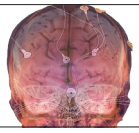
COMMENT

AI AlphaGo pioneer relishes Kasparov's intelligence book **p.413**



REPRODUCIBILITY The sins that led psychology off the righteous path **p.414**

PREDATORY PUBLISHING How to tell real journals from fakes now that Beall's list is gone? **p.416**

AGEING A 400-year-old theory of why Greek island of Ikaria is rich in centenarians **p.416**

ILLUSTRATION BY DAVID PARKINS

Blinkered by bibliometrics

Science panels still rely on poor proxies to judge quality and impact. That results in risk-averse research, say **Paula Stephan**, **Reinhilde Veugelers** and **Jian Wang**.

There is a disconnect between the research that reviewers purport to admire and the research that they actually support. As participants on multiple review panels and scientific councils, we have heard many lament researchers' reluctance to take risks. Yet we've seen the same panels eschew risk and rely on bibliometric indicators for assessments, despite widespread agreement that they are imperfect measures[1–6].

The review panels we observed last year were using bibliometrics in much the same way as they did before the 2015 Leiden Manifesto[4], the 2012 San Francisco Declaration on Research Assessment, which *Nature* is signing (see page 394), and similar exhortations against their use. After all, bibliometric measures offer a convenient way to help evaluate a large number of proposals and papers.

Although journal impact factors (JIFs) were developed to assess journals and say little about any individual paper, reviewers routinely justify their evaluations on the basis of where candidates have published. Panel members judge applicants by Google Scholar results and use citation counts to score proposals for new research. This practice prevails even at agencies such as the European Research Council (ERC), which instructs reviewers not to look up bibliometric measures.

As economists who study science and innovation, we see engrained processes working against cherished goals. Scientists we interview routinely say that they dare not propose bold projects for funding in part because of expectations that they will produce a steady stream of papers in journals with high impact scores. The situation may be worse than assumed. Our analysis of 15 years' worth of citation data suggests that common bibliometric measures relying on short-term windows undervalue risky research[7].

How can we move beyond declarations and wean reviewers off bibliometric indicators that bias decisions against bold work?

BACK-DOOR BIBLIOMETRICS

A few funding agencies in the Czech Republic, Flanders (northern Belgium) and Italy ask applicants to list JIFs alongside their publications, but such requirements are not the norm. The ERC, the National Natural Science Foundation in China, the US National Science Foundation and the US National Institutes of Health do not require applicants to report bibliometric measures.

They do anyway. Grant applicants to the Natural Science and Engineering Research Council of Canada (NSERC) may choose to list measures such as the number of citations of their publications along with JIFs and other metrics, such as the *h*-index, derived from citations. It is common for applicants to report JIFs and the citation counts of publications; they are often advised to do so by the institutions supporting their applications. When researchers are asked to select a handful of their most important papers, their justifications are often based on JIFs and short-term citation counts, not a more-nuanced assessment of the work's value. This is understandable; if bibliometric measures are not provided, reviewers often download them before making decisions.

When it comes to hiring and promotion, bibliometric indicators have an even ▶

larger, often formal, role. In Spain, the *sexenio* evaluation (a salary increase based on productivity[8]) depends heavily on rankings derived from JIFs. In Italy, a formal bibliometric profile of each candidate up for promotion is provided to reviewers. At many campuses in Europe, the United States and China, faculty members are given lists of which journals carry the most weight in assessing candidates for promotion. In some countries, notably China, bonuses are paid according to the prestige of the journal in which research is published[8].

At our own universities, it is standard for deans and department chairs to summarize candidates' citations and JIFs when committees discuss the merits of their work. Colleagues and external reviewers routinely refer to the bibliometric indicators of junior faculty members who are up for promotion. Hiring committees may also emphasize these indicators to select colleagues who are likely to attract funding for their institutions.

It is easy to see why. Public funders use these measures to allocate resources to universities. In turn, universities use them to distribute resources to departments. For instance, in Flanders, the formula used to allocate funds to universities includes publications weighted by JIFs. In Brazil, Qualis — an official system for classifying scientific production that is maintained by a government agency linked to the Brazilian Ministry of Education — uses JIFs to determine funding allocations. The UK Research Excellence Framework (REF) exercise is a rare exception in that it explicitly does not use JIFs.

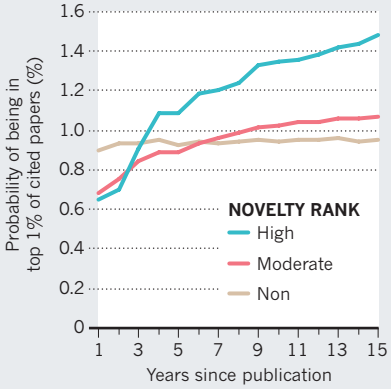
RESEARCH IMPACT

Much has been written about how the pursuit of flashy papers can push scientists to crowd into similar, competitive projects, cut corners or exaggerate the significance of their findings[4–6]. We think that the problem is more profound: popular short-term bibliometric measures discourage the kind of risky research that is likely to shift the knowledge frontier[7].

