Why Metrics Cannot Measure Research Quality: A Response to the HEFCE Consultation

written by Allegra
June, 2014

We re-post here the reponse to the Higher Education Funding Council for England written by Dr Meera Sabaratnam (SOAS) and Dr Paul Kirby (Sussex University), initially published on June 16 on The Disorder of Things Blog. Allegra stands in solidarity with all initiatives against 'impact' measurements in scholarship and wants to join forces with colleagues struggling against the

neoliberal forces attempting to transform our universities into corporatized knowledge factories.

Why Metrics Cannot Measure Research Quality: A Response to the HEFCE Consultation

The Higher Education Funding Council for England are reviewing the idea of using metrics (or citation counts) in research assessment. We think using metrics to measure research quality is a terrible idea, and we'll be sending the response to them below explaining why. The deadline for receiving responses is **12pm on Monday 30th June** (to metrics@hefce.ac.uk). If you want to add an endorsement to this paper to be added to what we send to HEFCE, ***please write your name, role and institutional affiliation below in the comments***, or email either ms140[at]soas.ac.uk or p.c.kirby[at]sussex.ac.uk before Saturday 28th June. If you want to write your own response, please feel free to borrow as you like from the ideas below, or append the PDF version of our paper available here.

Response to the Independent Review of the Role of Metrics in Research Assessment

June 2014

Authored by:

Dr Meera Sabaratnam, Lecturer in International Relations, SOAS, University of London

Dr Paul Kirby, Lecturer in International Security, University of Sussex

Summary

Whilst metrics may capture some partial dimensions of research 'impact', they **cannot be used as any kind of proxy for measuring research 'quality**'. Not

only is there **no logical connection** between citation counts and the quality of academic research, but the adoption of such a system could **systematically discriminate against less established scholars and against work by women and ethnic minorities.** Moreover, as we know, citation counts are highly vulnerable to **gaming and manipulation.** The overall effects of using citations as a substantive proxy for either 'impact' or 'quality' could be **extremely deleterious to the standing and quality of UK academic research as a whole.**

Why metrics? Why now?

1. The rationale for looking at metrics as a "potential method of measuring research quality and impact" (Consultation letter, section 1) is somewhat opaque in the consultation letter. This letter notes that some people may use metrics to assess research, and that the Secretary of State wishes to look at the issue again. The previous review on the matter in 2008/9 concluded that the 'data was insufficiently robust' to adopt their use.

2. To speak more precisely, we might consider the following underlying rationales as driving this general interest:

- The research assessment exercises conducted at a national level (RAE 2008; REF 2014) and at institutional levels are difficult, time-consuming, expensive and laborious because they consume large quantities of academic energy. Universities and academics themselves have complained about this.
- Ministers, civil servants, research administrators and managers might prefer modes of assessment that do not require human academic input and judgement. This would be cheaper, not require academic expertise and would be easier to administer. This would facilitate the exercise of greater administrative control over the distribution of research resources and inputs.

- Moreover, in an age of often-digitised scholarship, numerical values associated with citations are being produced – mostly by data from large corporate journal publishers – and amongst some scholarly communities at some times they are considered a mark of prestige.

3. This present consultation proposes to take views on the use of metrics – for the most part meaning citation counts – to prospectively incorporate these into mechanisms of research assessment once more. In particular, they want to look at 'research quality and impact' as areas in which research should be assessed.

4. We suggest that it is *imperative to disaggregate 'research quality' from 'research impact'* – not only do they not belong together logically, but running them together itself creates fundamental problems which change the purposes of academic research.

5. We also want to note a contradiction in different reasoning for using metrics. On the one hand, one position seems to be that we should be using metrics as a source of 'big data' we don't currently have to produce different judgements about what good academic research is. On the other hand, the argument is that metrics do actually replicate the outcomes of peer review processes so approximate a cheaper and quicker way of doing the same thing. There is an important tension here: the former reasoning implies we want to change what we think good academic research is and a downgrading of peer review processes; the latter implies that peer review is still the key standard for assessing research but we want to do it (or something like it) more quickly. The Review team need to make a clear determination on which of these objectives it is pursuing.

Using metrics for measuring impact: what are we actually measuring?

