Arch. Immunol. Ther. Exp., 2009, **57**, 1–11
PL ISSN 0004-069X
DOI 10.1007/s00005-009-0008-y

The use and misuse of journal metrics and other citation indicators

David A. Pendlebury

Thomson Reuters, Philadelphia, PA, USA

Received: 2008.12.05, **Accepted:** 2008.12.17, **Published online:** 2009.02.14

Abstract

This article reviews the nature and use of the journal impact factor and other common bibliometric measures for assessing research in the sciences and social sciences based on data compiled by Thomson Reuters. Journal impact factors are frequently misused to assess the influence of individual papers and authors, but such uses were never intended. Thomson Reuters also employs other measures of journal influence, which are contrasted with the impact factor. Finally, the author comments on the proper use of citation data in general, often as a supplement to peer review. This review may help government policymakers, university administrators, and individual researchers become better acquainted with the potential benefits and limitations of bibliometrics in the evaluation of research.

Key words: impact factor, bibliometric indicators, peer review, citation analysis.

Corresponding author: David A. Pendlebury, Thomson Reuters, Philadelphia, PA, USA,
e-mail: david.pendlebury@thomsonreuters.com

THE SCIENCE OF SCIENCE

Polish researchers have played a key role in the development of the science of science, specifically the idea of the use of quantitative measures to analyze science activity and its social structure. Recently, Paul Wouters, in The Citation Culture, described this movement in Poland among philosophers and sociologists during the 1920s and 1930s (Wouters 1999). One leader of this school, sociologist Florian Znaniecki, believed in an empirical approach to "wiedza o nauce," knowledge about science. Soon the term "naukoznawstwo," the science of science, appeared in the pages of the journal Nauka Polska. A new journal, Organon, appeared in 1936 and carried an outline of its focus and mission in a lead editorial by sociologists Maria Ossowska and Stanislaw Ossowski entitled "Nauka o Nauce", the science of science (Ossowska and Ossowski 1936). Sadly, this movement, and so much else in Poland's rich intellectual life, was extinguished by the Nazi regime, but the achievement of Polish researchers was not forgotten (Krauze 1977)

In 1955, Eugene Garfield described the idea of a citation index for the sciences (Garfield 1955). Between 1955 and the early 1960s, Garfield worked to make his idea a reality. As a practical matter, he had to decide which journals to index. He recognized that the selection of journals by total papers published or total citations received would not do: that would reward older journals or those that publish many papers or publish often. He understood that some recent and smaller journals had great influence. Working with Irv Sher, he designed the impact factor calculation to analyze and identify influential journals (Garfield and Sher 1963).

In 1963, the first Science Citation Index appeared, for the year 1961. In 1972, Garfield published "Citation analysis as a tool in journal evaluation" in Science (Garfield 1972). In 1976 he published "Significant journals of science" in Nature (Garfield 1976). These two articles had a profound influence on the way researchers viewed the structure of journal literature and the dynamics of its use. Also in 1976, Garfield published the first Journal Citation Reports (herein abbreviated as JCR) as part of the Science Citation Index (Bensman 2007a; Bensman 2007b). This presented the impact factors for 1975 as well as many other measures of journal use. The same year, Francis Narin of Computer Horizons Research, a US-based consulting firm, using publication and citation data from the Institute for Scientific Information (as Garfield's firm was then called), published Evaluative Bibliometrics, a key treatise on using these data for research evaluation, and not just for journals (Narin 1976). In 1978 a new journal,

Scientometrics, was established by Tibor Braun at the Hungarian Academy of Sciences in Budapest. In the following year Garfield published Citation Indexing, addressing the use of publication and citation data in information retrieval, evaluation, as well as in analyses of the history and sociology of research (Garfield 1979).

The 1980s and 1990s saw the establishment of research centers specializing in the use of bibliometrics, all using the ISI database, at the Science Policy Research Unit at the University of Sussex, at the Center for Science and Technology Studies (CWTS) at the University of Leiden, and of course in Budapest at the Hungarian Academy of Sciences, among others. Many specialized studies of various research phenomena and focused case studies were published in Scientometrics and other academic journals. National science funding agencies took notice and to varying degrees started to employ these data in their surveys of national performance in research. Many began to publish regular reports featuring bibliometric analyses. The U.S. government was an early adopter, as far back as the 1970s when the National Science Foundation began working with Narin; however, European governments not only caught up, but in some ways surpassed the United States in their exploration and exploitation of bibliometrics.

The field of bibliometrics, also called scientometrics, as the name of the journal indicates, grew yet more mature and the International Society for Scientometrics and Informetrics was founded in 1993. The past decade has witnessed much growth in many directions: new metrics have appeared, the internet has become the focus of potential new measures of information dissemination and use, and new mapping and visualization exercises have been offered up. A sea change occurred in 2004, when Elsevier introduced its own citation index, Scopus, and Google introduced Google Scholar. The ISI database of Thomson Reuters was no longer the sole multidisciplinary citation database, and bibliometricians hungrily devoured the possibilities of these new data sources. Another key development was the publication of Citation Analysis in Research Evaluation, by Henk F. Moed, of the CWTS in Leiden (Moed 2005b). This book, by an expert bibliometrician, provides an authoritative review of the field up to 2005.

THE JOURNAL IMPACT FACTOR

The journal impact factor, published each year in the JCR (Thomson Reuters, http://scientific.thomson.com/ /products/jcr/), has been the subject of much controversy and a good deal of misunderstanding. It is fair to say that the impact factor is now seen, over 40 years after the publication of the first JCR, as primarily a number for evaluating papers and their authors (Monastersky 2005). Such usage was never intended, and for all this time and in countless articles, essays, and speeches, Garfield has warned against employing the impact factor for the evaluation of individual articles and scientists (Garfield 1996; Garfield 1999; Garfield 2006). It is meant to be one, and only one, measure of *journal* performance.

