# Integrated Impact Indicators Compared With Impact Factors: An Alternative Research Design With Policy Implications

**Loet Leydesdorff**
*University of Amsterdam, Amsterdam School of Communication Research, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands. E-mail: loet@leydesdorff.net*

**Lutz Bornmann**
*Max Planck Society, Hofgartenstrasse 8, D-80539 Munich, Germany. E-mail: bornmann@gv.mpg.de*

**In bibliometrics, the association of "impact" with central-tendency statistics is mistaken. Impacts add up, and citation curves therefore should be integrated instead of averaged. For example, the journals *MIS Quarterly* and *Journal of the American Society for Information Science and Technology* differ by a factor of 2 in terms of their respective impact factors (IF), but the journal with the lower IF has the higher impact. Using percentile ranks (e.g., top-1%, top-10%, etc.), an Integrated Impact Indicator (*I3*) can be based on integration of the citation curves, but after normalization of the citation curves to the same scale. The results across document sets can be compared as percentages of the total impact of a reference set. Total number of citations, however, should not be used instead because the shape of the citation curves is then not appreciated. *I3* can be applied to any document set and any citation window. The results of the integration (summation) are fully decomposable in terms of journals or institutional units such as nations, universities, and so on because percentile ranks are determined at the paper level. In this study, we first compare *I3* with IFs for the journals in two Institute for Scientific Information subject categories ("Information Science & Library Science" and "Multidisciplinary Sciences"). The library and information science set is additionally decomposed in terms of nations. Policy implications of this possible paradigm shift in citation impact analysis are specified.**

## Introduction

Let us introduce the problem of defining impact by taking as an example the citation curves of two journals with very different impact factors (IFs). Figure 1 shows the citation curves of the 66 and 375 citable items published in *MIS Quarterly* and the *Journal of the American Society for Information Science and Technology* (*JASIST*), respectively, during 2007 and 2008.[1] These two journals are both attributed by Thomson Reuters—the present owner of the Institute for Scientific Information (ISI)—to the ISI Subject Category of "Library and Information Science" (LIS),[2] although they are very different in character (Nisonger & Davis, 2005; Zhao & Strotmann, 2008). Within this subject category, *MIS Quarterly* had the highest IF in 2009: 4.485. The IF of *JASIST* in 2009 is approximately half this size: 2.300. However, the 66 most highly cited publications of *JASIST* obtained 380 citations more than the 66 citable items published in *MIS Quarterly* (downloaded on February 17, 2011). The lower IF factor is entirely due to the tail of 300+ additional publications in *JASIST* with lower citation rates.

In our opinion, this confusion finds its origin in the definition of the "impact factor" as a 2-year average of "impact" (Garfield, 1972; Garfield & Sher, 1963; cf. Bensman, 2007; Rousseau & Leydesdorff, 2011).[3] Impact (as a variable), however, is not an average but the result of the sum of the momenta of the impacting units. For example, two meteors impacting on a planet can have a combined impact larger than that of each of them taken separately, but the respective velocities also matter.

In physics, momentum is defined as the vector of mass times velocity ($p = m\vec{v}$). Using the metaphor of impact, both the number of publications (the "mass") and their citation counts (the quality of "the velocity") matter for the impacts.

Received March 27, 2011; revised June 13, 2011; accepted June 13, 2011

 • Published online 8 August 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21609

[1]The *Journal Citation Reports* (*JCR*) 2009 lists 370 instead of 375 citable issues for 2007 plus 2008. This difference originates from the date in March that the *JCR* team at Thomson Reuters decides to use for producing the *JCR* of the year before (M. McVeigh, personal communication, April 7, 2010). IFs are notoriously difficult to reproduce using Web of Science data (e.g., Brumback, 2008a, 2008b; Pringle, 2008; Rossner et al., 2007, 2008).

[2]The ISI uses "Information Science & Library Science" as the name of this category.

[3]More recently, the ISI also introduced the 5-year IF in the *Journal Citation Reports* (*JCR*).

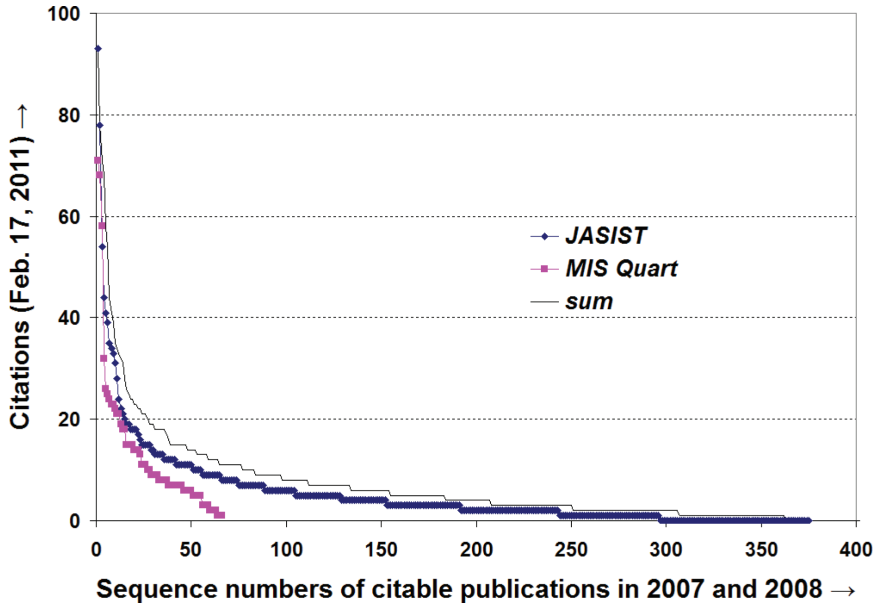


FIG. 1. Citation curves for *JASIST* ($n = 375$ publications) and *MIS Quarterly* ($n = 66$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Comparison of *MIS Quarterly* and the *Journal of the American Society for Information Science and Technology* (*JASIST*) in terms of citation rates to citable items in 2007 and 2008. Publication and citation data retrieved at the Web of Science on February 17, 2011.

| | IF 2009 | Publications in our data | Citations in our data | Citations/publications | *Mdn* |
|---|---|---|---|---|---|
| *MIS Quarterly* | 296/66 = 4.485 | 66 | 847 | 12.83 | 8 |
| *JASIST* | 851/370 = 2.300 | 375 | 1,975 | 5.27 | 3 |
| Total | 1,147/436 = 2.631 | 441 | 2,822 | 6.40 | 3 |

Because citations are scalar counts, one can disregard the direction of the vectors in the summation ($\Sigma m\vec{v}$). The research question is then how to operationalize $m$ in terms of the numbers of publications and $v$ in terms of citations to obtain a relevant measure of impact as a sum. The impact of each subset can be expressed as a percentage impact of the set.

It has been argued (e.g., Bornmann & Mutz, 2011; Leydesdorff & Opthof, 2011) that the median should be used in citation analysis instead of the mean because of the skewness of citation distributions (e.g., Seglen, 1992). For the two journals in the example presented earlier—and using the 2-year time window of the ISI-IF—Table 1 shows that the median is in this case even more sensitive to the tails of the distributions than is the mean. A more radical solution is therefore needed: Impact has to be defined not as a distribution but as a sum. Very different distributions can add up to the same impact. The number of citations can be highly skewed, and in this situation, any measure of central tendency is theoretically meaningless.

Whereas distributions of citations can be tested nonparametrically for the significance of the differences among them, impacts are sum values. These values can be tested against the expected values of the variables. For example, if the set of documents in one journal is twice as large as the set in another, the chance that it will contain a top-1% most highly cited document is twice as high. If the observed value, however, would be four times as high, this achievement above expectation may be statistically significant, but this also depends on the sample size ($N$).

Central tendency statistics cannot capture the increases in impact when two sets ("masses") are added, such as when two research groups join forces or two journals merge. We penciled the line for the sum total of *JASIST* and *MIS Quarterly* into Figure 1 to show that one has to sum surfaces, and thus the citation curves have to be *integrated* instead of averaged. Assuming the integrals, however, would lead to numbers equal to the "total citations" without qualifying the documents in terms of their citedness. To weigh the documents, we suggest to transform the citation curves first into curves of 100 percentiles, as in Figure 2.

