

Counting on citations: a flawed way to measure quality

Garry Walter, Sidney Bloch, Glenn Hunt and Karen Fisher

The journal Impact Factor and citation counts are misconstrued and misused as measures of scientific quality. Articles must be read in order to judge their quality. We have introduced a system, which may be easily replicated, to identify the best articles published in a journal. (MJA 2003; 178: 280-281)

GLOOM OR GLEE? Each September, journal editors and publishers anxiously await news of a particular figure from the Institute of Scientific Information (ISI) in Philadelphia, USA. The figure's value is promptly met with despair or delight. We are referring, of course, to the Impact Factor (IF) and the ritual surrounding its release that has come to dominate the editors' and publishers' calendar.

Two of us, as editors ourselves, participate in this practice, albeit reluctantly and with rising apprehension that scientific publishing is being undermined by "numerology". What was introduced as an aid to librarians more than four decades ago to guide their selection of scientific journals has become, in the view of some people, an inappropriate means of scrutinising an applicant's "track record" when allocating research funds or considering academic promotions.¹⁻⁴ Why inappropriate? Because the IF (defined as "the number of citations to a journal's articles published in the previous two years divided by the number of articles published by that journal during those two years"⁴⁻⁶) is conceptually and technically flawed, on a number of grounds:

- the quality of published material cannot be constrained by time — the two-year period set by the ISI for citations is arbitrary;⁷
- the number of journals in the ISI's database is a minute proportion of those published;⁴
- reviews are cited more frequently than original research, thus favouring journals that opt for these articles as part of a publishing strategy;¹
- the IF does not take into account self-citations, which amount to a third of all citations;⁸
- errors are common in reference lists (occurring in up to a quarter of references), inevitably affecting IF accuracy;⁹ and
- the assumption of a positive link between citations and quality is ill-founded, in that we cite articles for diverse reasons, including to refer to research judged suspect or poor.¹⁰

If these flaws were not enough to instil scepticism about the IF's validity and precision, then the arbitrary assumption that the quality of a specific article correlates with the IF of the journal in which it appears is entirely ill-founded and, in itself, sufficient to warrant concern about its continuing use. As editors and researchers, we are duty-bound to analyse this assumption rigorously. What do we find?

Consider two psychiatric journals published for a general readership: the *Australian and New Zealand Journal of Psychiatry* (ANZJP) and the *Canadian Journal of Psychiatry* (CJP). Applying the ISI's own "Web of Science" database,¹¹ we can examine the purported link between IF, citations and the worth of a particular article. For instance, if we calculate the proportion of citations to all articles in each journal that is accounted for by the most-cited 50% of papers, a striking pattern emerges. In the case of the ANZJP, the most-cited 50% of papers published between 1990 and 1995 account for 94% (range, 91% [1992] to 98% [1990]) of all citations to articles in that journal. Figures for the CJP are virtually identical: 94% (range, 91% [1991] to 96% [1994]). A blunt summary of these findings is that half the articles published in both journals receive virtually no citations. We can conclude from the data (and from comparable findings from other journals, such as those in cardiology¹⁰) that to determine the academic worth of a paper from the IF of the journal in which it appears is ill-conceived and misleading.

In an era of evidence-based medicine, its proponents avow that scientific progress can only be achieved by dint of diligent scrutiny of available data. Is it not incongruous, then, that the scientific community continues to cling to such an inadequate tool as the IF? What's more, the "parent" IF has spawned a range of flawed offspring, including "Scope-adjusted IF", "Discipline-specific IF", "Journal-specific influence factor", "Immediacy index" and "Cited half-life". As if that were not disconcerting enough, lo and behold, ISI recently faced a new rival, albeit short-lived (the venture collapsed in the wake of a threat from ISI to sue for violation of intellectual property rights). "PrestigeFactor.com" was launched in 2001, enticing us to ditch the IF and supplant it with another measure of journal quality, the "Prestige Factor" (PF).¹² Its proponents boldly asserted that the PF provided "truer value" than the IF.¹² There was, however, a fly in the ointment. Despite minor refinement (eg, the PF separated review articles from research reports and included citations to journal articles over the previous *three* years versus two), the underlying premise of both measures — that quality and number of citations are inextricably linked — was identical. It is also worth reporting the contemporary practice of open-access "e-journals" tracking their most popular articles through "hit rates".^{13,14} Again, we doubt that a popularity poll can indicate academic merit and fear that it may be misconstrued in this way.