The impact factor is defined as: Citation counts in Year 3 to a journal's contents in Years 1 and 2, divided by the number of so-called citable items in that journal in Years 1 and 2, where citable items are defined as original research reports and reviews. The denominator excludes editorials, letters to the editor, news items, tributes and obituaries, correction notices, and meeting abstracts. The exclusion of meeting abstracts, in particular, avoids a penalty in the impact factor that a journal would suffer if these generally little cited items were to be included in the denominator.

Thus the journal impact factor is a ratio of citations to citable items that approximates a mean. As an example one may consider the impact factor of the journal Neuron, as presented in the 2007 JCR for the Science Citation Index. The impact factor for this journal for 2007 is 13.410. It should be emphasized that the JCR tells us more about this title: it received 50,707 citations in 2007 to papers in the journal from any year, including 2007. It published 277 articles in 2007, a number that represents citable items, that is, regular articles and review articles. Another measure, the immediacy index (for this title 2.906) is a score of the number of citations to the journal in 2007 divided by the number of citable items published in the journal in 2007. Finally, in the JCR there appear cited half-life and citing half-life scores for this journal. The cited-half life score is the number of years one must travel back in time to account for half of the current-year citations to the journal, whereas the citing-half life concerns the time in years necessary to account for half of the references in the journal's current-year articles. The JCR contains other information as well, such as ranked lists by different measures and the citing and cited citation flows between journals. In short, the JCR offers much more data on journals than the impact factor alone.

When ranked by journal impact for 2007, in the category of neurosciences journals, Neuron is fifth among 211 titles. The Annual Review of Neuroscience ranks first, with a score of 26.077. Excellent review journals often appear in the top ranks of a field category owing to their influence and their papers being convenient sources of reference that attract many citations. The impact factor is calculated to three decimal points simply to break ties. There is little difference in real rank when numbers are close, of course.

To these figures on impact factors for neurosciences journals, one may compare those for mathematics journals. Whereas the highest impact factor in neurosciences was 26.077, the highest in mathematics, for Annals of Mathematics, was 2.739. The tenth ranked journal by impact factor in neurosciences had a 2007 score of 8.958, while the tenth ranked journal in mathematics exhibited an impact factor of 1.323. Plainly there are wide differences in impact factor scores from one field to another. These arise from

a complex of citation behaviors, prominent among which are the typical number of references in a field's journals and the "velocity" of citations to publications in a field.

OTHER JOURNAL PERFORMANCE MEASURES

The JCR is not the only product of Thomson Reuters that offers journal citation data. Another is its Journal Performance Indicators database, including publication and citation data on science, social science, and humanities titles, combined in one database and currently covering the period 1981–2007 (Thomson Reuters, http://scientific.thomsonreuters.com/products/jpi/). This product allows for the calculation of longer-term impact and is not restricted to citations to the previous two years only. For example, one can view a five-year impact for neuroscience titles. In this view, The Annual Review of Neuroscience achieved a score of 54.32 for 2003–2007. As expected, if one gives papers more time to collect citations, the average scores increase. These two calculations are not really comparable, however, since the data in the Journal Performance Indicators are calculated in a different way: citations are matched to specific papers (articles and reviews only) and the citing and cited windows of the calculation overlap (2003–2007 citations to 2003–2007 papers, not one year citing the previous two years). While differently calculated, the principle holds, however: the more time a journal is allowed to collect citations, the higher the average score in terms of citations per paper.

Yet another product is Thomson Reuters's Essential Science Indicators database (Thomson Reuters, http://scientific.thomson.com/products/esi/), which covers the sciences and social sciences during the last decade and which is updated every two months instead of annually, as is the case for the JCR and the Journal Performance Indicators. Again, only articles and reviews are tracked, and citations are linked to individual papers, not the journal title. In neurosciences, the 10-year impact for The Annual Review of Neuroscience is 146.19, a yet higher number as we might expect for a longer time frame. An important difference in this database is the selective categorization of papers from multidisciplinary journals. To do this, Thomson Reuters analyzes the journals cited by each paper in these titles and the journal titles that cite these papers and then arrives at a field categorization on the paper level. By 10-year impact for neurosciences, Science, Nature, and the Proceedings of the National Academy of Sciences of the USA rank second, third, and ninth. Of course, these are made-up versions of these titles, representing only the papers in them matched to this field. (Essential Science Indicators also permits one to explore data on scientists, institutions, nations, and papers, as well as journals.)

Simply stated, these three products, using different methods, produce different results, which are sometimes complementary and sometimes not.

JOURNAL IMPACT FACTOR: STRENGTHS AND LIMITATIONS

To return to the traditional journal impact factor published in the JCR – the focus of so much controversy – we note some of its positive features and also its limitations as described by critics.

Among its strengths:

- It provides a global view of internationally influential journals within the scope of a vetted corpus.
- It is a relatively simple calculation to understand.
- It does not reward journals because they have been publishing many years or because they publish many papers or are issued frequently.
- It gives insight into recent performance (citations in the current year to articles in the previous two years).
- In matching citations to a journal title, and not to individual papers, it can include many erroneous or variant citations, such as those showing a wrong first author or initial page number.
- Impact factors have been produced over many years in the same way, so one can view changes through time.
- The rankings by impact factor generally produce reasonable results.
- The impact factors are widely available and have been usefully employed over many years, denoting acceptance by users.

Frequent criticisms of the impact factor include:

- It is too simple a measure that does not capture enough of the multidimensional phenomena of a journal's influence.
- There is confusion and concern over the definition of citable items (the denominator in the calculation).
- The journal impact factor may be inflated in the numerator, in fact it is inflated, by "free citations," which are citations to article types, such as editorials or letters, not accounted for in the denominator.
- The impact factor is a ratio that is the mean of a skewed distribution. This is the familiar 80/20 rule, meaning that a small portion of the population of papers accounts for a very large portion of the influence, in this case citations; thus it is misleading concerning central tendency.
- Review journals often have high impact factors and thus have an advantage over non-review journals.
- The widely differing absolute impact factors from one field to another make cross-field comparisons difficult or meaningless.
- The two-year citation window is too short and penalizes some fields and their journals, such as mathematics, where citations typically peak beyond this citation window.
- By matching citation counts to journal titles, Thomson Reuters omits or mismatches citations, since the cited reference data carry only a 20-character field that makes it sometimes difficult or impossible to recognize a journal's identity accurately.

