OPINION

How to improve the use of metrics

Since the invention of the science citation index in the 1960s, quantitative measuring of the performance of researchers has become ever more prevalent, controversial and influential. Six commentators tell *Nature* what changes might ensure that individuals are assessed more fairly.

Get experts on board

Tibor Braun
Founder and editor-in-chief of *Scientometrics*, Hungarian Academy of Sciences

Basic research in scientometrics — the quantitative measurement and analysis of science — has boomed over the past few decades. This has led to a plethora of new measures and techniques (see page 864). Thanks in part to easy access to big, interdisciplinary publication and citation databases (such as Web of Science and Scopus), evaluative metrics can seem very easy to use. Because it is so easy to produce a number, people can be deluded into thinking that they have a thorough understanding of what those numbers mean. All too often, they don't know which database to use, how to clean raw data, which indicator to use or how to use it for the task at hand.

Many evaluators of tenure promotions and grants use evaluative metrics without this background knowledge. It is difficult to learn from mistakes made in such evaluations, because the decision-making processes are rarely transparent: most are not published except as internal reports or in the grey literature. Further, the most flawed metrics can be those that measure the academic performance of individual scientists (as opposed to the performance of a group, institution, nation or journal). In part, this is simply because statistical reliability decreases as the size of the data set decreases.

Anyone who uses metrics, or is simply interested, should read the classic texts[1,2] or browse the key journals: *Scientometrics*, *Research Evaluation*, the *Journal of Informetrics* and the *Journal of the American Society for Information Science and Technology*. Better still they could attend one of the international conferences about scientometrics and their use.

Every evaluating body at any (national, institutional or individual) level should incorporate a scientist with a good publication record in scientometrics. For example, Charles Oppenheim, emeritus professor of information science at Loughborough University, UK, was asked to help the Higher Education Funding Council for England to develop the bibliometric side of the country's Research Excellence Framework. That said, this will not be feasible for every tenure committee: there are, at a rough guess, only about 1,500 people worldwide who consider themselves primarily scientometricians.

The use of evaluative metrics and the science of scientometrics should be included in the curricula of major research universities. The graduate course on measuring science at the Centre for Science and Technology Studies at Leiden University in the Netherlands (go.nature.com/c5H3c7) is one good example. The European Summer School for Scientometrics in Vienna, inaugurated this week, will also help to educate those who use metrics in evaluation (www.scientometrics-school.eu). Finally, many people would benefit from an introductory book on how metrics can best be used to measure the performance of individual scientists. There are many good monographs available[3–4] but such a guidebook is sorely needed.

Use ranking to help search

Carl T. Bergstrom
Co-developer of Eigenfactor.org, University of Washington, Seattle

"Science is being killed by numerical ranking," a friend once told me, "and you're hastening its demise." He was referring to my work on network-based ranking systems with the Eigenfactor project, and I can see his point. When asking questions related to large collections of material, such as "How often do biologists draw on results from economics papers?", numerical evaluation makes sense. But all too often, ranking systems are used as a cheap and ineffective method of assessing the productivity of individual scientists. Not only does this practice lead to inaccurate assessment, it lures scientists into pursuing high rankings first and good science second. There is a better way to evaluate the importance of a paper or the research output of an individual scholar: read it.

"We need to stop misusing rankings and instead demonstrate how they can improve science."

My motivation for developing network-based ranking systems is not to say that Peter is better than Paul or Princeton better than Yale. Rather, the greatest value of ranking is in the service of search. Among the most important questions asked by those who do science (rather than those who evaluate it) is "What should I read to pursue my research?" It is in this domain that rankings help rather than harm the scientific enterprise.

