COMMENT

SUSTAINABILITY Data needed to drive UN development goals **p.432**

CONSERVATION Economics and environmental catastrophe **p.434**

GEOLOGY Questions raised over proposed Anthropocene dates **p.436**

HISTORY Music inspired Newton to add more colours to the rainbow **p.436**

The Leiden Manifesto for research metrics

Use these ten principles to guide research evaluation, urge **Diana Hicks**, **Paul Wouters** and colleagues.

Data are increasingly used to govern science. Research evaluations that were once bespoke and performed by peers are now routine and reliant on metrics[1]. The problem is that evaluation is now led by the data rather than by judgement. Metrics have proliferated: usually well intentioned, not always well informed, often ill applied. We risk damaging the system with the very tools designed to improve it, as evaluation is increasingly implemented by organizations without knowledge of, or advice on, good practice and interpretation.

Before 2000, there was the Science Citation Index on CD-ROM from the Institute for Scientific Information (ISI), used by experts for specialist analyses. In 2002, Thomson Reuters launched an integrated web platform, making the Web of Science database widely accessible. Competing citation indices were created: Elsevier's Scopus (released in 2004) and Google Scholar (beta version released in 2004). Web-based tools to easily compare institutional research productivity and impact were introduced, such as InCites (using the Web of Science) and SciVal (using Scopus), as well as software to analyse individual citation profiles using Google Scholar (Publish or Perish, released in 2007).

In 2005, Jorge Hirsch, a physicist at the University of California, San Diego, proposed the *h*-index, popularizing citation counting for individual researchers. Interest in the journal impact factor grew steadily after 1995 (see 'Impact-factor obsession').

Lately, metrics related to social usage ▶

ILLUSTRATION BY DAVID PARKINS

and online comment have gained momentum — F1000Prime was established in 2002, Mendeley in 2008, and Altmetric.com (supported by Macmillan Science and Education, which owns Nature Publishing Group) in 2011.

As scientometricians, social scientists and research administrators, we have watched with increasing alarm the pervasive misapplication of indicators to the evaluation of scientific performance. The following are just a few of numerous examples. Across the world, universities have become obsessed with their position in global rankings (such as the Shanghai Ranking and *Times Higher Education*'s list), even when such lists are based on what are, in our view, inaccurate data and arbitrary indicators.

Some recruiters request *h*-index values for candidates. Several universities base promotion decisions on threshold *h*-index values and on the number of articles in 'high-impact' journals. Researchers' CVs have become opportunities to boast about these scores, notably in biomedicine. Everywhere, supervisors ask PhD students to publish in high-impact journals and acquire external funding before they are ready.

In Scandinavia and China, some universities allocate research funding or bonuses on the basis of a number: for example, by calculating individual impact scores to allocate 'performance resources' or by giving researchers a bonus for a publication in a journal with an impact factor higher than 15 (ref. 2).

In many cases, researchers and evaluators still exert balanced judgement. Yet the abuse of research metrics has become too widespread to ignore.

We therefore present the Leiden Manifesto, named after the conference at which it crystallized (see http://sti2014.cwts.nl). Its ten principles are not news to scientometricians, although none of us would be able to recite them in their entirety because codification has been lacking until now. Luminaries in the field, such as Eugene Garfield (founder of the ISI), are on record stating some of these principles[3,4]. But they are not in the room when evaluators report back to university administrators who are not expert in the relevant methodology. Scientists searching for literature with which to contest an evaluation find the material scattered in what are, to them, obscure journals to which they lack access.

We offer this distillation of best practice in metrics-based research assessment so that researchers can hold evaluators to account, and evaluators can hold their indicators to account.

TEN PRINCIPLES

1 Quantitative evaluation should support qualitative, expert assessment. Quantitative metrics can challenge bias tendencies in peer review and facilitate deliberation. This should strengthen peer review, because making judgements about colleagues is difficult without a range of relevant information. However, assessors must not be tempted to cede decision-making to the numbers. Indicators must not substitute for informed judgement. Everyone retains responsibility for their assessments.

2 Measure performance against the research missions of the institution, group or researcher. Programme goals should be stated at the start, and the indicators used to evaluate performance should relate clearly to those goals. The choice of indicators, and the ways in which they are used, should take into account the wider socio-economic and cultural contexts. Scientists have diverse research missions. Research that advances the frontiers of academic knowledge differs from research that is focused on delivering solutions to societal problems. Review may be based on merits relevant to policy, industry or the public rather than on academic ideas of excellence. No single evaluation model applies to all contexts.

