Research Evaluation **21** (2012) pp. 183–188
Advance Access published on 24 July 2012
doi:10.1093/reseval/rvs012

An Integrated Impact Indicator: A new definition of 'Impact' with policy relevance

Caroline S. Wagner[1,*] and Loet Leydesdorff[2]

[1]*John Glenn School of Public Affairs, The Ohio State University, Columbus, OH 43210, USA and*
[2]*Amsterdam School of Communication Research (ASCoR), Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands*
**Corresponding author. Email: wagner.911@osu.edu*

Allocation of research funding, as well as promotion and tenure decisions, are increasingly made using indicators and impact factors drawn from citations to published work. A debate among scientometricians about proper normalization of citation counts has resolved with the creation of an Integrated Impact Indicator (*I3*) that solves a number of problems found among previously used indicators. The *I3* applies non-parametric statistics using percentiles, allowing highly cited papers to be weighted more than less-cited ones. It further allows unbundling of venues (i.e. journals or databases) at the article level. Measures at the article level can be re-aggregated in terms of units of evaluation. At the venue level, the *I3* creates a properly weighted alternative to the journal impact factor. *I3* has the added advantage of enabling and quantifying classifications such as the six percentile rank classes used by the National Science Board's *Science & Engineering Indicators*.

Keywords: impact; citation; normalization; percentile; statistics; indicator.

1. Introduction

A refereed exchange among scientometricians about appropriate normalization (Gingras and Larivière 2011) has resulted in the creation of a refined indicator that solves a number of problems that arise when assessing the citation impact of scientific articles and venues. Citation and publication distributions are well known to be heavily skewed (Seglen 1992, 1997). Following the prime example of the impact factors, however, scientometric indicators have been based on using averages. The impact factor, for example, was defined by Garfield (1972; cf. Sher and Garfield 1965) as the number of citations in a given year to the citable items in a venue during the two preceding years. Journals are then compared in terms of central-tendency statistics.

Using percentiles (deciles, quartiles, etc.) one is able to compare skewed distributions. It is possible to organize percentile rank classes such as the top-1%, top-5%, etc., the method used for more than a decade in the *Science & Engineering Indicators* of the U.S. National Science Board (2012; Bornmann and Mutz 2011). Non-parametric statistics make it possible to test whether the percentile scores are above or below expectation, and also to test whether differences among two units (journals, departments) are statistically significant (Bornmann et al. 2012; Leydesdorff and Bornmann 2012). The percentage of top-1% or top-10% most highly cited papers, for example, can also be considered as an Excellence Indicator (EI) (Tijssen et al. 2002; Waltman et al. 2012; cf. SCImago Institutions Rankings at http://www.scimagoir.com/pdf/sir_2011_world_report.pdf.)

The *Integrated Impact Indicator* (*I3*) provides a framework for organizing these percentile-based indicators.[1] *I3* can formally be written as follows: $I3 = \sum_i x_i * n(x_i)$, in which x_i denotes the percentile (rank) value i, and n the number of papers with this value. The ordering in terms of six percentile rank classes (*PR6*) such as the ones used by the National Science Foundation (NSF) or in terms of an EI follow from *I3* as aggregations. The top-10% most highly cited papers—used increasingly as an *EI*—can be considered as a special case of *I3* in which only two percentile rank

classes are distinguished and weighted with zero and one, respectively (Rousseau 2012).

This article provides examples of the application of *I3* and *PR6* at the researcher and venue levels. Assuming that a decision maker or research manager wishes to use publications and citations as output indicators (Donovan 2011; cf. Bornmann and Daniel, 2008), the changes in measurement and ranking come down to definitions and mathematical principles. It has long been known that publication rankings require normalization because of differences in publication numbers and citation practices across fields of science (Garfield 1979). Practitioners in all fields, however, acknowledge one another's work by citing influential papers. As a single paper accrues more citations, it is assumed to be higher in quality and thus of higher impact. (There are notable exceptions to this rule, such as the Fleishman–Pons claim for nuclear fusion at room temperature, but negative citations are the exception in science (Bornmann and Daniel 2008).)

Just as publication frequencies differ across fields, so do citing norms and patterns. To account for these differences, it is standard practice to normalize by field or at the venue level creating an average of relative citations. The average of citations per publication (c/p) has the obvious *disadvantage* that the total number of publications is the denominator, greatly watering down the impact factor for the few highly cited papers. (For example, when one adds to a principal investigator (PI) the less-cited papers of other members of his team, the average impact will go down because of the larger N in the denominator.) It is nearly always the case that citation distributions are skewed, with a few papers garnering many citations and most papers receiving one or none.