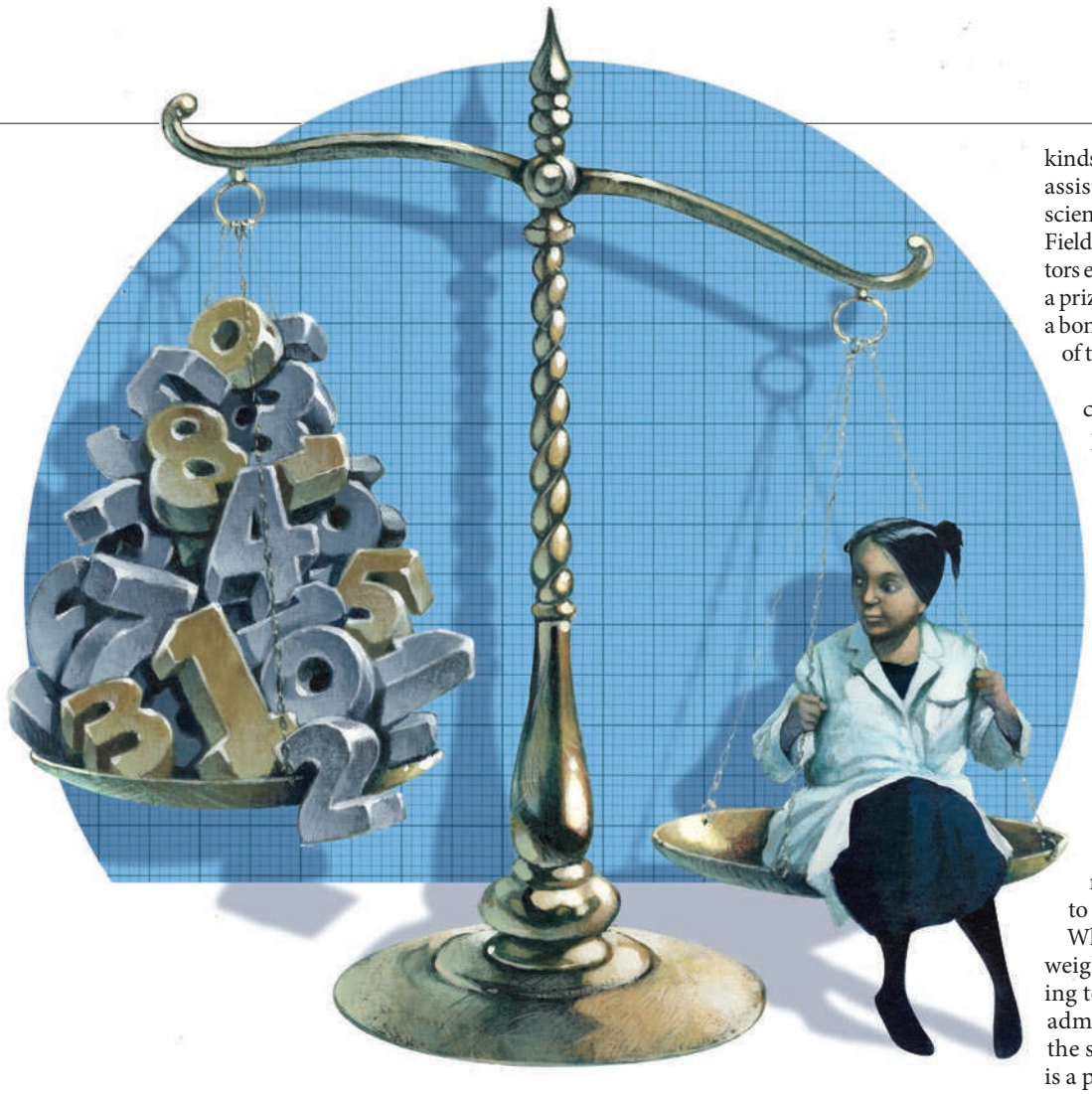
The Internet has shown the way. Although conceived as a tool for document delivery, the web's great power has turned out to be document discovery. It helps people to find objects that are relevant to their interests (a problem of matching) and of sufficiently high quality to merit attention (a problem of ranking). Google's PageRank algorithm demonstrates that the hyperlinked network structure of the web provides all of the information needed to solve the matching problem and the ranking problem simultaneously.

In the scientific literature, the cumulative process of knowledge construction leaves behind it a lattice of citations, analogous to the hyperlink structure of the web. At Eigenfactor.org, we have implemented a ranking algorithm similar to PageRank for scholarly journals, in which important journals are those that are frequently cited by important journals.

The next step is to develop article-level metrics that better map how and why papers are linked to one another other. Combining mapping with search will help scientists to navigate the literature, work out what to read and decide what background they need to understand that material. At Eigenfactor.org we are working on this now.