> *"Simplicity is a virtue in an indicator because it enhances transparency."*

3 Protect excellence in locally relevant research. In many parts of the world, research excellence is equated with English-language publication. Spanish law, for example, states the desirability of Spanish scholars publishing in high-impact journals. The impact factor is calculated for journals indexed in the US-based and still mostly English-language Web of Science. These biases are particularly problematic in the social sciences and humanities, in which research is more regionally and nationally engaged. Many other fields have a national or regional dimension — for instance, HIV epidemiology in sub-Saharan Africa.

This pluralism and societal relevance tends to be suppressed to create papers of interest to the gatekeepers of high impact: English-language journals. The Spanish sociologists that are highly cited in the Web of Science have worked on abstract models or study US data. Lost is the specificity of sociologists in high-impact Spanish-language papers: topics such as local labour law, family health care for the elderly or immigrant employment[5]. Metrics built on high-quality non-English literature would serve to identify and reward excellence in locally relevant research.

4 Keep data collection and analytical processes open, transparent and simple. The construction of the databases required for evaluation should follow clearly stated rules, set before the research has been completed. This was common practice among the academic and commercial groups that built bibliometric evaluation methodology over several decades. Those groups referenced protocols published in the peer-reviewed literature. This transparency enabled scrutiny. For example, in 2010, public debate on the technical properties of an important indicator used by one of our groups (the Centre for Science and Technology Studies at Leiden University in the Netherlands) led to a revision in the calculation of this indicator[6]. Recent commercial entrants should be held to the same standards; no one should accept a black-box evaluation machine.

Simplicity is a virtue in an indicator because it enhances transparency. But simplistic metrics can distort the record (see principle 7). Evaluators must strive for balance — simple indicators true to the complexity of the research process.

5 Allow those evaluated to verify data and analysis. To ensure data quality, all researchers included in bibliometric studies should be able to check that their outputs have been correctly identified. Everyone directing and managing evaluation processes should assure data accuracy, through self-verification or third-party audit. Universities could implement this in their research information systems and it should be a guiding principle in the selection of providers of these systems. Accurate, high-quality data take time and money to collate and process. Budget for it.

6 Account for variation by field in publication and citation practices. Best practice is to select a suite of possible indicators and allow fields to choose among them. A few years ago, a European group of historians received a relatively low rating in a national peer-review assessment because they wrote books rather than articles in journals indexed by the Web of Science. The historians had the misfortune to be part of a psychology department. Historians and social scientists require books and national-language literature to be included in their publication counts; computer scientists require conference papers be counted.

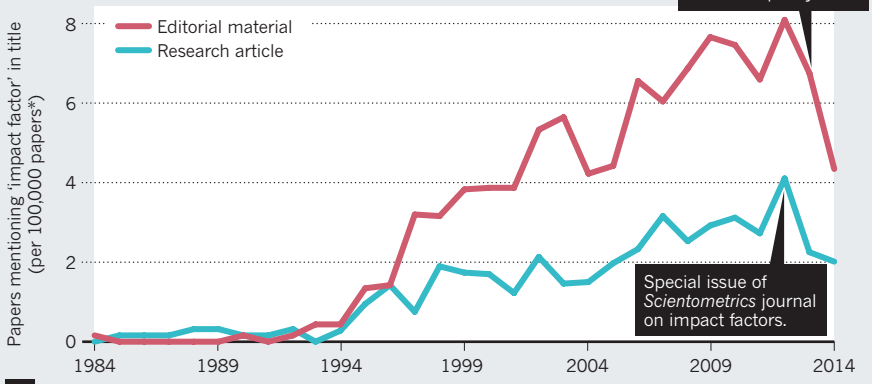
Citation rates vary by field: top-ranked journals in mathematics have impact factors of around 3; top-ranked journals in cell biology have impact factors of about 30. Normalized indicators are required, and the most robust normalization method is based on percentiles: each paper is weighted on the basis of the percentile to which it belongs in the citation distribution of its field (the top 1%, 10% or 20%, for example). A single highly cited publication slightly improves the position of a university in a ranking that

DATA SOURCE: THOMSON REUTERS WEB OF SCIENCE; ANALYSIS: D.H., L.W.

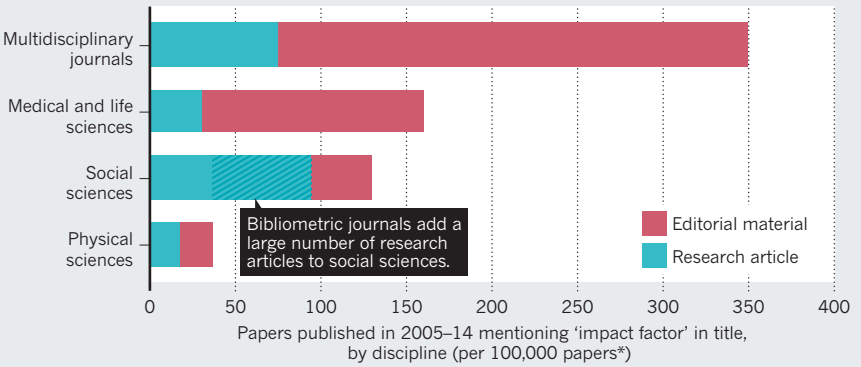
IMPACT-FACTOR OBSESSION

Soaring interest in one crude measure — the average citation counts of items published in a journal in the past two years — illustrates the crisis in research evaluation.

