Learned Publishing, 21, 85–91
doi: 10.1087/095315108X288901

Trends in the use of ISI citation databases for evaluation

James PRINGLE
Thomson Scientific

ABSTRACT. This paper explores the factors shaping the current uses of the ISI citation databases in evaluation both of journals and of individual scholars and their institutions. Given the intense focus on outcomes evaluation, in a context of increasing 'democratization' of metrics in today's digital world, it is easy to lose focus on the appropriate ways to use these resources, and misuse can result.

Measuring performance in an age of democratization

Today's researchers build their careers in an intensely outcomes-driven and metrics-oriented academic world, and it has never been more important to use care and good judgment in evaluation. A growing chorus of voices from the research community decries the over-use of research metrics and appeals for greater balance in evaluative judgments. Some strike a passionate tone of rage against 'fashion, the management cult and the politics of our time, all of which favor numerical evaluation of "performance" and reward compliance'.[1] Thomson Scientific, the producer of the ISI citation databases, continually joins this chorus and reminds the research community that citation metrics should support, but never substitute for, informed peer judgment in research evaluation. Yet such efforts often seem an exercise in sweeping back the tide.

The factors driving this trend are immensely powerful and global in nature. Among them are:

- Funding pressures, evidenced by increased rejection rates of grant proposals in some countries.
- Efforts by university administrators to derive objective measures for promotion and tenure decisions free of bias and 'old boy' networks.
- Pressures on these same administrators to upgrade the reputation and quality of their programs to compete effectively in global academic markets – and to demonstrate that they have done so.
- Global competition in the sciences, leading to national policies designed to increase global competitiveness and to demonstrate national achievement.
- Pressures in the journal publishing industry, where sluggish revenue growth rates and pricing pressures heighten the com-

James Pringle

petitive need to demonstrate journal reputations in order to protect subscriptions and to ensure continued participation by top authors.

The interplay of these forces is taking place in a context of increased availability of data with which to measure performance in the digital world. The first decade of the 21st century has seen what might be called the 'democratization' of citation metrics. Every researcher at virtually every research institution of significance has access to the ISI citation databases in the form of the *Web of Science®*. ISI citation data is complemented by a growing number of other citation databases, including some that are openly available over the Internet, such as the *Astrophysics Data System* (ADS) and *Google Scholar*. Many of these researchers are eager to test new ways of calculating citation metrics and are quick to comment about their findings globally in interconnected Web communities.

the citations found in the ISI citation databases represent 'community-generated' content created by the research community

In earlier decades, citation metrics were the preserve of a limited group of specialists in 'bibliometrics' and 'scientometrics', who shared their expertise with each other and with pockets of high-level analysts in government departments and major publishing houses. Today, citation metrics are widely used and debated by a broad public of administrators, analysts, editors, librarians, and individual researchers.

The power of the trends underlying outcomes evaluation and the ubiquity of citation data in a 'democratized' digital world portend that use of quantitative citation data to measure academic performance will only increase in the coming years. Appreciating these trends helps frame the issue of proper use of citation metrics today, and emphasizes that it is more critical than ever to understand what constitutes good use of citation data in research evaluation.

Citations as community-generated content

What attracts such a wide audience, and makes debates about citation metrics so lively, is the special character of citations. To describe this special character in a way that is compatible with today's 'Web 2.0' world, we might say that the citations found in the ISI citation databases represent 'community-generated' content created by the research community as it engages in formal scholarly communication via peer-reviewed journals.

As journals migrate to the Internet, they are becoming centers where new types of user-generated content can be captured, including usage data, social tags, inbound web links, and community votes. While these types of data may at some point prove to be complementary indicators to citation data, as yet their meaning is only beginning to be understood. Citation data as an indicator of influence in the scholarly communication process, in contrast, have been explored and studied for decades. For this reason, there is currently no comprehensive alternative to citation data for quantitative assessment of research outputs, and it is not surprising that all parties interested in metrics turn to these data.

The ISI citation databases are widely used for these purposes because their continuously growing corpus of journal literature based on careful selection criteria, their non-duplicative coverage, and their focus on serious scholarly journals represent a clearly defined universe within which to measure and study this formal community-generated content. Other citation resources may show similar citation patterns,[2] and indeed, given the many different types of documents indexed within them (articles, books, conference abstracts, working papers, preprints, student term papers), may eventually show higher absolute numbers of citations than does the *Web of Science*. However, such resources are more difficult to utilize for evaluative purposes because it is not always clear what is being counted, and their data structures may not lend themselves as effectively to extraction of comparative name, journal, and institutional address data. As a result, one cannot be sure what is actually being compared to what.

