Metrics in Research Evaluation

Sarah Huggett ***Georgin Lau***

Abstract

A recent rise in the accountability of research programmes, driven by economic circumstances, makes research evaluation increasingly needed. At the same time, computational advances and the advent of the Internet have given rise to a variety of metrics. If used appropriately, and in conjunction with peer evaluation and careful interpretation, these can inform and enhance research assessment through the benefits of impartiality, comparability, and scalability. Three initiatives (the DORA declaration, the HEFCE review of metrics, and the Snowball metrics project) have recently made the call for metrics to be appropriately used in research evaluation at various levels. Research assessment scenarios are multiple and complex; circumstances may shift for the same entities in different roles. A few easy tricks can help make the best use of metrics in research evaluation exercises: the use of trends, benchmarks and comparators, normalisation to size or resources, geographic or subject breakdowns, triangulation of indicators, and last but not least vigilant consideration of context.

Keywords: Research Evaluation, Research Performance, Research Assessment, Metrics, Bibliometrics, Altmetrics and Indicators

1. Introduction

In an era of tightening budgets and scarce funding, the relevance of research evaluation has increased, and there is growing demand for metrics to inform the research evaluation process. Computational advances and the rise of the Internet mean that is now easier than ever to track, measure, benchmark, and analyse. Metrics can be a useful tool for research evaluation, provided they are calculated, chosen, used, and interpreted appropriately.

Research evaluation may have different meanings and implications for a student, a postdoc, a tenured researcher, a librarian, a provost, a government official, or a funding body employee, for instance. Each needs to evaluate research for different purposes, to varying degrees, at diverse levels, and therefore does so in different ways. Research assessment is a directional process, and one can in turn find oneself in an assessor or assessee role; research evaluation practices may also vary further depending on context.

In this paper we present an overview of various recent research evaluation scenarios and describe how the suitable use of metrics can help inform each of these.

2. Calls for Metrics in Research Evaluation

Peer review has long been recognised as one of the best research assessment tools, but it can be time-consuming or subjective. Metrics offer the advantage of scalability and allow systematic and impartial comparisons. Therefore, it has been advocated that both should be used in conjunction, , especially as they tend to reinforce rather than contradict each other.

Three recent initiatives to such effect have received particularly far-reaching attention:

1.1. San Francisco Declaration on Research Assessment (DORA)

This set of 18 recommendations originated from a group of editors and publishers of scholarly journals at the 2012 Annual Meeting of The American Society for Cell Biology (ASCB). It advocates appropriate use of metrics in research evaluation, with a particular emphasis on using metrics at the right level and explicit, open, and transparent communication about criteria, sources, and methodologies.

1.2. Higher Education Funding Council for England (HEFCE) Independent review of the role of metrics in research assessment

The review is conducted by an independent steering group chaired by Professor James Wilsdon, from the Science Policy Research Unit (SPRU) of the University of Sussex. It aims to review how metrics should be integrated in research evaluation. HEFCE issued a call for evidence to gather information; this call received 153 responses, offering a variety of views on the issue from numerous stakeholders.

1.3. Snowball Metrics Project

This bottom-up initiative, owned by research-intensive universities around the globe, aims to define approved metrics for research assessment. "The universities aim to agree on methodologies that are robustly and clearly defined, so that the metrics they describe enable confident comparison of apples with apples. These metrics are data source- and system-agnostic, meaning that they are not tied to any particular provider of data or tools. The resulting benchmarks among research-intensive universities provide reliable information to help understand research strengths, and thus to establish and monitor institutional strategies. The aspiration is for these metrics to become global standards that enable institutional benchmarking, and to cover the entire spectrum of research activities."

These three initiatives reflect a growing awareness of the issue beyond the usual circle of informed bibliometricians and research evaluators, and the increasingly widespread call for metrics to be appropriately integrated into research assessment exercises.

3. Examples of Research Evaluation Scenarios

Research evaluation is multi-faceted and can have many purposes, which directly influences what should be measured and how the data is collected and calculated. Depending on the research evaluation scenario (see Table 1), the assessor and assessee roles are played by different entities and measured by different metrics.