- Multidisciplinary journals, which offer a mixed set of papers in terms of fields, produce a "mixed" impact factor, which is of little use.
- The definition of fields in the JCR is subjective and fuzzy, and even if it is more or less reasonable it does not take into account subfield variations.
- The measure is not useful in some fields such as the humanities, where citation practices are much different than in the sciences and where books are a main vehicle of communication (Thomson Reuters does not produce a JCR in the arts and humanities).
- If the journal is not indexed by Thomson Reuters, there is no impact factor available.
- Thomson Reuters's journal coverage is biased against certain nations and languages, or is biased in favor of certain nations and English-language journals. This affects the impact factor scores, especially of nationally influential (rather than internationally influential) non-English-language journals. Thus the Thomson Reuters's journal selection rewards some journals and hurts others.

The pluses and minuses of the impact factor, as viewed by various bibliometricians, are widely discussed in the literature (Leydesdorff 2008; Moed 2002; Moed 2005a; Moed and van Leeuwen 1995; Moed and van Leeuwen 1996; Moed et al. 1996; Moed et al. 1999; Moed et al. 2004; van Leeuwen et al. 1999; van Leeuwen and Moed 2002; van Leeuwen and Moed 2005; Zitt and Small 2008).

As mentioned, journal impact factors were not designed for or intended to be used as a measure or proxy for the performance of individual papers or researchers (Seglen 1997). The skewed nature of the citation distribution means that in most cases the impact factor overestimates the influence of particular papers or people. The literature is full of angry comments against Thomson Reuters and the impact factor. These comments are generally misdirected. Many attribute to Garfield and Thomson Reuters motives not intended and uses not endorsed (Garfield 1996; Garfield 1999; Garfield 2006). Particularly egregious are certain government evaluations that require publication in Thomson Reuters-indexed journals with an impact factor of, say, 1 or above to receive promotion or funding, and some of these reviews even sum the absolute impact factor scores, all this without regard to field differences in impact factors. As computer scientists say, "Garbage in, garbage out." Instead, a reviewer of specific publications or authors should, as a first step, look up the actual citations to these papers and people. Discussions and criticisms of the use of the impact factor in evaluating the journal literature are, however, fair game.

ALTERNATIVES TO THE JOURNAL IMPACT FACTOR

Alternatives to the impact factor often seek to weight citations or to relativize or normalize journal impact factor scores or variants of them, in the latter case especially to allow for cross-field comparisons (Banks and Dellavalle 2008; Egghe and Rousseau 2002). Shortly after the appearance of the JCR, researchers began to offer alternative calculations. Many were responding to what they saw as limitations or weaknesses in the formula for the impact factor's calculation. Concerns, as noted, included the use of a mean for a skewed distribution, the range of years used, the mismatch between the citations in the numerator and all publication types in the denominator, the different average rates of citation for different document types, the different citation rates by field as well as their different maturing and decaying rates for attracting citations, and the percentage of cited vs. uncited papers, among many others. The suggested alternatives attempted to address one or more of these concerns. Wolfgang Glanzel and Henk Moed provided a thorough review of this literature in Scientometrics in 2002 (Glanzel and Moed 2002).

The past few years have seen a number of new alternatives to the impact factor. Before reviewing these, it is important to note the work of Gabriel Pinski and Francis Narin from 1976. They introduced the notion that not all citations are created equal. Their Influence Weight is a "size independent measure of the weighted number of citations a journal receives from other journals, normalized by the number of references it gives to other journals" (Pinski and Narin 1976). In essence, it gives a higher value to citations received from influential journals. The reason for mentioning this older work in the context of recent research is that this method is having a renaissance. The method greatly influenced the founders of Google, who drew upon this methodology in constructing their PageRank algorithm (Brin and Page 1998). That, in turn, has ignited new interest in this methodology.

The review of research since 2004 given here is meant to provide a flavor of the newest journal performance measures. In 2004, Garfield himself with his colleague Alexander Pudovkin proposed a rank normalized impact factor (rnIF) (Pudovkin and Garfield 2004). Their formula takes the number of journals in a category and subtracts the rank of a journal in that category, adds 1, and divides this by the number of journals in the category. Here, cross-field comparisons of impact factors are addressed. Peter Vinkler devised an indicator called Specific Impact Contribution (SIC) to explore the contribution of a subset of articles or a journal, but the result was that normalized impact factors and normalized SIC indicators essentially were identical (Vinkler 2004). The same year, three Thai researchers introduced the Cited Half-Life Impact Factor (CHAL) (Sombatsompop et al. 2004). It is the ratio of the number of current-year citations to articles in the previous X years to that of articles published in the previous X years, where X is the cited-half life of the journal in the current year. They obtained differences in rankings and concluded that their method and results were "more suitable." This approach addresses differences in

journals' rates of citation decay. As noted, the appearance of Google Scholar and Elsevier's Scopus in late 2004 changed the landscape for bibliometrics and journal performance measures (Falagas et al. 2008; Harzing and van der Wal 2008; Jasco 2005; Meho and Yang 2007).

The year 2005 witnessed a number of contributions and new measures. Van Leeuwen and Moed published a study of the role of uncited papers and the citation distribution frequency and how both affect the journal impact factor (van Leeuwen and Moed 2005). While not a new method, it addresses traditional concerns and has been influential in formulating new measures. The Thai researchers Sombatsompop and Markpin introduced the Impact Factor Point Average (IFPA), another effort to normalize impact factors for different fields (Sombatsompop and Markpin 2005; Sombatsompop et al. 2005). Ronald Rousseau, building on the earlier (2004) work of the Thai researchers, renamed their CHAL impact factor the Median Impact Factor (MIF) and extended their idea to explore percentile impact factors (Rousseau 2002; Rousseau 2005). In this case, attention was paid to the form of the citation curve for a journal.