The absence of 'best practices'

ISI citation data are most conspicuously used for journal ranking. They are also used for many other evaluative purposes, involving individuals, universities and university programs, national assessment programs and inter-country comparisons. At the level of national and international policy, such organizations as the National Science Foundation (NSF), the European Union (EU), and the Observatoire des Sciences et des Techniques (OST) have used ISI citation data for many years to chart the progress of the sciences in their regions. A recent study by the NSF, available online, provides a good example of this type of use.[3] Where and how citation metrics are used in national assessment programs generally follows the structure of scientific funding on a country basis. Where central funding of universities is more important, as in a number of European countries, citation metrics are widely used at the national level. In the USA, where funding patterns are more diverse, there is less emphasis on citations in national assessment and more activity at the local university level. Assessments of individuals for promotion and tenure decisions may be conducted as an extension of a national program or as an individual departmental activity, with many variations in practice.

In this period of experiment within a more democratic digital world, there are no recognized 'best practices'. Some analysts are seeking to develop such guidelines, but these are as yet not widely employed or agreed upon.[4] In the absence of clear guidelines, it is no wonder that citation metrics are sometimes subject to misuse.

in the absence of clear guidelines, it is no wonder that citation metrics are sometimes subject to misuse

What constitutes misuse, however, is not always black and white. Citation data can be used effectively in many different ways to answer many different questions. It is important to use such data in a way that contributes to a deeper understanding and better judgment about a question of interest. Because the nature of the questions asked can be so varied, the charge of 'misuse' must be leveled with an understanding of the specific use to which the data are being applied.

Several years ago, an analyst for Thomson Scientific proposed a set of rules for applying citation analysis to evaluation.[5] This presentation remains unpublished, but I will extract from it two key 'rules' that particularly stand out as relevant to the present analysis, and I will call them the 'golden rules'.

- First, 'consider whether the available data can address the question'. Too often, a user of citation data, attracted to their ease of quantification, treats citation analysis like a hammer in search of nails, rather than determining whether the data are available and, to rephrase slightly, appropriate to the question asked.
- Second, 'compare like with like'. This fundamental rule of citation analysis pervades every evaluative exercise, but is often honored only in the breach. Whenever one hears the question 'is a high journal impact factor always better?', or when an assistant professor in economics is judged less worthy because she 'has fewer citations' than her colleague in immunology, this rule has been violated.

The journal impact factor debate

Both rules are in play in the most active area of debate in citation metrics, that concerning the journal impact factor (JIF). The JIF is the best-known example of citation metrics used for evaluation. It is a journal-level metric designed for one purpose – to compare the citation impact of one journal with other journals. Because it has become so well known, 'impact factor' has almost come to stand as shorthand for any use of citation metrics of which one disapproves. Yet most of the issues raised concerning the JIF in evaluation, when these are not simply data issues affecting a limited number of cases, result from a failure to follow the two rules

The journal impact factor

The journal impact factor (JIF), often known simply as the impact factor (IF), is a ratio between citations and recent citable items published. It is calculated as follows:

A = total cites in 2006
B = 2006 cites to articles published in 2004–5 (a subset of A)
C = number of articles ('citable items', excluding editorials, letters, news items, and meeting abstracts) published in 2004–5
D = B/C = 2006 impact factor

listed above, and concern uses of the JIF for purposes other than that for which it was designed.

There is no shortage of criticism of this particular metric – criticism largely born of its success in becoming widely used and available – and thus part of 'democratization' in the digital world. While the JIF is a good general comparative measure of journal performance, it has limitations, as does any statistical measure, which have been voluminously documented in recent years.

