Influencing the changing world of research evaluation

This article provides an overview of the current research evaluation landscape, the problems inherent within it and some of the ways a newly formed INORMS (International Network of Research Management Societies) working group on research evaluation is seeking to address them. These include developing a way of rating university ranking bodies and developing a set of briefing materials on responsible research evaluation for senior managers.

Keywords
Research evaluation; bibliometrics; responsible metrics; INORMS

Whatever scholarly forum you find yourself in at the moment, the critical issues usually boil down to problems with research evaluation. Want to encourage greater take-up of open access? Change the way research is evaluated to prize open scholarship. Want to improve the reproducibility and rigour of research? Change the way research is evaluated to value high quality data and negative results as much as sensational findings more likely to be welcomed by ‘top’ journals. Want to improve equality and diversity in academia? Change recruitment and promotion criteria to include a wider range of qualities, skills and outcomes – not primarily cash and citations. These are all sensible conclusions. The way research is evaluated has a significant effect on the whole of the scholarly endeavour. This is called Campbell’s Law: often summarized as ‘the way you measure someone affects the way they behave’. However, identifying the problem is far more straightforward than solving it – especially when it is so deeply embedded in existing, global, scholarly cultures. This short commentary seeks to provide an overview of the current research evaluation landscape, the problems inherent within it and some of the ways a newly formed INORMS (International Network of Research Management Societies) working group on research evaluation is seeking to address them.

More evaluation

The first thing to say about the changing world of research evaluation is that there is just more and more of it. Not a week goes by without another global, national or subject ranking being published and these are mainly based on research indicators as they are the only consistent, global data sources we have. There is also an increase in national research evaluation exercises – both in the volume of them (most EU countries now have some form of national exercise) and the sophistication (read cost). An increasing number are also linked with funding.

An increase in external evaluation leads predictably to an increase in internal evaluation. We see universities’ research ambitions stated and evaluated from whole university key performance indicators through to departmental targets and, of course, most controversial of all, individual-level evaluations through annual appraisal, recruitment and promotion. Although academics complain most bitterly about the latter, they are often equally complicit in the evaluation culture by engaging in what I call ‘measure for pleasure’ activities: watching their h-index creep up and comparing themselves to peers – or, should I say, competitors?
More competition

An evaluation culture necessarily leads to an increase in competition. As soon as you can put a number on something, you can rank it. And, where there is a ranking, you can bet a university, department, or academic will want to compete to get to the top of it. Indeed, they may feel as though they have little choice – their finances and reputation depend on it. Of course, once you have made it to the top, a whole new level of anxiety kicks in as you seek to stay there. It is one thing to be a university on the rise, quite another to be a university in decline.

National research evaluation exercises also keep raising the bar as to what level of brilliance is needed to win prizes. After the 2014 UK Research Excellence Framework (REF), four-star (world leading) research was funded at a ratio of 4:1, with three-star (internationally excellent) research perhaps inadvertently sending a message that ‘excellence is not good enough’. For REF 2021, due to the change in methodology (submit all papers, but with an average of 2.5 papers per person), there is speculation that four-star research will get a much higher proportion of the available funding and that perhaps three-star research will not get funded at all.

In other forms of competition, we repeatedly hear in the research press that there are more and more researchers chasing fewer and fewer funding opportunities, and that there are too many PhDs for the academic jobs available. Indeed, in a blog post earlier this year, Imperial College professor Chris Jackson stated that were he attempting to move into academia from industry now, ‘there is no way I’d get a job’.

More at stake

Not surprisingly, with more competition, there is more at stake. A university’s ranking position has always (inevitably) been linked to its reputation and indirectly to funding through students who may or may not choose to attend that university based on their ranking. However, rankings are now directly linked to funding as more and more global funders are using university ranking position as a barrier to entry. Financial institutions also seem to be using a university’s ranking position as shorthand for credit-worthiness.

There is also more at stake for academics now with, at best, recruitment and promotion based heavily on measurable publication performance (how many, what journals, who co-authored, how cited) and, at worst, direct cash prizes associated with publications in ‘top journals’. According to a People’s Daily news story last year, some 90% of Chinese universities have policies of rewarding publications, and the practice is far from unique to China. (However, some Chinese funders are now seeking to back Plan S, which may change this situation).

Of course, this sort of activity puts pressure on journals to become ‘top journals’ and we are now seeing artificial intelligence services such as Meta Bibliometric Intelligence offering publishers the opportunity to predict ‘high impact’ publications based on historical publication and citation data in order ‘to control the future impact factor of the journal’.