In late 2005, physicist Jorge E. Hirsch introduced the h-index (Hirsch 2005; Hirsch 2007). He stated that a scientist has an index of h if h of his or her papers have at least h citations. In other words, a researcher with 20 papers that each has 20 or more citations has an h-index of 20. He or she may have published many more papers. Hirsch argued that "h is preferable to other single-number criteria commonly used to evaluate scientific output of a researcher." The method can be applied to any set of papers; thus, it can be applied to a journal. It is a number that combines publication activity and citation influence. It puts emphasis on the top of the citation distribution while ignoring the bottom. Of course, the h--index is also subject to different rates of citation in different fields. Braun, Glanzel, and Schubert have explored its application to journals (Braun et al. 2005). Jayant S. Vaidya responded to the h-index by suggesting the v-index as fairer, since it compensates for the academic age of a researcher (Vaidya 2005).

Leo Egghe's work of 2005 on fractional relative impact factors attempted a more dynamic view of a journal's citation influence (Egghe 2005), but perhaps more influential was his introduction of the g-index in 2006. The g-index is a modification of the h-index that takes into account the presence of highly cited papers beyond the h value. It is defined as "the (unique) largest number such that the top g articles received (together) at least g^2 citations" (Egghe 2006).

In 2006, Johan Bollen and colleagues introduced the y-index, specifically for journals. Unlike the h- and g-indexes, the y-index combines the journal impact factor and a weighted PageRank algorithm. The authors believe that the y-index, a prestige measure, is superior to the impact factor, called a "metric of popularity" (Bollen et al 2006).

In 2007, biologist Carl T. Bergstrom and colleagues introduced the Eigenfactor score as the measure of a journal's importance (Bergstrom 2007). He uses Thomson Reuters's JCR data for 2001–2006. The Eigenfactor algorithm, he writes, "estimates the relative influence of reference items based on cross-citation tables. Like Thomson Scientific's Impact Factor metric, Eigenfactor measures the number of times that articles published during a census period provide citations to papers published during an earlier target window. While Impact Factor has a one year census period and uses the two previous years for the target window, Eigenfactor has a one year census period and uses the five previous years for the target window." Journal self-citations are excluded. Unlike the impact factor, the Eigenfactor score is a measure of the total influence of a journal. To obtain an impact factor-like score, Bergstrom and colleagues divide their measure of total journal influence by the number of articles published and call this the Article Influence score. The Eigenfactor method is similar to the weighted approach of Pinski and Narin, and that of Google.

In February 2008, three scientists from Northwestern University, in Evanston, Illinois, Michael J. Stringer, Marta Sales-Pardo, and Luis A. Nunes Amaral, published their journal ranking methodology in the online publication PLoS One (Stringer et al. 2008). Using Thomson Reuters's citation data, the authors analyzed 19.4 million articles published in 2267 journals. To rank the journals they employed the Probability Ranking Principle, also known as the multi-class "area under the curve" (AUC) statistic. Their article includes a table of journal rankings in Ecology, in which the AUC score and the journal impact factor score for each title are given. Indeed, the two scores give somewhat different ranks.

In May 2008, the Thai research group proposed another new index, termed the "Article-Count Impact Factor" (ACIF), consisting of a ratio of the number of items cited in the current year to the source items published in the journal during the previous two years (Markpin et al. 2008). This is another way to look at cited/uncitedness in the short term. As mentioned, this is a phenomenon examined by van Leeuwen and Moed in their April 2005 study (van Leeuwen and Moed 2005).

In June 2008, Anne-Wil K. Harzing and Ron van der Wal described the use of Google Scholar as an alternative source of data to Thomson Reuters's Web of Science database and its JCR. They use a software program, which they call Publish or Perish, to obtain citations from Google Scholar. It also calculates an h-index, a g-index, and citations per paper (CPP). This article provides a table of scores (impact factor, h-index, g--index, and citations per paper) and ranks for 20 management journals. The authors state that "for the field of management, the various Google Scholar citation metrics provide a more comprehensive picture of journal impact than the ISI JIF" (Harzing and van der Wal 2008).

Matthew E. Falagas and colleagues recently made use of the SCImago journal rank indicator (SJR), developed in Granada, Spain. In a paper in the FASEB Journal, they compare the SJR indicator with the impact factor (Falagas et al. 2008). The SJR indicator uses data from Scopus. Unlike the journal impact factor, the SJR measure uses weighted citations (applying the PageRank algorithm), excludes journal self-citations, and includes all documents types in the denominator of its calculation, not just citable items. The methodology produces, as one might expect, different rankings.

In August 2008, in the pages of the Journal of the American Society for Information Science and Technology, Henry Small of Thomson Reuters and Michel Zitt described a new approach to field normalization of the impact factor which they called the "audience factor." This measure examines the propensity of one journal to cite another. It uses the mean number of references of each citing journal and fractionally weights the citations from the citing journals. As Zitt and Small report, "a comparison with the standard journal impact factors from Thomson Reuters shows a more diverse representation of fields within various quintiles of impact, significant movement in rankings for a number of individual journals, but nevertheless a high overall correlation with standard impact factors" (Zitt and Small 2008).

WHICH MEASURE?

H-index, v-index, g-index, y-index, Eigenfactor, audience-factor: What is the non-bibliometrican to think of this mélange of measures? It is important to recognize that different measures attempt to answer different questions and that each will emphasize or highlight certain aspects and nuances of a phenomenon. This is not to deny that some measures may be better, in general terms, than others. There is certainly room for advancement in terms of new and better measures. But it is also necessary to point out that there is a fallacy in demanding a single-number metric or just one approach to analysis. Regarding new metrics, bibliometricians themselves have a challenge: to do much more research to gauge these different measures against some standards for accuracy and acceptance, which means against the opinion of field experts and policymakers, to test superiority (Butler 2008; Harnad 2008; Zitt and Bassecoulard 2008).

