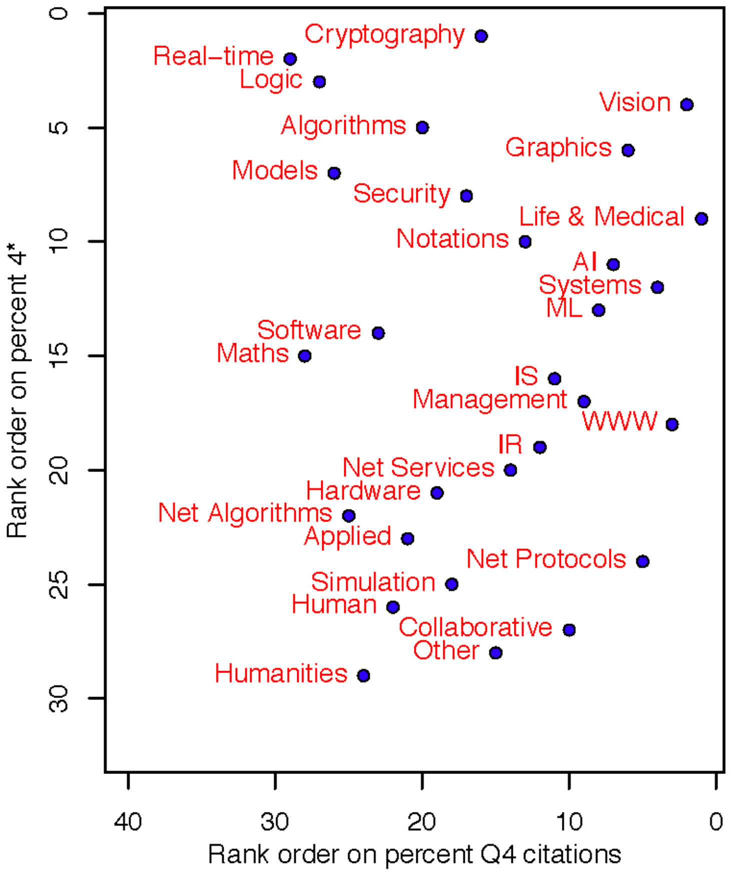


Figure 5. Rank order Q4 citations vs 4* REF

We do not expect an exact relationship. Indeed, the reason for not using pure metrics is that we believe that there are aspects that human peer review will see that bare numbers will not. However, we do expect a level of consistency when averaged. While this appears to be the case for institutional comparisons (by some measures) it is not so between topic areas.

10. Variations, but one theme

Seven different variants were mentioed:

1. using Scopus citations for the entre period

2. using Scopus citations for 2008--2011

3. using Scopus citations 'corrected' for topic area by the REF contextual data [REFb].

4. using Google citations for entire period, but ignoring outputs not in Google scholar

5. as 4, but for 2008-2011 only

6. as 4, but treating missing Google scholar entries and no citations

7. as 7, but for 2008-2011 only

This range of indicators investigated was in case some topic areas had unusual characteristics, for example more recent outputs, which meant that one indicator gave an unreliable measure. However, there was little difference between the variants, except, if anything the more reliable (but slightly more complex) variants (2 and 3) were if anything more extreme than Figure 4. The results are all in spreadsheet [Dix15].

The only major qualitative difference is that applications in life sciences show up stronger in Scopus that in Google scholar, and Web (which is strong on all metrics) gets an extra fillip in Google scholar citations.

While 1,2,4,5,6 and 7 are all internal comparisons within the REF outputs, that is looking just at UK computing. Metric 3 allows one to say with reasonable reliability that a certain proportion of outputs are in the top 1% of their area worldwide. In some areas an output needs to be not only in the top 1%, but the top 0.6% to obtain a 4*, while in others it is sufficient to be in the 4 or 5 %.

No matter how you do the analysis, the overall message is the same. Some areas (e.g. vision) do well in both REF scores and metrics. Some areas (e.g. logic) do a lot better on REF scores, than they do when you look at citation analysis. Some areas (e.g. collaborative computing and web) do a lot worse under REF than citation analysis.

11. Which is right?

As noted earlier, there are good reasons to use human rather than citation or other metric-based evaluation. That is, as an academic community, across all fields, not just computing, we believe that human-based evaluations do a better job.

Furthermore, as a member of the panel the author can attest that the panel strove to be as fair and even handed as possible during the process.