The distributions of percentile ranks can be compared fairly across document sets, and these linearly transformed distributions can be integrated. The integrals in this stepwise function are equal to $\sum_i x_i * f(x_i)$, in which $x$ represents the percentile rank and $f(x)$ the frequency of that rank; $i$ is 100 when using percentiles, but, for example, 4 when using quartiles as percentile rank classes, and so on. One also can consider 100 percentiles as a continuous random variable and


DOI: 10.1002/asi

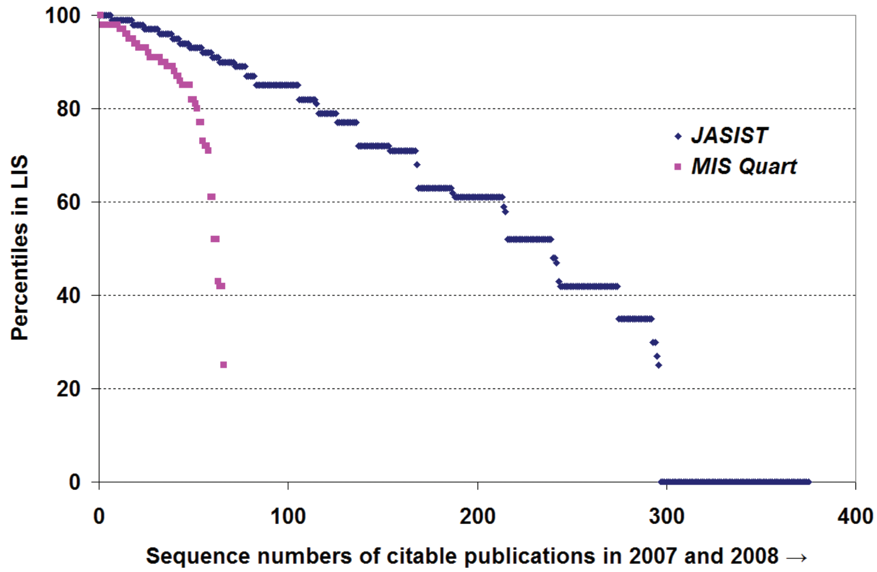


FIG. 2. Distributions of 100 percentile ranks of *JASIST* and *MIS Quarterly* with reference to the 65 journals of the ISI Subject Category LIS. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

sum these values, as we shall explain in more detail later. The function integrates both the number of papers (the "mass") and their respective quality in terms of being cited normalized as percentiles with reference to a set.

The idea of using percentile rank classes was first formulated in the discussion about proper normalization of the citation distribution that took place last year in the *Journal of Informetrics* (e.g., Bornmann, 2010; Opthof & Leydesdorff, 2010; van Raan, van Leeuwen, Visser, van Eck, & Waltman, 2010; cf. Gingras & Larivière, 2011). In this context, Bornmann and Mutz (2011) proposed to assess citation distributions in terms of six percentile rank classes (*6PR*): the top-1%, top-5%, top-10%, top-25%, top-50%, and bottom-50%. This (normative!) evaluation scheme accords with those currently used in the biannual *Science and Technology Indicators* of the National Science Board (2010, Appendix Table 5–43). Each publication would then be weighted in accordance to its class as a 6 for the top-1% category and a 1 for the bottom-50% category. Leydesdorff, Bornmann, Mutz, and Opthof (2011) extended this approach to hundred percentiles, which also can be weighted as classes from 1 to 100 (*100PR*).

The advantage of using percentile ranks is that one is thus able to compare distributions of citations across unequally sized document sets using a single scheme for the evaluation of the shape of the distribution. However, Bornmann and Mutz's (2011) approach remained sensitive to the central-tendency characteristic discussed earlier because these authors *averaged* over the percentile ranks using the following formula: $R_i = \sum_i x_i * p(x_i)$. In this formula, $x$ is the rank class and $p(x)$ its *relative* frequency (or proportion). However, this probabilistic approach implies a division by the number of cases and thus normalization to the mean (albeit of the distribution of the percentiles). Leydesdorff et al. (2011) therefore called this method the "mean percentile rank" approach.

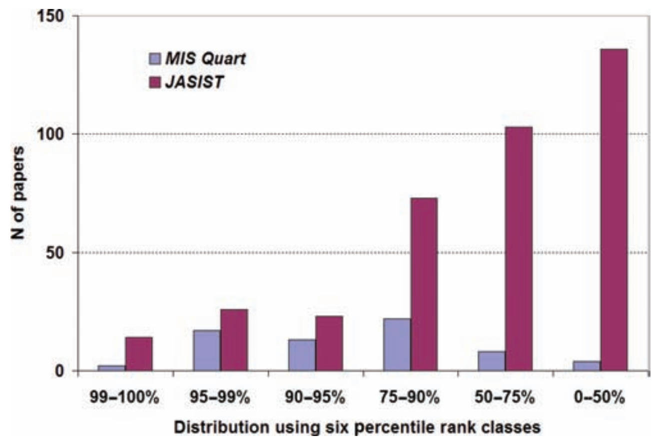


FIG. 3. Distributions of the six percentile ranks of publications in terms of citations to *JASIST* and *MIS Quarterly* (with reference to all 65 of LIS). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Means cannot be added whereas impacts are additive. Bensman and Wilder (1998) concluded on the basis of validation studies that the prestige of journals in chemistry is correlated with the total number of citations more than with the IFs of journals. As noted earlier, however, total citations do not yet qualify the shape of the distributions since every citation is then counted equally. IFs only qualify the distribution in terms of the mean, but deliberately abstract from size (Bensman, 2007). Using the sum total of the frequencies ($f$) in each percentile, however, accounts for both the size and shape of the distribution. The citations are weighted in accordance with the percentile rank class of each publication in an Integrated Impact Indicator (*I3*): $I3 = \sum_i x_i * f(x_i)$.

Figure 3 shows the distribution of the six percentile ranks of the National Science Board (2010) for *MIS Quarterly* and *JASIST*. Both Figures 2 and 3 show that *JASIST* has an impact


DOI: 10.1002/asi

TABLE 2. Mean (±*SEM*), median, and total in the case of 100 or 6 percentile ranks.

| | *100PR* | | | *6PR* | | |
|---|---|---|---|---|---|---|
| | *M* | *Mdn* | Total | *M* | *Mdn* | Total |
| *MIS Quarterly* | 84.57 ± 1.98 | 90 | 5,581.4 | 3.56 ± 0.15 | 3 | 235 |
| *Journal of the American Society for Information Science and Technology* | 55.50 ± 1.76 | 61 | 20,811.3 | 2.31 ± 0.07 | 2 | 867 |

