

The evolution of journal assessment

snip & sjr
new perspectives in journal metrics
www.journalmetrics.com

July 2011

There is no single “best” indicator that could accommodate all facets of the new reality of the journal metrics

Wolfgang Glänzel, of the Expertise centrum O&O Monitoring (Centre for R&D Monitoring, ECOOM)

Abstract

Journal metrics are central to most performance evaluations, but judging individual researchers based on a metric designed to rank journals can lead to widely recognized distortions. In addition, judging all academic fields and activities based on a single metric is not necessarily the best basis for fair comparison

Bibliometricians agree that no single metric can effectively capture the entire spectrum of research performance because no single metric can address all key variables. In this white paper, we review the evolution of journal metrics from the Impact Factor (IF) until today. We also discuss how research performance assessment has changed in both its scope and its objectives over the past 50 years, finding that the metrics available to perform such increasingly complex assessments have not kept pace, until recently.

There has been rapid growth in the field of citation analysis, especially journal metrics, over the last decade. The most notable developments include:

- Relative Citation Rates (RCR) / Journal to Field Impact Score (JFIS)
- The h-index
- Article Influence (AI)
- SCImago Journal Rank (SJR)
- Source-Normalized Impact per Paper (SNIP)

Each metric has benefits and drawbacks, and it is our intention to analyze each metric's strengths and weaknesses, to encourage the adoption of a suite of metrics in research-performance assessments and, finally, to stimulate debate on this important and urgent topic.

The evolution of journal assessment

Most people in the academic world are affected by performance and ranking assessments to some degree. Every level of academic life – from individual researchers and research groups to universities, and even entire countries or regions – is increasingly benchmarked and ranked to measure return on investment.

While no one doubts the need to review and assess research performance, there is intense debate across academia on how these assessments are performed and what they are used for.

Contents

The evolution of journal assessment	3
Benchmarking performance	4
Citations as proxies for merit	4
Rewarding speed and quantity	5
Natural evolution	6
Demand for more choice	7
Tailoring the metric to the question	8
Weighing the measures	8
A balanced solution	9
Comparison table of SNIP, SJR, IF, AI and JFIS	10-11
References	12
Featured metrics	12

Benchmarking performance

Research funding comes from governments, funding agencies, industry, foundations and not-for-profit organizations, among others, and each naturally wants the best outcome for their investment. Governments in particular are increasingly viewing their country's research and development capabilities as key drivers of economic performance.

[1] These stakeholders use various performance indicators to determine where the best research is being done, and by whom, but are also demanding increasingly detailed reporting on how funds are used. Consequently, decision-makers are seeking more and better methods of measuring the tangible social and economic outcomes of specific areas of research.

In response, those tasked with administering a university's finances, allocating its resources and maintaining its competitiveness are increasingly managing their universities like businesses, and are demanding indicators that give insight into their performance compared with their peers.

On the individual level, many academics also use indicators to measure their own performance and inform their career decisions. In journal management, the IF is used by editors to see whether their journal is improving relative to its competitors, while researchers take note of these rankings to ensure they derive maximum benefit from the papers they publish.

According to Janez Potocnik, European Commissioner for Science and Research: "Rankings [of institutes] are used for specific and different purposes. Politicians refer to them as a measurement of their nation's economic strength, universities use them to define performance targets, academics use [them] to support their professional reputation and status, students use [them] to choose their potential place of study, and public and private stakeholders use [them] to guide decisions on funding. What started out as a consumer product aimed at undergraduate domestic students has now become both a manifestation and a driver of global competition and a battle for excellence in itself." [2]

Citations as proxies for merit

In most fields, research results are communicated via papers in scholarly journals, although conference proceedings, books and patents may also play a role.

These publications cite earlier research the authors have found useful or wish to respond to in some way and, in turn, attract citations from researchers who find their work worth citing. It is, therefore, possible to gauge a publication's impact by counting the number of citations it attracts. And when Eugene Garfield first introduced the idea of using citations to create a journal impact factor in 1955, followed by publication of the Science Citation Index in 1961, the world of academic assessment changed forever. [3]

The Impact Factor (IF) counts citations in a given year to any item published in a journal during the previous two years, and divides this by the total number of articles and reviews published in the same two-year period.

Editors have grown to rely on the IF as a key indicator of their journals' performance, and, when used sensibly to track performance over time and within context, it is an excellent indicator.