That is, one way to read the disparity on a topic-by-topic level is simply as evidence that this is the right approach: a metric-based approach would have given the wrong answer.

However, we also expect that these differences between metrics and human evaluation should 'average out' over substantial areas, and this does not seem to have happened.

Let's look again at some major reasons for disparity:

1. outputs that cannot be sensibly assessed using citations either because they are (a) unusual (patents, software, etc.) or (b) because they are too recent for citations to be meaningful

2. outliers such as grand opus work that means individuals have low output for periods before making breakthroughs (e.g. Fermat's Last Theorem), papers that are cited because they are wrong (e.g. cold fusion), papers that are very 'sexy' but lack depth.

3. outputs in areas that are valuable, but may systematically attract fewer citations than others (e.g. Engineering and paper that build bridges)

4. outputs of institutions that have less research experience and so may well be placed in poorer, less visible venues and hence attract lower citations

Going through these. (1a) are rare. (1b) Is the reason for doing years 2008-2011 analysis as well as all years, and, if anything, the disparities are greater in this (probably more accurate) variant. (2) Are important when looking at an individual paper and maybe a small research centre, but will be insignificant in larger groupings. The last two, (3) and (4), are particularly important as they represent systematic effects that will not be 'averaged out'. The alternative analysis using the REF contextual data [REFb] attempts to remove the bias in (3), but again, this tends to amplify the same disparities as the simpler analyses. This is because the areas that you might worry would be disadvantaged by (3) in a metrics-based approach, actually fair even worse under REF. Finally, on (4) if you repeat the topic-by-topic analysis for intra-institutional comparisons, the relationship between REF scores and citation metrics is far closer. However, if anything those that tend to do less well under REF than a purely-metrics-based approach, are precisely the overall 'weaker' institutions (with some exceptions see later).

In summary, where there are systematic effects that might mean citation metrics may disadvantage certain kinds of outputs, these very kinds of outputs do even worse under human evaluation.

The REF contextual data analysis is particularly telling here. It gives an indication of which outputs are in the top of their own field worldwide. Some of the mappings between ACM and Scopus classifications are not clear, but overall, and in particular in those where the mapping is reliable, we see disparities such that in some areas a paper in the top 1% of its area is up to five times more or likely to be assigned a 4* than the top 1% of another area.

There appear to only two conclusions (or a mix of the two), either:

(i) The REF human evaluation is right and there are some areas where the worldwide research in the area is 5-10 times less good than others. That is, the REF analysis could be used to create a global league table of subject areas.

(ii) The citation metrics are right and our human evaluation systematically undervalued certain areas.

I don't think anyone involved in the REF process would publicly support claim (i). Even if there is an element of truth in it, the massive difference between REF's evaluation of international areas is hard to justify. Furthermore is explicitly counter to the assumptions made in inter-panel validation with REF, and indeed the very definition of 4* quality.

Therefore it appears that, however unpalatable, the differences are due to (ii).

12. Does this matter - sub-areas?

In section 6, we noted that analysis of previous iterations of RAE/REF process had a strong correlation between metric-based scoring and human analysis, and figure 2 showed similar correlation of REF2014 figures, especially when it comes to then 'big hitters' at the top right.

This is what REF is designed to do, to allocate money to institutions, not to cross-evaluate areas within a discipline.

However, the publication of the sub-area analysis in figure 1 runs the danger that this will be interpreted, not merely as interesting data, but as a comparison of the strengths of UK areas of research. For example, it is very important that the research councils do not look at these as a simple measure of UK research health in the areas.

In addition, if REF has undervalued certain areas, then this will also have affected the narrative feedback given to institutions. This is poorly understood anyway, but certainly the fact that an area is not mentioned as strong will be affected by these between-area differences. This is then likely to affect individual institutions' decisions about hiring and investment within the UoA.

Arguably these hiring and investment decisions are economically defensible, if the effects seen in figure 1 are likely to be repeated and not simply a temporary artefact of REF2014's processes. It doesn't matter how good or bad an area or group 'really is' if it doesn't attract REF 4*s, it doesn't bring in money.

Of course, while economically rationale for each institution, such a policy outcome would be disastrous for UK computer science and against the whole ethos of REF, the Research *Excellence* Framework.

13. Does this matter - institutions?

As has been noted several times, the purpose of REF is to help the funding agencies such, as HEFCE in England, to allocate monies, it is not designed to compare sub-areas.