But what is the appropriate procedure if two PIs have different publication and citation profiles? Can two papers in the 39th percentile be considered as equivalent to one in the 78th or is a non-linearity involved (as in the case of the six percentile rank classes)? In Figure 1, we compare the citation curves of two PIs of the Academic Medical Center of the University of Amsterdam. In this academic hospital, the c/p-ratios are used in a model to allocate funding, raising the stakes for methods of assessing impact and inciting the researchers to question the exactness of the evaluation (Opthof and Leydesdorff 2010). The *I3* quantifies the skewed citation curves by normalizing the documents first in terms of percentiles. The question of the normative scheme used for the evaluation can then be considered as the specification of an aggregation rule for the binning and weighting of these scores.

2. Impact at the Level of the Individual Researcher

Figure 1 shows the outputs of two PIs: PI 1 has 1,623 citations from 23 papers and PI 2 has 1,578 citations

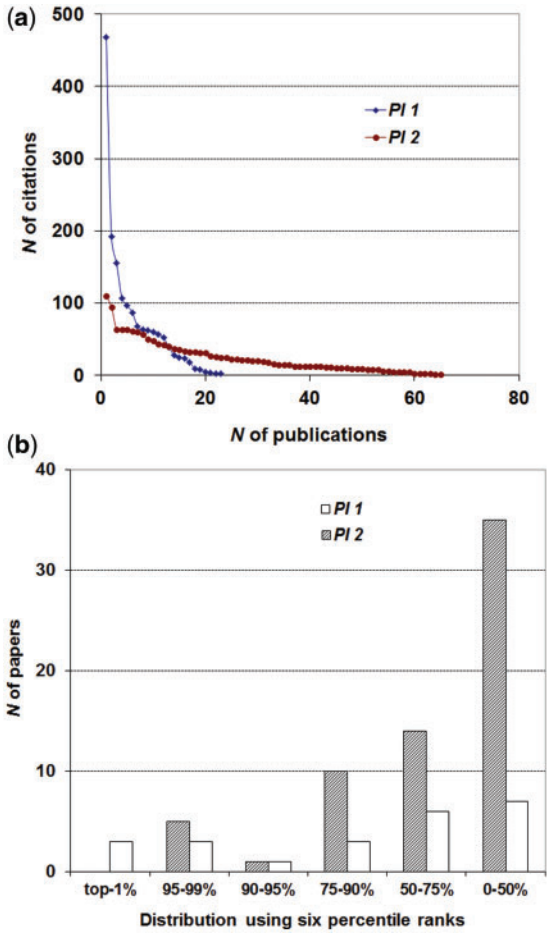

Figure 1. Citation curves and percentile ranks for 23 publications of *PI 1* and 65 publications of *PI 2*, respectively.

from 65 papers. For analytical reasons, integration of the surfaces underneath the citation curves in the top figure provides the total numbers of citations. Whereas the average c/p ratio of PI 1 is $1{,}632/23 = 70.96$ against $1{,}578/65 = 24.28$ for PI 2, the total numbers of citations are not so different. However, the alternative of using the total number of citations without normalization does not yet qualify highly cited papers as different from less highly cited ones.

As the right-hand figure shows, normalization of each paper in terms of the percentile ranks obtained in the different journals in which they are, respectively, published (after proper control for the same publication year and document type) changes the picture. The integration of the normalized citation distributions provides the *I3* and shows that PI 2 has a higher overall impact (Table 1).

The difference between the *I3*-scores for these two PIs is statistically significant (Leydesdorff et al. 2011). Normalization in terms of percentiles greatly improves comparisons across articles at the level of individual researchers and research groups. Using this normalization, for example, a group of researchers has a citation impact equal to the sum of the impacts of the group members.