More pressure on the academy

Inevitably, all this evaluation-based competition has led to more pressure on the academy. A frequently cited case is that of Professor Stefan Grimm, formerly of Imperial College London, who committed suicide allegedly after he was told he was not meeting the expected income-generation target of £200,000 p.a. Over the summer of 2018 another academic, at Cardiff University, took his own life due to pressure of work. Higher education is a sector with a known, significant mental health problem, and when we consider appropriate forms of research evaluation, this can never be far from our minds.
More calls for ‘responsible metrics’

Having outlined all that, it is hardly surprising that there have been increasing calls for so-called ‘responsible metrics’. In 2012 a group established by the American Society of Cell Biology wrote the Declaration on Research Assessment (DORA). This is principally a backlash against the use of Journal Impact Factors (JIF) to evaluate individual scholars and individual articles. Both individuals and institutions are invited to sign the declaration. In April 2015 a group of scholars from the CWTS research group at Leiden University published The Leiden Manifesto. This is formed of ten principles for the responsible use of bibliometrics in research evaluation, so a much broader focus than DORA. Perhaps of most significance to the UK was the publication of The Metric Tide in July 2015, which was an independent review of the use of metrics in UK research assessment. This established five principles for the responsible use of all metrics (not just bibliometrics) and called on universities to establish their own principles for the use of research metrics.

The Lis-Bibliometrics Forum has run annual ‘state-of-the-art’ surveys around responsible metrics since 2015 to map the sector’s engagement with these calls. Figures 1 and 2 show the uptake of DORA and the development of responsible metrics principles by higher education institutions (HEIs) over time. They show fairly equal levels of engagement with both. Thirty have signed and six are likely to sign DORA (36 in total), whilst 11 already had, and 24 were currently developing, their own responsible metrics principles (35). It is interesting to note that more had actually signed DORA than had already developed their own principles. This may be because it is easier to sign an existing protocol than develop your own. However, the tide appears to be turning, with fewer actively considering DORA in 2018 (19) than actively considering their own principles (33). Similarly, a higher proportion had rejected DORA by 2018 (11) than had rejected the idea of developing their own principles (3).

Figure 1. Engagement by HEIs with DORA over time

Figure 2. HEI development of responsible metrics principles over time
Under Plan S it may be that funders seek to push this agenda even further by mandating engagement with responsible metrics. Indeed, the Wellcome Trust has recently announced that as part of its new open access policy guided by Plan S, from 1 January 2020 it will only fund organizations that have made a public commitment to responsible research evaluation. It is unclear as yet how other funders will interpret the cOAlition S requirement for members ‘to sign DORA and implement those requirements in their policies’.

**Who is responsible for responsible metrics?**

It should be noted that most of the pressure to deliver on responsible metrics seems to be focused on universities. However, the truth is that there are many other players in this space. Universities measure because they are measured.

In the UK, one of the biggest external evaluators is the REF. The REF prides itself on being a peer-review exercise, and not a controversial metrics-based exercise such as that in Italy. Citation metrics are supplied to some panels, but essentially it is based on peer review. However, whilst peer review is held up as the ‘gold standard’ of research evaluation, evidence suggests that peer review can be just as problematic as metrics in terms of bias and poor practice. After the 2014 REF, Professor Alan Dix, who sat on the Computing panel, ran a citation-based correlation with the peer-review ratings in his panel and found significant bias against female authors and applied papers. He concluded that ‘metrics are rubbish, but people are worse’. Perhaps the focus on responsible metrics should expand to include responsible peer review?

Of course, when it comes to reporting the findings of this big qualitative peer-review exercise, the outcomes are 100% quantitative: sheets and sheets of numbers – 120,000 data points, to be precise. Thus, if anyone wants to assess the quality of civil engineering research at Loughborough, for example, they can only find out in a quantitative way. Is that responsible?

Secondly, we have research funders. A 2017 study sought to understand what the level of understanding was of bibliometrics within National Institute for Health Research panels. Responses included ‘I am not at all an expert in bibliometrics; I just have a general idea of what it is’. In fact, most panel members described their understanding of bibliometrics in similar terms, as ‘rudimentary’, ‘cursory’ and ‘limited’, even though they were using bibliometrics to inform funding decisions. Interestingly, even if funders sign up to Plan S and seek to evaluate outputs based only on their intrinsic merit rather than by journal brand, unless they eradicate the use of bibliometrics completely we may still see situations like this. Is that responsible?

**Tackling the world university rankings**

Then we have ranking organizations. There is no shortage of evidence for the lack of methodological credibility for the rankings. Indeed, any research study which claimed to be able to identify the ‘top’ or ‘best’ universities would be rejected at peer review for a badly formed research question. (Top at what?) The truth is that there is no governance around the activity of rankings and, considering the power they have (as outlined above), is that responsible? Indeed, this is one of the greatest challenges facing research evaluation simply because of their reach. National research evaluation exercises can be tackled at national level. Funders are also, to an extent, regional. However, the global university rankings are, well, global.