Even if there are systematic effects whereby some sub-areas are undervalued, this is may not affect overall income distribution as institutions typically have a mix of types of work. The only systematic effects would be if institutions were particularly skewed towards areas more or less favoured.

Unfortunately, the size of the sub-area differences is large, and so this averaging effect may not be sufficient to prevent a level of bias in grading and funding of individual UoAs.

Figure 2 shows research power (GPA * FTE) as worked out using REF start scores and simple citations. It shows a good correlation. Note particularly that towards the bottom left are the smaller institutions, where the averaging effects for metric-based analysis are less likely to hive an accurate measure. The median number of outputs assessed is around 80 per institution, and the average % of 4* outputs was 22%. Based on these numbers we would expect variation of up to a quarter of this figure, so it is reasonable to believe that the difference here really is to do with the human evaluation giving a more accurate picture than metrics.

The weightings used for GPA are 4:3:2:1 for 4*, 3*, 2*, and 1* respectively; that is a 4* is counts twice as much as a 2* and just 1/3 more than a 3*. However, funding in England, which is formula based, ramps much more steeply. For the previous RAE this was approximately 7,3,0,0; that is 2* and 1* outputs attract no funding at all and 4* outputs attract more than twice as much as 3*. This ramping is to be higher for RAE2014 [El15]

Figure 6 shows the equivalent of figure 2, but this time using the 7:3:0:0 weighting to obtain a measure, called GPA# here, multiplied by FTE, which is a closer match to money allocation. Note that the spread at the lower end is greater, but this is not unexpected as the increased weighting of 4* means that a few papers make a big difference, precisely where averaging does not even things out and hence metrics are likely to give poor results. The unexpectedly high second and third institutions (top right) are maybe a little more surprising, but the real outlier is not an English university anyway and is funded non-algorithmically, so maybe doesn't matter too much!

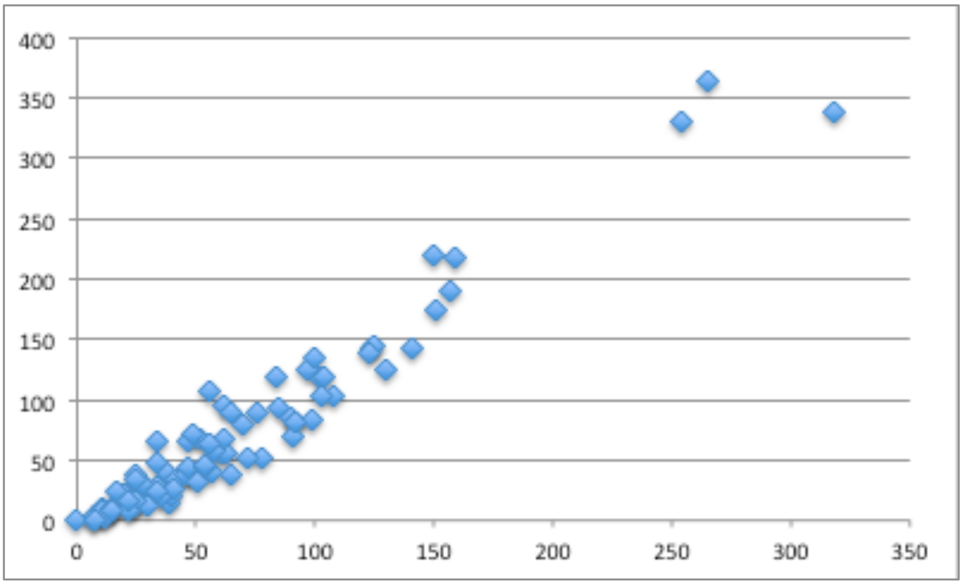


Figure 6. GPA # Power: citations vs REF

Figure 7 shows the plain funding-weighted GPA# score, which is somewhat spread, but shows clear correlation. As noted we would expect a lot of variation given the heavy focus on a few outputs: he median number of outputs is around 80, and so the median number of 4* outputs around 16.

What is more interesting is looking to see if there are common characteristics to those getting higher or lower GPA# based on REF scores vs citations.