The IF has an unprecedented level of influence in performance evaluation, and some editors tend to focus on increasing citations more than on any other criterion of quality. Like many indicators, manipulation is possible. For instance, it is widely known that review articles are generally cited more than research papers, and review journals often have the highest IFs in their respective fields. Therefore, journals wishing to push up their IF can include more reviews. It is questionable, however, how much this actually serves the community or science.

In addition, because the IF only counts articles and reviews in its denominator, any citations to other published items, such as editorials or letters, are shared among a smaller number of citation targets. This can reduce the ability to make reliable judgments of quality using the IF.

Rewarding speed and quantity

The IF also receives criticism because it favors fields that cite heavily and rapidly: journals serving fields in the life sciences, for example, tend to have higher IFs than those in mathematics or the social sciences.

In a report on Citation Statistics, the International Mathematical Union (IMU), in cooperation with the International Council of Industrial and Applied Mathematics and the Institute of Mathematical Statistics, says: "The two-year period used in defining the impact factor was intended to make the statistic current. For some fields, such as bio-medical sciences, this is appropriate because most published articles receive most of their citations soon after publication. In other fields, such as mathematics, most citations occur beyond the two year period. Examining a collection of more than three million recent citations in mathematics journals (the Math Reviews Citation database)

one sees that roughly 90% of citations to a journal fall outside this two-year window. Consequently, the impact factor is based on a mere 10% of the citation activity and misses the vast majority of citations." [4]

In addition, any publications not covered by Thomson Reuters Web of Science do not receive IFs and are simply not ranked. As the IMU et al. note: "Thomson Scientific indexes less than half the mathematics journals covered by Mathematical Reviews and Zentralblatt, the two major reviewing journals in mathematics." [4]

Natural evolution

The IF's success in measuring journal prestige, as well as its universal acceptance and availability, has led to a great number of "applications" for which it was never conceived. And Garfield himself has acknowledged: "In 1955, it did not occur to me that 'impact' would one day become so controversial." [3]

The IF represents the average impact of an article published in that journal. However, it is not possible to ascertain the actual impact of a particular paper or author based on this average: the only way to do this is to count citations received by the paper or author in question (see case study 1). Yet, the IF is often used to do exactly this. Article values are often extrapolated from journal rankings, which are also being used to assess researchers, research groups, universities, and even entire countries.

Professor David Colquhoun of the Department of Pharmacology, University College London, voiced his concerns about this in a letter to Nature. [5] "No one knows how far IFs are being used to assess people, but young scientists are obsessed with them. Whether departments look at IFs or not is irrelevant, the reality is that people perceive this to be the case and work towards getting papers into good journals rather than writing good papers. This distorts science itself: it is a recipe for short-termism and exaggeration." [6]

Case study 1: Journal metrics cannot predict the value of individual articles

Suppose I have two journals, journal A and journal B, and the IF of A is twice that of B; does this mean that each paper from A is invariably cited more frequently than any paper from B? More precisely: suppose I select one paper each from B and A at random, what is the probability that the paper from B is cited at least as frequently as the paper from A?

The Transactions of the Mathematical Society has an IF twice that of the Proceedings of the Mathematical Society (0.85 versus 0.43). The probability that a random paper from the Proceedings is cited at least as frequently as one from the Transactions is, perhaps surprisingly, 62%. [4]

Professor Henk Moed carried out the same type of analysis on a large sample of thousands of scientific journals with IFs of around 0.4 and 0.8 in some 200 subject fields, and found that the International Mathematical Union (IMU) case is representative. He obtained an average probability of 64% (with a standard deviation of 4) that the paper in the lower-IF journal was cited at least as frequently as the paper in the higher-IF journal.

In its report on Citation Statistics, the IMU, in cooperation with the International Council of Industrial and Applied Mathematics and the Institute of Mathematical Statistics, says: "Once one realizes that it makes no sense to substitute the impact factor for individual article citation counts, it follows that it makes no sense to use the impact factor to evaluate the authors of those articles, the programs in which they work, and (most certainly) the disciplines they represent. The impact factor, and averages in general, are too crude to make sensible comparisons of this sort without more information. Of course, ranking people is not the same as ranking their papers. But if you want to rank a person's papers using only citations to measure the quality of a particular paper, you must begin by counting that paper's citations." [4]

Demand for more choice

There is clearly a market need for additional metrics that can reliably compare authors, institutions and countries, and in response, many new metrics have been developed over the past decade.