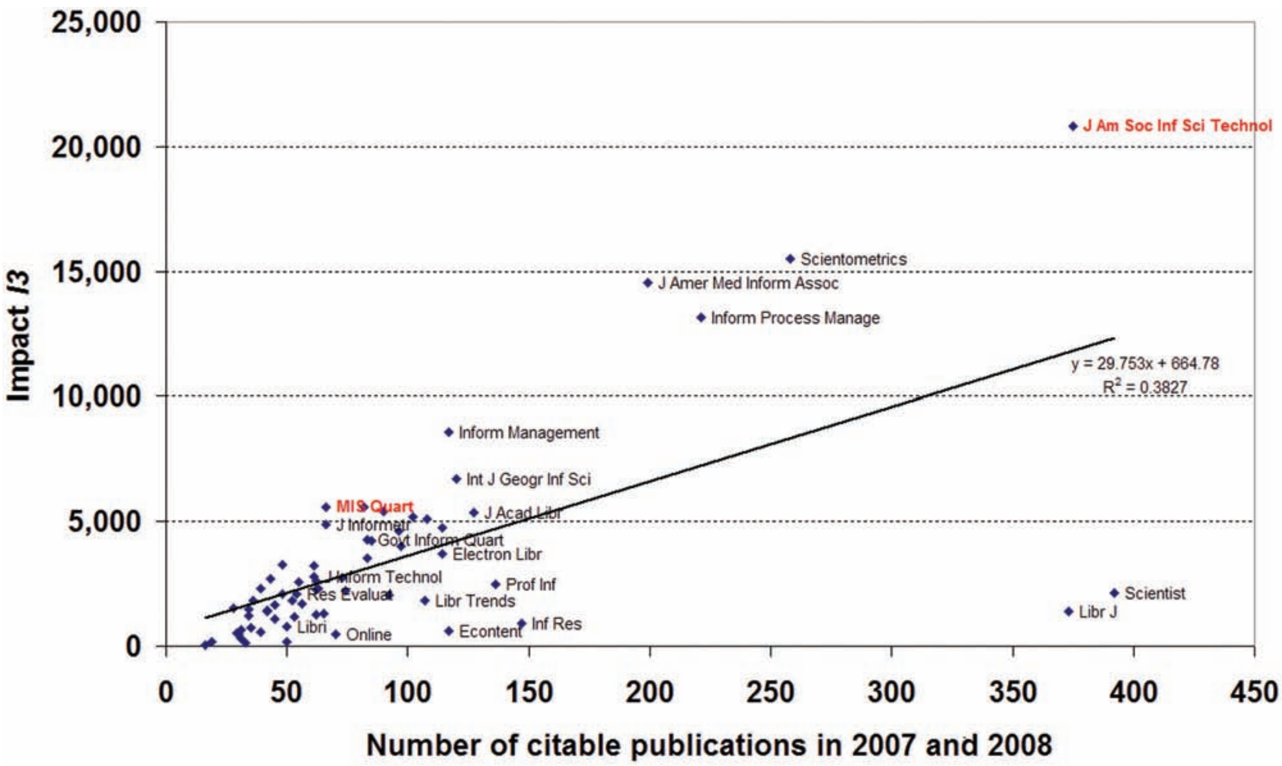


FIG. 4. Regression of impact (*I3*) against number of citable publications in 2007 and 2008 for the 65 journals of LIS. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

higher than that of *MIS Quarterly* using these normalized curves. Table 2 shows that this higher value can be captured by the sum, but not by the means or medians, of the percentile distributions.

The sum values can be added (and subtracted), for example, for the purpose of the aggregation or decomposition (e.g., in terms of contributing nations), and they also can be expressed as percentages of the total integrated impact of an ISI Subject Category (e.g., the 65 journals in the LIS category). For example, in 2009 the sum total of the impact of all 5,737 citable items in the LIS category was 213,906.2.[4] *MIS Quarterly* contributed 2.61% to this total impact of the set when using *100PR*, and 2.34% in the case of *6PR*. These percentages were 9.73% and 8.63%, respectively, for *JASIST*. *JASIST* is thus to be considered as the journal with the highest impact in the set of 65 journals subsumed under LIS among the ISI Subject Categories (Nisonger, 1999).

[4]This sum total is equal to 39.8% of the maximally possible impact of $(100 \times 5,737 =)$ 537,700 in the case of *100PR*. In the case of *6PR*, the total is 10,049, or 29.2% of the maximally possible impact of $(6 \times 5,737 =)$ 34,422.

Figure 4 further elaborates this example by showing the regression line of the impacts thus calculated against the number of citable publications in 2007 and 2008. Unlike dividing sums to obtain average values—which was the core issue in the previous controversy about the Leiden crown indicator (Gingras & Larivière, 2011)—this regression informs us that the journal set under study is heterogeneous ($r^2 = 0.38$), both in size and functions. *The Scientist* and the *Library Journal*, for example, are grouped in the bottom-right angle of this figure because they function as newsletters more so than scholarly journals. Among the journals at the top-right, the label for *JASIST* was colored red to show its leading position in this set. *MIS Quarterly* is a journal above the regression line in the set of specialized journals at the bottom-left, but also colored red for the purpose of this comparison.

## Methods

Data were harvested from the Web of Science (WoS) in February 2011. Because we wished to compare our results for *I3* with the latest available IFs 2009, we downloaded citable items in 2007 and 2008 in two ISI Subject Categories, namely, the one for LIS containing 65 journals and the category "Multidisciplinary Sciences" (MS) containing 48 journals, but including important journals such as *Science*, *Nature*, and *Proceedings of the National Academy of Sciences, USA* (*PNAS*). The delineation of the ISI Subject Categories is beset with error (Rafols & Leydesdorff, 2009). Within this context, however, we use them pragmatically as reference sets because it is beyond our capacity—and perhaps illegal—to download the entire database. Only articles, proceedings papers, reviews, and letters are included because these categories are indicated by Thomson Reuters as citable items (cf. Moed & Van Leeuwen, 1996). Note that *I3* is by no means confined to this definition of impact in terms of 2 previous years but can be used for any document set and with any citation window.

The citation of each paper is rated in terms of its percentile in the distribution of citations to all items with the same document type and publication year in its ISI Subject Category as the respective reference sets. The percentile is determined by using the counting rule that the number of items with lower citation rates than the item under study determines the percentile. Tied citation numbers are thus provided with the highest values, and this accords with the idea of providing all papers with the highest possible ranking (i.e., we wish to give the papers "the benefit of the doubt"). Other schemes also are possible. For example, Pudovkin and Garfield (2009) first averaged tied ranks.

The percentiles can be considered as a continuous random variable in the case of 100 percentiles. In the case of six percentile ranks, one has to round off. Differently from Leydesdorff et al. (2011), the rounding off in this study will be based on adding 0.9 to the count (i.e., count +0.9) because otherwise one can expect undesirable effects for datasets that are smaller than 100. For example, if a journal with many articles publishes only 10 reviews each year, the highest possible percentile within this set would be the 90th (i.e., 9 of 10) whereas this could be the 99th (i.e., 9.9 of 10), and thus included in the top-1% percentile rank (with a value of 6 in the mentioned evaluation scheme of the NSF).

As shown in Figures 2 and 3 above comparing *MIS Quarterly* with *JASIST*, the percentiles provide us with a scale that can be compared across document sets with different sizes. In the case of a normative evaluation scheme such as that of the NSF, the percentiles are binned in six percentile rank classes. This transformation is nonlinear and one looses information, but an evaluator may gain clarity in the distinctions from a policy perspective (Leydesdorff et al., 2011). We shall use this second set of values throughout this study for the comparison, but distinguish it from *I3* by denoting this measure as *I3*(*6PR*). The formula is then specified as follows: $I3(6PR) = \sum_{i=1}^{6} x_i * PR_i$, in which $PR_i$ is the frequency value in the respective class. Other evaluation schemes also are possible, but in our opinion, this is a normative discussion which can be expected to change with the policy context.

The set based on the 65 journals of LIS contained 5,737 citable items published in 2007 or 2008. The set indicated by the ISI as journals in MS was much larger in terms of the number of documents ($n = 24{,}494$ citable items) despite the smaller number of journals ($n = 48$). The two sets were brought under the control of relational database management, and when necessary, dedicated routines were written to format the data for analysis in SPSS (Version 18) and Excel. Using the WoS, the numbers of citations were determined at the date of downloading—in our case, February 2011.

The most relevant routine in SPSS is "Compare Means" using the journals (in each set, respectively) as the independent (grouping) variable and the percentiles (or the six classes, mutatis mutandis) as the dependent variables. This routine allows for determining the mean, the sum, the standard error of the mean, confidence levels, and other statistics in a single pass. Since we are mainly interested in the sum, the mean, and the standard error of the mean, this routine is sufficient for our purpose. Correlation analysis (both Pearson's $r$ and Spearman's rank-order correlation $\rho$) also will be pursued using SPSS to compare the new indicators with IFs.

In addition to analyzing the impact of each journal, the question can be raised about whether the citation distributions also are significantly different. Nonparametric statistics enable us to answer this question using the citation distributions (as depicted in Figure 1) without averaging or first transforming them into percentile ranks. Among the routines available for multiple comparisons in SPSS (with Bonferroni correction), Dunn's test can be simulated by using least significant differences with family-wise correction for Type I error probability. In the case of $N$ groups to be compared, the number of comparisons is $N \times (N-1)/2$. For example, in the case of 50 journals, $50 \times 49/2 = 1{,}225$ comparisons are pursued, and the significance should hence be tested at the 5% level using $0.05/1{,}225 = 0.000041$ instead of 0.05 (Levine, 1991, pp. 68 ff.).