Table 1. PI 1 and PI 2 compared in terms of the six percentile classes used by NSB

Percentile rank	Weight of rank (x_i)	PI 1		PI 2	
		(n_i)	(n_i * x_i)	(n_i)	(n_i * x_i)
top-1%	6	3	$3 \times 6 = 18$	0	$0 \times 6 = 0$
95–99%	5	3	$3 \times 5 = 15$	5	$5 \times 5 = 25$
90–95%	4	1	$1 \times 4 = 4$	1	$1 \times 4 = 4$
75–90%	3	3	$3 \times 3 = 9$	10	$10 \times 3 = 30$
50–75%	2	6	$6 \times 2 = 12$	14	$14 \times 2 = 28$
0–50%	1	7	$7 \times 1 = 7$	35	$35 \times 1 = 35$
Total		23	$\sum_i^6 x_i n_i = 65$	65	$\sum_i^6 x_i n_i = 122$

Furthermore, an impact measure, in our opinion, should correlate strongly with both the number of publications and citations. When one averages and thus divides the number of citations by the number of publications, one can expect to lose the correlations with each of these two indicators in the numerator and denominator, respectively (Leydesdorff 2009).

In addition to using hundred percentiles (as a continuous random variable), the six classes (top-1%, top-5%, top-10%, top-25%, top-50%, and bottom-50%) can be obtained by simple aggregation of the weighted rank classes as provided, for example, in the National Science Board's *Science and Engineering Indicators* (2010, Appendix Tables 5–43). However, it is also possible to use deciles or quartiles once the percentile values are known at the article level. Thus, the choice of a normative framework for the evaluation is not pre-determined by the analysis.

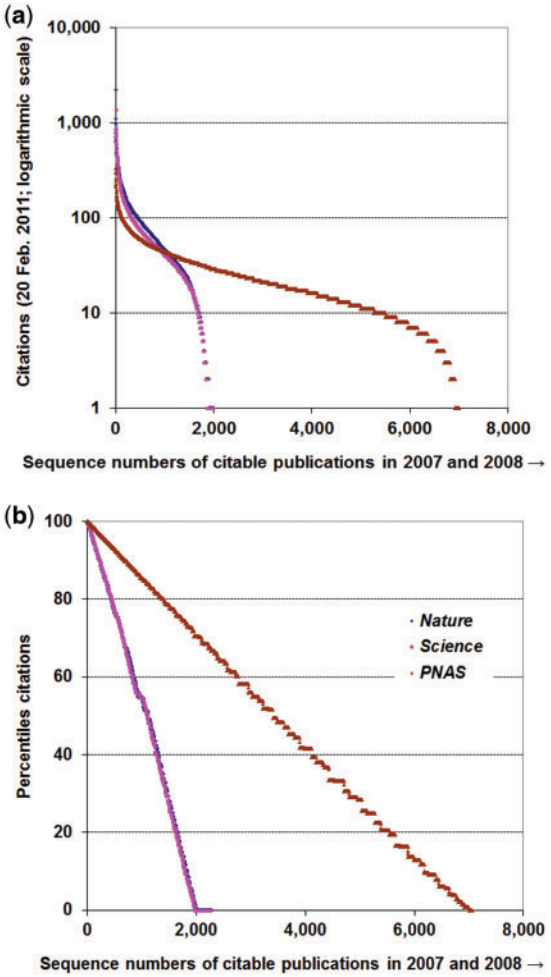

Figure 2. Citation rates and percentiles for *Nature* (♦), *Science* (■), and *PNAS* (▲), respectively; using 48 'multidisciplinary' journals in the *Science Citation Index* as the reference set. (Source: Leydesdorff and Bornmann 2011, p. 2141.).

3. Impact at the Venue Level

Scientometricians often normalize at the venue level (i.e. journals or sets of journals) using the field classification systems in the *Science Citation Index* and *Scopus*. Publishers regularly advertise their 'journal impact factor' (JIF) to improve the quality of submissions. Impact factors, however, can be considered as two-year averages over skewed distribution and therefore one can expect problems of unfair evaluations similar to the problems with c/p ratios of individual researchers and research groups.

In Figure 2, the 48 journals classified in the *Science Citation Index* as 'multidisiciplinary' are used as the reference set to compare the three leading journals in this category (*Science*, *Nature*, and *Proceedings of the National Academy of Sciences PNAS*). The top panel shows the raw citation curves for 2009 of the articles in the three journals during 2007 and 2008; the c/p values are then by definition equal to the JIFs 2009. The visual shows *Science* and *Nature* competing for first place. Then—using the same data—the lower panel shows results with the normalization in terms of percentiles. *Science* and *Nature* still have nearly identical curves, but *Proceedings of the National Academy of Sciences PNAS* stands out as having significantly higher impact. (The values for the IFs, *I3*, and six percentile ranks (*PR6*) are summarized in Table 2.)