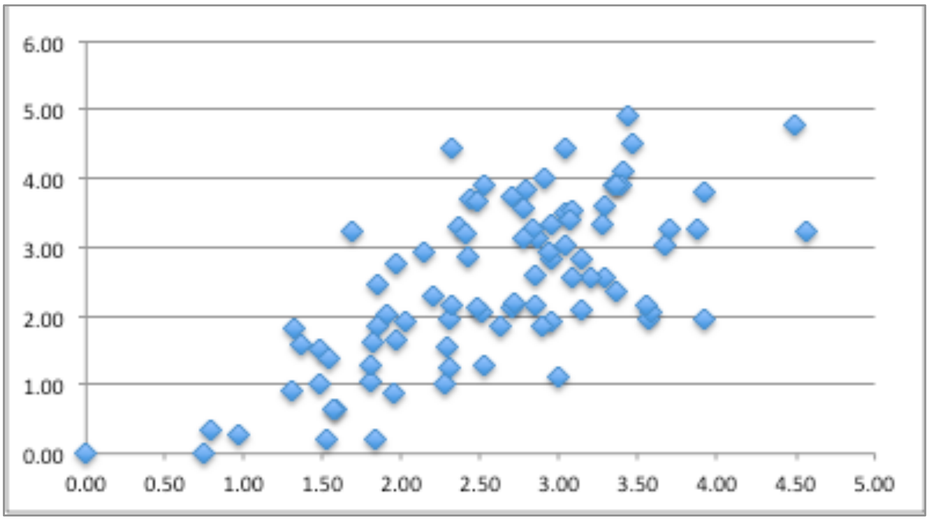


Figure 7. GPA # (approx. money per FTE): citations vs REF

As hinted earlier, the newer post-1992 universities seem to do less well under REF than under metrics, even though you might imagine they would find it harder to establish the reputational factors that help attract citations. If you take the ratio between GPA# based on REF vs citations, of the 27 getting less than 75% of the citation-based prediction, 22 are post-1992, and of the 18 getting more than 125% of the citation prediction, only 1 is post-1992.

There have been accusations in the past that the process is biased towards the 'obvious candidates', but past studies have refuted this in specific areas, [CP11], and certainly, as previously stated, the author can attest to the panel's efforts made to be impartial, as is also evident in the minutes of the sub-panel meetings [REFc].

The difference between sub-areas might be a reason for these variations as, anecdotally, post-1992 universities often work in these more applied, areas. The fact that the sub-area differences seem to be particular sensitive at the 3*/4* boundary would heighten this effect. The impression is further strengthened by the fact that, the institutions that seem to have most positive REF gains compared with citation metrics, are also ones with substantial strengths in the sub-areas that REF seems to favour.

There are alternatives to a sub-area effect. It may be that the lower-ranked institutions were less good at writing the 100 word statements that helped focus panellists on the significance of the output, leading to lower scores. Another reason might be that some institutions publish in 'poorer' venues, and may then be attracting citations from these poorer venues, that is the citations are less 'good' citations. More sophisticated metric based evaluation (e.g. [ZT15] or second order metrics) might be able to determine if this is indeed the case.

Indeed, either of these latter effects could even go some way to explaining sub-area differences. However, given the volume in these lower-end submissions tended to be low, this would not be a complete explanation, at best offering a contribution to a long-tail effect, but not the systematic devaluing of high-citation outputs in some areas.

14. Appealing broadly

Although, I've tried to be even handed with different potential explanations, it is evident that citations are measuring something different from the REF star scores.

One way this is definitely true is in who is doing the assessment. For citation-based metrics it is effectively the judgement of people within the sub-area that counts. However, the nature of REF process in computing meant that panellists were assessing work quite far from their own specialism. In the computing sub-panel, there were three readers for each output, more than many panels, which used just two. However, typically that would be one relative expert and two more broad readers.

Other sub-panels took different approaches to reviewer allocation, but within computing a 4* is more about the broad appeal of an output to computer scientists in general and less about the way it is valued within its own area.

As has been alluded to earlier, it is easy to see the rigour of a mathematical proof, or numerical evaluation of an algorithm, but harder for an outsider to assess the wider set of methods of development and evaluation found in more applied areas. It is therefore not unsurprising to see some variation of scoring.

The level of difference is surprising, as every effort was made to be even-handed, but this is most marked in the proportions of 4*, where it is perhaps especially hard to say something is excellent where one is uncertain of methodologies.