Many scientists are justifiably concerned that ranking has detrimental effects on science. To allay this concern and reduce the hostility that many scholars feel towards ranking, we need to stop misusing rankings and instead demonstrate how they can improve science.

© 2010 Macmillan Publishers Limited. All rights reserved



ILLUSTRATIONS BY DAVID PARKINS

Motivate people with prizes

Bruno S. Frey & Margit Osterloh
University of Zurich, Switzerland

Pay levels and pay rises in some academic institutions — such as the University of Western Australia in Perth and the Vienna University of Economics and Business — are based heavily on metrics such as numbers of publications and citations. This is not a sensible policy.

The primary motivation of scholars is not money. They are driven by curiosity, autonomy and recognition by peers; in exchange, they accept lower pay[5].

Giving pay rises on the basis of simple measures of performance means that the inducement to 'beat the system' can get the upper hand. Research reverts to a kind of 'academic prostitution', in which work is done to please editors and referees rather than to further knowledge[6]. Motivation to do good research is crowded out[7]. In Australia, the metric of number of peer-reviewed publications was linked to the funding of many universities and individual scholars in the late 1980s and early 1990s. The country's share of publications in the Science Citation Index (SCI) increased by 25% over a decade, but its citation impact ranking dropped from sixth out of 11 OECD countries in 1988 to tenth by 1993 (ref. 8).

The factors measured by metrics are an imperfect indicator of the qualities society values most in its scientists. Even the Thomson Reuters Institute for Scientific Information (ISI) uses citation metrics only as one indicator among others to predict Nobel prizewinners. Of the 28 physics Nobel prizewinners from 2000 to 2009, just 5 are listed in ISI's top 250 most-cited list for that field.

> "If journals listed the papers that they had rejected alongside the published science it could form the basis of a demerit system."

An incentive system for scholars has to match their main motivating factors. Prizes and titles are better suited for that purpose than citation metrics. Honorary doctorates, different kinds of professorships and fellowships (from assistant to distinguished), membership of scientific academies and honours such as the Fields Medal or Nobel prizes are great motivators even for those who do not actually win such a prize. The money attached to such rewards is a bonus, but less important than the reputation of the award-giving institution[9].

If academic rewards are linked to overall contributions to research as reflected in prizes, scientists will pursue their work driven more by research agendas than by simple metrics.

Learn from game theory

Jevin D. West
University of Washington, Seattle

Giving bad answers is not the worst thing a ranking system can do — the worst thing is to encourage bad science. The next generation of scientific metrics needs to take this into account.

When scientists order elements by molecular weight, the elements do not respond by trying to sneak higher up the order. But when administrators order scientists by prestige, the scientists tend to be less passive. There is a powerful feedback between the ranking systems used to assess scientific productivity and the actions of scientists trying to further their careers via these ranking systems.

If tenure committees value quantity over quality, faculty members have strong incentives to churn out large numbers of lower-quality papers. Some advisers even encourage young academics to publish the smallest possible slivers of their work to raise self-confidence and satisfy bean counters — from deans to department heads to those in charge of handing out grants. Sadly, this is probably good advice given the current reward systems.

Because of this feedback, the problem of ranking scholarly output cannot be viewed simply as a problem in applied statistics, in which we wish to extract maximal information from a data set. Instead it is a game-theoretic problem in mechanism design.

The first step in addressing any mechanism-design problem is to identify the desired outcomes. Two objectives the community might set its sights on are alleviating

© 2010 Macmillan Publishers Limited. All rights reserved

the increasing burden on the peer-review system and remediating the growing tendency of authors to break up their work into 'least publishable units' — small and possibly overlapping papers.

If journals listed the papers that they had rejected alongside the published science, it could form the basis of a kind of demerit system. This, in turn, would encourage scientists to send a paper to an appropriate journal on first submission, rather than shooting for the top every time. In addition, tenure committees could permit faculty members to submit only their five best papers when being assessed, and not take into account the total tally of publications — much as committees should ignore ethnicity, gender and age. Scientists would then have the incentive to write higher-quality papers with fuller narratives.

Both of these rules would alter the motivations of researchers (probably for the betterment of science). The publishers and grant-givers in the game of science have the incentive and the power to implement such rules. What sort of behaviours should be encouraged, and how best to do that, remains very much an open question.