BEYOND JOURNAL MEASURES

Bibliometric analyses generally, beyond assessments of journal performance, deserve consideration (Bar-Ilan 2008; Borgman and Furner 2002; Bornmann et al. 2008; Butler 2008; Joint Committee on Quantitative Assessment of Research 2008; Nicolaisen 2007). A central question is "What do citations measure?" This question falls under the rubric "the theory of citation." It is a vast area of research with a large bibliography, usefully summarized by Moed (Moed 2005b). Generally, citations represent the notions of use, reception, utility, influence, significance, and the somewhat nebulous word "impact." Citations do not, however, represent measures of quality. Quality assessments require human judgment.

Quality is the standard policymakers and funders would like to use in making their decisions. Those decisions are difficult. There is an obvious need, and it is keenly felt today, to be selective, to highlight significant or promising areas of research, and to manage better investments in science. Resources have not grown as fast as science, which demands hard decisions about what should be supported and what should not, or which research projects and researchers should receive more support than others.

Instead of despairing that resources are limited, policymakers and research funders have taken the positive step of trying to put such decisions on a more rational footing (Bornmann et al. 2008; Butler 2008). And so they have turned the main tool of science, quantitative analysis, on science itself. It is perhaps unnecessary to note how much of modern life revolves around characterizations of human activity in terms of statistics. But, as stated, citations are only indicators of performance, and of a particular kind. They are not direct measures of quality (Moed 2005b).

Until recently, peer review has been the main route by which science policymakers and research funders have coped with decisions on what course to set for science. Peer review still represents the standard approach to research evaluation and decisions about allocating resources. Experts reviewing the work of their colleagues should rightly be the basis of research evaluation. However, the daunting, even overwhelming, size and specialized nature of research today, mentioned already, makes it difficult for a small group of experts to evaluate fully and fairly a bewildering array of research, both that accomplished and that proposed. Moreover, bias in peer review, whether intentional or inadvertent, is widely recognized as a confounding factor in efforts to judge the quality of research.

Thomson Reuters has never advocated that quantitative analysis supersede or replace peer judgments. Rather, publication and citation analysis is meant to be, and only in certain cases, a complement or supplement to peer review. It is the two together, peer review and quantitative analysis, which better informs peer review. that holds the best promise for research evaluation.

PEER REVIEW AND QUANTITATIVE ANALYSIS

One may compare and contrast both approaches and at the same time highlight certain advantages of quantitative analysis with respect to peer review. Quantitative analysis is global in perspective, a "top-down" review, whereas peer review is essentially "bottom-up." It collects data on all activity in an area, summaries these

data, and obtains a comprehensive perspective on activity and achievements. Weighted quantitative measures, such as papers per researcher or citations per paper, can remove the advantage of size, which strongly colors human perceptions of quality. When one thinks of the best, it is hard not to think automatically of the biggest producers, whether individuals, labs, universities, or the like. Quantitative analysis can focus on recent contributions and ignore those of the distant past. Again, it is difficult for senior scientists, those most often involved in peer review, not be influenced by what they recall from their earlier days to be the top performers, but their perceptions may be, let us admit, based on work and reputations of a decade or more ago.

THE "TEN COMMANDMENTS" OF CITATION ANALYSIS

For those who need to analyze research performance using publication and citation data, the following 10 guidelines, or "commandments", may prove helpful.

No. 1: Consider whether available data can address the question

Before even beginning an analysis, ask if the data available, whether from the Thomson Reuters database or other databases, are sufficient to analyze the research under review. A general observation: the analysis will usually be more reliable at face value the more basic the research and the larger the dataset. There are exceptions. One should also explore field definitions, which are inherently fuzzy. National publication patterns as well as language use should also be considered.

No. 2: Choose publication types, field definitions, and years of data

These are technical matters. In terms of publication types, the standard practice is to use journal items that have been coded by Thomson Reuters as regular discovery accounts, brief communications (notes), and review articles; in other words, those types of papers that contain substantive scientific information. Traditionally left to the side are meeting abstracts (generally not much cited), letters to the editor (often expressions of opinion), correction notices, and other marginalia. Already mentioned is the problem of field definitions and field categories. Finally, one must decide which years of publications and citations to use. These do not have to be the same. Generally, when citations are to be used to gauge research impact, at least five years of publications and citations are recommended, since citations take some time to accrue to papers. In the fastest moving fields, such as molecular biology and genetics, this might take 18 months to two years, whereas in others, such as physiology or mathematics, the peak in citations might be, on average, three or more years.

No. 3: Decide on whole or fractional counting

This is another technical matter, but an important one. The question is: Should each author or institution listed on a paper receive whole or a fractional, or proportionate, share of the paper and for that matter the citations that it attracts? (Thomson Reuters records all authors and addresses listed on a paper, so these papers can be attributed to all producers.) For example, let us take a paper by three scientists at three different universities which has been cited 30 times. Should each receive credit for one-third of the paper and, say, 10 citations (one-third of the citations)? Or should each receive a whole publication count and credit for all 30 citations? Another possibility would be to use fractional publication counts but whole citation counts, so that in our example, each researcher or university would receive credit for one-third of paper but all 30 citations. Thomson Reuters almost always uses whole publication and citation counts. The reasoning? If one examines the original paper, the researchers themselves fail to distinguish who is responsible for how much of the work reported. Even when there is a lead author indicated (in some fields traditionally the first listed, but in other fields the last), there is never a quantitative accounting of credit. The presentation of a research paper suggests that all are equal in their authorship and contributions, although this must be a fiction. There is also honorary authorship. Many believe that anyone appearing as an author should be able to explain and defend fully the contents of the paper, but career concerns and the simple approach of using length of one's list of publications as a measure of achievement has brought things to the point where there is much unwarranted authorship. Two fields in particular may demand fractional counting: high-energy physics and large-scale clinical trials. These fields often produce papers presenting hundreds of authors and almost as many institutions. Faced with such papers, one wonders if these reports are really attributable to any scientists or any institutions.

