emerald PUBLISHING

Performance Measurement and Metrics

Principles to guide reliable and ethical research evaluation using metric-based indicators of impact

Journal:	*Performance Measurement and Metrics*
Manuscript ID	PMM-06-2016-0025
Manuscript Type:	Research Paper
Keywords:	altmetrics, impact, metrics, metric-based indicators, academic pratice, scholarly impact

Principles to guide reliable and ethical research evaluation using metric-based indicators

Bibliometrics can be defined as a set of tools and techniques which enable quantitative analyses of scholarly literature. The analyses can be conducted for a variety of purposes—from collection development and evaluation in research libraries, to tracking changes in scholarly disciplines, to studying social and organizational structures of science or assessing the role and/or contribution of individual researchers, research groups, institutions or nations. These studies were traditionally the domain of bibliometric researchers and scholars. They had access to often expensive data sets and worked to develop and validate bibliometric methods as well as to understand their limitations. This community created formal structures of discourse including specialized peer-reviewed journals (e.g., *Scientometrics, Research Evaluation, Journal of the Association for Information Science and Technology,* and *Journal of Informetrics),* conferences (e.g., the Science and Technology Indicators Conference and the Conference of the International Society for Scientometrics and Informetrics), and societies (e.g., the International Society for Scientometrics and Informetrics).

More recently, bibliometric tools have been employed to inform assessment of the quality and impact of research, either in an attempt to replace or to serve alongside the peer review process. We can consider citations as a form of peer review if we subscribe to the Mertonian theory of the normative structure of science. This theory proposes that scientific progress is possible only if scientists follow certain accepted norms of behavior, including organized skepticism [1].

Keen to understand more about the research they fund and conduct, research institutions, funders and government agencies turned to bibliometrics for a bias-free and inexpensive assessment

method. With the development of widely available on-line commercial tools, bibliometrics has become a method available to all, not just a select group of expert scholars.

The journal impact factor (JIF) and h-index are familiar indicators to thousands of researchers around the world who need to demonstrate the impact of their past work. Promotion and Tenure Committees and organizations that fund research regularly ask applicants to provide the computed values of these indicators, and then use them in ways for which they were never intended. For example, the JIF tells us nothing about the quality of an individual paper in that journal. Anthony van Raan, former director of the Centre for Science and Technology Studies at Leiden University in the Netherlands, noted: "If there is one thing every bibliometrician agrees, it is that you should never use the journal impact factor to evaluate research performance for an article or for an individual—that is a mortal sin." [2] Equally, comparing h-indices of researchers at different stages in their careers or across different disciplines can result in grossly misleading findings, which vary depending on the underlying data [3].

The research community affected by these practices is increasingly concerned about the inappropriate use of bibliometrics for evaluation and decision-making. Critiques have led, among others, to the Declaration on Research Assessment (DORA) declaration [4], which called for the abandonment of the use of JIF for evaluation of individuals. Others have claimed that the current evaluation regimes in many countries, such as the United Kingdom [5,6,7], Brazil [8], and Australia [9], may lead to distortions in scholars' behaviors, leading to decreased creativity, risk aversion and less willingness to undertake interdisciplinary research.

Against this backdrop, research libraries are taking an increasingly proactive role in either supporting organization-wide evaluation efforts or developing bibliometric services to help researchers navigate the maze of research metrics and use them effectively. For instance, in Australia and the United Kingdom, libraries are funded to collect and verify research outputs submitted by institutions for national assessment. In the United States and elsewhere, libraries are launching new services to help researchers and their organizations understand the impact of their research. The Association of Research Libraries (ARL) SPEC survey [10] reported that in January 2015 there were 79 ARL-member libraries[1] with such services either already developed or being developed. Developments continue, such as at the University of Waterloo in Canada, which recently released a guide on bibliometric measures, together with a description of available tools [11]. With very few exceptions, these services are not run by bibliometrics researchers. A panel entitled "How to Deal with Unsettling Realities of Bibliometric Services in Universities" addressed this very issue at the 2014 Science and Technology Indicators Conference in Leiden [12].