The routine for multiple comparisons in SPSS is limited to 50 groupings at a time. In the case of the MS set, 48 journals are involved; however, in the case of LIS, 65 journals can be compared. We perform the analysis in this study using the 50 journals with the highest IFs among these 65 journals (The IF was chosen as criterion to not bias our results in favor of the proposed measure.) Alternatively, one can test any two journals against each other using the Mann–Whitney $U$ test with Bonferroni correction. However, in the case of 65 journals, these would be 2,080 one-by-one comparisons. This did not seem necessary for the purpose of this study.

Furthermore, the algorithm of Kamada and Kawai (1989)—as available, for example, in Pajek—provides us with a means to visualize groups of journals as *not* significantly different—that is, homogenous—in terms of their citation distributions (cf. Leydesdorff & Bornmann, 2011, p. 224). Journals which can be considered similar in this respect were linked in the graphs whereas in the case of significant differences the grouping links were omitted. The $k$-core sets which are most homogeneous in terms of citation distributions thus can be visualized.

In a final section, we return to the issue of performance measurement of institutional units, individuals, and/or nations (but using this same data). Since the attribution of the percentile rank is done at the paper level, one can aggregate and decompose subsets in terms of their contribution to the reference set. We shall use country names in the address field as an example. Each contribution to *I3* also can be expressed as a percentage.

The observed contributions can be tested against the expected ones on the basis of the distribution of citable items across units of analysis (e.g., journals or nations). In a larger set, for example, one can expect more highly cited papers for stochastic reasons. Whether a difference is statistically significant can be assessed for each case using the binomial *z* test or the standardized residuals of the $\chi^2$. We use the latter measure $\left[Z = \frac{observed-expected}{\sqrt{expected}}\right]$ because this test is simpler and less conservative than is the binomial test. Expected values below 5 are discarded as unreliable.

A *z* value of 1.96 (i.e., almost 2 *SD*s) can be considered as significant at the 5% level, and similarly $z_{0.01} = 2.576$. The notation of SPSS will be followed in this study using two asterisks for significance at the 1% level and a single asterisk for the 5% level. However, we use the signs of the differences (++, +, −−, or −) when relevant in the tables to indicate whether the observed values are significantly above or below the expected values, and at which level of significance.

In summary, we distinguish between testing (a) observed impacts as sum values of percentiles against expected impacts of units of analysis (e.g., journals, nations, etc.) using *Z* statistics and (b) differences in the citation distributions, for example, in terms of Dunn's test. The latter test provides us with a nonparametric alternative to comparing these distributions in terms of their arithmetic averages, as is done in the case of comparing among IFs (cf. Leydesdorff, 2008).

## Results

### *I3 for the 65 journals of LIS*

Table 3 provides the Pearson correlations (in the lower triangle) and the Spearman rank order correlations (upper triangle) between the various indicators under discussion. Most of the correlation coefficients are high (≥0.7), and significant at the 1% level. Using Pearson correlation coefficients, however, the numbers of publications (*N*s) in journals of this set are not correlated to the IFs ($r = 0.151$, n.s.) or the mean values of *100PR* ($r = 0.042$, n.s.) and of *6PR* ($r = 0.134$, n.s.). These indicators have in common that they are based on averages and therefore division by the number of publications in each set.

The value of *I3*, however, is correlated significantly to both the number of publications ($r = 0.619$, $p \leq 0.01$) and the total number of citations ($r = 0.963$, $p \leq 0.01$). These correlations are higher than the correlation between the numbers of citations and publications ($r = 0.555$, $p \leq 0.01$), which is largely a spurious correlation caused by size differences among the 65 journals.

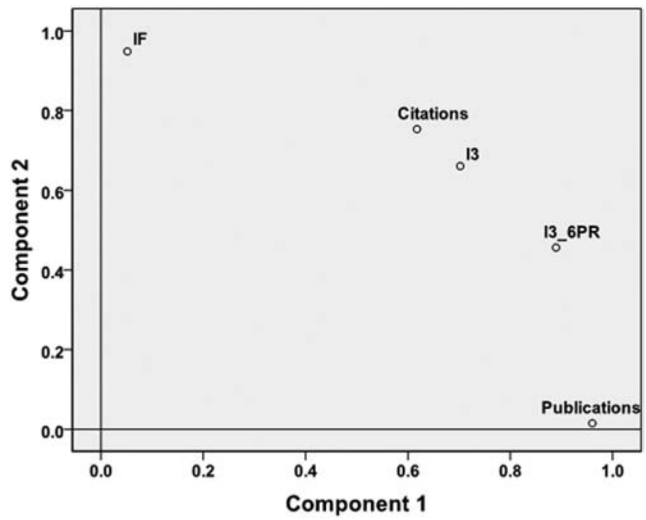


FIG. 5. Varimax rotated two-factor solution of the variables IF, *I3*, *I3*(*6PR*), number of publications, and citations.

The correlations with both the number of publications and citations could be expected because of the definition of *I3*. Like the h-index (Hirsch, 2005), *I3* takes both dimensions—the number of publications and their citation rates—into account in the definition of impact, but differently from the h-index, the tails of the distributions are not discarded as irrelevant. Ceteris paribus in terms of the top-segment (e.g., $h = 10$), the number of publications with less impact matters for the overall impact of two otherwise comparable documents sets.

Figure 5 shows the plot of the (varimax-rotated) two-factor solution of the variables that chiefly interest us here. As can be expected, the number of publications and the IF span orthogonal coordinates (Leydesdorff, 2009). The *I3* values are closest to total cites because the transformation is linear whereas a nonlinearity is involved in the case of *I3*(*6PR*). Unlike the total number of citations, however, the new indicator takes the shapes of the distributions into account by normalizing in terms of percentiles. The number of publications has an effect on *I3* independent of the latter's correlation with the total number of citations.

This can be shown as follows: the partial correlation between *I3* and the number of publications controlled for the number of citations ($r_{IF,N|TC}$) is 0.391 ($p = 0.01$) whereas $r_{I3,N|TC} = -0.373$ ($p \leq 0.05$). For the *I3*(*6PR*), this partial correlation $r_{I3(PR6),N|TC} = 0.985$ ($p \leq 0.01$), indicating that the binning into six percentile rank classes uncouples relatively from the citation rates and makes the publication rates therefore more important. The 100 percentile ranks provide a finer grained and therefore more precise indicator of citation impact than does the ranking in six classes.

The correlations in Table 3 were calculated at the journal level. Although the correlation between *I3* and the sum of total citations is very high, $r = 0.963$, $p \leq 0.01$, the underlying data also allow us to consider the correlation between the times cited and the percentiles at the level of the 5,737

TABLE 3. Rank order correlations (Spearman's $\rho$; upper triangle) and Pearson correlations $r$ (lower triangle) for 65 journals in library and information science.

| Indicator | IF 2009 | *I3*(*100PR*) | *M 100PR* | *I3*(*6PR*) | *M 6PR* | No. of publications | Total citations |
|---|---|---|---|---|---|---|---|
| IF 2009 | | 0.804** | 0.924** | 0.582** | 0.936** | 0.263* | 0.893** |
| *I3*(*100PR*) | 0.591** | | 0.843** | 0.875** | 0.862** | 0.670** | 0.974** |
| *M 100PR* | 0.839** | 0.651** | | 0.571** | 0.983** | 0.238 | 0.907** |
| *I3*(*6PR*) | 0.479** | 0.924** | 0.417** | | 0.608** | 0.907** | 0.817** |
| *M 6PR* | 0.893** | 0.648** | 0.950** | 0.506** | | 0.271* | 0.931** |
| No. of publications | 0.151 | 0.619** | 0.042 | 0.861** | 0.134 | | 0.562** |
| Total citations | 0.685** | 0.963** | 0.631** | 0.894** | 0.713** | 0.551** | |

*$p = 0.05$ (two-tailed). **$p = 0.01$ (two-tailed).