It is for this reason that the INORMS Research Evaluation Working Group has decided to try and tackle this problem through one of its work packages. Attempts have been made before to expose the flaws in the world rankings, but these have tended to be done by national groups, not by an international collective. The Group plans to develop a scoring mechanism for ranking
organizations – perhaps treating them to a ranking of their own. It is hoped that by joining forces to call for more critical understanding of the rankings, we may see fewer universities outsourcing their values to them.\textsuperscript{10}

**More data and indicators**

Given the growth in evaluation activity, it should not come as a surprise that the metrics industry has grown alongside it. In addition to the standard range of bibliographic databases (Web of Science and Scopus), all of which now include citation metrics, we have altmetric services (Altmetric.com and Plum Analytics) and dedicated citation benchmarking tools (Incites and SciVal). In 2018 alone we also saw three new tools: Dimensions,\textsuperscript{31} 1finder\textsuperscript{32} and Wizdom,\textsuperscript{33} all of which contain citation indicators and, increasingly, other forms of data such as grants, patents and clinical trials.

In addition to the increase in data sources, we are seeing bibliometric scholars develop more and more indicators. In 2014 Lorna Wildegaard and colleagues analysed 108 individual author metrics.\textsuperscript{34} The vast majority of these had been developed post-2000. We see a similar picture with journal metrics. The 2016 *Handbook of bibliometric indicators* lists in excess of 50 journal citation metrics.\textsuperscript{35} However, whilst we are seeing huge growth in the number of metrics being developed, we are seeing only a small increase in the number of metrics available to us in our citation tools. This might be a good thing in terms of overload, but all of these indicators are the result of many research projects, and who knows how much funding, to try and create something better than we have. Some of them may be better than the indicators we have, but we would never know about it because they are not adopted by our citation tools. There is a question then as to how metrics get into our products. Indeed, a recent study into bibliometric scholarship concluded that ‘commercial providers have gained a powerful role in defining de-facto standards of research excellence without being challenged by expert authority’.\textsuperscript{36} There is also a question as to how these metrics get designed, and whether academics are ever involved in the development of the sticks used to measure them.

**More cost**

Of course, with more products comes more cost. It is difficult to get any figures from suppliers on the take-up of their products. However, I know from keeping a watching brief on this area that we are seeing a growth in subscriptions to citation benchmarking tools. Data from a survey of 114 ACRL libraries in the US found that 80% had a ‘faculty activity reporting (FAR)’ tool.\textsuperscript{37} We are also seeing growth in the number of research impact or analytics-type posts in universities (mostly in libraries)\textsuperscript{38} and more recently (as universities are struggling to recruit) I have observed an increase in the seniority of these posts. Again, the ACRL library survey found that 70% had an individual or group assigned to research impact.\textsuperscript{39}

**More skills**

With all these tools and people, we are seeing more and more sophisticated analyses. A Lis-Bibliometrics survey which asked the community what messages they would like to send to suppliers reported many calls for open data, increased download limits and standard identifiers.\textsuperscript{40} This was to enable them to combine data from their citation tools with other data in visualization tools such as Tableau and PowerBI and network analysis tools such as VosViewer. They are also keen to run their own statistical analyses and calculate confidence intervals (perhaps because suppliers don’t do it for them). All of these analyses require more skills. However, a survey in 2017 found that only 29% of librarians who were supporting their institutions with bibliometrics had actually covered bibliometrics in their LIS diploma or degree.\textsuperscript{41} One of the outcomes of that project was a set of bibliometric competencies by which practitioners can assess their own skill levels.
and seek to improve on any gaps. Perhaps not surprisingly, at a ‘Statistics for Responsible Bibliometrics’ course run by the Lis-Bibliometrics group, the largest tranche of delegates (50%) described their skills as ‘entry level’.

**Skilling the decision makers**

Of course, it is not just the analysts that need upskilling. As we saw from the NIHR analysis of grant reviewers, there is a distinct lack of skill amongst those relying on bibliometric data for decision-making and policy setting. Arguably, it is this group that has the most power to do harm through a lack of understanding. It is for this reason that the INORMS Research Evaluation Working Group has decided to make this the focus of its second work package and to develop a set of briefing materials for senior managers. The plan is to develop a standard, globally relevant, slide set that gives our policymakers a better understanding of the strengths and weaknesses of various research evaluation methods. Ultimately, responsible metrics is not just about our duty of care but about good decision-making. Decisions made using poor indicators, incomplete data, or inappropriate methods are simply not going to lead to good outcomes. All senior managers should care about that, and it is this group that the responsible metrics lobby most urgently needs to reach. As I have said before, although we use the expression ‘responsible metrics’, metrics cannot be responsible, only people can be responsible. Ensuring we skill up responsible people is our best bet for engendering a responsible research evaluation culture in the future.

You can keep up to date with the work of the INORMS Research Evaluation Working Group by joining their mailing list at: [https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=INORMS-RES-EVAL](https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=INORMS-RES-EVAL)

**Abbreviations and Acronyms**

A list of the abbreviations and acronyms used in this and other Insights articles can be accessed here – click on the URL below and then select the ‘full list of industry A&As’ link: [http://www.uksg.org/publications#aa](http://www.uksg.org/publications#aa)

**Competing interests**

The author has declared no competing interests.
References


8. Rauhvargers, Global University Rankings ref. 4.