THE LEIDEN MANIFESTO

At that same conference, bibliometric researchers and practitioners from around the globe issued the so-called Leiden Manifesto [13]. This document, published in *Nature*, laid down ten principles intended to guide best practices for bibliometric-based research assessment. These principles are summarized below. Although they are well understood within the expert bibliometric community, they need to be adopted by others seeking to implement or understand bibliometric-based evaluation.

[1] The Association of Research Libraries (ARL) is a group of 125 research libraries across the United States and Canada.

Principle 1. Metrics can provide additional dimensions to the assessment process, but should never be used in isolation from qualitative assessment (e.g., peer-review). Metrics-based evaluation can supplement and provide additional dimensions to qualitative assessment, but should never replace it.

Principle 2. Metrics used to evaluate research performance should reflect the research objectives of the institution, research groups or individual researchers. Individual indicators often provide a one-dimensional view of research impact while intended research goals of the evaluated units or individuals may be multi-dimensional. For example, they may include advances of science or improvements of social outcomes and may be aimed at differing audiences—from researcher, to industry, to policy makers. No single metric or evaluation model can apply in all contexts.

Principle 3. Measure locally relevant research using appropriate metrics, including those that build on journal collections in local languages or that cover certain geographic locations. Big international citation databases (used most frequently to derive data used for constructing indicators) still mostly focus on English-language, western journals.

Principle 4. Metrics-based evaluation, to be trusted, should adhere to the standards of openness and transparency in data collection and analysis. What data are collected? How is it collected? How are citations captured? What are the exact methods and calculations used to develop indicators? Is the process open to scrutiny by experts and by the assessed?

Principle 5. Those who are evaluated should be able to verify data and the analyses used in the assessment process. Are all relevant outputs identified, captured and analyzed?

Principle 6. Just as all metrics are not suitable for assessing all aspects of scholarship (see Principle 2) neither can they be applied equally across all disciplines. We know that disciplines vary in their publication and citation practices, and these need to be taken into consideration when selecting metrics to compare disciplines. For instance, a bibliometric profile of a researcher studying causes of lung disease will be rather different from that of a researcher studying the social effects of smoke cessation programs. Health policy research tends to behave more in line with the "softer" disciplines, with fewer citation counts, a more diffuse set of outlets, and top journals with lower impact factor values. In contrast, biomedical research tends to behave more in line with the "harder" sciences (i.e., with higher average citation rates). For instance, the top-ranking health policy journal has a JIF of 4.9 while the top respiratory system journal has a JIF of 12.6[2].

If comparisons across disciplines are called for, the most suitable metrics are those that statistically normalize for disciplinary differences. They should compare to the discipline baselines. For instance, a researcher might ask, "Do my hematology publications have more citations than an average hematology publication of the same age? In a percentile distribution of all hematology publications, based on citation counts, are my publications in the 99th, 95th or 90th percentile?"

Principle 7. Do not rely on a single quantitative indicator when evaluating individual researchers. The h-index, currently the most popular author-level indicator, favors older researchers with longer publication lists. Moreover, it does not adjust for disciplinary differences and ignores the

[2] The 2014 Journal Citation Report shows *Health Affairs* with a JIF of 4.966 and *American Journal of Respiratory and Critical Care Medicine* a JIF of 12.993.

impact of highly cited papers. The signatories of the Leiden Manifesto state that: "Reading and judging a researcher's work is much more appropriate than relying on one number." [13]

Principle 8. Sets of indicators can provide a more reliable and multi-dimensional view than a single indicator. The Manifesto authors give an example of a set of impact factors shown to three decimal places, creating a false impression that journals can be reliably ranked even if small differences in scores are observed. It is better to consider a range of indicators to identify differences.