TABLE 4. Rankings between 15 journals of LIS with highest values on *I3* (expressed as percentages of the sum) compared with IFs, total citations, and %*I3*(*6PR*).

| Journal | No. of papers (a) | %*I3* (b) | IF 2009 (c) | Total citations (d) | %*I3*(*6PR*) (e) |
|---|---|---|---|---|---|
| *Journal of the American Society for Information Science and Technology* | 375 | 9.73 [1]$^{++}$ | 2.300 [7] | 1,975 [1] | 8.63 [1]$^{++}$ |
| *Scientometrics* | 258 | 7.24 [2]$^{++}$ | 2.167 [10] | 1,336 [3] | 6.37 [2]$^{++}$ |
| *Journal of the American Medical Informatics Association* | 199 | 6.80 [3]$^{++}$ | 3.974 [2] | 1,784 [2] | 6.15 [3]$^{++}$ |
| *Information Processing & Management* | 221 | 6.14 [4]$^{++}$ | 1.783 [15] | 921 [4] | 4.90 [4]$^{++}$ |
| *Information Management* | 117 | 4.01 [5]$^{++}$ | 2.282 [8] | 822 [6] | 3.35 [5]$^{++}$ |
| *International Journal of Geographical Information Science* | 120 | 3.14 [6]$^{++}$ | 1.533 [17] | 446 [9] | 2.55 [6]$^{++}$ |
| *MIS Quarterly* | 66 | 2.61 [7]$^{++}$ | 4.485 [1] | 847 [5] | 2.34 [7]$^{++}$ |
| *Journal of Management Information Systems* | 82 | 2.60 [8]$^{++}$ | 2.098 [11] | 496 [8] | 2.31 [8]$^{++}$ |
| *Journal of Health Communication* | 90 | 2.52 [9]$^{++}$ | 1.344 [22] | 380 [10] | 2.04 [10a]$^{++}$ |
| *Journal of Academic Librarianship* | 127 | 2.51 [10]$^{++}$ | 1.000 [26] | 252 [19] | 2.05 [9] |
| *Journal of Information Science* | 102 | 2.43 [11]$^{++}$ | 1.706 [16] | 355 [13] | 1.98 [11] |
| *Journal of Computer-Mediated Communication* | 108 | 2.37 [12]$^{++}$ | 3.639 [3] | 374 [11] | 2.04 [10b]$^{++}$ |
| *Journal of Informetrics* | 66 | 2.28 [13]$^{++}$ | 3.379 [4] | 598 [7] | 2.04 [10c] |
| *Journal of the Medical Library Association* | 114 | 2.21 [14]$^{++}$ | 0.889 [31] | 248 [20] | 1.93 [12] |
| *Telecommunications Policy* | 96 | 2.15 [15]$^{++}$ | 0.969 [27] | 264 [17] | 1.80 [13] |

*Note.* Ranks are added between brackets.
$^{++}p \leq 0.01$; above the expectation.

documents. The Pearson correlation in this case is only 0.639 ($p \leq 0.01$).[5]

In summary, *I3* provides us with an indicator which takes both the number of publications and their citations into account. The normalization to percentile ranks appreciates the shape of the distribution; the transformation of the citation curve is linear. No parametric assumptions (e.g., averages and standard deviations) are made. The definitions are sufficiently abstract so that impact is no longer defined in terms of a fixed citation window; any document set can be so evaluated. Different from the h-index, the full citation curve is weighted into these nonparametric statistics.

Table 4 provides the rankings for the 15 journals with the highest values for *I3* in comparison to rankings of the IFs 2009, the total citations, and *I3*(*6PR*) as an alternative classification scheme. One can see that on all measures except the IF, *JASIST* is ranked in first place. *MIS Quarterly* holds the seventh position in terms of both *I3* and *I3*(*6PR*).

The highly skewed citation distribution (in Column *d*) cannot prevent the *Journal of the American Medical Informatics Association* with 1,784 citations and an IF of 3.974 from ranking below *Scientometrics* with only 1,336 citations and the lower IF of 2.167, but nevertheless occupying the second position behind *JASIST*. Below the top segment, the six classes become less fine-grained than are the 100 percentiles. This is visible in Table 4, as the *Journal of Health Communication*, the *Journal of Computer-Mediated Communication*, and the *Journal of Informetrics* are tied for the 10th position (within this set of 65 journals). In the case of the *Journal of Informetrics*, however, the *I3*(*6PR*) value is no longer significantly different from the expectation.

We have argued that one needs a statistic for testing the differences among citation distributions for their relative significance beyond testing impacts as integrated values against expected impacts. Using the Kruskall–Wallis rank variance test, the null hypothesis that citation distributions are the same

[5]As could be expected, the correlation between times cited and the binned values of *I3*(*6PR*) is much higher ($r = 0.815$, $\rho = 0.911$) because the binning reduces the variance.


DOI: 10.1002/asi

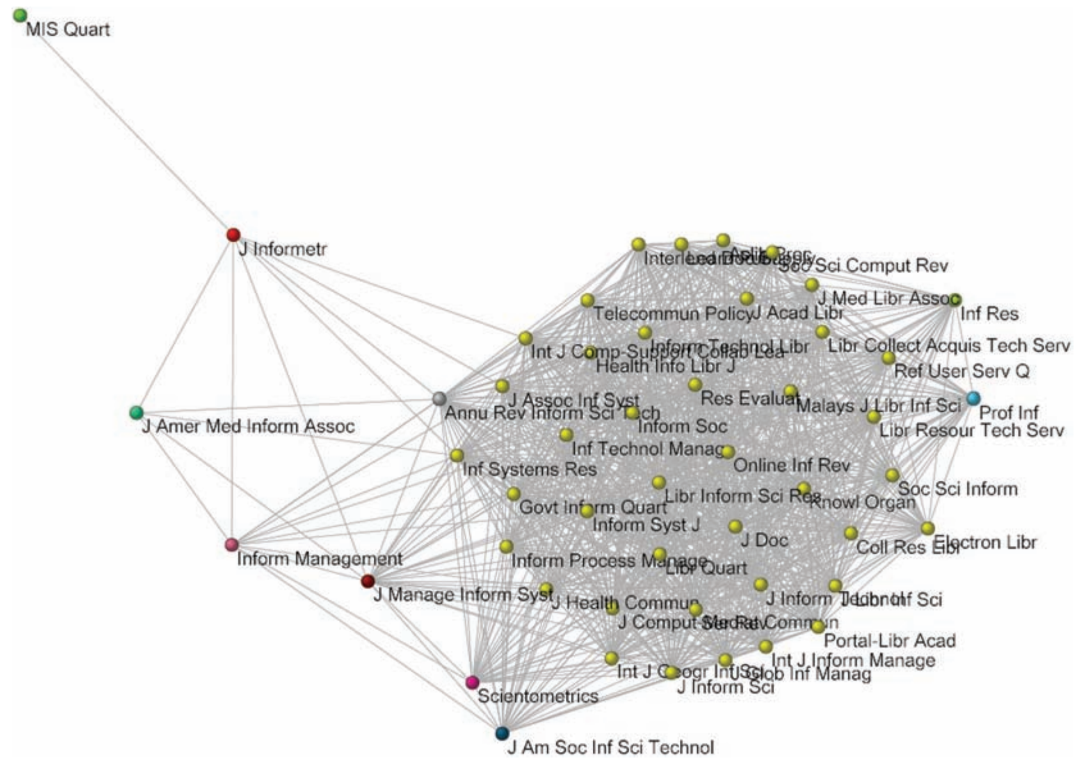


FIG. 6. Fifty journals of LIS organized according to (dis)similarity in their being-cited patterns to 5,125 publications in 2007 and 2008. Note that Dunn's test for multiple comparisons ($\alpha \leq 0.000041 = \{0.05/[(50 \times 49)/2]\}$; Kamada and Kawai (1989) used for the visualization. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

across these 65 journals was rejected at the 1% level. Given this result, we may further test between any two journals whether their citation distributions are significantly different. As noted, we used Dunn's test for the comparison among the citation distributions of the 50 journals with the highest IFs (among the 65 in the LIS category); the results are summarized in Figure 6.

