Five recommendations for using alternative metrics in the future UK Research Excellence Framework

blogs.lse.ac.uk/impactofsocialsciences/2014/10/23/alternative-metrics-future-uk-research-excellence-framework-thelwall/

10/23/2014

Although many are excited by the possibilities for using alternative metrics to supplement research assessment, others are concerned about the ease with which the figures can be gamed. It is clear that there is already gaming within traditional citation impact metrics in peer reviewed journals and without quality control mechanisms, social media metrics would be susceptible to the same. **Mike Thelwall** *provides his recommendations for uses of alternative metrics in national research evaluation exercises.*

In the previous UK Research Excellence Framework (REF), due to be reported in December 2014, individual sub-panels (i.e., subject areas) had the option to request citation counts for submitted articles, together with field and time normalization information. Most did not and there was no option to ask for any alternative types of metrics. Although it is not clear yet how the different REF sub-panels have used their citation counts in practice, one way in which they could be useful is an indicator (not measure) of the likely extent of impact of an article within the scholarly community.

Although an article can have a high impact and few citations or even a negative impact and high citations, in general peer review scores tend to correlate with appropriately normalised citation counts in many disciplines. Citation counts may therefore be useful to compare with peer judgements of articles: if the two disagree then the sub-panel expert reviewer could perhaps spend some additional time to double-check their original judgement. This initial judgement may have taken only 10-15 minutes due to the volume of submissions assessed, and so it is likely that mistakes are frequently made.

Image credit: FutUndBeidl (Flickr, CC BY 3.0)

Now that academics need to demonstrate the wider impacts of their research in the REF there have been calls to assess whether there are other metrics that could be used to assess this. For example, since a significant proportion of the public use social media, such as Twitter and Facebook, would it be helpful to count how often academic articles are tweeted or posted in Facebook in case this reflects some kind of wider research impact? Similarly, should we count how many times each publication submitted to the RAE is blogged about? Here is a partial list of alternative metrics that have been proposed for academic articles or books.

- Tweet citations
- Facebook post citations
- Blog citations
- General web citations
- Grey literature citations
- Online syllabus citations
- Online presentation citations
- Discussion forum citations
- Mainstream media citations
- Mendeley, CiteULike or Zotero bookmark counts
- Library holdings (of books)
- Citations from Google Books

A significant amount of research has given empirical evidence that many alternative metrics tend to give higher scores for articles that are more cited, demonstrating that these scores are not random and probably have some relationship with traditional scholarly impact. Nevertheless, there is still little real evidence that any of them reflect particular types of wider impact (exception: online syllabus mentions clearly reflect educational impact) and no direct evidence that any of them would reflect peer judgements of any type of wider impact.

Although many are excited by the possibilities for using alternative metrics, others are concerned about the ease with which the figures can be gamed, since most are not subject to quality control. It is clear that there is already gaming within the peer reviewed publications (for example, leading to Thomson Reuters withdrawing impact factors from some journals) and it seems obvious that this will be enormously worse in social media without quality control if there is a financial incentive, such as REF funding, to be able to report high scores.

Here are my personal recommendations for uses of alternative metrics, including altmetrics, in national research evaluation exercises, such as the REF. Please leave comments if you see problems with these recommendations. Otherwise I will argue for them in future meetings of the HEFCE independent review of the role of metrics in research assessment led by James Wilsdon.

1. **Alternative metrics should not be provided routinely for all articles**. Almost all alternative metrics seem currently to be highly susceptible to gaming and spam and most seem to give little added information for typical articles. It is likely to be damaging to routinely collect alternative metrics for articles for evaluation purposes because this will push academics and research support offices towards wasting their time trying to attract tweets etc. to their work. Of course, a certain amount of self-publicity is a good thing and should not be discouraged but if it is measured then it will automatically lead to gaming and wasted time, rendering the results useless.
2. **Exception**: Book-specific metrics have not been used in the REF before and I think that there are two that might be helpful and that might be difficult to game. Counts of the number of libraries holding a book can be obtained from worldcat.org (sometimes called libcitations) and this can be an indicator of the reach of a book.

Of course this number should be interpreted in the context of the type of book and whether it is aimed at a wide audience or not – low numbers are not automatically bad and high numbers are not automatically good. Similarly, counts of citations from books to books can be extracted from Google Books and these could give an indicator of book impact within the scholarly community. Google Books can probably be spammed to some extent, for example by publishing internal reports as books or perhaps even by publishing student projects as books. Both of these book impact indicators are probably only relevant to arts and humanities researchers and are much more crude indicators of impact than are citations because of differing citing cultures. Nevertheless, assessing hundreds of books on highly varied and specialist topics must be a hugely challenging task for REF panel members and anything that can help them identify books that have apparently had an unexpectedly high impact may be useful – but the assessors should interpret the indicators in the context of the type of book, its subject area and year of publication and then apply their own expert judgement after reading ("scanning bits of" is probably more realistic?) the books.

