STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics?

Anne-Wil Harzing*

*anne@harzing.com
Middlesex University Business School, London, The Burroughs, Hendon, London NW4 4BT

Introduction

This paper examines the feasibility of replacing the resource-intensive and contested expert review panels in the British national research evaluation – called REF (Research Excellence Framework) – with a much simpler way to distribute Quality Related (QR) research funding, based on metrics. I am by no means the first to argue ditching peer review in current REF for this purpose.

> *"If metrics-based measures can produce much the same results as those arrived at through an infinitely more costly, laborious and time-consuming process of 'expert review' of individual outputs, there is a compelling reason to go with the metrics; not because it is necessarily a valid measure of anything but because it as reliable as the alternative (whose validity […] is no less dubious for different reasons) and a good deal more cost-efficient." [Sayer, 2015:89]*

Neither am I the first to conduct empirical research on the extent to which the results of metrics match those of peer review in the REF. For a detailed literature review see the Metrics Tide report (Wouters et al. 2015); a few key examples can be found in *Scientometrics* (Mryglod et al. 2013) and *British Journal of Management* (Taylor, 2011). These studies, however, typically used limited samples and employed fairly time-consuming data collection methods. They also relied on subscription-based databases such as the Web of Science or Scopus that have limited coverage in the Social Sciences and Humanities and, to a lesser extent, Engineering (Harzing & Alakangas, 2016).

Most recently Mingers, O'Hanley & Okunola (2017) used a freely accessible database – Google Scholar – to create a university ranking based on citations drawn from Google Scholar Citation Profiles. Although inventive, their method is still fairly time-consuming and relies on academics having created a Google Scholar Citation Profile. Uptake of these profiles is by no means universal and might be heavily dependent on university policies promoting their usage, thus leading to unbalanced comparisons between universities.

In this study, I therefore use Publish or Perish (PoP) (Harzing, 2007) to source citation data from another freely accessible database – Microsoft Academic (MA) – which has been shown to have a comprehensive coverage across disciplines (Harzing & Alakangas, 2017) and – unlike Google Scholar – allows for searching by affiliation and research field. Citation data from MA were subsequently compared with the 2014 REF power rating of universities.

Methods

My aim was to assess the use of metrics as an alternative to peer review to distribute QR research funding. I thus chose a size-dependent dependent variable, i.e. the REF power rating, rather than a size independent dependent variable, such as the REF Quality rating.[1] The 2014 REF power rating used Research Fortnight's calculation. Research Fortnight first created a quality index, reflecting the funding formula with 4* research counting three times as much as 3* research, which was then weighted by staff numbers. They then created a power ranking by converting the quality index with the best performing university (Oxford) scored 100.

Citation data were collected through a PoP MA search in July 2017[2], including the name of the university in the Affiliation field and restricting the publication years to 2008-2013 [the period covered for the 2014 REF exercise]. As Business Schools often have a name that is distinct from the university, I constructed OR queries such as "University of Oxford OR Said Business School". Universities that were entered in only one Unit of Assessment (UoA) as well as specialised institutes such as the Institute of Cancer Research were excluded. This left me with 118 universities.

Although PoP calculates a wide variety of metrics, including various metrics dependent on the number of publications, I decided to focus on citation-based metrics. Publications are a sign of productivity, but this doesn't necessarily translate into research quality. Although citations are not a perfect measure of research quality, they are a generally accepted proxy. I thus collected all citations for the 2008-2013 period, a fairly time-consuming process [app. 10 hours] as the number of publications amounted to well over a million. Minimal data cleaning was conducted, concentrating on the most cited publications. This involved removing book publications that had inflated citation counts, as all citations had been attributed to the latest edition, and removing two duplicated highly cited publications for a few universities.

However, as citation data tend to be highly skewed, I also collected data for each university's top 1,000 publications. Once queries had been defined, this process took little more than half an hour! For completeness sake, I use the total number of publications in this study; obviously focusing on the top 1,000 publications only would significantly reduce the time spent on data collection. Regardless of whether I used only the top 1,000 publications or the total number of publications for each university, the correlations reported below were almost identical.

Methodological differences between REF peer review and MA citation ranking

Both rankings evaluate publications. However, while the REF ranking does so through peer review of these publications, my MA ranking focuses on citations to these publications instead. There are eight other ways in which the two ranking approaches differ.