Initially, critics pointed to factors affecting the calculation of the JIF itself, such as a difference between items counted in numerator and denominator and the possibility that self-citation may artificially inflate the JIF of particular journals.[6,7] These limitations exist, but appear to affect only a very small percentage of journals in any year.[8,9]

Other critics point out structural limits such as the perceived short time window of the calculation. It is readily demonstrable that journals in different disciplines have different citation patterns. This difficulty is easily overcome by comparing like with like – comparing journals with others in the same field, rather than with journals from another field. In this vein, refinements of the JIF have been proposed for cross-disciplinary studies that take into account these differing patterns. Calculations such as 'rank-normalized' impact factors,[10] or including weightings based on citation time periods,[11] can be used to correct for this variation. The number of such 'corrective' measures is constantly growing.

some new journal metrics employ complex methodologies for uncertain results

Another line of criticism charges that over-use of 'impact factors' has shifted behavior among authors and publishers in undesirable ways. This line of criticism has become a popular journalistic theme.[12] However, given the fundamental trends already mentioned, it is likely that this criticism actually applies to any possible quantitative ranking system used to compare journals. Pressure to publish in top journals derives from an intensifying outcomes-based culture, not from a particular measure. It is highly debatable, for example, whether *Science* and *Nature* would have lower rejection rates if the JIF did not exist. Furthermore, studies of actual editorial behavior undertaken to influence the JIF for particular journals show that most editors sought to raise their journals' impact factors over time by applying sound editorial practices and cultivating top authors, not by editorial manipulation of citing behaviors.[13]

Variants on the JIF, and proposed alternatives to it, are appearing at an increasing rate, which a statistical study, if one were conducted, might possibly find to be running as high as 1–2 new journal metrics per month. Each such new metric tends to be hailed as the ultimate solution to the problems of journal ranking. It is questionable, however, whether the sheer multiplication of metrics will somehow lead us to a fundamental change in practice, or whether the adoption of new metrics will fundamentally change our view of the overall ranking of journals.

Some new journal metrics employ complex methodologies for uncertain results. For example, the Eigenfactor is a metric that uses the journal citation matrices presented in the *Journal Citation Reports®* to rank journals based on their position within the overall network of journal citation relationships.[11,14] It relies on a PageRank-like algorithm that weights citations based on the citation counts of the citing journal in a complex vector algorithm. Such a mathematical approach has been proven to achieve more relevant web searching, but whether it also generates better evaluative rankings is only beginning to be tested. Allowing a citation from a highly cited journal to 'count more' than a citation from a less-cited journal may prove to be highly controversial in principle, once it is more widely understood.

Likewise, the prospect of usage metrics seems to promise a 'counterweight' to the JIF.[12,15] This idea merits consideration, but the meaning of a download currently remains much less well understood than that of a citation, and so the meaning of such metrics is still to be determined until they have been similarly battle-tested. For some disciplines, in which publications are oriented toward practitioners who read but rarely cite, usage may in fact prove to be a useful complement to citation in journal ranking. Elsewhere, its role is less clear.

These new approaches are oriented towards specific issues in journal ranking. However, they do not address the issue of broader use of the JIF in research evaluation. It is here that the two rules listed above apply most directly. The use of the JIF in these contexts tends to take one of two main forms:

- An indicator of success achieved in having an article accepted by a prestigious journal.
- A surrogate for a more carefully derived direct measure of citation impact.

While the first use may have some utility, the second appears difficult to justify.

The first use – as an indicator that the author has succeeded in having an article accepted in a prestigious journal – has some justification. There is a hierarchy of journals within subject areas, and this hierarchy broadly corresponds to impact rankings, so in a formal communication system based on peer review, acceptance of an article for publication in one of these journals is an important scholarly achievement. Rewards based on this achievement can be seen as encouraging scholars to aim high in their publishing goals. The data may be appropriate to the question being asked, if one is comparing achievement within a given field. However, comparing like with like, our second rule, is more difficult to apply across disciplines since JIFs vary greatly across disciplines. Rank in category is very important and can be addressed in several ways. One way is to create a quadrant division among journals within disciplines, and then compare publication in quadrants across disciplines.[16]

But the use of the JIF does not end there. It is also used as a surrogate for more direct measures in such concepts as a 'total impact factor' or other calculations, in which the JIF stands in for article performance and further calculations are performed on lists of JIFs. It is very hard to see how such data, so manipulated, are appropriate for any question related to evaluation or comparison.

In response to issues raised in the perceived misuse of the JIF in evaluation, calls for the use of 'article-level metrics' have been made, and the most popular of such metrics at the moment is the 'h-index'.[17] This metric, and an increasing number of other article-level metrics, are readily calculated in the *Web of Science*, and even in some of the openly accessible citation resources.[18]

h-index

The h-index was first proposed by J.E. Hirsch in 2005. He defines it as follows: 'A scientist has index h if h of his or her N_p papers have h citations and the other $(N_p - h)$ papers have $\leq h$ citations each.'[17] This metric is useful because it discounts the disproportionate weight of highly cited papers or papers that have not yet been cited.