3. **Units of assessment (e.g., departments) should be given the option to provide their own choice of alternative metrics in special cases**. Research may have impacts that are not obvious from reading it or even from citation counts. For example, a research group may maintain a website that is popular with schools, host key software with a substantial uptake or produce books that are in reading lists around the world. To give a specific case, the point of many blogs is to attract a wide audience but how else can you prove that a blog is widely read than by reporting how many readers or visitors it has? You can immediately tell that you are reading something special when you get to Stephen Curry's blog but its real value comes from the number of other people who have come to the same conclusion. Researchers should have the opportunity to present data to support their claim of having a non-standard impact. For units of assessment that do not allow the routine use of citation counts, I think that citation counts should be allowed in special cases (I did this last time). For all units of assessment, I think that alternative metrics should also be allowed in special cases. I think that they will be particularly valuable for social impact case studies but can also be useful to demonstrate educational impacts for research. I suspect that this has already happened to some extent in the impact case studies, but we will see when they are published next year.
4. **Assessors should be cautioned to not interpret alternative metrics at face value but to take them as approximate pointers to the potential impact of the research**. There are two important problems with interpreting alternative metrics. First, in most cases it is impossible to effectively normalise them for field and discipline and so it is hard to be sure whether any particular number is good or not. Second, this is exacerbated because an alternative metric could partly reflect "important" impact and partly reflect irrelevant impact, such as derision. For example, an article with a humorous title could be tweeted for amusement or for the value of its findings. In practice, this means that assessors should use the alternative metrics to guide them to a starting position about the impact of the activity but should make their own final judgement, taking into account the limitations of alternative metrics. I see alternative metrics as being essentially qualitative indicators in this context. For example, a huge number of people have watched Lady Gaga's "Bad Romance" played on the Iowa State University carillon at YouTube (790+ thousand) which might be claimed as wider impact for music scholarship but this cannot be directly compared with the original (600+ million) or with unrelated activities, such as a scientist picking worms out of sewage with his bare hands (110+ thousand). All of these numbers are evidence of huge audiences for the type of video but it makes no sense to compare the numbers and disciplinary experts will need to judge the merits of the type of impact generated. This basic caution, the "Gaga test", is that altmetrics should not be compared against each other unless the sources are clearly similar in type, date and scope. Although social media statistics can be easily gamed, (a) the incentive will be lower because fewer such figures will be reported and so gaming will be easier to detect, and (b) the numbers can only ever be very crude so a large amount of gaming would be needed to have any real affect, and this should also make the gaming easier to detect.
5. **Units of assessment submitting any alternative metrics should include a declaration of honour to state that they have not attempted to game the metrics and to declare any unintentional manipulation of the metrics**. Unintentional manipulation might include librarians teaching students how to tweet articles

with examples of the university's REF publications. The declaration of honour and the alternative metrics will be made fully public and it is likely that future researchers in computer science will develop algorithms to detect manipulation of REF metrics and so it will be highly embarrassing for anyone that has submitted manipulated data, even if unintentionally. It is likely that this process will be effective because individuals may well gain access to the raw data used in services like Twitter and Mendeley. This makes it reasonably likely that someone could discover, for example, whether the IP addresses of tweeters clustered within a university, or detect abnormal patterns of the accumulation of a metric over time. Hence, even sophisticated manipulation strategies would have a chance of detection. This is the kind of "adversarial information retrieval" problem that computer scientists love and are good at, after years of developing anti-spam software. A declaration of honour would give a non-trivial degree of risk to the submitting unit of assessment and should act as a deterrent to using metrics in all except the most important cases.

Finally, I hope that if it is clear that alternative metrics will not be routinely used for articles then academics would not be pressured to engage in social media for all of their articles in order to generate high altmetric scores. I also hope that if it is recognised that it is possible to use alternative metrics to support individual cases for the wider impact of research then those who are good at engaging with the public will be given the freedom to blog, tweet or make videos in the knowledge that their work can be valued in the REF. I imagine a researcher having their annual review, admitting that they hadn't published any journal articles that year but mentioning that they had continued to maintain their blog and expecting to be fired. Instead, their manager asks: So what exactly is your blog about? How many people read it? Can we make the case that it is disseminating scholarship in some way?

Thank you to David Colquhoun for comments on an earlier version of this post: answers to his criticisms have been incorporated above. Many of the ideas here (including the discovery of the two scholarly videos mentioned) are due to the work of Kayvan Kousha and others on developing and evaluating alternative metrics. I have not attempted to inflate any of the alternative metrics used in the article (YouTube view counts) although Kayvan and I have written about the two (wonderful!) scholarly videos before, when they already had gained unusually many views.

Top left Image Credit: Tape measure by Simon A. Eugster (Wikimedia, CC BY 3.0)

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Mike Thelwall *is head of the Statistical Cybermetrics Research Group at the University of Wolverhampton. He has written 200 refereed journal articles and two books about research-related indicators from the web and has developed the free program Webometric Analyst to extract data for these indicators. He is part of the independent review of the role of metrics in research assessment commissioned by HEFCE. http://www.scit.wlv.ac.uk/~cm1993/mycv.html*