No. 4: Judge whether the data require editing to remove "artifacts"

"Artifacts" are aspects of the data that may confound the analysis or mislead the analyst. Just mentioned are papers with hundreds of authors and institutional addresses. In a small to medium-sized dataset, the inclusion of such papers and whole counting could and sometimes do have this effect. To touch on more mundane matters, there is the frequent necessity of unifying names of authors in a single form if they have presented their names on papers in several ways. So, too, for institutions, one must generally unify variant designations so that the statistics for an institution appear under one preferred name. The Thomson Reuters database records the institutional name as given on the original paper, but authors are not uniform in how they list their own institution. This is dreary but important work.

There are three objections frequently heard to the use of citation counts in evaluating research, and all come under the rubric of possible artifacts in the data. They are: negative citations, the "over-citation" of review articles and methods papers, and self-citations. Analysts at Thomson Reuters generally do not concern themselves with any of these for reasons stated below. All, however, could be conflating factors in very small datasets.

Negative citations are few in number. They are rare events, statistically speaking. Scientists typically cite for neutral or positive reasons, to note earlier work or to agree with and build upon it. Assessing whether a citation is positive or negative requires a careful, informed reading of the original paper, and this obviously cannot be attempted with more than a few hundred papers at the most. Of several articles published in which this sort of analysis was attempted, always dealing with a sample of papers in a single field that could be controlled by the analyst, outright negative citations were few, on the order of 10% or less. Naturally, one can recall notorious examples, such as the case of the Cold Fusion papers, but these and other rare cases are the so-called exceptions that prove the rule. Frequency of citation, many studies have shown, correlates positively with peer esteem. Negative citations, to the degree they appear, are little more than background noise and do not materially affect the analyses.

To the complaint that methods papers and reviews are over-cited, the first because a method may be used by many and acknowledged perfunctorily and the second because a review offers a convenient and concise way of recognizing and summarizing the previous literature in an area, a reply might be, "Yes, but only useful methods and only good reviews are highly cited" (Ketcham and Crawford 2007). In the end, if these types of publications are a concern, they can be removed from the analysis.

Finally, a comment on self-citation is in order. Self-citation is a normal and normative feature of publication, with 25% self-citation not uncommon or inordinate in the biomedical literature. It is only natural that when a researcher works on a specific problem for some time, that researcher would cite his or her earlier publications. If someone set out on a strategy to boost their citation counts through self-citation, there would be several obstacles to overcome. The first is peer review, objections from reviewers and the journal editor that there were unnecessary citations and perhaps the absence of citations to appropriate work. The author would then perhaps aim to publish in lower-impact journals with looser standards of review, but in this case fewer people would see and cite these articles, so one would lose citations from others to some degree. It seems a self-defeating strategy.

These so-called myths of bibliometrics, and others, have been recently summarized and addressed by Glanzel (Glanzel 2008).

No. 5: Compare like with like

The methodology for a bibliometric analysis must always compare like with like, or "apples with apples, not apples with oranges." Different fields of research, as noted, exhibit quite different citation rates or averages, and the difference can be as much as 10:1. Even within the same field, one should not compare absolute citation counts of an eight-year-old paper with those of a two-year-old paper, since the former has had more years to collect citations than the latter. Likewise, there is little sense in comparing the thick publication dossier of a researcher who has been publishing for 30 years and runs a large laboratory with the handful of recently published papers from a newly minted Ph.D. in the same field. This is all really no more than common sense. Still, comparing like with like is the "golden rule" of citation analysis.

No. 6: Use relative measures, not just absolute counts

This applies to citation counts rather than to publication counts, since there is very little data collected on average output for a researcher by field and over time; the problem is attributing papers to unique individuals in order to calculate typical output. The Thomson Reuters database carries no marker for unique individuals, only unique name forms.

Absolute citation counts do have their place. Garfield has said that he thinks there is no better indicator of status and peer regard in science than total citations, and his research has demonstrated the frequent correlation between scientists with the most citations in their field and those who are chosen for the Nobel Prize. But this is a very select and statistically atypical group of researchers, with thousands or tens of thousands of citations. For mere mortals, however, the situation is different. Most claim hundreds, not thousands, of citations. As one deals with smaller numbers, it is important not to put too much weight on minor differences in total citations. It again should be recognized that the citation totals of a researcher likely reflect the number of papers produced, the field of research, and how many years the papers have had to collect citations.

To begin to make distinctions among individuals with a normal, or more typical, number of citations, the following can be recommended, among other measures:

Absolute counts:
- papers in Thomson Reuters indexed journals
- papers per year on average
- papers in top journals (various definitions)
- number of total citations

and relative measures:
- citations per paper compared with citations per paper in the field over same period
- citations vs. expected (baseline) citations
- percent papers cited vs. uncited compared with the field average
- rank within the field or among the peer group by papers, citations, or citations per paper.

Field averages are usually generated using journal sets that serve to define the field. This is not always optimal, and everyone has a different idea about such journal-to-field schemes. Although imperfect, such field definitions offer the advantage of uniformity and of measuring all within the same arena, although it is an arena with irregular outlines.

Expected or baseline citations are geared to a specific journal, a specific year, and a specific article type (such as review, note, meeting abstract, letter, etc.) The expected citation score is an attempt to gauge relative impact as precisely as possible based on these three attributes of a paper:

- the year the paper was published, since, as mentioned, older papers have had more time to collect citations than younger ones
- the journal in which the paper appeared, since different fields exhibit different average citation rates, and even in the same field there are high- and low-impact titles, and
- the type of article, since articles and reviews are typically more cited than meeting abstracts, corrections, and letters.

The expected citation score is an attempt to come as close as possible to the peers for the paper under review, i.e. to compare like with like as closely as possible. It is an effective measure for assessing a paper, multiple papers by a researcher, those of a team, and even those of an entire institution.