Figure 6 shows that *MIS Quarterly* is significantly different in terms of its citation distribution from all other journals in this group, except for the *Journal of Informetrics* (This exceptional distribution leads to, among other things, the high IF of this journal.) Using measures for interdisciplinarity, Leydesdorff and Rafols (2011) showed that these two journals can be considered as relatively monodisciplinary specialist journals within this set. *JASIST* and *Scientometrics*—both high on interdisciplinarity relative to this (!) set—are positioned at another corner of the figure (at the bottom) as significantly different from a number of journals in a major group of 37 journals that form a $k = 25$ core set. The *Journal of Computer-Mediated Communication*, for example, is part of this core set with an IF 2009 of 3.639 while at the lower end, *Interlending & Document Supply* has an IF 2009 of 0.403. Differences in IFs of an order of magnitude do not inform us about the significance of differences in citation distributions nor in terms of citation impact unless, of course, one defines "citation impact" in these terms (e.g., Garfield, 1972). Note that *I3* is an indicator of the impact of document sets in terms of citations, and thus the semantics are somewhat different.

### Multidisciplinary Sciences

The ISI Subject Category MS contains a heterogeneous set of 48 journals, ranging from *Science* and *Nature* with IFs of 29.747 and 34.480, respectively, to *R&D Magazine* with an IF = 0.004 in 2009. However, 65.2% of all citable publications in this set during 2007 and 2008 ($n = 24{,}494$) were published in six major journals: *PNAS* (7,058; 28.8%), *Nature* (2,285; 9.3%), *Science* (2,253; 9.2%), *Annals of the New York Academy of Science* (1,996; 8.2%), *Current Science* (1,271; 5.2%), and the *Chinese Science Bulletin* (1,115; 4.6%).

Among these journals, *Science* and *Nature* seem to have a very similar profile (Figure 7). For example, the number of not-cited papers in this set is 279 for *Science* and 282 for *Nature*. In both cases, this is more than 10% of all citable items in the journal. The largest among these journals, *PNAS*, however, has a very different profile: Only 58 (≤1%) of its 7,085 citable publications had never been cited by the date of the download (February 20, 2011).

In Figure 7, the numbers of citations are log-scaled to make the differences in these skewed distributions more visible. The citation curve for *Nature* remains consistently above the one for *Science*, but the one for *PNAS* is very differently shaped. This journal has in total 27,419 more citations

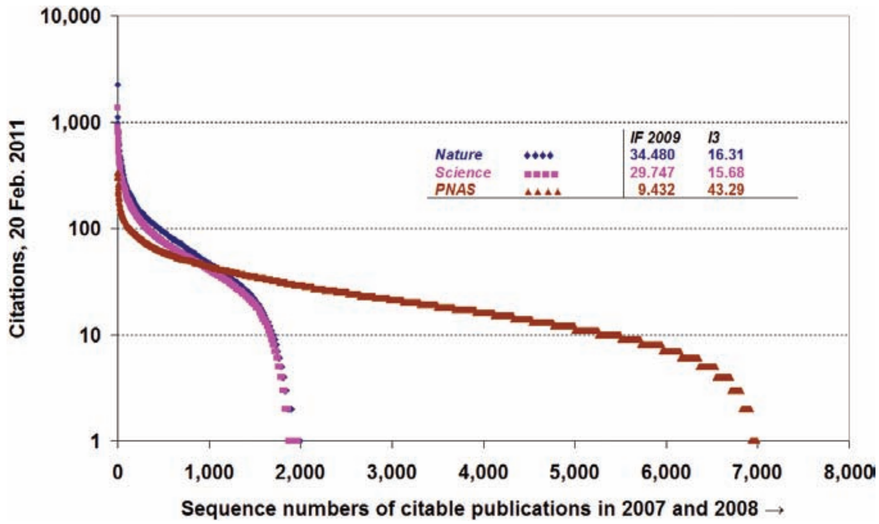


FIG. 7. Log-scaled citation distributions for the citable publications in 2007 and 2008 in *Nature*, *Science*, and *PNAS*; downloaded at the Web of Science on February 20, 2011. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 5. Fifteen MS journals with highest values on *I3* compared in ranking with IFs, total citations, and *I3*(*6PR*).

| Journal | No. of papers (a) | %*I3* (b) | IF 2009 (c) | Total citations (d) | %*I3*(*6PR*) (e) |
|---|---|---|---|---|---|
| *Proceedings of the National Academy of Sciences, USA* | 7,058 | 43.29$^{++}$ | 9.432 [3] | 178,137 | 33.64 [1]$^{++}$ |
| *Nature* | 2,285 | 16.31$^{++}$ | 34.480 [1] | 150,718 | 16.46 [2]$^{++}$ |
| *Science* | 2,253 | 15.68$^{++}$ | 29.747 [2] | 126,230 | 15.27 [3]$^{++}$ |
| *Annals of the New York Academy of Sciences* | 1,996 | 9.33$^{++}$ | 2.670 [5] | 14,284 | 8.29 [4]++ |
| *Current Science* | 1,271 | 2.33$^{--}$ | 0.782 [22] | 1,551 | 3.40 [5]$^{--}$ |
| *Chinese Science Bulletin* | 1,115 | 2.11$^{--}$ | 0.898 [20] | 2,239 | 2.55 [6]$^{--}$ |
| *Philosophical Transactions of the Royal Society A* | 451 | 1.78$^{--}$ | 2.295 [9] | 3,065 | 1.46 [7]$^{--}$ |
| *Journal of the Royal Society Interface* | 257 | 1.18$^{++}$ | 4.241 [4] | 2,772 | 0.83 [13]$^{--}$ |
| *International Journal of Bifurcation and Chaos* | 553 | 1.09$^{--}$ | 0.918 [17] | 1,204 | 1.34 [8]$^{--}$ |
| *Naturwissenschaften* | 288 | 1.01$^{--}$ | 2.316 [8] | 1,697 | 0.75 [14]$^{--}$ |
| *TheScientificWorld JOURNAL* | 358 | 0.82$^{--}$ | 1.658 [10] | 1,183 | 0.91 [11]$^{--}$ |
| *Progress in Natural Science* | 463 | 0.67$^{--}$ | 0.704 [24] | 710 | 1.07 [9]$^{--}$ |
| *Scientific American* | 370 | 0.37$^{--}$ | 2.471 [7] | 585 | 0.87 [12]$^{--}$ |
| *Comptes rendus de l' académie bulgare des Sciences* | 427 | 0.33$^{--}$ | 0.204 [37] | 234 | 0.95 [10]$^{--}$ |
| *South African Journal of Science* | 191 | 0.33$^{--}$ | 0.506 [28] | 220 | 0.47 [15]$^{--}$ |

$^{++}p \leq 0.01$ above the expectation. $^{--}p \leq 0.01$ below the expectation.

than does *Nature* whereas the latter has 24,488 more citations than does *Science* (as of February 20, 2011), yet the IF of *PNAS* is less than one third (IF 2009 = 9.432). The large tail in the distribution of moderately cited papers works as discussed earlier to the disadvantage of the larger journal. Note that such a tail would similarly disadvantage a highly productive research team or university.

Table 5 provides the data for the 15 journals with the highest value of *I3* among the 48 journals of this subject category in a format similar to that of Table 4 (for LIS), and Table 6 provides the correlation coefficients (as in Table 3). Although the two measures of *I3* and IF significantly correlate again over the set (Table 6), the order of the journals as indicated in Table 5 is very different.

For example, *Current Science* and the *Chinese Science Bulletin* were ranked at the 22nd and 20th places, respectively, in this set, with IFs of 0.782 and 0.898, respectively, but are now rated as the fifth- and sixth-largest impact journals, respectively. The scoring in terms of *I3*(*6PR*) in the rightmost column of Table 5 shows that this is not only an effect of the large tails in the distribution with infrequently cited papers but is consistent when using this evaluation scheme of the six classes which rewards excellence (top-1%, etc.) disproportionally.