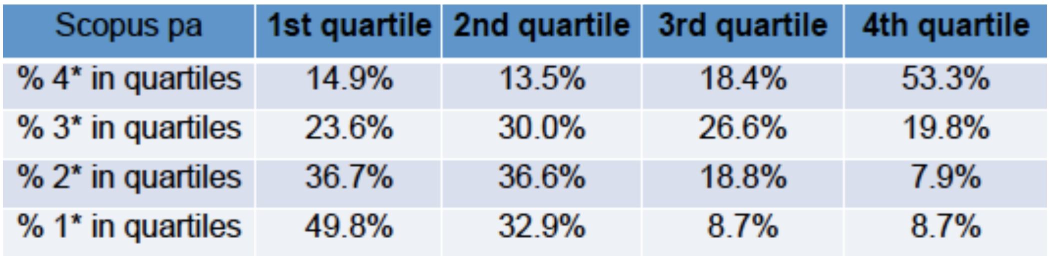
15. Should we use citations more?

So should we use the citation data more?

We have already discussed several of the arguments and counter arguments, and even during the REF submission and assessment period the debate has continued. Oswald and Sgroi argued for the use of journal impact factors for more recent outputs combined (using Baysian methods) with citations as a paper ages [OS13]. However, others critique any use of impact factors, as at best simplistic and poor statistics [Mo13]. I should not that not only did no REF panel use impact factors, in the computing sub-panel noted that, based on the outputs submitted, "the reputation of a journal is becoming even less of an indicator of quality than hitherto". Others are arguing for greater use of more modern web and social media metrics such as Mendeley, particularly for recent work where traditional citations are not usable [Cu12].

Before the REF, the author would have been very strongly with those against the use of metrics, particularly because of the danger of them distorting institutional policy, but now less sure. It appears that many of the distortions of metrics based assessment may be worse for human assessment. There is perhaps potential for a mixed approach, which either uses citations in a more structured way as part of a human assessment process.

Figure 8 shows another table from Sloman's analysis of REF results [Sl15]. This shows Scopus citation quartiles vs star ratings, but was performed before the output scores were destroyed and is on an output-by-output basis. Whilst the above analysis shows that if you drill into individual sub-areas, there are some systematic deviations, when aggregated together figure 8 shows there is a strong correlation between citation and REF score. In particular note that around 50% of all 4* papers are in the top quartile for citations, and similarly for the lowest quartile and 1* outputs.



Scopus pa	1st quartile	2nd quartile	3rd quartile	4th quartile
% 4* in quartiles	14.9%	13.5%	18.4%	53.3%
% 3* in quartiles	23.6%	30.0%	26.6%	19.8%
% 2* in quartiles	36.7%	36.6%	18.8%	7.9%
% 1* in quartiles	49.8%	32.9%	8.7%	8.7%

Figure 8. Overall citation vs REF star rating (from Sloman [Sl15])

At the top left we have outputs that had low citations, but high REF scores. There are various reasons for this, as discussed previously: non-article outputs, maybe papers published in poorly selected venues and so not attracting the interest they deserve.

However the bottom right is more worrying. These are papers with high citations but low REF scores. Again, there are valid reasons for this, which we have discussed, from the papers cited because they are wrong, to those that are sexy but shallow; however, the number does seem high. Possibly future exercises could adopt an 'innocent until proven guilty approach' to outputs with high citations, allocating them 3*/4* scores automatically and then requiring a strong argument that they should exceptionally be reduced.

Alternatively, or as well as this, mid-way through the output scoring process it would be possible to run an analysis such as reported here to act as a check and provoke panel discussions.

16. Should we use citations more?

As noted multiple times, the purpose of REF is the allocate money between institutions, *not* to compare sub-areas. As the sub-panel 11 overview report says regarding the data in figure 1, "*These data should be treated with circumspection*" [REFc, p.48]

The author would urge both policy makers and also individual institutions to heed this warning and to use other measures instead or in addition to REF data when assessing different sub-areas within computing.

Furthermore, when reading textual feedback from the panel, heads of department and others in the institutions, should be aware that the comparison with citation data suggests that the processes in the computing sub-panel may undervalue certain areas.

The data for all but Sloman's analysis tables is available in the public domain. The reader is encouraged to repeat this analysis with more sophisticated methods, and maybe to repeat citation analysis after a few years when all ordinary outputs are able to gather sufficient citations for a more robust long-term analysis.

It may be that this may show that, for example, the human assessment in fact does match more closely to more sophisticated long-term analysis. However, on present data, this seems unlikely. Certainly, more sophisticated and long-term analysis may suggest ways in which we could improve intra-discipline calibration in the REF2020.

