Evaluating research assessment: Metrics-based analysis exposes implicit bias in REF2014 results.

The recent UK research assessment exercise, REF2014, attempted to be as fair and transparent as possible. However, **Alan Dix***, a member of the computing sub-panel, reports how a post-hoc analysis of public domain REF data reveals substantial implicit and emergent bias in terms of discipline sub-areas (theoretical vs applied), institutions (Russell Group vs post-1992), and gender. While metrics are generally recognised as flawed, our human processes may be uniformly worse.*

Thursday 24 March, 2016 is the deadline for submitting evidence to the independent review of the Research Excellence Framework (REF) led by Lord Nicholas Stern.

University research assessment has become a normal part of academic life in the UK, but is also occurring or planned across the world. Crucially we need to be sure that whatever methods are used are as impartial and fair as possible. REF2014, the most recent round of research assessment in the UK was a massive exercise, based primarily on expert panels reading and assessing many thousands of outputs (papers, books, patents, performances, etc.). Every effort was made by the organisation and all panel members to make this process as fair as possible.

This blog reports a post-hoc metrics-based analysis of the REF results for the computing sub-panel (SP11), where there is particularly rich information available in the public domain. Despite the best efforts of all involved, this does reveal substantial apparent bias both against individual areas (theoretical/applied) and institutions (Russell Group / pre-1992 / post-1992), and also potential implicit gender discrimination.

REF 2014 Key Facts

4 main panels
36 sub-panels
~200K outputs

sub-panel 11: computer science and informatics

21 panelists
~7000 outputs

The REF assessment included individual research outputs, and narrative elements covering policy and practices for encouraging research quality and impact outside academia. Elements were graded from 4* (top) to 1* to obtain a profile for each institution. While the individual assessments have been destroyed for privacy reasons, most of the original submission data is available in the public domain. This includes narrative elements about policy and practices and also more numerical data about individual research 'outputs', including, for panels where this was used, Scopus citation data.

The narratives are a wonderful source of best practice from innovative research supervision, to techniques to

encourage societal impact. The numerical data is interesting from a theoretical point of view allowing rich analyses, but of course are also of great practical interest, as the results of REF lead directly to research funding over the next five ears and indirectly to a shaping of the UK research landscape

The computing sub-panel (SP11: computer science and informatics) has particularly rich data as those submitting were asked to add a code indicating the sub-area of computing. This was initially intended purely as a way to aid the allocation of outputs to panellists, however, after each output had been graded this was also used to create sub-area profiles, giving for each sub-area the proportion of 4*, 3*, 2* and 1* outputs. These varied greatly, with some areas having up to 40% of work graded 4*, while for others it was well below 10%.

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 / Met	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 / t	10	455	5.9%	27.3%	51.4%	20.7%	0.7%	3.1
Security services / hardw	19	207	2.7%	26.1%	40.1%	29.5%	4.3%	2.9
Information Retrieval / Document management, text processing	17	155	2.0%	17.6%	45.1%	33.5%	3.9%	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					38.1%	43.2%	13.6%	2.3
Total					47.2%	25.7%	4.7%	

theoretical areas
30-40% 4*

applied/human areas
10-20% 4*

These sub-area profiles were released with a strong ‘health warning’, in particular the grades are averaged over all institutions and tell you nothing about the strength of individual research groups. However, they are widely interpreted as a sort of league table between areas and are already affecting institutional recruitment policies.

Given the effect this is having, it seemed important to verify via some other means how accurate a picture this represented, and to disentangle the various possible explanations for the observed differences between subareas. While there is substantial distrust of the use of citation metrics for assessment itself, the volume of results overall mean that is a widely acknowledged that this is an effective means of validating outcomes, where the numbers of outputs are high enough that exceptional cases are averaged out statistically. Indeed HEFCE commissioned their own metrics-based analysis, “The Metric Tide“.