6. Why do academics cite each others' work? This is a core question to answer if we want to know what citation count metrics actually tell us, and what they can be used for. Possible answers to this question include:

- It exists in the field or sub-field we are writing about
- It is already well-known/notorious in our field or sub-field so is a useful reader shorthand
- It came up in the journal we are trying to publish in, so we can link our work to it
- It says something we agree with/that was correct
- It says something we disagree with/that was incorrect
- It says something outrageous or provocative
- It offered a specifically useful case or insight
- It offered a really unhelpful/misleading case or insight

7. As an example, an extremely widely cited piece in the field of International Relations is Samuel Huntington's book on 'The Clash of Civilizations'. This has been one of the most controversial pieces in the discipline, and has probably been cited for all of the reasons above (he initially published a short version in *Foreign Affairs* journal). As of today, GoogleScholar lists 22,353 citations to the book or article. Amongst these citations are an extremely large number of 'negative' ones criticising the research and critiquing the piece for its gross simplifications, inflammatory political claims, selective and problematic reading of the historical record, cultural essentialism and neglect of multiple other issues such as the global economy. After 9/11 however, various non-academic readers seized on some of the broad arguments to suggest a perennial struggle between Christianity and Islam, as validated by a famous Harvard professor (with no academic background on either of these religions). This no doubt has contributed to a political climate which has facilitated military interventions in the Middle East and more aggressive attitudes towards religious diversity from members of different religions. On the other hand, much more detailed and nuanced work exists based on solid historical evidence and knowledge of contemporary relations, which will have many fewer citations due to publishing outlet, the profile of the author, and the less outrageous, if much more rigorous, findings. These accumulated citations to Huntington clearly indicate that the texts have been central to networks of scholarly argument about world politics in recent

decades, and we might learn much from that fact. But this is no measure of quality, not even one of 'popularity' (if we understand that to carry positive connotations).

Metrics and the measurement of impact

8. Based on the analysis in points 6 and 7 above, it is clear that citation counts can be one way of thinking about the generic 'impact' of an academic piece on a field. However, in their current form they cannot properly differentiate between 'positive' impact or 'negative' impact within a field or sub-discipline – i.e. work that 'advances' a debate, or work that makes it more simplistic and polarised. Even where there is some inclusion of 'positive' or 'negative' evaluation, such crude forms of voting miss the complexities of much scholarly work (such as where others might find the empirical discussion useful, but reject the theoretical framing or inferences drawn). Without such fine-grained information on the actual contribution of a piece to a debate, it would be very short-sighted to suggest that aggregate citations are any grounds for awarding further funding or prestige. Indeed, the overall pressure it creates is simply to get cited at all costs. This might well lead to work becoming more provocative and outrageous *for the sake of citation*, rather than making more disciplined and rigorous contributions to knowledge.

9. Moreover, we must be clear to differentiate between this kind of academic 'impact' and the public 'impact' sought in terms of the present REF case studies. Citations can tell us about academic citations – themselves a mixture of good and bad – but they *can tell us very little about the public engagement and contribution made by particular pieces of work for non-academic communities in society.* To the extent that the 'impact case studies' in the REF genuinely seek to open the door for academic work to better engage with the society in which it is embedded, citation counts cannot be used as a way of judging this at all. This is especially the case where academics are trying to work with small-scale and grassroots

organisations rather than governments or international organisations. Wider forms of alternative metrics like number of social media shares extend the definition of impact, but are also likely to be driven by controversy, and are even less likely to reflect the underlying *academic* quality of pieces (since the audience is generally less expert that for scholarly citations).

Metrics and the measurement of research quality

10. It should be further evident that because of what citation counts actually measure, these are not an appropriate proxy for research quality. The current REF asks its panel members to apply criteria of 'originality, significance and rigour'. These are broadly the same kind of criteria that expert peer reviewers apply when reviewing book manuscripts or journal articles.

11. On 'originality' – work *may* be cited because it is original, but it may also be cited because it is a more famous academic making the same point. Textbooks and edited collections are widely cited because they are accessible – not because they are original. Moreover, highly original work may not be cited at all because it has been published in a lower-profile venue, or because it radically differs from the intellectual trajectories of its sub-field. There is absolutely no logical or necessary connection between originality and being cited.

12. On 'significance' – 'significance' also seems to imply the need for broad disciplinary recognition of the contribution. To this extent, we might expect 'significant' work to have a high citation count; *however having a high citation count does not mean that the work is 'significant'.* In addition, using citation counts will systematically under-count the 'significance' of work directed at more specialised sub-fields or technical debates, or that adopts more dissident positions. Moreover, when understood through the problems discussed in point 8, it becomes clear that 'significance' can be a distinctly ambiguous category for evaluating research quality. If we understand 'significance' as 'academic fame' then there is some kind of link with citation counts. However, if we understand

‘significance’ as ‘the development of the intellectual agenda of the field’ (REF panel criteria), then citation counts are not an appropriate proxy. In addition, as is well-known, in fields with long citation ‘half-lives’ – particularly arts and humanities, present research assessment cycles are far too short for the ‘significance’ of the work to emerge within citation counts, if it was going to do so.