Table 6 shows that in this case, the differences in size among these 48 journals are so important that all correlations are much higher. In other words, these correlations are spurious since they are caused by differences in size more than in the previous case since there are six major journals and 42 small ones. Despite the higher correlations, the same effects as discussed previously for the case of LIS can be found. The partial correlation between *I3* and the number of publications controlled for the number of citation $r_{I3,N|TC} = 0.850$ ($p \leq 0.01$) whereas $r_{I3(6PR),N|TC} = 0.982$ ($p \leq 0.01$) is again


DOI: 10.1002/asi

TABLE 6. Rank order correlations (Spearman's $\rho$; upper triangle) and Pearson correlations $r$ (lower triangle) for the 48 journals of MS.

| Indicator | IF 2009 | *I3* | *M 100PR* | *I3 (6PR)* | *M 6PR* | No. of publications | Total citations |
|---|---|---|---|---|---|---|---|
| IF 2009 | | 0.798** | 0.884** | 0.517** | 0.844** | 0.479** | 0.840** |
| *I3* | 0.590** | | 0.777** | 0.854** | 0.837** | 0.829** | 0.986** |
| *M 100PR* | 0.775** | 0.691** | | 0.408** | 0.817** | 0.364* | 0.838** |
| *I3*(*6PR*) | 0.660** | 0.987** | 0.706** | | 0.646** | 0.996** | 0.801** |
| *M 6PR* | 0.956** | 0.716** | 0.875** | 0.775** | | 0.605** | 0.887** |
| No. of publications | 0.492** | 0.953** | 0.605** | 0.967** | 0.635** | | 0.772** |
| Total citations | 0.841** | 0.922** | 0.756** | 0.945** | 0.884** | 0.839** | |

$^*p = 0.05$ (two-tailed). $^{**}p = 0.01$ (two-tailed).

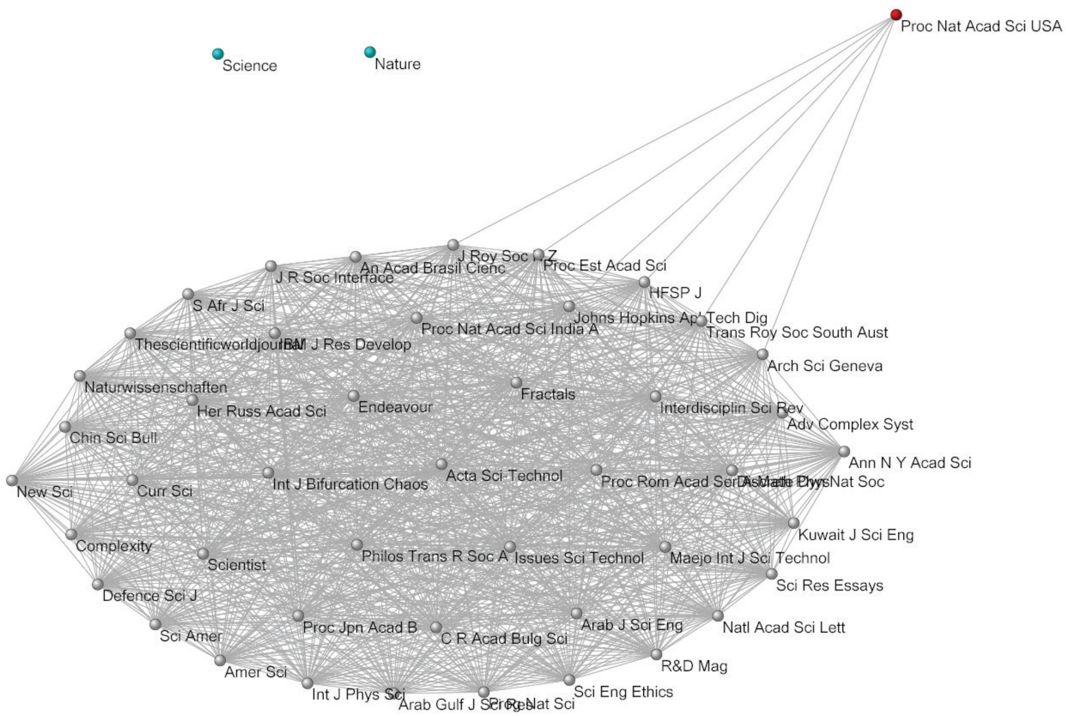


FIG. 8. Forty-eight journals in Multidisciplinary Sciences organized according to (dis)similarity in their being-cited patterns to 24,494 publications in 2007 and 2008. Note. Dunn's test for multiple comparisons, $\alpha \leq 0.000044 = \{0.05/[(48 \times 47)/2]\}$; Kamada and Kawai (1989) used for the visualization. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

higher. As in the previous case, $r_{IF,N|TC}$ is negative (−0.724; $p \leq 0.01$) because the IF is based on dividing by the number of publications.

Both *I3* and *I3*(*6PR*) significantly correlate again with the number of citable publications and citations. This follows from the definition of impact by analogy to the product of "mass" (number of publications) and "velocity" (quality of each publication). Both terms thus contribute to the impact, but with qualification of the citedness of each publication in terms of its percentile rank.

Dunn's test applied to the citation patterns of these 48 journals is visualized in Figure 8. The figure shows that 45 of the 48 journals form a $k = 25$ core set of journals. Both *Science* and *Nature* are significantly different in their citation patterns from all other journals in this set, and—perhaps counterintuitively—from each other. The citation distribution of *PNAS*, however, is not significantly different from a number of other journals in the set.

*Current Science* (Curr Sci) and the *Chinese Science Bulletin* (Chin Sci Bull), however, are positioned to the left within the core set, in the neighborhood of the *New Scientist* (New Sci). This latter journal has an IF of only 0.333 and a contribution to the impact in terms of *I3* of only 0.17% of the total impact of the set (1,035,332.14).

In summary, these results show that on the one hand, two journals with a similarly high IF, such as *Science* and *Nature*, nevertheless can differ significantly in their citation distributions. Note that Dunn's test is performed directly on the raw citation scores; that is, before the transformation into percentile ranks. On the other hand, a journal with a very high impact such as *PNAS* may not differ much in its citation pattern from a journal such as the *Proceedings of the Estonian*

TABLE 7. Percentage share of publications (2007 and 2008) and percentage *I3* in the set of 65 journals of LIS (sorted by the ratio of percentage of *I3*/percentage share of publications in Column c).

| | Percentage publications (a) | %*I3* (b) | Ratio of (b) and (a) (c) | %*I3*(*6PR*) (d) | Ratio of (d) and (a) (f) |
|---|---|---|---|---|---|
| The Netherlands | 2.23 | 3.75 | $1.68^{++}$ | 3.23 | $1.45^{++}$ |
| Switzerland | 0.77 | 1.24 | $1.61^{++}$ | 1.21 | $1.57^{++}$ |
| Belgium | 1.15 | 1.81 | $1.57^{++}$ | 1.56 | $1.36^{++}$ |
| South Korea | 1.45 | 1.88 | $1.30^{++}$ | 1.58 | $1.09^{+}$ |
| Taiwan | 2.23 | 2.87 | $1.29^{++}$ | 2.52 | $1.13^{++}$ |
| Peoples Republic of China | 2.35 | 2.97 | $1.26^{++}$ | 2.63 | $1.12^{++}$ |
| Italy | 0.83 | 1.02 | $1.23^{++}$ | 0.86 | $1.04^{++}$ |
| EU-27 | 24.67 | 30.21 | $1.22^{++}$ | 28.20 | $1.14^{++}$ |
| Canada | 3.46 | 4.08 | $1.18^{++}$ | 3.73 | $1.08^{++}$ |
| Australia | 2.19 | 2.58 | $1.18^{++}$ | 2.37 | $1.08^{+}$ |
| Singapore | 1.16 | 1.34 | $1.16^{++}$ | 1.17 | $1.01^{++}$ |
| United Kingdom | 8.82 | 10.18 | $1.15^{++}$ | 9.49 | $1.08^{++}$ |
| Sweden | 0.96 | 1.06 | 1.10 | 1.04 | $1.08^{++}$ |
| USA | 40.36 | 42.91 | $1.06^{++}$ | 41.52 | $1.03^{++}$ |
| France | 1.10 | 1.16 | $1.05^{++}$ | 1.13 | $1.03^{++}$ |
| Finland | 1.09 | 1.10 | 1.01 | 1.05 | $0.96^{++}$ |
| Spain | 4.31 | 4.22 | $0.98^{-}$ | 4.07 | $0.94^{++}$ |
| Germany | 2.51 | 2.38 | $0.95^{--}$ | 2.58 | $1.03^{++}$ |
| %accounted | 88.72 | 97.58 | $1.10^{++}$ | 92.91 | $1.05^{++}$ |

$^{++}p < 0.01$ above the expectation. $^{--}p < 0.01$ below the expectation. $^{+}p < 0.05$ above the expectation. $^{-}p < 0.05$ below the expectation.