To contrast the results obtained when using the new *I3* and the JIFs, Table 2 compares six prestigious journals that target a broadly interested readership and with highest citation impacts using the percentile ranking. Since absolute values of *I3* and *PR6* are based on summations, we use their relative values as percentages for clarity. When using these new indicators (columns c and d), *PNAS* has a much higher impact than would be derived by using an average-based JIF (shown in column e). Indeed, the finding of much higher impact factors for *Science* and *Nature* (column e) are an artifact of the smaller numbers of publications (column a) in these two journals rather than higher citation counts at the top end. All three

Table 2. 'Multidisiciplinary' journals with highest values for *I3* and six percentile ranks (PR6) compared in rankings (between brackets) with JIFs and total citations

Journal	*N* of papers (a)	*N* of citations (b)	% I3 (c)	% PR6 (d)	JIF 2009 (e)
Proc Natl Acad Sci USA	7,058	178,137	43.29 [1][a]	33.64 [1][a]	9.432 [3]
Nature	2,285	150,718	16.31 [2][a]	16.46 [2][a]	34.480 [1]
Science	2,253	126,230	15.68 [3][a]	15.27 [3][a]	29.747 [2]
Ann NY Acad Sci	1,996	14,284	9.33 [4][a]	8.29 [4]	2.670 [5]
Curr Sci	1,271	1,551	2.33 [5][b]	3.40 [5][b]	0.782 [22]
Chin Sci Bull	1,115	2,239	2.11 [6][b]	2.55 [6][b]	0.898 [20]

Source: Leydesdorff and Bornmann 2011. [a]Above expectation at $p < 0.01$; [b]below expectation at $p < 0.01$ (using the *z*-test).

journals, however, have an impact significantly above expectation ($p < 0.01$).

In contrast, consider the next three journals in Table 2. The *Annals of the New York Academy of Sciences* follows at the fourth position in terms of *I3*, but if the six percentile ranks of the NSB are used, *Annals* no longer scores significantly above expectation. *PR6* gives more weight to top-cited papers than *I3*. The two Asian journals in the category—the *Chinese Science Bulletin* and the Indian journal *Current Science*—are ranked at the fifth and sixth positions among this group of 48 'multidisciplinary' journals, while they were ranked much lower—20th and 22nd, respectively—using JIFs, as can be seen in column (e). The citation rates of these two journals, however, are still below expectation.

If the comparison is made among JIFs with *I3* and *PR6* values for the full set of 48 journals in this set (with *N* of documents is equal to 24,494), the Pearson correlations are 0.590 and 0.660 ($p < 0.01$), respectively (Table 3). As can be expected *I3* and *PR6* are highly correlated between them ($r = 0.987$), as they are both referencing citation and publication rates. All these correlations with productivity and impact are larger than 0.9. However, JIF correlates 0.49 with the number of publications and 0.84 with the number of citations. In other words, the division by the number of publications makes *average* impact different from impact, and this change in the semantics matters in evaluation scenarios.

4. An Essential Change: Impacts Add Up Instead of Averaging Out

Before this change in the definition of impact, it was common to use two conventions for normalization: (1) normalization in terms of fields of science, and (2) comparison of a paper's or journal's citation rate to the world average. Both of these conventions raise problems when assessing impact.

The sources of error in the first practice—normalizing in terms of a field—come from using journals such as *Nuclear Physics B* or *Cell* as the units for normalization: a number of studies have demonstrated that specialist journals do not necessarily represent a single discipline (Pudovkin and Garfield 2002; Rafols and Leydesdorff 2009; Boyack and Klavans 2011). Therefore, even if the *Science Citation Index* and *Scopus* refined the journal classifications, it is not clear that this would solve the problem of field delineation or improve the quality of rankings (Leydesdorff 2006).

The second convention—comparison to the world average—was defined early in the scientometric enterprise by Schubert and Braun (1986) who proposed to compare the mean observed citation rate (*MOCR*) within a database (representing, for example, a field of science) with the corresponding mean *expected* citation rate (*MECR*) as the average citation rates of papers of the same datatype (reviews, articles, or letters) and publication year. The Relative Citation Rate ($= MOCR/MECR$) is thus normalized to the world average.

The combination of these measures caused the problems: The division of two means results in mathematical inconsistency because the order of operations says that one should divide first and then sum, not sum and then divide the averages. As this error became clear (Opthof and Leydesdorff 2010; cf. Lundberg 2007), the renowned Leiden Centre for Science and Technology Studies (CWTS) changed its main ('crown') indicator (van Raan et al. 2010). CWTS called this 'new crown indicator' the Mean Normalized Citation Score (*MNCS*). One advantage of the new indicator is that the mean is a statistic with a standard deviation, and consequently a standard error of the measurement can be defined and can be published as an error bar in relevant assessments. Waltman et al. (2011) showed that this new indicator is mathematically consistent.