One of the most successful of these is the h-index, which was originally created to compare individual researchers. Conceived by Jorge Hirsch in 2005, the h-index is the number of papers by a particular author that receive h or more citations. While the h-index has proved highly popular, it too suffers from some of the same issues as the IF: bias (see case study 2) and misuse.

To many critics, bibliometrics is the source of this "worrying" trend that fails to fully take into account how research is conducted in the various disciplines, but that is not the whole picture.

Bibliometricians caution against over-reliance on, and misuse of, their tools. Wolfgang Glänzel says: "Uninformed use of bibliometric indicators has brought our field into discredit, and has consequences for the evaluated scientists and institutions as well. Of course, this makes us concerned." [8]

Most in the bibliometric community agree that the solution is to have a broad selection of metrics and to use different combinations and weightings to answer specific questions. In this White Paper we focus on measures of journal impact.

Professor Henk Moed explains: "All indicators are weighted differently, and thus produce different results. This is why I believe that we can never have just one ranking system: we must have as wide a choice of indicators as possible. No single metric can do justice to all fields and deliver one perfect ranking system." [9]

Professor Félix de Moya agrees: "Ideally, whenever a quantitative measure is involved in research-performance assessment, it should be always supported by expert opinion. In cases where the application of quantitative metrics is the only way, efforts should be made to design fair assessment parameters." [10]

Numerous alternative journal-ranking metrics to the IF have entered the field in the past few years. Many of these new metrics aim to correct for the IF's bias in favor of fields with high and rapid citation rates. Most notable among these are: Relative Citation Rates (RCR), Article Influence (AI), SCImago Journal Rank (SJR), and Source- Normalized Impact per Paper (SNIP).

Case study 2: All h-indices are not equal

The h-index is highly biased towards "older" researchers with long careers, and towards those active in fields with high citation frequencies, and, if used alone as the sole metric in evaluations, it can provide an incomplete picture of a researcher's actual citation impact.

This table presents the publication lists of three different authors, and ranks their papers (P) by how many citations they have received:

- Author 1 (A1) has published seven papers, and five of these have been cited at least five times. A1, therefore, has an h-index of five.
- Author 2 (A2) has also published seven papers with the same citation frequency as A1, but also has a large number of additional papers that have been cited four or fewer times. A2 has performed consistently over a larger body of work than A1, yet A2's h-index is also five.
- Author 3 (A3) has published seven papers, two of which are very highly cited; however, A3's h-index is also five.

This simple case shows that distinct citation distributions can generate the same h-index value, while it is questionable whether they reflect the same performance. It highlights the value of using multiple ways of looking at someone's performance in giving a comprehensive picture of their contributions. [7]

Author 1		Author 2		Author 3	
Citations	Papers	Citations	Papers	Citations	Papers
30	P1	30	P1	300	P1
10	P2	10	P2	100	P2
8	P3	8	P3	8	P3
6	P4	6	P4	6	P4
5	P5	5	P5	5	P5
1	P6	4	P6	1	P6
0	P7	4	P7	0	P7
		4	P8		
			
		4	P16		
h=5		h=5		h=5	

Publication lists of three different authors, ranking their papers (P) by received number of citations.

Tailoring the metric to the question

Source Normalized Impact per Paper (SNIP) was developed by Professor Henk Moed, previously working at the Centre for Science and Technology Studies, Leiden University, to correct for IF's bias towards fields with rapid and high citation rates. SNIP is the ratio of a source's average citation count per paper and the "citation potential" of its subject field. Citation potential is an estimate of the average number of citations a paper can be expected to receive relative to the average for its subject field. [12]

SCImago Journal Rank (SJR) is a prestige metric inspired by Google's PageRank™, whereby the subject field, quality and reputation of the journal have a direct effect on the value of its citations. Developed by Professors Félix de Moya at Consejo Superior de Investigaciones Científicas and Vicente Guerrero Bote at the University of Extremadura, SJR weights citations according to the SJR of the citing journal; a citation from a source with a relatively high SJR is worth more than a citation from a source with a relatively low SJR. [13]

Also inspired by PageRank, Article Influence (AI) is a derivative of the Eigenfactor, and is calculated by dividing the journal's Eigenfactor by the number of articles published in it. Originally developed by Carl Bergstrom at the University of Washington to

analyze journal economics, Eigenfactor uses a "random walk" model and reflects the percentage of time you would spend reading each journal if you followed random citations through a journal citation network. [14]

Relative Citation Rates (RCR) are calculated by dividing a journal's average citations per paper by the world's citation average in that journal's subject field. Unlike SNIP, the subject field is a predefined classification of journals, usually determined by the database provider. There are several variants of this idea, and in this paper we will use the example of Journal to Field Impact Score (JFIS), proposed by Van Leeuwen et al. [11] RCRs account for different citation rates between document types – for example, original research articles and reviews.