Using the Web of Science database, we analysed citations in more than 660,000 research articles published in 2001 to see how risky research tracks with JIFs and citation counts. Our proxy for risky research is whether a paper cites new combinations of journals in its reference list, taking into account the likelihood of such combinations. For example, a paper[9] in the *Journal of Biological Chemistry* revealed which protein a known antipsychotic drug interacted with, and used this to identify other biological effects. Its reference list is the first ever to pair the journal *Gene Expression* with others such as the *Journal of Clinical Psychiatry* and *Neuropsychopharmacology*. In fact, of the 861 journal pairs that could be created from its 42-item reference list, 9 are new.

NOVELTY NEEDS TIME

Highly original papers are more likely to be highly cited after three or more years.



SOURCE: REF. 3

We found that 89% of all papers made no new pairings of cited journals. Of the 11% that made new combinations, most (54%) made only one new combination.

We used these analyses to classify papers as 'non-novel', 'moderately novel' and 'highly novel', and compared how papers were cited from 2001 to 2015, a much longer period than is generally used to calculate JIFs. More-novel papers were more likely to be either a big hit or ignored compared with non-novel papers in the same field. The ones that became big hits took time to be recognized (see 'Novelty needs time').

For the first three years after publication, the probability that a highly novel paper was among the top 1% of highly cited papers was below that of non-novel papers; beyond three years, highly novel papers were ahead. We are not saying that non-novel papers cannot be important or influential, but that current systems of evaluation undervalue work that is likely to have high, long-term impact. Fifteen years after publication, highly novel papers are almost 60% more likely to be in the top 1% of highly cited papers. Highly novel papers also tend to be published in journals with lower impact factors.

In a nutshell, our findings suggest that the more we bind ourselves to quantitative short-term measures, the less likely we are to reward research with a high potential to shift the frontier — and those who do it. We hope our findings will resonate with the scientific community and encourage change.

NOW WHAT?

Boosting scientists' appetite for taking risks means shrinking the use of short-term bibliometric indicators. Our prescriptions are familiar, but our experience shows that they cannot be repeated enough.

Researchers: stop relying exclusively on short-term citation counts and JIFs in guiding the choice of topics and where to submit. Stop including them in funding applications.

Funders: insist on multiple ways to assess applicants' and institutions' publications. Stop grant applicants from providing short-term citation counts and JIFs. Scrub them from proposals and ban them from being discussed in reviews. Include experts outside the main field on review panels, and periodically examine the performance of grant applicants using five- or even ten-year windows. The ERC, although still too young to implement such a long-term window for evaluation purposes, aspires to do precisely this in the future.

Reviewers: resist seeking out and relying on metrics, especially when calculated over less than a three-year window.

Editors: reject shoddy metrics used to evaluate journals. Advocate for metrics to be assessed over longer time spans, as done by *Research Policy*[5] and recommended in a joint proposal from editors at several high-profile publishers, including *Nature*[10].

Universities: adopt as standard practice a requirement that committees actually read candidates' research, as is done in the REF exercise. Emphasize how researchers approach questions when evaluating candidates.

If we are to truly create more 'objective' assessments, all of us — from early-career researchers to the heads of funding agencies — need to use quantitative and qualitative tools responsibly. If we care about pushing the knowledge frontier forward, we must avoid indicators that penalize the types of researcher and project that have the greatest potential to push boundaries. ■ SEE EDITORIAL P.394

Paula Stephan *is a professor of economics at Georgia State University, Atlanta, Georgia, USA.* **Reinhilde Veugelers** *is a professor, and* **Jian Wang** *a postdoctoral researcher, at the Department of Managerial Economics, Strategy and Innovation (MSI) and Center for R&D Monitoring (ECOOM), University of Leuven (KU Leuven), Belgium.*
e-mail: pstephan@gsu.edu

1. Alberts, B., Kirschner, M. W., Tilghman, S. & Varmus, H. *Proc. Natl Acad. Sci. USA* **111,** 5773–5777 (2014).
2. Petsko, G. A. *Genome Biol.* **13,** 155 (2012).
3. Stephan, P. *How Economics Shapes Science* (Harvard Univ. Press, 2012).
4. Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. *Nature* **520,** 429–431 (2015).
5. Martin, B. R. *Res. Pol.* **45,** 1–7 (2016).
6. Monastersky, R. *Chron. High. Educ.* **52,** 14 (2005).
7. Wang, J., Veugelers, R. & Stephan, P. *Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators* NBER Working Paper No. 22180 (2016).
8. Franzoni, C., Scellato, G. & Stephan, P. *Science* **333,** 702–703 (2011).
9. Phiel, C. J. *et al. J. Biol. Chem.* **276,** 36734–36741 (2001).
10. Lariviere, V. *et al.* Preprint at bioRxiv http://dx.doi.org/10.1101/062109 (2016).