13. With regard to ‘rigour’, there is also no necessary connection between citation counts and this aspect of research quality. To the extent that citation counts in part depend on how widely-read a journal is, and to the extent that widely-read journals may apply exacting peer review standards, and to the extent that these peer reviews are focused on the ‘rigour’ of a piece, there is again a potential or hypothetical link between a citation count and ‘rigour’. However, there are a *lot* of intervening variables within here, not least those discussed in point 6, which would disrupt the relationship between the number of times a piece is cited and how rigorous it is. To the extent that more ‘rigorous’ pieces may be more theoretically and methodologically sophisticated – and thus less accessible to ‘lay’ academic and non-academic audiences, there are reasons to believe that the rigour of a piece might well be *inversely* related to its citation count. To summarise, citation counts are not a reliable indicator of rigour.

14. Overall then, upon close examination the relationship between citation counts and our historic and current definitions of academic research quality is *extremely* weak in logic, and problematic in practice. Notwithstanding that in certain disciplines the practice of using citations as a proxy for quality as taken hold, the practice is itself fundamentally flawed and should not be encouraged, much less institutionalised within national, international or institutional research assessment contexts.

15. That REF panellists and other academics may informally use the reputation of a journal as a quick means of judging a piece on which they are unable or unwilling to provide detailed expert opinion does not mean that this is a good idea. One argument for the use of metrics has been that quantitative and

qualitative measures sometimes mirror each other. However, this may be explained often by the fact that qualitative assessments often themselves take place under flawed conditions which do not entail double-blind peer review; rather the review of pieces in which one already knows the author and the publishing outlet tends in practice to lead to shortcut decisions which confirm prejudices – and not academic judgements – reflected in citation counts.

Potential consequences of using citation metrics as an indicator for research impact and/or quality in research assessment

16. Whilst our concerns are with the basic logic of attempting to use citation counts as a proxy for research quality and impact, there are also a number of troubling potential consequences of research assessment of moving in this direction as a widespread practice. We focus here on problems of inherent conservatism, structures of academic discrimination and emerging practices of gaming/manipulation, although this is a non-exhaustive list.

17. If we use metrics as a mode of assessing research quality or impact, we potentially introduce a further *conservative bias into the field by favouring the work of already-famous scholars.* Whilst they may be famous for an ongoing and productive research agenda, they may also be famous generally for work produced many years ago which has generated a lot of citations. This is an indication of 'reputation' in general, but for the purposes of choosing who/what to fund or one's professional contribution, this introduces further prejudices against less established scholars, who really do need to compete on a level playing-field in terms of the quality of their ideas and findings. This will over time lead to a greater concentration of research funding and prestige in a smaller circle of people – not the most innovative researchers.

18. This problem of conservatism is compounded when we look at the systematic under-citation of women and minority groups. Recently, the large international TRIPS survey found evidence of a massive bias against citing women in the field

of International Relations.[1] We also know that academics more generally carry sexist and racist biases as evidenced through experiments for hiring processes and judging academic quality.[2] Reasonably assuming that these prejudicial attitudes drive citation count differences as well, the move to metrics and away from peer review processes would compound (or at the very least mirror) the effects of these prejudices and embed them into research assessment.