Although individual output grades are no longer available, it is possible to use the Scopus citation data and Google scholar data (gathered after REF assessment) to create their own profiles and ‘league tables’. Seven different analyses were performed, all gave broadly similar results, suggesting that, despite the best efforts of the panellists, the overall process has emergent bias giving up to ten fold advantages to some sub-areas and some kinds of institutions.

As an illustration, one of the analyses used Scopus data to position each output within its own area worldwide: is an output in the top 1%, 5%, 10%, etc. This yields a ‘league table’ for each topic area showing how well it is performing compared with other countries’ research, with some areas having more than 30% of work in the top 1% worldwide,

and others less than 5%. Given the REF definition of 4* is ‘world-leading’, one might expect this to correlate well with star ratings, but this is far from the case.

	Topic	%_1	%_5
1	(Applied computing) Life and medical sciences	33.1%	26.2%
2	Computer vision	31.1%	26.5%
3	Networks (protocols)	29.0%	22.6%
4	Computer systems organization	27.8%	31.3%
5	World Wide Web	25.9%	33.3%
6	Collaborative and social computing	24.7%	30.3%
7	Computer graphics	24.2%	32.6%

22	Cryptography	9.4%	34.4%
23	Applied computing: law, humanities, education, a	9.3%	20.4%
24	Software notations & tools / Parallel programmin	8.7%	18.4%
25	Real-time and fault-tolerant systems	6.7%	20.0%
26	Models of computation / formal languages / com	6.0%	21.7%
27	Mathematics of computing	5.4%	21.2%
28	Logic	4.1%	24.5%

Many of the areas right at the bottom of this table were towards the top of the REF assessment. In general, if the actual REF grades are compared with those predicted based on citations, there is no apparent correlation. However, there is an evident trend that more theoretical/mathematical areas are favoured under REF compared with metric predictions, whilst more applied and human-centric areas are disfavoured.

In terms of the world rankings, an output in a more applied area, on average, needed to be in the top 0.5% (top 1 in 200) of its discipline to obtain a 4*, whereas in theoretical areas it was sufficient to be in the top 5% (1 in 20). That is, our panel scores did not at all reflect the assessment of global peers.

A level of bias is inevitable in any assessment process, but the particular practices of the computing sub-panel may have exacerbated this. An algorithm was used after all outputs were scored to ‘normalise’ between more or less generous panellists. This required substantial overlaps between panellists allocations, and hence panellists had to ‘spread’ their expertise, meaning that each output was reviewed by at best one expert and two non-experts. Reviewing far from one’s core area, it is inevitable that surface judgements become more likely.

When considering institutional profiles, the picture is also deeply worrying. The proportion of 4* is particularly important as this most heavily affects funding (HEFCE use approximately 4:1 weighting for 4* and 3* with 2* and 1* obtaining no money). If citation data is used to predict profiles for each institution and then this and the actual REF profiles are used to obtain funding per staff figures, one can look at winners and losers (using here 25% above or below prediction as threshold).

The winners are predominantly old (pre-1992) universities and especially Russell Group, whereas the losers are predominantly new (post-1992) universities. Unlike the situation with sub-areas, it is not that there are major reversals, the strongest institutions tend to score well both under REF and also metrics, it is just they get an extra, but very large additional fillip. That is, outputs that would appear equivalent based on external citations are scored far more highly if it comes from a known ‘good’ institution. In terms of money, it suggest that new universities may be awarded up to two thirds less research funding than might have received under a blind system.

Again, REF processes may have exacerbated existing institutional bias, this time affecting all panels. Outputs were not anonymised, and panellists' scoring spreadsheets were, by default, ordered by institution, which both made the institution very obvious and also likely to create anchoring effects.

Finally, both the sub-areas of computing and the institutions disadvantaged by REF are those that tend to have a higher proportion of female academics. That is, the apparent disciplinary and institutional bias would be likely to create implicit gender bias. Indeed, amongst other things, HEFCE's "The Metric Tide" revealed that the computing sub-panel awarded substantially more 4*s (top grade) to male authors than female ones, an effect that persists even once other explanatory factors considered.