Rivendell Unit, Child Adolescent and Family Psychiatric Services, Thomas Walker Hospital, Concord West, NSW.

Garry Walter, PhD, FRANZCP, Director; and Editor, *Australasian Psychiatry*; Karen Fisher, MB BS, Research Officer, Central Sydney Area Health Service.

Department of Psychiatry, University of Melbourne, Fitzroy, VIC.

Sidney Bloch, PhD, FRANZCP, Professor; and Editor, *Australian and New Zealand Journal of Psychiatry*.

Department of Psychological Medicine, Research Unit, Rozelle Hospital, Rozelle, NSW.

Glenn Hunt, MSc, PhD, Senior Research Fellow, University of Sydney.

Reprints will not be available from the authors. Correspondence: Dr Garry Walter, Thomas Walker Hospital ("Rivendell"), Hospital Road, Concord West, NSW 2138. gwalter@mail.usyd.edu.au

A watershed

We have reached a watershed — either we persevere with the notion that citations lie at the heart of scientific quality or we make a clean break. The latter option is attracting growing support. For instance, Richard Frackowiak, Dean of the Institute of Neurology in London, asserts that current measures are crude and reliance on them in making hiring-and-firing decisions is counterproductive.¹⁵ Zach Hall, a leading figure in US research, sees numerical methods as “excuses for not thinking”.¹⁵ David Adam, a writer for *Nature*, highlights the absurdity of the situation in Finland, where government funding of university hospitals utilises a sliding scale corresponding to the IF of journals in which researchers publish their work.¹⁶ A notable development is a similar questioning among scientific bodies. The *Deutsche Forschungsgemeinschaft*, Germany’s central research organisation, has promulgated innovative guidelines emphasising qualitative criteria in evaluating published material.¹⁷ As they posit, “Publications must be read and critically compared with the relevant state of the art and to the contributions of other individuals and working groups”.¹⁷ Admirable but vexing. How are we to judge quality objectively?

A more appropriate option?

We have grappled with this challenge and devised an option for the ANZJP. Five international members of the journal’s advisory board were invited in 2001 to identify that year’s “top” articles. The quintet, selected on the basis of scholarship, professional integrity and knowledge of scientific psychiatry, were asked to select three publications per issue that satisfied one or more of the following criteria:

- adds consequentially to the field through original, innovative research findings;
- expands or challenges current knowledge;
- opens additional areas for new research activity;
- opens a pathway to advance knowledge;
- integrates discoveries obtained by different approaches and/or disciplines through creative synthesis, thus bringing new insights to bear on original research; and
- reflects critically on research findings to guide the direction of further research.

The data were collated, and the titles of the nine articles gaining the most votes were announced in the April 2002 issue of the journal and posted on the websites of the Royal Australian and New Zealand College of Psychiatrists and Blackwell Publishing. This procedure is familiar in that it is, in essence, an extension of peer review. While by no means foolproof, we proffer this approach as a fresh way to establish the quality of the individual article. We sought feedback from the judges and learned that the task is feasible and the clarity and utility of the assessment criteria are satisfactory. The judges also found the assignment personally rewarding, even enjoyable! We are examining the method’s reliability. Testing validity, of course, is more taxing given the lack of an objective yardstick.