Weighing the measures

All metrics have their strengths and weaknesses, and when sweeping judgment calls are based on just one of them, accurate assessment of quality is impossible.

All new metrics correct for context: where SNIP and JFIS do this deliberately, SJR and AI also achieve this by virtue of being calculated relative to the underlying database, allowing better comparison between research areas. In addition, SJR and SNIP have increased the citation window from the IF's two years to three years, while five years is used for AI and JFIS.

Both SNIP and SJR are calculated across the entire breadth of the Scopus database while AI and JFIS use data from Journal Citation Reports. Using the Scopus database offers additional benefits: the metrics will be refreshed twice a year, and will reflect retrospective changes in the database as content is added. This also means that users have access to the entire underlying dataset for their own analyses.

Moed says: "Scopus is a most interesting database because of its wide coverage. And, because it did not yet include journal metrics, it was a challenge to come up with something new and useful. I was also very happy that Scopus did not go for just one metric, but launched two really complementary ones."

In addition, unlike JFIS, SNIP and SJR do not base delimitation of a journal's subfield on a fixed and imposed classification of journals. Rather, the sub-field depends on the origin of the citations, and so accounts dynamically for changing scope. This ensures that each journal is considered within its proper current subject field.

All new metrics also counter any potential for accusations of editorial manipulation by taking this into account in their calculations (see Table 1).

Finally, SNIP, SJR and AI are all publicly and freely available (see Featured metrics, page 16) and their respective methodologies have been published. [12, 13, 14]

A balanced solution

It is clear that the many uses of bibliometric indicators cannot be covered by one metric, and especially not by one designed to rank journals. As discussed above, using a single metric to rank journals is fraught with problems because each has a natural bias. Extending this practice to assess researchers, universities and countries can lead to even more serious distortions, and puts the validity of such assessments into doubt. Undue emphasis on increasing citations to a journal is already causing some journals to put sensation before value, but setting up a system of evaluation that encourages researchers to place more importance on where they publish than what they are publishing cannot be in the best interests of science.

The problem is that the current system of research assessment has largely evolved from journal assessment, with little effort to adjust the metrics or take specific contexts into account. This requires two broad solutions: first, more varied algorithms should be widely available, together with information about their responsible use for specific assessment purposes. Secondly, those using such metrics – from governments and research-performance managers to editors and researchers – should start using these metrics more responsibly.

SNIP and SJR were launched concurrently in Scopus as complementary metrics. Together, they provide part of the solution, and they underscore the fact that one metric can never be enough.

While metrics can provide excellent shortcuts for gauging influence in diverse aspects of academic activity, this is always at the expense of comprehensiveness. Therefore, the only way to truly understand research performance in all its aspects is to look at it from as many perspectives as possible. Scientific research is too multifaceted and important to do anything less.