To further refine the indicator for broad application, Leydesdorff and Bornmann (2011) elaborated this approach by defining the *I3*. *I3* leaves the parametric domain of working with averages behind and moves to non-parametric statistics using percentiles. Rousseau (2012) discussed the mathematical properties of the new indicator in more detail. Unlike the *h*-index, the tails of

Table 3. Rank-order correlations (Spearman's ρ; upper triangle) and Pearson correlations r (lower triangle) for the 48 journals attributed to the WoS Subject Category 'multidisciplinary sciences' in 2009

Indicator	IF-2009	I3	PR6	Number of publications	Total citations
IF-2009		0.798[a]	0.517[a]	0.479[a]	0.840[a]
I3	0.590[a]		0.854[a]	0.829[a]	0.986[a]
PR6	0.660[a]	0.987[a]		0.996[a]	0.801[a]
N of publications	0.492[a]	0.953[a]	0.967[a]		0.772[a]
Total citations	0.841[a]	0.922[a]	0.945[a]	0.839[a]	

Source: Leydesdorff and Bornmann, 2011, p. 2142. Note: [a]Correlation is significant at the 0.01 level (2-tailed).

the distribution are also weighted in *I3*, and nonparametric statistics (as available, for example, in SPSS) can be used.

When creating the ISI-impact factor in the early days of the *Science Citation Index*, Eugene Garfield (e.g. 1972) deliberately chose to normalize by dividing by N in order to prevent the larger journals from overshadowing the smaller. Bensman (2007) found 'Total Citations' to be more closely correlated than the JIFs with subjective appreciation of faculty. Even so, Total Citations and Impact Factors were crude (first-generation?) measures. The percentile approach allows a user both to account for the skewed distribution of citations and appreciate differences among highly cited papers and less highly cited ones. As said, the user can then aggregate the percentiles in a normative evaluation scheme (e.g. quartiles or the six classes of the NSF).

5. Normative Implications and Policy Relevance

Research funds are often allocated based upon these measures. Fields of science, institutions, and nations are increasingly ranked based upon the recognition bestowed by citation counts. Policy decisions about spending the incremental research dollar, euro, or yen often rest upon the excellence of research units in terms of citation counts and impacts (Hicks 2012). Thus, these distinctions are important to users across a wider spectrum of research evaluation and policy.

Indicators clearly matter to the individual researchers, as discussed in the comparison of PI 1 and PI 2 above; research units and nations are also served by the improvement offered by *I3*. Using a refined indicator can improve the efficiency of research spending by increasing the likelihood that the most relevant or appropriate researcher or research unit receives funding. In times of difficult budget choices, it is even more important to ensure the accuracy of the basic measures of the research system and its components—the benefit offered by using the I3.

Additional information and free software for the automatic analysis of document sets in terms of the percentile values can be found online at http://www.leydesdorff.net/software/i3/index.htm.

Note

1. The percentiles can be considered as a continuous random variable (quantiles; Rousseau 2012).

References

Bensman, S. J. (2007) 'Garfield and the Impact Factor', *Annual Review of Information Science and Technology*, 41/1: 93–155.

Bornmann, L. and Daniel, H. D. (2008) 'What do Citation Counts Measure? A Review of Studies on Citing Behavior', *Journal of Documentation*, 64/1: 45–80.

Bornmann, L. and Mutz, R. (2011) 'Further Steps Towards an Ideal Method of Measuring Citation Performance: The Avoidance of Citation (Ratio) Averages in Field-normalization', *Journal of Informetrics*, 5/1: 228–30.

Bornmann, L., de Moya-Anegón, F. and Leydesdorff, L. (2012) 'The New Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011', *Journal of Informetrics*, 6/3: 333–5.

Boyack, K. W. and Klavans, R. (2011) 'Multiple Dimensions of Journal Specifity: Why Journals Can't be Assigned to Disciplines'. In: Noyons, E., Ngulube, P. and Leta, J. (eds) *The 13th Conference of the International Society for Scientometrics and Informetrics*, Vol. I, pp. 123–33. Durban, South Africa: ISSI, Leiden University and the University of Zululand.