1. REF assesses not just publications, but also non-academic impact – through impact case studies – and the research environment, through a combination of metrics and narrative. My approach only looks at [citations to] publications.
2. REF requires academics to be submitted in a particular disciplinary area (UoA). My university-wide citation ranking doesn't necessitate any disciplinary choices as it evaluates all of the university's output.

[1] Another reason for this choice was that the Quality rating was subject to a substantial level of "gaming": some universities only submitted an infinitely small proportion of their staff to maximise their Quality rating. Even so, correlations between the Quality rating and another, slightly less size-dependent, citation metric (the hI,annual) are very high at 0.89.

[2] I repeated the data collection mid July 2018 for the top 1,000 publications only; the results in terms of correlations and mean difference in rank were virtually identical.

3. REF only includes publications by academics selected by their institution for submission to the REF. My approach includes publications by *all* academics in the institution.
4. REF only includes academics employed at the institution at the census date. My approach includes all publications that carry the university's affiliation. If academics move institutions, or are given a fractional appointment at an institution just before the census date, this influences the REF ranking, but not my citation ranking.
5. REF includes a maximum of four publications for each academic. My approach includes *all* publications for each academic.
6. REF output included mostly journal publications. My approach includes *all* publications covered in the MA database. This means a non-negligible number of books, book chapters, conference papers, and software are included (Harzing & Alakangas, 2017).
7. REF allowed submission of publications that were accepted in 2013 (or earlier), even if they were published after 2013. My approach only includes articles that were actually published between 2008 and 2013.
8. REF was conducted in 2014. My approach used citation counts as of July 2017.

REF 2021 will differ from REF 2014 in that it will include *all* academics [point 3] and that publications will stay with the institution, though with some transitional arrangements [point 4]. As a result, the rules for REF 2021 are closer to my citation-based approach than the rules that were applied for REF 2014. Thus, we can expect any positive correlations between the REF peer review ranking and a MA citation ranking to be stronger for REF 2021 than for REF 2014.

Results

The correlation between the REF power rating and the total number of MA citations is 0.9695. This correlation, however, might have been partly driven by the extremely high scores at the top of the ranking. Using rank correlations instead, the correlation *does* decline, but only by 0.001 to 0.9685. Figure 1 shows the regression plot for the correlation between the REF Power Rank and the MA Citation Rank. Most universities cluster around the regression line and the average difference in rank is only 6.8 places (on a ranking of 118 universities).