Governments need to evaluate their country's research at a national level (e.g. the International Comparative Performance of the UK Research Base – 2013 report, commissioned by the UK government Department of Business, Innovation, and Skills), or at institutional levels (e.g. the Research Excellence Framework (REF) in the UK or Excellence in Research for Australia (ERA) in Australia).

Universities need to evaluate the impact of their research programmes to inform resource allocation. They need to make recruitment and promotion decisions. Various metrics at researcher or aggregated levels can help inform these choices.

Researchers may use metrics as an additional criterion in their reading selection, in job or grant applications, and/or when searching for talent or collaborative partners.

Funders, recruiters, and evaluators can all use metrics to supplement their assessments. Librarians can use them to inform collection decisions.

Table 1: Examples of research evaluation scenarios

Scenario Examples	Assesse Type	Assessor	Purpose of assessment
National-level research assessments, for example, Research Excellence Framework (REF) and Excellence in Research for Australia (ERA)	Institutions	REF: The Higher Education Funding Council for England (HEFCE), the Scottish Funding Council (SFC), the Higher Education Funding Council for Wales (HEFCW) and the Department for Employment and Learning, Northern Ireland (DEL) ERA: Australian Research Council (ARC)	REF: Assess the quality of research and produce outcomes for each submission made by institutions[1] ERA: Evaluate the quality of the research undertaken in Australian universities against national and international benchmarks[2]
Grant application	Principal Investigator (PI) / Project Leader, and team members	Funding agencies, for example, the National Institutes of Health (NIH) [3]	Determine to which research proposal to award grants or other forms of financial support
Budgeting / resource allocation exercise within an institution	Departments / research groups	Office of the Provost / Evaluation committee	Determine research fund allocation based on the institution's strategy
Faculty performance appraisal	Individual researchers	Office of the Provost / Evaluation committee	Provide a basis for awarding of merit and other recognition
Library collection decisions	Publishers / journals	Librarian	Help inform which (bundles of) journals to purchase, retain, or discard
Evaluation of journal as possible publication venue	Journal	Prospective author	Help decide to which journal to submit one's manuscript
Evaluation of journal for inclusion in indexing database	Journal	Expert committee (e.g. LSTRC for MEDLINE, Scopus Content Selection & Advisory Board, Thomson Reuters editors)	Help choose which journals to select for inclusion into a particular A&I service or database
Evaluation of paper for potential reading	Paper	Researcher / Student	Help decide which papers to read, download, reference, etc.

3

Regardless of the scenario, both qualitative and quantitative aspects of assessment are important during research evaluation. For the quantitative aspect, analyses often revolve around the measure of scholarly output (using publications) and of impact and excellence (using citation counts and its derivations). Owing to the different purposes of each research evaluation, care must be taken to ensure that each metric is selected appropriately to measure the objective, based on the context of the research scenario.

4. Appropriate Use of Metrics

The number and type of metrics available have reached unprecedented levels – it is now possible to attempt to evaluate, directly or as a proxy, a staggering variety of aspects, such as financial inputs, scholarly output, impact, growth, excellence, scholarly and public engagement, knowledge transfer between sectors, collaboration, mass media mentions, etc. , , All of these indicators can help provide information on the various stages of the research workflow (see Figure 1).

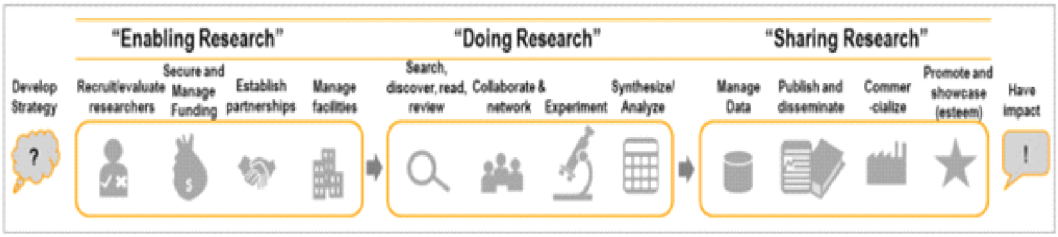

Figure 1: researcher workflow; source: Elsevier's Response to HEFCE's call for evidence: independent review of the role of metrics in research assessment.