One final note is that the public domain data does not include the staff member with whom the output is associated. This makes it hard to perform diversity metrics form the public domain data. However, from the author's experience, the areas that appear to be undervalued by REF are precisely those with a higher proportion of female researchers. That is, the apparent sub-area bias in computing may systematically undervalue the work of female researchers. It may be possible to check this by using the DOI, ISBN or Google scholar links to obtain authors' names and automatically match first names against gender-lists.

References

[Cu12] Curry, A. (2012) Sick of Impact Factors. Blog Post. August 13, 2012. http://occamstypewriter.org/scurry/2012/08/13/sick-of-impact-factors/

[CP11] Clerides, S., Pashardes, P. and Polycarpou, A. (2011) 'Peer review vs metric-based assessment: testing for bias in the RAE ratings of UK economics departments', *Economica*, vol. 78(311), pp. 565–83. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=982219

[Dix15] REF topic analysis compared with citations, ver.5, (spreadsheet) 9th May 2015.

[El15] Else, H. (2015). Research funding formula tweaked after REF 2014 results. *Times Higher Education*, 20 February 2015. http://www.timeshighereducation.co.uk/news/research-funding-formula-tweaked-after-ref-2014-results/2018685.article

[HC03] Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. http://www.ariadne.ac.uk/issue35/harnad

[HO01] Holmes, A. & Oppenheim, C. (2001) Use of citation analysis to predict the outcome of the 2001 Research Assessment Exercise for Unit of Assessment (UoA) 61: Library and Information Management.

[LT10] Levitt, J., Thelwall, M., and Oppenheim, C. (2010). Preliminary findings that can be used when assessing the advantages and limitations of using bibliometric data in the assessment of Economics research *Proceedings of the American Society for Information Science and Technology* 11/2010; 46(1):1 - 7. DOI: 10.1002/meet.2009.1450460349

[Li07] Lipsett , A. (2007). Institute criticises proposed RAE replacement. *The Guardian*, Thursday 13 December 2007. http://www.theguardian.com/education/2007/dec/13/researchassessmentexercise.highereducation

[Mo13] Moriarty, P (2013). Not Everything That Counts Can Be Counted. *Physics Focus*. April 19, 2013. http://physicsfocus.org/philip-moriarty-not-everything-that-counts-can-be-counted/

[Op95] Oppenheim, C. (1995) The correlation between citation counts and the 1992 Research Assessment Exercises ratings for British library and information science departments, Journal of Documentation, 51:18-27.

[Op98] Oppenheim, C. (1998) The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology, Journal of Documentation, 53:477-87. http://dois.mimas.ac.uk/DoIS/data/Articles/julkokltny:1998:v:54:i:5:p:477-487.html

[OS08] Oppenheim, C. and Summers, M. (2008). Citation counts and the Research Assessment Exercise, part VI: Unit of assessment 67 (music). *Information Research*, 13(2), June 2008. http://www.informationr.net/ir/13-2/paper342.html

[OS13] Oswald, A. and Sgroi, D. (2013). REF panels must blend citation and publication data. Times Higher Education, 19 September 2013 http://www.timeshighereducation.co.uk/comment/opinion/ref-panels-must-blend-citation-and-publication-data/2007402.article

[REFa] Research Excellence Framework 2014: Overview report by Main Panel B and Sub-panels 7 to 15, January 2015 http://www.ref.ac.uk/panels/paneloverviewreports/panelminutes/

[REFb] REF contextual data 2013 (spreadsheet)

[REFc] REF2014 Sub-panel 11 - Collated Minutes (December 2013 – November 2014) http://www.ref.ac.uk/panels/paneloverviewreports/panelminutes/

[REFd] Download submission data, 11 – Computer Science and Informatics. http://results.ref.ac.uk/DownloadSubmissions/ByUoa/11

[Sl15] Sloman, M. (2015). Sub-panel 11 Computer-Science and Informatics REF Analysis.

[SM02] Smith, A., and Eysenck, M. (2002) "The correlation between RAE ratings and citation counts in psychology," June 2002 http://psyserver.pc.rhbnc.ac.uk/citations.pdf

[ZT15] Zhu, X. Turney, P., Lemire, D., and Vellino, A. (2015). Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology, 66(2):408-427. doi: 10.1002/asi.23179 http://dx.doi.org/10.1002/asi.23179