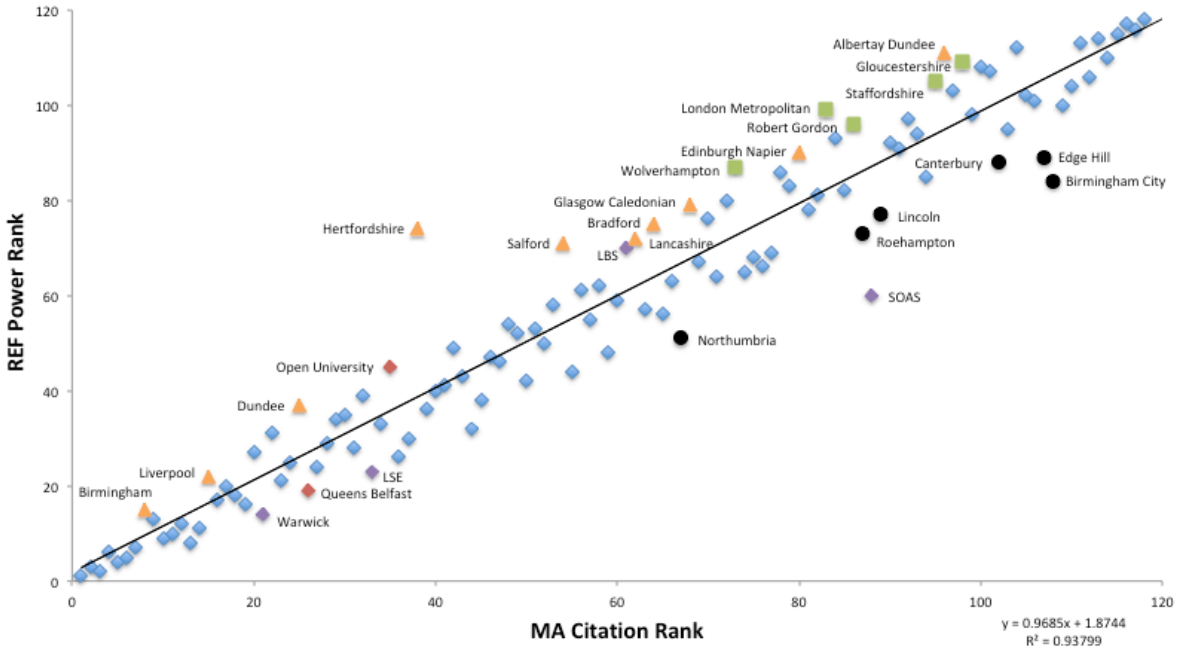
Individual outliers

Although most universities cluster around the regression line, there are some notable deviations. Apart from the possibility that either peer review or metrics provides a superior way to measure the underlying construct, these deviations are caused by three key categories of problems. The first relates to problems with the MA data, which in turn lie in three areas.

- First, MA doesn't correctly identify the publications of two universities in our sample: Queens University Belfast and the Open University. The former sees quite a lot of their highly-cited papers attributed to Queens University (Canada). The latter experiences the opposite problem, a substantive portion of their highly-cited papers were in fact produced by the Dutch or Israeli Open University.
- Second, although the MA attribution of affiliation to academics has improved tremendously since my first test in February 2017, it is not perfect. There are still papers where some of the authors do not have an associated affiliation.
- Third, MA – like Google Scholar – generally aggregates citations for books to the latest edition of a book and sometimes attributes citations to a review of the book. For the most-cited books this was resolved manually by removing these publications.

Whilst these problems are unfortunate, they are eminently solvable. Since its launch, MA has been actively improving their matching algorithm. Hence, I do expect these MA-related problems to become less and less prominent.

Figure 1: REF power rank by MA citation rank



A second category of problems mainly involves post-92 universities, composed of two distinct groups. First, there is a group of universities (marked with a circle in Figure 1) that is fairly highly ranked in the REF without having a correspondingly high level of citations. These universities might have had relatively high scores for (societal and policy) impact thus leading to a better REF ranking than citation ranking. The fact that they have generally improved their REF ranking substantially since 2008 (when impact was not included) seems to point in that direction. A second group of universities (marked with a square in Figure 1) is ranked higher on citation count than on the REF rating. These discrepancies might have been caused by a "small numbers game" in terms of staff, publications, and citations. As a result, individual idiosyncrasies – such as highly cited textbooks or one highly-cited academic – have a disproportionate impact and might lead to substantial volatility in the rankings, especially given that differences between universities in the lower regions are very small.

The third category relates to universities that employ one, or a small group of, academic(s) participating in huge consortia, doing research in for instance particle physics or gene technology. Publications resulting from these collaborations might have over a thousand authors and a large number of citations. Although these publications might have been highly ranked in the REF, they would have made up only a small proportion of the institution's REF submission. In contrast, they are likely to represent a disproportionate share of citations, especially for smaller institutions. Universities marked with a triangle in Figure 1 all share this problem to varying extents. A solution might be to remove mega-authored papers from the comparison. THE decided to do so after seeing a fairly obscure university storm up their rankings, simply because of one individual staff member's involvement in this type of papers.

A related problem is the more general issue that citation practices differ by discipline. Citation levels tend to be much higher in the Sciences and Life Sciences than in the Social Sciences and Humanities, with Engineering falling between these two extremes (Harzing & Alakangas, 2016). Thus, universities that have a heavy concentration in the (Life) Sciences (marked with a triangle in Figure 1) are likely to have higher citation levels of than universities who have a strong presence in the Social Sciences and Humanities. The School of Oriental and African Studies (SOAS) and to a lesser extent the London School of Economics (LSE) clearly suffer from this, as does Warwick, a university whose strong presence in Physics is counteracted by a strong presence in the Social Sciences. To appreciate the effect this might have, consider for instance Warwick and Birmingham, who are ranked similarly (#14 and #15 respectively) on REF rank. On citation rank, however, Birmingham (#8) substantially outranks Warwick (#21). In their study, Mingers, O'Hanley & Okunola (2017) signalled the same problem and applied a correction for disciplinary composition; this catapulted LSE to the top of their ranking. In the next section, I will suggest an alternative correction.