To examine the publication record of an individual, one may sum all the actual citation counts to their papers and then sum all the expected citation scores for each paper. Then one can make a ratio of the two to gauge better than average, average, or lower than average performance, and by how much.

In following this methodology it is as if one is creating an exact double of the researcher under review. This double would be the average researcher. Every time the real researcher published a paper, the double would publish one in the same journal, in the same year, and of the same article type. The papers of the double would always achieve the exact average in terms of citations for such papers. The real researcher would not, of course, but the comparison of the two is often enlightening.

No. 7: Obtain multiple measures

From multiple measures a kind of mosaic of research influence may be seen. The use of multiple measures is a kind of insurance policy against drawing false conclusions from one or two measures alone.

No. 8: Recognize the skewed nature of citation data

Whether one is reviewing the papers of an individual researcher, those of a research team, papers in a single journal or group of journals, those of a specific field in a given year, or those of a university or an entire nation, the citation distribution of the dataset will be highly skewed. That is, as noted earlier, a small number of papers in the population will be highly cited and the large majority will be cited relatively little or not at all. This should not cause surprise; it is the nature of these data at every level of analysis.

No. 9: Confirm that the data collected are relevant to the question

and

No. 10: Ask whether the results are reasonable

These two can be treated together. They represent no more than double-checking that the data collected are relevant to the question one originally set out to answer and that the data should be viewed as scientists approach any data, or should, with skepticism. Good scientists do this and even try to refute the conclusions they obtain from their data, and they do not draw conclusions that go beyond the limits of the data collected. Bibliometricians and those who use these data for science policy and research funding decisions should do no less.

CONCLUSION

One can be an advocate of quantitative analysis for research evaluation and simultaneously a critic of naïve methodologies and the misleading uses to which the data are sometimes put. The consequences of such misuse can be profound – for individuals, research groups, institutions, journal publishers, and even nations and their national research programs. Any quantitative analysis should be straightforward in its methodology, simple to explain, and the results should be presented openly so that others can understand and check them. Such transparency will help demystify this type of research evaluation. When, however, the purpose for pursuing quantitative analysis of research is for "window dressing" or to prove to policymakers, administrators, or funding agencies something decided upon even before the data are collected and analyzed, this works against acceptance of this approach and the true goal of the analysis. That goal is to discover something, to obtain a better, more complete understanding of what is actually taking place in research. This deeper understanding can better inform those charged with making their difficult choices about allocating resources, generally in the context of peer review.

Finally, it is important to recognize that numbers can be dangerous because they have the appearance of being authoritative. In the face of statistics, many discussions stop. That is unfortunate, since they should fuel discussions and illuminate features in the research landscape that might otherwise be overlooked. It is only nat-

ural in light of the ever growing complexity of science and the challenge of rationing resources, that government policymakers, managers, and others would turn to quantitative analysis to help make their task easier, but the truth is that the collection and use of quantitative indicators adds an extra burden, in fact requires more work and thought on the part of policymakers, managers, and analysts. This extra effort is often worthwhile both for the greater understanding and practical help the data offer as well as for its beneficial effect of adding fairness to evaluation by helping to prevent abuses that may arise from small-scale closed peer review.

REFERENCES

Banks MA, Dellavalle R (2008) Emerging alternatives to the impact factor. OCLC Systems & Services 24(3). Available via: http://eprints.rclis.org/archive/00014614/01/OCLC%2BPaper%2BFinal-v2Feb2008.pdf

Bar-Ilan J (2008) Informetrics at the beginning of the 21st century: a review. J Informetrics 2:1–52

Bensman SJ (2007a) Garfield and the impact factor. Annu Rev Inf Sci Technol 41:93–155

Bensman SJ (2007b) The impact factor, total citations, and better citation mouse traps: a commentary. J Am Soc Inf Sci Technol 58:1904–1908

Bergstrom CT (2007) Eigenfactor: measuring the value and prestige of scholarly journals. C&RL News 68:5. Available via: http://www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may07/eigenfactor.cfm, also see: http://www.eigenfactor.org//whyeigenfactor.htm

Bollen J, Rodriguez MA, Van de Sompel H (2006) Journal status. Scientometrics 69:669–687

Borgman C, Furner J (2002) Scholarly communication and bibliometrics. Annu Rev Inf Sci Technol 36:3–72

Bornmann L, Mutz R, Neuhaus C et al (2008) Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics Sci. Environmental Politics 8:93–102

Braun T, Glanzel W, Schubert A (2005) A Hirsch-type index for journals. Scientist 19:8

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Networks ISDN Systems 30:107–117

Butler L (2008) Using a balanced approach to bibliometrics: quantitative performance measures in the Australian Research Quality Framework. Ethics Sci Environmental Politics 8:83–92.

Egghe L, Rousseau R (2002) A general framework for relative impact factors. Can J Inf Library Sci 21:29–48

Egghe L (2005) Continuous, weighted Lorenz theory and applications to the study of fractional relative impact factors. Inf Processing Management 41:1330–1359

Egghe L (2006) Theory and practice of the g-index. Scientometrics 69:131–152

Falagas ME, Kouranos VD, Arencibia-Jorge R et al (2008) Comparison of SCImago journal rank indicator with journal impact factor. FASEB J 22:2623–2628

Garfield E (1955) Citation indexes for science. A new dimension in documentation through association of ideas. Science 122:108–111

Garfield E (1972) Citation analysis as a tool in journal evaluation. Science 178:471–479

Garfield E (1976) Significant journals of science. Nature 264:609–615

Garfield E (1979) Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. John Wiley & Sons, New York

Garfield E (1996) How can impact factors be improved? BMJ 313:411–413

Garfield E (1999) Journal impact factor: a brief review. CMAJ 161:979–980

Garfield E (2006) The history and meaning of the journal impact factor. JAMA 295:90–93

Garfield E, Sher IH (1963) New factors in the evaluation of scientific literature through citation indexing. Am Documentation 14:195–201