REF2014 has determined current funding and is shaping the development of computing as an academic discipline. Based on this analysis the grounds for both may be fundamentally flawed. While there are specific aspects of computing that may exacerbate these effects, it may well be that other panels have similar issues.

This Thursday is the deadline for submitting independent review of the Research Excellence Framework (REF) led by Lord Nicholas Stern. The review is examining how university research funding can be allocated more efficiently so that universities can focus on carrying out world-leading research. More here on the open consultation.

Note: This article gives the views of the author, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below. The author was a member of the REF computing sub-panel, but the analysis reported is based on public domain data only. On a personal note, having been a part of REF, I can attest to the desire by all concerned to make the process fair and transparent. However, based on my experiences and especially the analysis of the data summarised here, I now believe that, despite the best efforts of all involved, the REF output assessment process is not fit for purpose.

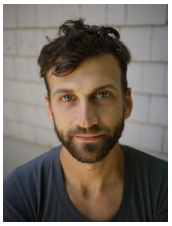
In a first blog of this series, Brett Buttlier quoted Curt Rice's concern that 'systems based on counting can be gamed', and until being a REF panellist I would have echoed this sentiment and rejected metrics-based evaluation. However, in my own advice to those preparing now for REF2020, and listening to that of other past REF panellists, we are continually saying "do good research … but do this to make it look good in REF" – that is, gaming. It seems that while metrics are bad, our human process was uniformly worse in all respects.

About the Author

Alan Dix is a Processor in Human Computer Interaction at the University of Birmingham and Senior Researcher at Talis. His research includes many aspects of the way people interact with technology from formal methods to physicality, creativity and the social and political implications of computation. He co-authored one of the main text books on HCI, runs a bi-annual maker/meeting workshop Tiree Tech Wave on a remote Scottish Island, co-invented intelligent fairylights, and in 2103 walked a thousand miles around the periphery of Wales as a personal and research exploration.

This is part of a series of pieces from the Quantifying and Analysing Scholarly Communication on the Web workshop. More from this series:

Context is everything: Making the case for more nuanced citation impact measures.

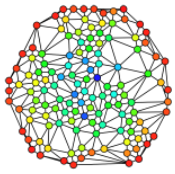


Access to more and more publication and citation data offers the potential for more powerful impact measures than traditional bibliometrics. Accounting for more of the context in the relationship between the citing and cited publications could provide more subtle and nuanced impact measurement. **Ryan Whalen** *looks at the different ways that scientific content are related, and how these relationships could be explored further to improve measures of scientific impact.*

The ResearchGate Score: a good example of a bad metric

According to ResearchGate, the academic social networking site, their RG Score is "a new way to measure your scientific reputation". With such high aims, **Peter Kraker**, **Katy Jordan** *and* **Elisabeth Lex** *take a closer look at the opaque metric. By reverse engineering the score, they find that a significant weight is linked to 'impact points' – a similar metric to the widely discredited journal impact factor.Transparency in metrics is the only way scholarly measures can be put into context and the only way biases – which are inherent in all socially created metrics – can be uncovered.*

Bringing together bibliometrics research from different disciplines – what can we learn from each other?



*Currently, there is little exchange between the different communities interested in the domain of bibliometrics. A recent conference aimed to bridge this gap.***Peter Kraker**, **Katrin Weller**, **Isabella Peters** *and* **Elisabeth Lex** *report on the multitude of topics and viewpoints covered on the quantitative analysis of scientific research. A key theme was the strong need for more openness and transparency: transparency in research evaluation processes to avoid biases, transparency of algorithms that compute new scores and openness of useful technology.*

We need informative metrics that will help, not hurt, the scientific endeavor – let's work to make metrics better.

Rather than expecting people to stop utilizing metrics altogether, we would be better off focusing on making sure the metrics are effective and accurate, argues **Brett Buttliere.** *By looking across a variety of indicators, supporting a centralised, interoperable metrics hub, and utilizing more theory in building metrics, scientists can better understand the diverse facets of research impact and research quality.*