Accentuate the positive

David Pendlebury
Citation Analyst, healthcare and science division, Thomson Reuters

There has always been push-back against metrics. No one enjoys being measured — unless he or she comes out on top. That's human nature. So it is important to remind scientists that metrics can be a friend, not a foe.

Importantly, publication-based metrics provide an objective counterweight in tenure and promotion discussions to the peer-review process, which is prone to bias of many kinds[10]. Research has become so specialized over the past few decades that it's often hard to have a panel of peer reviewers who are expertly informed about a given subject. And then there are the overt biases of academic politics, personality conflicts and prejudice against gender or race. Objective numbers can help to balance the system.

That said, there are dangers. Numbers look very authoritative, and people can put too much faith in them. A quantitative profile should always be used to foster discussion, rather than to end it. It is also misguided to expect one metric to explain everything. The *h*-index, for

example, has become so popular that it now seems as if every other bibliometrics article is about the *h*-index and its proposed variants. This creates an unfortunate impression, through the sheer quantity of research, that this is an ideal or all-purpose measure. It isn't.

Metrics are an aid to decision-making, not a shortcut to it. Their use demands more work in collecting, analysing and considering the data, but offers the prospect of a more thorough, informed and fairer review of research performance in concert with traditional peer review.

Reward effort, not luck

Jennifer Rohn
Wellcome Trust Fellow, University College London, UK

The current method of assessing scientists is flawed. The metrics I see being used by many evaluators are skewed towards outcomes that rely as much on luck as on skill and talent — such as hitting on the right place, time and trend to achieve a top-tier publication. In many professions, one's output is directly proportional to the amount of effort put in. Not so in science.

A promising new group leader might found a lab on an excellent, well-funded research plan, and work diligently for several years, only to discover quite late in the game — as commonly happens — that the project is doomed to failure. Or to have his or her project 'scooped' by a competing team. In these unfortunate situations, all this work will leave not a ghostly trace on the cited scientific record — and therefore in the eyes of most assessors, the person ceases to exist. Meanwhile, this group leader might have generated all sorts of helpful negative data, established a useful database used by the community and set up a complex experimental system co-opted by others to greater effect.

The efforts of such valuable but unlucky investigators need to be brought to light and rewarded. Giving credit for non-research activities — such as sitting on committees, public engagement, reviewing manuscripts, being a 'team player', proofing grants, raising crucial questions at seminars and otherwise enriching the community — is always going to be difficult. But there are ways to help make research success more proportional to the effort put in.

One solution is to establish more journals (or other formats) in which researchers can quickly and easily publish negative data, solid-but-uncelebrated results, raw data sets, new techniques or experimental set-ups, and even 'scooped' data. Publications such as *PLoS ONE* and *The Journal of Negative Results in BioMedicine* are rare steps in the right direction. In parallel, such 'lower-end' publications should be valued more when the time comes to recruit, fund or promote. We can't all be lucky enough to get *Nature* papers — but many of us make, through persistence and hard work, more humble cumulative contributions that in the long run may well be just as important. ■

1. de Solla Price, J. D. *Little Science, Big Science ... and Beyond* (eds Garfield, E. & Merton, R.) (Columbia Univ., 1986).
2. Nalimov, V. V. & Mul'chenko, Z. M. *Measurement of Science. Study of the Development of Science as an Information Process* (Foreign Technology Division, US Air Force Systems Command, 1971).
3. Moed, H. F. *Citation Analysis in Research Evaluation* (Springer, 2005).
4. Vinkler, P. *The Evaluation of Research by Scientometric Indicators* (Chandos, 2010).
5. Stern, S. *Manage. Sci.* **50,** 835–853 (2004).
6. Frey, B. S. *Public Choice* **116,** 205–223 (2003).
7. Deci, E. L., Koestner, R. & Ryan, R. M. *Psychol. Bull.* **125,** 627–668 (1999).
8. Butler, L. *Res. Policy* **32,** 143–155 (2003).
9. Frey, B. S. & Neckermann, S. *J. Psychol.* **216,** 198–208 (2008).
10. Langfeldt, L. *Res. Evaluat.* **15,** 31–41 (2006).

See Editorial, page 845, metrics special at www.nature.com/metrics, and comment on these articles at go.nature.com/tMSFQC.

© 2010 Macmillan Publishers Limited. All rights reserved