Research evaluation typically focuses on output metrics (e.g. scholarly output and impact, scholarly or public engagement). Input or process metrics (e.g. funding, collaboration) can be used to inform interpretation or allow systematic comparisons across various entities. There are simple ways to ensure metrics analysis yields valid and meaningful results, such as:

4.1. Trends

A data point in isolation may not mean much. For instance, how can one interpret the fact that China published nearly 400,000 papers in 2012? Looking at a trend over time is more informative, and gives a sense of the growth of China's recent output (see Figure 2).

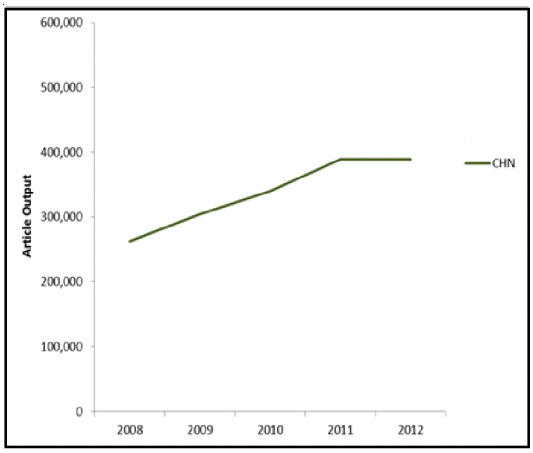

Figure 2: China's scholarly output 2008-2012; source: International Comparative Performance of the UK Research Base – 2013, Appendix F: Supplementary Data

4.2. Benchmarks and Comparators

A trend line in isolation also yields little information. Comparing China's scholarly output to that of other countries allows one to see how it performs against them: is it more or less productive? Are the differences large or small, and how are they evolving? For instance in Figure 3, we see that China is second only to the USA in scholarly output, and furthermore has been rapidly closing the gap in recent years due to its fast growth.

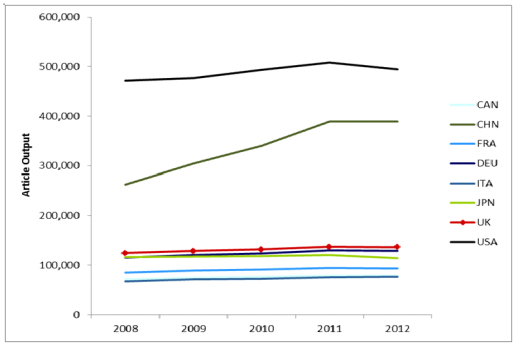

Figure 3: Eight countries' scholarly output 2008-2012; source: International Comparative Performance of the UK Research Base – 2013, Appendix F: Supplementary Data

Using different benchmarks (e.g. regional versus global benchmark) also allows one to make different observations about the same dataset. For instance, in Figure 4, the output (as share of the world) and impact (as rebased Field-Weighted Citation Impact) of Brain Research for specific countries can be compared to both that of the world and that of the EU41 group of countries.

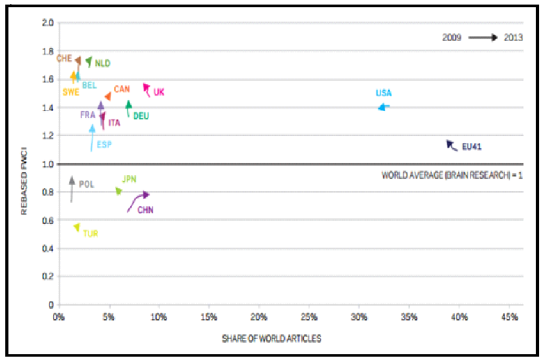

Figure 4: Rebased Field–Weighted Citation Impact versus share of the world's research output for brain and neuroscience research for comparator countries, 2009–2013; source: Brain Science: Mapping the Landscape of Brain and Neuroscience Research

4.3. Normalisation

This is especially important when comparing entities. For instance, normalising by size, output, or level of funding allows us to meaningfully compare entities of different size. Figure 5 shows scholarly output per million dollar GERD (Gross Expenditure on Research & Development) and reveals that the most prolific countries are not necessarily the most productive; China also presents a decreasing rather than increasing trend.