Disciplinary analysis

So far, my analysis has been at the level of the university, whereas the REF was conducted by discipline. QR funding, however, goes to universities, not to disciplines, and certainly not to individual researchers. If the outcome variable is defined at the university level, why do we go through the tremendous effort of evaluating universities by discipline? Even worse, why do we insist on evaluating individual academics for REF entry? As Sayer argues:

> *Metrics will not allow us to drill down to individuals (or possibly even to UoAs) with any degree of validity, but they do not need to if the REF is merely a funding mechanism. Any such additional information – and the time, money and effort spent in gathering it – would be superfluous to what is required. (Sayer, 2015: 89)*

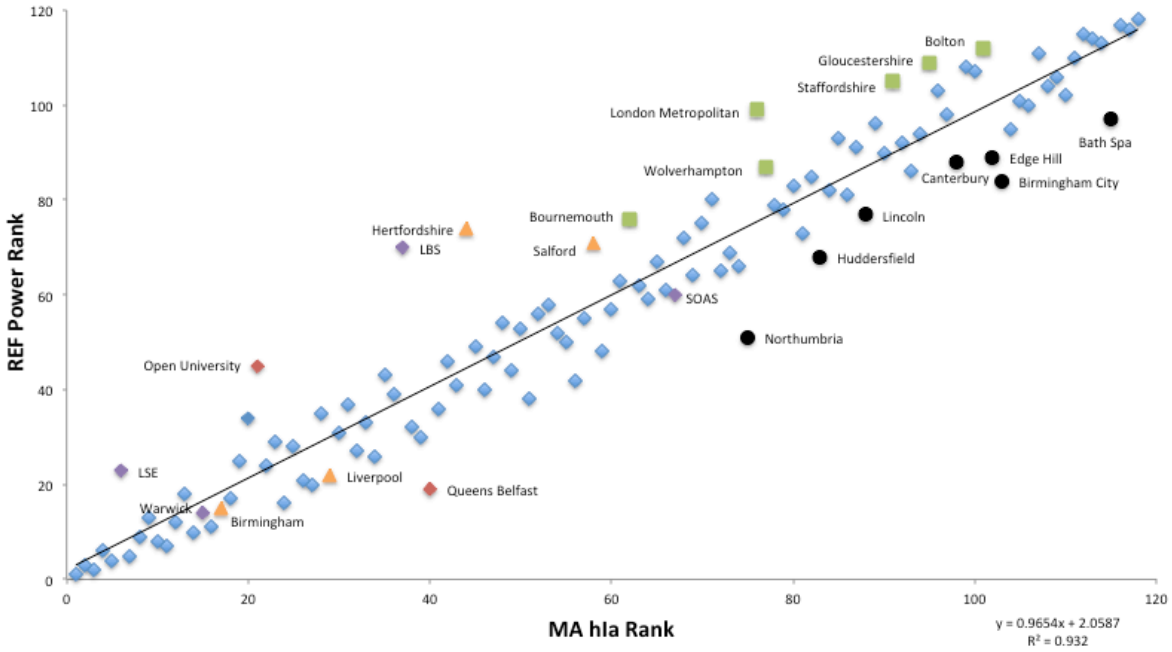
However, if we do want to conduct a disciplinary analysis, this would be as easy as adding a field label to the analyses conducted above. To test the feasibility of this suggestion, I conducted field-level analyses for Business & Management, Chemistry, and Computer Science. Rank correlations for Chemistry and Computer Science, for which I only conducted marginal cleaning, were 0.94. In my own discipline (Business & Management), for which I conducted more substantive cleaning, the rank correlation was 0.97. So clearly running this analysis at a disciplinary level - if so desired - is feasible.