Donovan, C. (2011) 'State of the Art in Assessing Research Impact: Introduction to a Special Issue', *Research Evaluation*, 20/3: 175–9.

Garfield, E. (1972) 'Citation Analysis as a Tool in Journal Evaluation', *Science*, 178/4060: 471–9.

——. (1979) 'Is Citation Analysis a Legitimate Evaluation Tool?', *Scientometrics*, 1/4: 359–375.

Gingras, Y. and Larivière, V. (2011) 'There are Neither "King" Nor "Crown" in Scientometrics: Comments on a Supposed "Alternative" Method of Normalization', *Journal of Informetrics*, 5/1: 226–7.

Hicks, D. (2012) 'Performance-based University Research Funding Systems', *Research Policy*, 41/2: 251–61.

Leydesdorff, L. (2006) 'Can Scientific Journals be Classified in Terms of Aggregated Journal-Journal Citation Relations using the Journal Citation Reports?', *Journal of the American Society for Information Science & Technology*, 57/5: 601–13.

——. (2009) 'How are New Citation-Based Journal Indicators Adding to the Bibliometric Toolbox?', *Journal of the American Society for Information Science and Technology*, 60/7: 1327–36.

Leydesdorff, L. and Bornmann, L. (2011) 'Integrated Impact Indicators (I3) Compared with Impact Factors (IFs): An Alternative Design with Policy Implications', *Journal of the American Society for Information Science and Technology*, 62/11: 2133–46.

——. (2012) 'Testing Differences Statistically with the Leiden Ranking', *Scientometrics*, <http://arxiv.org/abs/1112.4037> accessed 1 June 2012.

Leydesdorff, L., Bornmann, L., Mutz, R. and Opthof, T. (2011) 'Turning the Tables in Citation Analysis One More Time: Principles for Comparing Sets of Documents', *Journal of the American Society for Information Science and Technology*, 62/7: 1370–81.

Lundberg, J. (2007) 'Lifting the Crown—Citation z-score', *Journal of informetrics*, 1/2: 145–54.

Merton, R. K. (1968) 'The Matthew Effect in Science', *Science*, 159/3810: 56–63.

National Science Board. (2012) *Science and Engineering Indicators*. Washington DC: National Science Foundation, <http://www.nsf.gov/statistics/seind12/> accessed 1 June 2012.

Opthof, T. and Leydesdorff, L. (2010) '*Caveats* for the Journal and Field Normalizations in the CWTS ("Leiden") Evaluations of Research Performance', *Journal of Informetrics*, 4/3: 423–30.

Pudovkin, A. I. and Garfield, E. (2002) 'Algorithmic Procedure for Finding Semantically Related Journals', *Journal of the American Society for Information Science and Technology*, 53/13: 1113–9.

Rafols, I. and Leydesdorff, L. (2009) 'Content-based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects', *Journal of the American Society for Information Science and Technology*, 60/9: 1823–35.

Rousseau, R. (2012) 'Basic Properties of Both Percentile Rank Scores and the I3 Indicator', *Journal of the American Society for Information Science and Technology*, 63/2: 416–20.

Schubert, A. and Braun, T. (1986) 'Relative Indicators and Relational Charts for Comparative Assessment of Publication Output and Citation Impact', *Scientometrics*, 9/5: 281–91.

Seglen, P. O. (1992) 'The Skewness of Science', *Journal of the American Society for Information Science*, 43/9: 628–38.

——. (1997) 'Why the Impact Factor of Journals Should Not be Used for Evaluating Research', *British Medical Journal*, 314: 498–502.

Sher, I. H. and Garfield, E. (1965) *New tools for improving and evaluating the effectiveness of research*, Paper presented at the Second conference on Research Program Effectiveness, July 27–9, Washington, DC.

Tijssen, R. J. W., Visser, M. S. and Van Leeuwen, T. N. (2002) 'Benchmarking International Scientific Excellence: Are Highly Cited Research Papers an Appropriate Frame of Reference?', *Scientometrics*, 54/3: 381–97.

Van Raan, A. F. J. *et al*. (2010) *The New Set of Bibliometric Indicators of CWTS*, Paper presented at the 11th International Conference on Science and Technology Indicators, September 9–11, 2010, Leiden.

Waltman, L. *et al*. (2011) 'Towards a New Crown Indicator: Some Theoretical Considerations', *Journal of Informetrics*, 5/1: 37–47.

——. (2012) 'The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation', <http://arxiv.org/abs/1202.3941> accessed 1 June 2012.