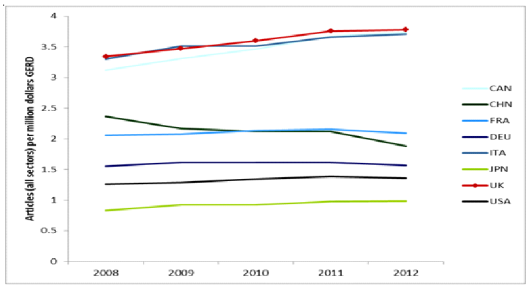

Figure 5: Eight countries' scholarly output per million dollar GERD 2008-2012; source: International Comparative Performance of the UK Research Base – 2013, Appendix F: Supplementary Data

Field normalisation accounts for the various practices in different fields (for instance, citations tend to accrue at a faster and greater rate in genetics relative to mathematics), so that one can compare the citation impact of institutes with different output subject area distributions. Figure 6 shows the Activity Indices of China and Japan, revealing how different their output distribution is to the world and to each other. For instance, China has high levels of activity in Engineering and Mathematics, and low levels in Clinical Sciences and Biological Sciences, while Japan shows activity levels closer to world average in Engineering, Clinical Sciences, and Biological Sciences, and lower levels in Mathematics. To factor in the different citation practices between different fields, field normalisation is needed when measuring an entity's overall citation impact, which is achieved here by using Field-Weighted Citation Impact (see Figure 7), which corrects for field variations by normalising impact against that of similar articles (of the same age, scope, and type).

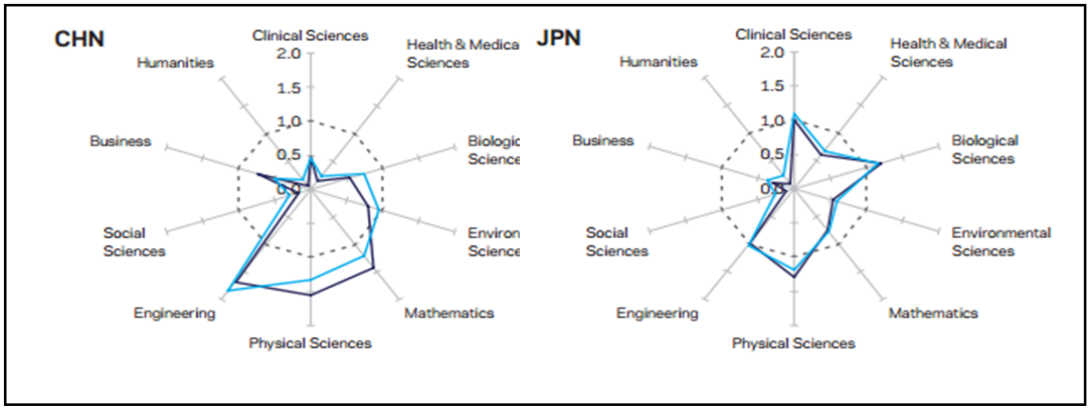

Figure 6: Activity Indices of China and Japan 2002 (purple) vs 2012 (blue); source: International Comparative Performance of the UK Research Base – 2013

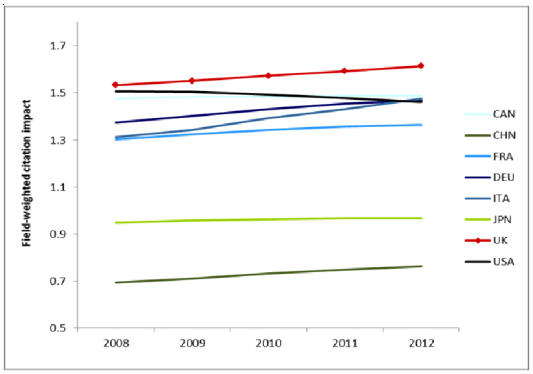

Figure 7: Eight countries' Field-Weighted Citation Impact 2008-2012; source: International Comparative Performance of the UK Research Base – 2013, Appendix F: Supplementary Data