1 ARTICLES MENTIONING 'IMPACT FACTOR' IN TITLE



2 WHO IS MOST OBSESSED?



*Indexed in the Web of Science. †DORA, San Francisco Declaration on Research Assessment.

is based on percentile indicators, but may propel the university from the middle to the top of a ranking built on citation averages[7].

7 Base assessment of individual researchers on a qualitative judgement of their portfolio. The older you are, the higher your *h*-index, even in the absence of new papers. The *h*-index varies by field: life scientists top out at 200; physicists at 100 and social scientists at 20–30 (ref. 8). It is database dependent: there are researchers in computer science who have an *h*-index of around 10 in the Web of Science but of 20–30 in Google Scholar[9]. Reading and judging a researcher's work is much more appropriate than relying on one number. Even when comparing large numbers of researchers, an approach that considers more information about an individual's expertise, experience, activities and influence is best.

8 Avoid misplaced concreteness and false precision. Science and technology indicators are prone to conceptual ambiguity and uncertainty and require strong assumptions that are not universally accepted. The meaning of citation counts, for example, has long been debated. Thus, best practice uses multiple indicators to provide a more robust and pluralistic picture. If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values. If this is not possible, indicator producers should at least avoid false precision. For example, the journal impact factor is published to three decimal places to avoid ties. However, given the conceptual ambiguity and random variability of citation counts, it makes no sense to distinguish between journals on the basis of very small impact factor differences. Avoid false precision: only one decimal is warranted.

9 Recognize the systemic effects of assessment and indicators. Indicators change the system through the incentives they establish. These effects should be anticipated. This means that a suite of indicators is always preferable — a single one will invite gaming and goal displacement (in which the measurement becomes the goal). For example, in the 1990s, Australia funded university research using a formula based largely on the number of papers published by an institute. Universities could calculate the 'value' of a paper in a refereed journal; in 2000, it was Aus$800 (around US$480 in 2000) in research funding. Predictably, the number of papers published by Australian researchers went up, but they were in less-cited journals, suggesting that article quality fell[10].

10 Scrutinize indicators regularly and update them. Research missions and the goals of assessment shift and the research system itself co-evolves. Once-useful metrics become inadequate; new ones emerge. Indicator systems have to be reviewed and perhaps modified. Realizing the effects of its simplistic formula, Australia in 2010 introduced its more complex Excellence in Research for Australia initiative, which emphasizes quality.

NEXT STEPS

Abiding by these ten principles, research evaluation can play an important part in the development of science and its interactions with society. Research metrics can provide crucial information that would be difficult to gather or understand by means of individual expertise. But this quantitative information must not be allowed to morph from an instrument into the goal.

The best decisions are taken by combining robust statistics with sensitivity to the aim and nature of the research that is evaluated. Both quantitative and qualitative evidence are needed; each is objective in its own way. Decision-making about science must be based on high-quality processes that are informed by the highest quality data. ■

Diana Hicks *is professor of public policy at the Georgia Institute of Technology, Atlanta, Georgia, USA.* **Paul Wouters** *is professor of scientometrics and director,* **Ludo Waltman** *is a researcher, and* **Sarah de Rijcke** *is assistant professor, at the Centre for Science and Technology Studies, Leiden University, the Netherlands.* **Ismael Rafols** *is a science-policy researcher at the Spanish National Research Council and the Polytechnic University of Valencia, Spain.*
e-mail: diana.hicks@pubpolicy.gatech.edu

1. Wouters, P. in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. *Learned Publ.* **24,** 95–97 (2011).
3. Seglen, P. O. *Br. Med. J.* **314,** 498–502 (1997).
4. Garfield, E. *J. Am. Med. Assoc.* **295,** 90–93 (2006).
5. López Piñeiro, C. & Hicks, D. *Res. Eval.* **24,** 78–89 (2015).
6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. *J. Informetrics* **4,** 431–435 (2010).
7. Waltman, L. *et al. J. Am. Soc. Inf. Sci. Technol.* **63,** 2419–2432 (2012).
8. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102,** 16569–16572 (2005).
9. Bar-Ilan, J. *Scientometrics* **74,** 257–271 (2008).
10. Butler, L. *Res. Policy* **32,** 143–155 (2003).