Principles 9 and 10. Goodhart's Law is evident in research evaluation; it states that, "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes".[3] [14] Every evaluation system creates incentives (intended or unintended) and these, in turn, drive behaviors. Use of a single indicator (like JIF) opens the evaluation system to such undesirable behaviors like gaming or goal displacement.[4] To mitigate against these behaviors multiple indicators should be used. Furthermore, indicators should be reviewed and updated in line with changing goals of assessment, and new metrics should be considered as they become available.

While I believe that libraries in research institutions are well placed to provide institutional support for metric-based evaluation, I also strongly believe that librarians should become advocates for the responsible and ethical use of these metrics.

[3] A more popular version of the Goodhart's Law reads 'when a measure becomes a target, it ceases to be a good measure".

[4] A good example of gaming are so-called "citation clubs" designed to artificially increase citation rates to publications, while goal displacement refers the behaviors in which the measurement becomes a goal (e.g., the only criterion for the selection of a publication outlet is its impact factor).

REFERENCES

1. Merton, R.K. The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press, 1973. Chapter 13, The Normative Structure of Science, p. 267-81.

2. Van Noorden R. Metrics: A profusion of measures". Nature 2010 June 16; 465(7300):864–866. doi:10.1038/465864a

3. Bar Ilan, J. Which h-index? — A comparison of WoS, Scopus and Google Scholar. Scientometrics 2008 74(2):257-271. doi:10.1007/s11192-008-0216-y

4. San Francisco Declaration on Research Assessment (DORA) http://www.ascb.org/dora/

5. de Rijcke, S., Woulters, P., Rushforth, A.D., Franssen, T.P., Hammarfelt, B. Evaluation practices and effects of indicator use—a literature review. Res Eval. 2015. Advanced access 19 Dec. 2015 doi:10.1093/reseval/rvv038

6. Moriarty, P. Addicted to the brand: the hypocrisy of a publishing academic. LSE Impact Blog. 2016 March 16. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2016/03/14/addicted-to-the-brand-the-hypocrisy-of-a-publishing-academic/

7. Shaw. C. Research that does not belong to a single subject area deemed too risky. The Guardian (US edition). 2013 Nov 21 Available at: http://www.theguardian.com/higher-education-network/blog/2013/nov/21/interdisciplinary-research-ref-submission-university

8. Ferreira R.C., Antoneli, F., Briones, M.R.S. The hidden factors in impact factors: A perspective from Brazilian science. Front Genet. 2013 July 11; 4(art.130). doi:10.3389/fgene.2013.00130

9. Kwok, J.T. Impact of ERA Research Assessment on University Behaviour and their Staff. Melbourne: NTEU National Policy and Research Unit, 2013. Available at: http://apo.org.au/files/Resource/nteu_impactofera_april2013.pdf

10. SPEC Kit 346: Scholarly Output Assessment Activities. May 2015. Available at: http://publications.arl.org/Scholarly-Output-Assessment-SPEC-Kit-346/

11. University of Waterloo Working Group on Bibliometrics. White Paper on Bibliometrics, Measuring Research Productivity and Impact Through Bibliometrics. 2015 Waterloo, Ontario: University of Waterloo. Available at: https://uwaterloo.ca/institutional-analysis-planning/sites/ca.institutional-analysisplanning/files/uploads/files/white_paper_on_bibliometrics_draft_for_consultation_07oct2015_0.pdf

12. Drenthe, G. How to Deal with Unsettling Realities of Bibliometric Services in Universities. Special Panel during 2014 STI Conference, Leiden.

13. Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., Rafols, I. The Leiden Manifesto for Research Metrics. Nature 2015 April 23; 520(7548):429-31. Available at: http://www.nature.com/polopoly_fs/1.17351!/menu/main/topColumns/topLeftColumn/pdf/520429a.pdf

14. Goodhart, C.A.E. Monetary Theory and Practice: The UK experience. London: The Macmillan Press, 1984. Chapter 3 Problems of Monetary Management: The UK Experience, p. 91-116.