Although article-level metrics appear to be more appropriate for questions about individual performance than the JIF, they may easily violate the rule that calls for us to 'compare like with like', at least until these metrics are better understood. A simple list of h-indexes, without reference to time periods or the consistency of the underlying data, may prove to be much more dangerous and deceptive than any proposed misuse of the JIF. Examples of this unbounded and uncorrected use are easily found. See, for example, a recent list published in *Retrovirology*.[19] Furthermore, in today's highly collaborative research environments, even the concept of deriving a metric reflecting individual performance from a set of coauthored papers, without some assignment of shares or roles, raises the issue of whether the data are in fact appropriate to the questions being asked of them.

comparing like with like is more difficult across disciplines since JIFs vary greatly across disciplines

'Horses for courses' in research evaluation

Evaluative uses of citation data are most effective when the citation data and metrics are carefully designed for the particular purpose and rely on appropriate benchmark and baseline data for comparisons. For this reason, Thomson Scientific provides many different metrics for different purposes and continues to develop new ones as specific new evaluative needs arise. The many partners and consultants who add value and resell ISI citation data generally do likewise.

Cited half life

The number of publication years from the current *Journal Citation Reports* year that account for 50% of the citations received by the journal. This metric is useful because it helps show the speed with which citations accrue to the articles in a journal and the continued use and value of older articles.

A few examples from the Thomson portfolio will illustrate these points.

In terms of journal level metrics, Thomson's *Journal Citation Reports*® annually provides many different metrics: the JIF, a journal-level Immediacy Index to show how rapidly journals accumulate citations in their first year, cited and citing 'half life' indicators and 10-year charts of citations to show how citations accumulate over periods of time, a chart of self-citations to show the percentages of citations to the same journal, and a *Journal Relatedness Index*,[20] to show the journals most closely associated with a given journal in terms of citations, plus basic statistics and rank in category and the full set of article, article type, and citation counts that support these calculations. Each of these metrics helps answer specific questions about the journal, and the user is encouraged to explore the data and use it appropriately, instead of simply grabbing the JIF and using it as the universal answer to any evaluative question.

citation data can be used productively and helpfully in evaluative processes

More advanced journal-level analysis is supported by more customized and flexible products, most notably the *Journal Performance Indicators*™, which allow calculation of JIF-like metrics for the user's chosen time periods and a more fine-grained analysis by document types and field-specific baselines. More locally focused, workflow-oriented analysis is supported by the *Journal Use Reports*™, which enable users to compare journal-level citation metrics with the publication, citation, and usage data for their own institutions.

For the most part, these tools do not support research evaluation, because of the limitations of journal-level metrics outlined earlier. For research evaluation, Thomson Scientific recognizes the need for custom data sets and for comparative data appropriate to the specific evaluative task and provides these for both standard and custom data sets as needed. Such metrics are built up from the individual article level and include, among many others:

- Simple counts of cites, papers, and cites per paper, providing basic statistics and baselines over a variety of time periods, as well as standard statistical measures such as mean, median and mode.
- Second-generation citation counts (citations from articles that cite the 'citing articles' of the original article), normalized by field, which measure of the long term impact of a paper.
- Time series data covering both standard citing and cited periods (one-year, five-year, 25-year, etc.) and user-defined time periods.
- Percent shares of papers or citations within a broader category (such as a field, country, or institution).
- Field baselines, representing the mean impact (cites per paper) for papers in a field (usually a journal set) defined for a specific cited/citing time window.
- Expected citation counts, which predict how often a paper is expected to be cited based on its year, journal, and article type.

The variety of these metrics and data sets reflects the recognition of the two 'golden rules' of citation data as applied to evaluation – that data be appropriate to the questions asked, and that all comparative questions rely on comparisons of like with like. When these rules are kept in mind, and those who use citation data take the time and patience to apply them carefully, citation data can be used productively and helpfully in evaluative processes.