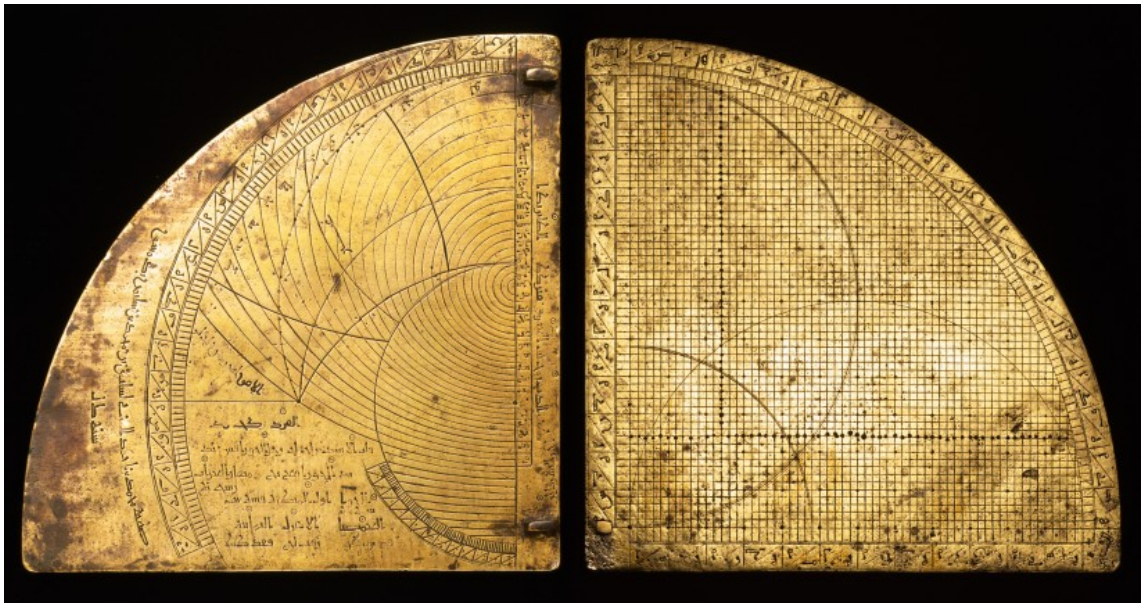
19. The last issue to consider is the gaming of citation counts. It has been demonstrated already effectively that GoogleScholar can be gamed with ease and with dramatic effects.[3] One counter-argument is that other metrics are harder to game, and that companies like Thomson Reuters police issues such as self-citation in their journal rankings. We do not argue that existing methods for measuring research quality are pure, or desirable. However, once systematic gaming sets in, it is increasingly difficult for any ranking system to keep itself 'clean'. As long as GoogleScholar remains game-able – and Google have not shown any interest in trying to change that, following its commercial model – then it will also affect any 'clean' rankings, as people using Google to look for references will be presented with Google's top articles and works first. In turn, this is likely to generate more 'real' citations for a piece based on its gaming of the GoogleScholar rankings. The closer the link between citations or altmetrics and assessments of quality, the greater the incentive for academics and their managers to game those metrics. In and of itself this should be a huge reason against using citation counts as a means of assessing research in any meaningful and serious way.

Conclusions

20. Overall, the academic community as a whole should resist the adoption of citation metrics as a means by which to make conclusions about either research impact or research quality. They are not logically connected to either issue, contain systematic biases against different researchers and are all too easily

manipulated, particularly by corporate rankings providers. They should certainly not become institutionalised in national, international or institutional practices.

21. It is, of course, difficult and time-consuming to assess academic research by having experts read it and carefully evaluate it against complex and demanding criteria, ideally under conditions of anonymity. That is as it should be. That is the whole point about good academic work and this cannot be automated or captured by present, or even future, citation counts. Simply because the market produces products, and because some people use them, does not mean that these are the things that we actually want or need for the purposes we have in mind. If we really are committed to using research assessment practices to fund the best quality, most innovative and most publicly engaged work, then citation counts are not the way to do it. Rather, we will end up funding not just those whose work is genuinely transformative, original and field-defining (assuming these qualities earn them high citations), but those who are best at self-promotion and rankings manipulation, and who are privileged by existing structures of prejudice.



[1] Daniel Maliniak, Ryan Powers and Barbara F. Walter (2013). The Gender Citation Gap in International Relations. *International Organization*, 67, pp 889-922. doi:10.1017/S0020818313000209. See open version here: http://politicalviolenceataglance.files.wordpress.com/2013/03/the-gender-citation-gap-in-ir.pdf

[2] Ilana Yurkiewicz, (2012) 'Study Shows Gender Bias in Science is Real. Here's Why It Matters', *Scientific American*, available at http://blogs.scientificamerican.com/unofficial-prognosis/2012/09/23/study-shows-gender-bias-in-science-is-real-heres-why-it-matters/; April Corrice (2009) 'Unconscious Bias in Faculty and Leadership Recruitment: A Literature Review', Association of American Medical Colleges, available at this location.

[3] Phil Davis, (2012), 'Gaming Google Scholars Citations: Made Simple and Easy', *Scholarly Kitchen* blog. Available at: http://scholarlykitchen.sspnet.org/2012/12/12/gaming-google-scholar-citations-made-simple-and-easy/