4.4. Breakdowns

Breaking down the data may yield further insights, revealing strengths or weaknesses in specific subjects or geographical areas. For instance, Figure 8, shows India's Field-Weighted Citation Impact by subject area, and reveals that the country is most impactful in Engineering, in which it nearly reaches world average. Figure 9 represents the distribution of sedentary researchers by US state or European countries, showing a low proportion of sedentary researchers for Switzerland, for example.

Figure 8: India's Field-Weighed Citation Impact, 2002 (purple) vs 2012 (blue); source: International Comparative Performance of the UK Research Base – 2013

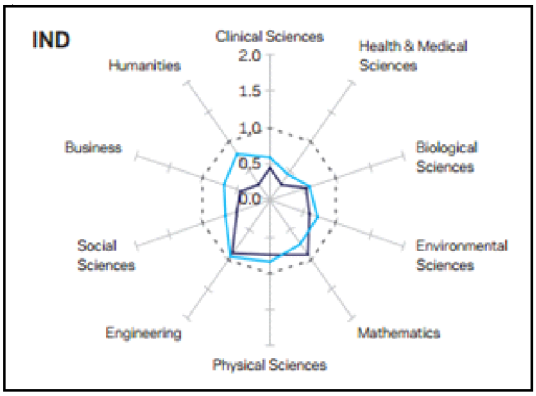

Figure 9: Sedentary researcher distributions by US State and European Country, 1996–2011; source Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility

4.5. Triangulation

Using several metrics in conjunction allows more wholesome and meaningful comparisons, as triangulation facilitates multidimensional research evaluation. For instance, in Figure 10 we see that although France has the largest share of internationally co-authored papers amongst the comparator countries on this chart, in terms of the Field-Weighted Citation Impact of these papers it ranks sixth. We also see that its Field-Weighted Citation Impact is lower than expected given the relationship between the two indicators observed for all comparators on the chart.

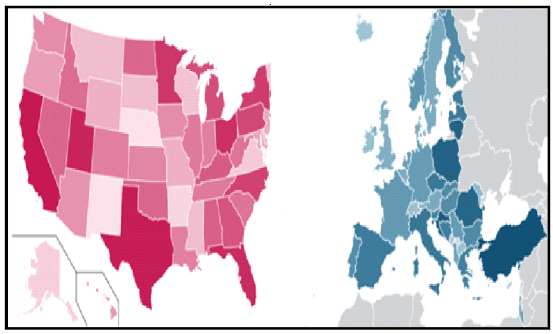

Figure 10: Eight countries' international co-authorship share versus internationally co-authored papers' Field-Weighted Citation Impact; source: International Comparative Performance of the UK Research Base – 2013

4.6. Context

Context is key to both posing the right question and producing the right interpretation.

For example, if a university wishes to assess the productivity of its researchers in a specific department to inform internal resource allocation, metrics may be normalised to service time to avoid favouring those with longer service. For an assessment of researcher impact across the whole university, metrics may be field normalised or benchmarked against field averages. If the university is keen to evaluate how different programmes compare, metrics may be normalised by resources, such as funding, number of researchers/PhDs etc. (e.g. Figure 11).

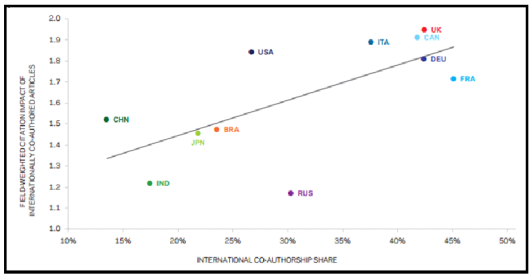

Figure 11: Eight countries' scholarly output per researcher 2008-2012; source: International Comparative Performance of the UK Research Base – 2013, Appendix F: Supplementary Data

Context can also help interpret data. For instance, a sudden rise in the output of a department could be linked to a previous rise in resources, bearing in mind the length of the publication process. On a small corpus, a spike in impact could be caused by one particularly highly-cited paper. For a female researcher, a gap in publication record could be due to maternity leave.