As Eugene Garfield , the founder of ISI and inventor of the ISI citation databases, has remarked, 'In 1955 it did not occur to me that "impact" would become so controversial.'[9] The pressures of outcomes-based funding decisions and personal rewards systems, along with the ready availability of citation data in a more democratized digital world, will continue to fuel this controversy. In view of these powerful trends, it is more important than ever to remind ourselves that citation metrics must be used carefully, and should only be used to inform, not to supplant, peer review and other forms of qualitative evaluation.

citation metrics must be used carefully, and should only be used to inform, not to supplant, peer review

References

1. Lawrence, P. 2007. The mismeasurement of science. *Current Biology* 17: R583–5. http://dx.doi.org/10.1016/j.cub.2007.06.014
2. Pauly D. and Stergiou K. 2005. Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics* 2005: 33–5. Available at http://www.int-res.com/articles/esep/2005/E65.pdf
3. Hill, D., Rapoport, A., Lehming, R. and Bell, R. *Changing US Output of Scientific Articles: 1988–2003*. Arlington, VA, National Science Foundation, Division of Science Resources Statistics, 2007. Available at http://www.nsf.gov/statistics/nsf07320/
4. Moed, H. *Citation Analysis in Research Evaluation*. Amsterdam, Springer, 2005.
5. Pendlebury, D. *The ISI Database and Bibliometrics: Uses & Abuses in Evaluating Research*. Unpublished presentation at Thomson ISI Symposium, Tokyo, 2002.
6. Moed, H. and Vanleeuwen, T. 1995. Improving the accuracy of Institute for Scientific Information's Journal Impact Factors. *Journal of the American Society for Information Science* 46: 461–7. http://dx.doi.org/10.1002/(SICI)1097-4571(199507)46:6<461::AID-ASI5>3.0.CO;2-G
7. Seglen, P. 1997. Citations and journal impact factors: questionable indicators of research quality. *Allergy* 52: 1050–6. http://dx.doi.org/10.1111/j.1398–9995.1997.tb00175.x
8. McVeigh, M. *Journal Self-Citation in the Journal Citation Reports – Science Edition (2002)*. Philadelphia PA, Thomson Corporation, 2004. Available at http://www.thomsonscientific.com/media/presentrep/essayspdf/selfcitationsinjcr.pdf
9. Garfield, E. *The Agony and the Ecstasy – The History and Meaning of the Journal Impact Factor*. International Congress on Peer Review And Biomedical Publication, Chicago, 16 September 2005. Available at http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf
10. Pudovkin A. and Garfield, E. Rank-normalized impact factor: a way to compare journal performance across subject categories. *Proceedings of the 67th ASIS&T Annual Meeting*. Providence, RI, American Society for Information Systems & Technology, 2004. Available at http://www.garfield.library.upenn.edu/papers/asistranknormalization2004.pdf.
11. Sombatsompop, N., Markpin, T. and Premkamolnetr, N. 2004. A modified method for calculating the impact factors of journals in ISI Journal Citation Reports: Polymer Science Category in 1997–2001. *Scientometrics* 60: 217–35. http://dx.doi.org/10.1023/B:SCIE.0000027794.98854.f6
12. Monastersky, R. 2005. The number that's devouring science. *Chronicle of Higher Education* 52: A12–17. Available at http://chronicle.com/free/v52/i08/08a01201.htm.
13. Chew, M., Villanueva, E. and Van Der Weyden, M. 2007. Life and times of the impact factor: retrospective analysis of trends for seven medical journals (1994–2005) and their Editors' views. *Journal of the Royal Society of Medicine* 100: 142–50. http://dx.doi.org/10.1258/jrsm.100.3.142
14. Bergstrom, C. http://www.eigenfactor.org/methods.htm.
15. Shepherd, P. *Final Report on the Investigation into the Feasibility of Developing and Implementing Journal Usage Factors*. UK Serials Group, 2007. http://www.uksg.org/usagefactors/final.
16. Magri, M.-H. and Solari, A. 1996. The SCI journal citation reports: a potential tool for studying journals? *Scientometrics* 35: 93–117. http://dx.doi.org/10.1007/BF02018235
17. Hirsch, J. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science*, 102: 16569–72. http://dx.doi.org/10.1073/pnas.0507655102
18. Harzing, A.-W. http://www.harzing.com/resources.htm#/pop.htm. Accessed October 2007.
19. Jeang, K. 2007. Impact factor, H index, peer comparisons, and *Retrovirology*: is it time to individualize citation metrics? *Retrovirology* 4: 42. http://dx.doi.org/10.1186/1742–4690–4–42
20. Pudovkin, A. and Garfield, E. 2002. Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology* 53: 1113–19. http://dx.doi.org/10.1002/asi.10153

James Pringle
Vice President, Product Development
Thomson Scientific
3501 Market Street
Philadelphia, PA, 19104, USA
Email: james.pringle@thomson.com