Glanzel W (2008) Seven myths in bibliometrics: about facts and fiction in quantitative science studies. Proceedings of the WIS 2008, Berlin. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting. Available via: http://www.collnet.de//Berlin-2008/GlanzelWIS2008smb.pdf

Glanzel W, Moed HF (2002) Journal impact measures in bibliometric research. Scientometrics 53:171–193

Harnad S (2008) Validating research performance metrics against peer rankings. Ethics Sci. Environmental Politics 8:103–107

Harzing A-WK, van der Wal R (2008) Google Scholar as a new source for citation analysis. Ethics Sci. Environmental Politics 8:61–73. Available via: http://www.harzing.com/resources.htm#/pop_hindex.htm

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 102:16569–16572

Hirsch JE (2007) Does the h-index have predictive power? Proc Natl Acad Sci U S A 104:19193–19198

Jasco P (2005) As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation--based and citation-enhanced databases. Curr Sci 89:1537–1547

Joint Committee on Quantitative Assessment of Research (2008) Citation Statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS), by Robert Adler, John Ewing (Chair), and Peter Taylor. Available via: http://www.mathunion.org/fileadmin/IMU//Report/CitationStatistics.pdf

Ketcham CM, Crawford JM (2007) The impact of review articles. Lab Invest 87:1174–1185

Krauze T (1977) The sociology of science in Poland. In:

Merton RK, Gaston J (eds) The Sociology of Science in Europe. Southern Illinois University Press, Carbondale, Illinois, pp. 193–223

Leydesdorff L (2008) Caveats in the use of citation indicators in research and journal evaluations. J Am Soc Inf Sci Technol 59:278–287

Markpin T, Boonradsamee B, Ruksinsut K et al (2008) Article-count impact factor of materials science journals in SCI database. Scientometrics 75:251–261

Meho LI, Yang K (2007) A new era in citation and bibliometric analyses: Web of Science. Scopus, and Google Scholar. J Am Soc Inf Sci Technol 58:2105–2125

Moed HF (2002) The impact-factors debate: the ISI's uses and limits. Nature 415:731–732

Moed HF (2005a) Citation analysis of scientific journals and journal impact measures. Curr Sci 89:1990–1996

Moed HF (2005b) Citation analysis in research evaluation. Springer, Dordrecht

Moed HF, Glanzel W, Schmoch U (eds) (2004) Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Kluwer Academic Publishers, Dordrecht

Moed HF, van Leeuwen TN (1995) Improving the accuracy of Institute for Scientific Information's journal impact factors. J Am Soc Inf Sci 46:461–467

Moed HF, van Leeuwen TN (1996) Impact factors can mislead. Nature 381:186

Moed HF, van Leeuwen TN, Reedijk J (1996) A critical analysis of the journal impact factors of Angewandte Chemie and Journal of the American Chemical Society: inaccuracies in published impact factors based on overall citation counts only. Scientometrics 37:105–116

Moed HF, van Leeuwen TN, Reedijk J (1999) Towards appropriate indicators of journal impact. Scientometrics 46:575–589

Monastersky R (2005) The number that's devouring science. The Chronicle of Higher Education 52:A12.

Narin F (1976) Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity. Computer Horizons Inc., Cherry Hill, New Jersey

Nicolaisen J (2007) Citation analysis. Annu Rev Inf Sci Technol 41:609–641

Ossowska M, Ossowski S (1936) The science of science. Organon 1:1–12

Pinski G, Narin F (1976) Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. Inf Processing Management 12:297–312

Pudovkin AI, Garfield E (2004) Rank-normalized impact factor: a way to compare journal performance across subject categories. ASIST 2004: Proceedings of the 67th ASIS&T Annual Meeting 41:507–515

Rousseau R (2002) Journal evaluation: technical and practical issues. Library Trends 50:418–439

Rousseau R (2005) Median and percentile impact factors. Scientometrics 63:431–441

Seglen PO (1997) Why the impact factor of journals should not be used for evaluating research. BMJ 314:498–502

Sombatsompop N, Markpin T (2005) Making an equality of ISI impact factors for different subject fields. J Am Soc Inf Sci Technol 56:676–683

Sombatsompop N, Markpin T, Premkamolnetr N (2004) A modified method for calculating the impact factors of the journals in ISI Journal Citation Reports: polymer science category in 1997–2001. Scientometrics 60:217–235

Sombatsompop N, Markpin T, Yochai W et al (2005) An evaluation of research performance for different subject categories using impact factor point average (IFPA) index: Thailand case study. Scientometrics 65:293–305

Stringer MJ, Sales-Pardo M, Nunes Amaral LA (2008) Effectiveness of journal ranking schemes as a tool for locating information. PLoS One 3:e1683. Available via: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0001683

Thomson Reuters, http://scientific.thomson.com/products/jcr/

Thomson Reuters, http://scientific.thomsonreuters.com/products/jpi/

Thomson Reuters, http://scientific.thomson.com/products/esi/

Vaidya JS (2005) V-index: a fairer index to quantify an individual's research output capacity. BMJ Rapid Response. Available via: http://bmj.bmjjournals.com/cgi/eletters/331//7528/1339-c#123188

van Leeuwen TN, Moed HF (2002) Development and application of journal impact measures in the Dutch science system. Scientometrics 53:249–266

van Leeuwen TN, Moed HF (2005) Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics 63:357–371

van Leeuwen TN, Moed HF, Reedijk J (1999) Critical comments on Institute for Scientific Information impact factors: a sample of inorganic molecular chemistry journals. J Inf Sci 25:489–498

Vinkler P (2004) Characterization of the impact of sets of scientific papers: the Garfield (impact) factor. J Am Soc Inf Sci Technol 55:431–435

Wouters P (1999) The Citation Culture. University of Amsterdam, Amsterdam, pp. 82–83

Zitt M, Bassecoulard E (2008) Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. Ethics Sci. Environmental Politics 8:49–60

Zitt M, Small H (2008) Modifying the journal impact factor by fractional citation weighting: the audience factor. J Am Soc Inf Sci Technol 59:1856–1860