If we wanted a comparison across universities as a whole, but take the differential disciplinary mix into account in order to provide a fairer comparison, I suggest using a metric that corrects for the number of co-authors, a crude, but quite effective, way to address disciplinary differences in citation counts. One of these metrics, the hIa (Harzing, Alakangas & Adams, 2014) reports the average number of impactful single-author equivalent publications a university publishes a year. As Figure 2 shows, a ranking based on the hIa-index correlates with the REF power rank at 0.9654. For many universities a hIa ranking reduces deviation from the REF rank when compared with a rank based on raw citations; the average difference in rank declines from 6.8 to 6.4.

Using the hIa, universities with a higher rank for citations than for REF power, with high citation levels caused primarily by their concentration in the (Life) Sciences move closer to the regression line. For instance, Warwick and Birmingham now score very similarly on REF power (#14 and #15) and on metrics (#15 and #17). The disciplinary correction has brought their metric rank very close to their REF rank. LSE and Liverpool, however, show a complete

reversal of fortunes. Whereas for raw citations LSE ranked 18 places *below* Liverpool, on discipline corrected citations it ranks 23 places *above* Liverpool. The actual size of the disciplinary correction is up for discussion. The current default setting for PoP is to take up to 50 authors into account, but this can easily be changed in the preferences. Most importantly, our results show that – if so desired – a disciplinary correction can be applied easily.

Figure 2: REF power rank by MA hIa rank



Conclusion: peer review or metrics?

Whenever metrics are proposed as an alternative to peer review, academics are quick to point out the manifold flaws of metrics. In principle, I agree with nearly all of their reservations. However, as I have argued before, and with me many others, these arguments usually compare a "reductionist" version of metrics with an *idealised* version of peer review rather than a "responsible and comprehensive" version of metrics with the *reality* of peer review, i.e. hurried semi-experts, with assessment potentially influenced by journal outlet, institutional affiliation and demographic characteristics.

When using citation analysis our publications are evaluated by thousands of peers, including international peers. Collecting citations in the way proposed in this study is also transparent and replicable in less than an hour by everyone with a computer, Internet access and the free PoP software. In contrast, in the current REF process, our academic work is evaluated by a select group of academics: those that have volunteered to serve on the REF panels. Some of these volunteers have had to read more than a 1,000 articles, often outside their immediate area of expertise, collectively spending over a 1,000 years of productive research time. They also have to "burn their papers" after the panel meeting.

I thus suggest that metrics should be considered seriously as an alternative to peer review, a conclusion supported by a very recent study by Pride & Knoth (2018) compared institutional GPA (app. Quality rating) with citations at the UoA level and concluded that *"citation-based indicators are sufficiently aligned with peer review results at the institutional level to be used to lessen the overall burden of peer review on national evaluation exercises leading to consi-*

derable cost savings". This doesn't mean that we should use metrics to evaluate individuals, although even there they are sometimes preferable to peer review, especially in countries suffering from academic nepotism. Neither does it mean that we should give up completely on evaluating research quality through peer review. But letting metrics do the "heavy lifting" of allocating research funding frees up a staggering amount of time and resources that would allow us to come up with more creative and meaningful ways to build in a *credible* quality component in the British national research assessment.

References

Harzing, A.W. (2007). *Publish or Perish*, available from http://www.harzing.com/pop.htm

Harzing, A.W., Alakangas, S. & Adams, D. (2014). hIa: An individual annual h-index to accommodate disciplinary and career length differences, *Scientometrics*, 99(3), 811-821.

Harzing, A.W. & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, *Scientometrics*, 106(2), 787-804.

Harzing, A.W. & Alakangas, S. (2017). Microsoft Academic is one year old: the Phoenix is ready to leave the nest, *Scientometrics*, 112(3), 1887-1894.

Mingers, J., Hanley, J.R.O. & Okunola, M. (2017). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*, 113(3), 1627-1643.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence, *Scientometrics*, 97(3), 767-777.

Pride, D., & Knoth, P. (2018). Peer review and citation data in predicting university rankings, a large-scale analysis. *arXiv preprint arXiv:1805.08529*.

Sayer, D. (2015). *Rank hypocrisies: The insult of the REF*, Sage Publications.

Taylor, J. (2011). The Assessment of Research Quality in UK Universities: Peer Review or Metrics?, *British Journal of Management*, 22(2), 202–217.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The metric tide. Literature review. Supplementary report I to the independent review of the role of metrics in research assessment and management*. HEFCE, London.