Interestingly, our experiment has been echoed by another initiative to highlight meritorious papers, namely the “Faculty of 1000” (F1000). Launched in November 2001 by the publishers of BioMed Central (a collection of wholly electronic biomedical “journals”),^{18,19} F1000 aims to identify the best papers in the basic biological sciences through the eyes of a “faculty” of over 1000 selected scientists who are experts in their fields. Thus, for

instance, cell biology is divided into 18 categories, and the faculty members for each category select two to four papers each month from any journal, ranking them as “recommended”, “must read” or “exceptional”. The experts also briefly explain their choices. We applaud this initiative and see our own effort as complementary. Indeed, it is commonsensical that more than one method of identifying outstanding papers should be instituted, as there cannot be an absolute consensus.

An invitation

We invite editors, publishers and authors to consider trying our experiment. We selected a new set of judges, again all distinguished figures in international psychiatry, to undertake a similar task for articles published in 2002. After this replication, we hope to be well placed to determine whether the method needs change. It would be disingenuous of us to conceal our fantasy that the annual “gloom or glee” ritual will be supplanted by published lists of “best quality articles”, with their authors duly acknowledged. We may even witness the ISI collating the results, including the names of the judges and the criteria applicable for participating journals (consensually agreed criteria across all journals would be ideal). The implications are clear: successful authors will be able to cite articles that have made it to the “top” when documenting their academic track record, whatever the purpose (eg, applying for research grants). Depending on a measure devoid of any rational link to the appraisal of academic worth will be but a hazy memory.

Competing interests

None identified.

References

1. Seglen PO. Why the impact factor of journals should not be used for evaluating research. *BMJ* 1997; 314: 498-502.
2. Hecht F, Hecht BK, Sandberg AA. The Journal Impact Factor: a misnamed, misleading, misused measure. *Cancer Genet Cytogenet* 1998; 104: 77-81.
3. Coleman R. Impact Factors: use and abuse in biomedical research. *Anat Rec* 1999; 257: 54-57.
4. Bloch S, Walter G. The Impact Factor: time for change. *Aust N Z J Psychiatry* 2001; 35: 563-568.
5. Garfield E. Journal Impact Factor: a brief review. *CMAJ* 1999; 161: 979-980.
6. Institute of Scientific Information. ISI Web of Knowledge. Available at: <http://www.isinet.com> (accessed Jul 2002).
7. Hansson S. Impact factor as a misleading tool in evaluation of medical journals [letter]. *Lancet* 1995; 346: 906.
8. Seglen PO. Citation and journal impact factors: questionable indicators of research quality. *Allergy* 1997; 52: 1050-1056.
9. Seglen PO. Citation and journal impact factors are not suitable for evaluation of research. *Acta Orthop Scand* 1998; 69: 224-229.
10. Opthof T. Sense and nonsense about the Impact Factor. *Cardiovasc Res* 1997; 33: 1-7.
11. Institute of Scientific Information. Web of Science. Available at: <http://wos.isiglobalnet2.com> (accessed Dec 2001).
12. Prestige factor. Available at: <http://www.prestigefactor.com> (accessed Nov 2001).
13. Journal of Medical Ethics online. Available at: <http://www.jmedethics.com> (accessed Dec 2002).
14. Medical Science Monitor online. Available at: <http://www.medscimonit.com> (accessed Dec 2002).
15. Citation data: the wrong impact? [editorial]. *Nat Neurosci* 1998; 1: 641-642.
16. Adam D. The counting house. *Nature* 2002; 415: 726-729.
17. Deutsche Forschungsgemeinschaft. Recommendations of the Commission on Professional Self Regulation in Science. Proposals for safeguarding good scientific practice. January 1998. Available at: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf (p.8) (accessed Sep 2002).
18. Faculty of 1000. Available at: <http://www.facultyof1000.com> (accessed Jul 2002).
19. Recommended reading? [editorial]. *Nat Med* 2002; 8: 1.

(Received 26 Aug 2002, accepted 12 Dec 2002)