Comparison table of SNIP, SJR, IF, AI and JFIS

	Publication window	Citation window	Inclusion of journal selfcitations?	Subject field normalization	Subject field delimitation	Document types used in numerator	Document types used in denominator	Role of 'status' of citing source	Effect of including more reviews	Underlying database	Effect of extent of database coverage	
SNIP Source-Normalized Impact per Paper	3 years	One year	Yes	Yes, based on citations originating from citing journals	The collection of journals citing a particular target journal, independent of database categorization	Articles, conference papers and reviews	Articles, conference papers and reviews	No role	Reviews tend to be more cited than normal articles, so increasing the number of reviews tends to increase indicators' value	Scopus	Corrects for differences in database coverage across subject fields	SNIP
SJR SCImago Journal Rank	3 years		Percentage of journal selfcitations limited to a maximum of 33%	Yes, distribution of prestige accounts for different numbers of references in different fields	Not needed by methodology	Articles, conference papers and reviews	Articles, conference papers and reviews	Weights citations on the basis of the prestige of the journal issuing them	Moderated by the SJR values of the journals the reviews are cited by	Scopus	Prestige values are redistributed, with more prestige being represented in fields where the database coverage is more complete	SJR
IF Impact Factor	2 years		Yes	No	No	All items	Articles and reviews	No role	Reviews tend to be more cited than normal articles, so increasing the number of reviews tends to increase indicators' value	Journal Citation Reports	Does not correct for differences in database coverage across subject fields	IF
AI Article Influence	5 years		No	Yes, distribution of prestige accounts for different numbers of references in different fields	Not needed by methodology	All items	Articles, letters and reviews	Weights citations on the basis of the prestige of the journal issuing them	Moderated by the Eigenfactor values of the journals the reviews are cited by		Prestige values are redistributed, with more prestige being represented in fields where the database coverage is more complete	AI
JFIS Journal to Field Impact Score	5 years		Yes	Yes, based on citations received by subject field	Predetermined classification of journals into subject categories, usually defined by database provider	Articles, letters, notes and reviews	Articles, letters, notes and reviews	No role	No effect, JFIS accounts for differences in expected citations for particular document types	Web of Science	Does not correct for differences in database coverage across subject fields	JFIS

References

1. Tassef, G. (2009) Annotated Bibliography of Technology's Impacts on Economic Growth, NIST Program Office.
2. Potocnik, J. (2010) Foreword: Assessing Europe's University-Based Research, European Commission (Directorate General for Research), Expert Group on Assessment of University-Based Research. Luxembourg: Publications Office of the European Union.
3. Garfield, E. (2005) "The agony and the ecstasy: the history and meaning of the Journal Impact Factor", International Congress on Peer Review and Biomedical Publication, Chicago, September 16, 2005. (www.garfield.library.upenn.edu/papers/jifchicago2005.pdf)
4. Adler, R., Ewing, J. (Chair), Taylor, P. (2008) Citation Statistics, Joint Committee on Quantitative Assessment of Research: A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (CIAM) and the Institute of Mathematical Statistics (IMS). (www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf)
5. Colquhoun, D. (2003) "Challenging the tyranny of impact factors", Nature, Correspondence, issue 423, pp 479. (www.vetscite.org/publish/items/001286/index.html)
6. Colquhoun, D. (July 2008) "The misuse of metrics can harm science", Research Trends, issue 6. (www.info.scopus.com/researchtrends/archive/RT6/exp_op_6.html)
7. Moed, H.F. (2009) "New developments in the user of citation analysis in research evaluation", Archivum Immunologiae et Therapiae Experimentalis (Warszawa), issue 17, pp. 13–18.
8. Geraeds, G.-J. & Kamalski, J. (2010) "Bibliometrics comes of age: an interview with Wolfgang Glänzel", Research Trends, issue 16. (www.info.scopus.com/researchtrends/archive/RT15/re_tre_15.html)
9. Pirotta, M. (2010) "A question of prestige", Research Trends, issue 16. (www.info.scopus.com/researchtrends/archive/RT15/ex_op_2_15.html)
10. Pirotta, M. (2010) "Sparking debate", Research Trends, issue 16. (www.info.scopus.com/researchtrends/archive/RT15/ex_op_1_15.html)
11. Van Leeuwen, Th.N. and Moed, H.F. (2002) "Development and application of journal impact measures in the Dutch science system", Scientometrics, issue 53, pp. 249–266.
12. Moed, H. (2010) "Measuring contextual citation impact of scientific journals", Arxiv. (arxiv.org/abs/0911.2632)
13. De Moya, F. (2009) "The SJR indicator: A new indicator of journals' scientific prestige", Arxiv. (arxiv.org/abs/0912.4141)
14. Berstrom, C. (2007–2008) Collected papers on Eigenfactor methodology. (www.eigenfactor.org/papers.html)

Featured metrics

SNIP and SJR at Journal Metrics (www.journalmetrics.com)

SCImago Journal Rank (SJR) (<http://www.scimagojr.com/>)

Source-Normalized Impact per Paper (SNIP) (www.journalindicators.com)

Impact Factor (IF) (http://thomsonreuters.com/products_services/science/free/essays/impact_factor)

h-index (http://help.scopus.com/robo/projects/schelp/h_hirschgraph.htm)

Article Influence (AI) (www.eigenfactor.org)

Relative Citation Rates (RCR)/Journal to Field Impact Score (JFIS)



SNIP
Universiteit Leiden

The evolution of journal assessment