Resources such as the Snowball metrics recipe book or the SciVal metrics guidebook can be useful reference points to consider when embarking on a research assessment exercise.

5. Conclusion

Metrics have been increasingly called upon to inform research evaluation, and have been used in recent research evaluation exercises, supplementing or informing qualitative evaluation. Provided that simple methodological guidelines are followed, such as the right selection of metrics to measure a certain facet of research, that appropriate benchmarks and triangulation are applied, and that they are interpreted in the right context, metrics are a useful addition to research evaluation that allow scalable, systematic, and impartial analyses.

References

1. Brain Science: Mapping the Landscape of Brain and Neuroscience Research http://www.elsevier.com/online-tools/research-intelligence/resource-library/resources/brain-science-mapping-the-landscape-of-brain-and-neuroscience-research
2. Carpenter, T.A. (2014), "Comparing digital apples to digital apples: Background on NISO's effort to build an infrastructure for new forms of scholarly assessment", Information Services and Use, Volume 34, Number 1-2.
3. Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility http://www.elsevier.com/online-tools/research-intelligence/resource-library/resources/comparative-benchmarking-of-european-and-us-research-collaboration-and-researcher-mobility
4. Derrick, G.E., Haynes, A., Chapman, S., Hall, W.D. (2011), "The Association between Four Citation Metrics and Peer Rankings of Research Influence of Australian Researchers in Six Fields of Public Health", PLOS ONE
5. Elsevier's response to HEFCE's call for a review of the role of metrics in research evaluation http://www.elsevier.com/__data/assets/pdf_file/0015/210813/Elsevier-response-HEFCE-review-role-of-metrics.pdf
6. Gunn, W. (2014), "On Numbers and Freedom", El profesional de la información, 2014, septiembre-octubre, volume 23, issue 5.
7. HEFCE's Independent review of the role of metrics in research assessment http://www.hefce.ac.uk/whatwedo/rsrch/howfundr/metrics/
8. International Comparative Performance of the UK Research Base – 2013 http://www.elsevier.com/online-tools/research-intelligence/resource-library/resources/international-comparative-performance-of-the-uk-research-base-2013
9. Kostoff, R.N. (1997), "Use and Misuse of Metrics in Research Evaluation", Science and Engineering Ethics, issue 3.

10. Kostoff, R.N. (2007), "The unintended consequences of metrics in technology evaluation", Journal of Informetrics, volume 1, issue 2.

11. Moed, H. and Halevi, G. (2014), "Research assessment: Review of methodologies and approaches", Research Trends, issue 36

12. San Francisco Declaration on Research Assessment (DORA)http://www.ascb.org/dora-old/files/SFDeclarationFINAL.pdf

13. Snowball metrics recipe book http://www.snowballmetrics.com/wp-content/uploads/snowball-metrics-recipe-book-upd.pdf

14. Snowball response to HEFCE's call for a review of the role of metrics in research evaluation http://www.snowballmetrics.com/wp-content/uploads/Snowball-response-to-HEFCE-review-on-metrics-300614F.pdf

15. SciVal metrics guidebook http://www.elsevier.com/__data/assets/pdf_file/0006/184749/scival-metrics-guidebook-v1_01-february2014.pdf

16. Taylor, M. (2013), "The Challenges of Measuring Social Impact Using Altmetrics", Research Trends, issue 33

About Authors

Ms. Sarah Huggett, Elsevier Analytical Services, 3 Killiney Road #08-01, Winsland House I, Singapore, 239519, Singapore.
Email: s.huggett@elsevier.com

Ms. Georgin Lau, Elsevier Analytical Services, 3 Killiney Road #08-01, Winsland House I, Singapore, 239519, Singapore.
Email: g.lau@elsevier.com