*Academy of Science,* although both their respective *I3*s and IFs differ by orders of magnitude.

Recall that the IF was designed precisely with the objective to correct for these size differences between otherwise similar journals such as *PNAS* and the *Proceedings of the Estonian Academy of Science* (Bensman, 2007; Garfield, 1972). It completely fails to do so because of the parametric assumption involved in using an arithmetic average (cf. Rousseau & Leydesdorff, 2011).

## Performance Measurement

Because percentiles are attributed at the paper level, the datasets enable us to perform aggregations and decompositions other than in terms of journals or journal sets. Documents and document sets can be analyzed both in terms of disciplinary structures and also be considered as products of authors, institutions, nations, and so on (Narin, 1976; Small & Garfield, 1985). On one hand, one refers to what a journal accepts as worthy of publication whereas on the other hand, one refers to how a person, for example, performs and communicates scientific work. Whereas journals do not produce scientific knowledge the way people/institutions do and, of course, are evaluated, *I3* is so general that one is able to combine the two different types of evaluations (Leydesdorff, 2008). One can, for example, compare the productivity and impact of an institution or nation in two different journal categories (and test the difference for its significance).

As an example, let us recompose the 5,737 citable documents of the LIS set using the country addresses provided in the bylines of these papers. Fractional counting of the addresses will be used to keep consistent the total numbers. In other words, if a paper is coauthored between two authors from Country A and one from Country B, the attribution is one third to Country B and two thirds to Country A.

Table 7 provides the results; the table is composed by first using only players in the field with at least a 1% contribution to *I3* and then by sorting them in Column c using the ratio of this share divided by the percentage share of publications as the expected distribution (Column *a*) (The regression line is, in this case, less interesting since it is overdetermined by the outliers for the USA and EU-27.)

Only 5,090 (88.72%) of the 5,737 records contained ($n = 8{,}510$) addresses with country names. These records are cited more often than are records without addresses, so that the "world average" given the set of 65 LIS journals would be 1.10 using *I3*, or 1.05 in the case of *I3*(*6PR*). Because this indicator is based on a summation, the value for the European Union (EU-27) is equal to the sum value of the 27 nations composing the European Union (Similarly, the value for the United Kingdom is constructed by adding records with England, Scotland, Wales, and Northern Ireland as country indicators in the ISI set.)

Using *I3*, The Netherlands scores highest with 1.68, but using *I3*(*6PR*), Switzerland, which was second on the scale of 100, is now highest with 1.57. As noted, *I3*(*6PR*) is sensitive to an above-average representation in the top segments of the percentile distribution. Switzerland is known to be well represented in these segments (e.g., King, 2004). The USA outperforms all other nations (including a constructed EU-27) in terms of absolute numbers on both scales. The low contribution of the People's Republic of China in this set is notable. Important countries such as India, Russia, and Brazil are not listed as contributing because of the threshold used of more than a 1% contribution to the total impact.


DOI: 10.1002/asi

In summary, percentile ranks are defined at the level of each individual paper in a set. How the set is composed (e.g., in terms of publications in 2 years in this study, but for the purpose of a comparison with the IFs) still can be decided on the basis of a research question. For example, one may wish to compare the impact of publications of rejected versus granted principal investigators (PIs) in a competition (Bornmann, Leydesdorff, & Van den Besselaar, 2010; Van den Besselaar & Leydesdorff, 2009). In each such study, one can determine percentiles, test citation curves against one another for the statistical significance of differences (using Dunn's test), and test for each subset whether the impact is significantly above of below the expectations (using the *Z* test). Our method is thus most general and avoids parametric assumptions.

## Discussion and Conclusions

We elaborated on the *I3* values using the 2-year set of citable items to facilitate—in this study—the comparison with the IF. However, as shown earlier, this indicator is not restricted to journals, document sets, time periods, and so on but is more general: Only the specification of a reference set is required from which the samples under study are drawn (Bornmann, Mutz, Neuhaus, & Daniel, 2008). Earlier, we used two ISI Subject Categories as reference sets, but one also could use the entire *Science Citation Index*, Scopus data, data from Google Scholar, or patent databases that contain citations. One even can apply this to a citation count in "grey literature." *I3* provides a general measure of citation that can be applied across samples of different sizes; the nonparametric statistics account for the typically highly skewed citation distributions.

Our first point was that impact is not captured correctly using central-tendency statistics such as the mean or the median. Lundberg (2007, p. 148) noted that one does not have to average the (field-normalized) citation scores but also can use their sum values as a "total" field-normalized citation score. Using the Leiden Rankings, the Center for Science and Technology Studies (CWTS) multiplied the product of the number of publications *P* with the "old" crown indicator *CPP*/*FCSm* to obtain as a result what was called the "brute force indicator." In the new set of Leiden indicators, analogously, a "total normalized citation score" was proposed (Van Raan Eck, van Leeuwen, Visser, & Waltman, 2010, p. 291). However, all these indicators are based on the parametric assumption (of the central limit theorem) that one is allowed to compute with the mean as a summary statistic given a sufficiently large number of observations (e.g., Glänzel, 2010).

As with the IFs, citation analysis has hitherto been caught in the paradigm of parametric statistics, although this approach is mostly not fruitful for bibliometrics (cf. Ahlgren, Jarneving, & Rousseau, 2003). Changing to the median, however, is not sufficient because the median as a central-tendency measure is as sensitive—and sometimes even more so—to the tails of the distributions. A finer grained scheme of 100 percentiles can be envisaged. Actually, we used the percentile ranks presented earlier as a continuous random variable which can be specified to any desirable degree of precision in terms of decimal numbers. Thus defined, the percentile ranks are attributes of the publications which can be added to perform integration along the qualified citation curve.

In addition, we showed that one can vary the evaluation scheme using the six percentile ranks that are used in the *Science and Engineering Indicators* (National Science Board, 2010, Appendix Table 5–43; Bornmann & Mutz, 2011). The emphasis on the more highly cited publications in this scheme enhances the distinctions as more significant (e.g., in Tables 5 and 7), but one may lose some information such as fine-grained distinctions between units of analysis with tied ranks. *I3* for 100 percentiles provides the general scheme from which others can be derived given different policy contexts. As noted, 100 percentiles can be considered as a continuous variable, and one thus can provide the degree of precision in decimals.

In the meantime, the percentile rank approach also is used by the new *InCites* database of Thomson Reuters that functions as an overlay to the WoS. Unfortunately, the percentile ranks are averaged in this case, and one cannot escape from the scheme of ISI Subject Categories as the reference sets for determining the percentiles (cf. Pudovkin & Garfield, 2002). Using percentile ranks, however, the classification into categories in the future also can be paper-based, such as using the Medical Subject Heading (MeSH) in the MEDLINE database of the National Institutes of Health (Bornmann et al., 2008) or using the keywords of dedicated databases such as *Chemical Abstracts* (Bornmann, Schier, Marx, & Daniel, 2011). We expect the state of the art to change rapidly in this respect.

Our suggestion to use summations for the impact may raise the question of whether impact per paper was not defined earlier as rate of summations rather than as a summation of rates. Last year's debate about normalization was about using "rates of averages" versus "averages of rates," as Gingras and Larivière (2011) succinctly summarized the crucial issue of the controversy. However, percentile ranks are rates, albeit nonparametric ones. As shown earlier (e.g., in Figure 4), the resulting sums for different units of analysis can be regressed upon the number of publications, and thus the impact/paper can be indicated. This impact/paper can be tested for its significance against the distribution of papers under study (using $\chi^2$ statistics). The differences in underlying citation distributions can be tested for their significance using, for example, Dunn's test. Using percentiles, the evaluation scheme for both the performance of authors and institutes and the quality of journals can be brought methodologically into a single framework (Leydesdorff et al., 2011). Since the *I3* measure is fully decomposable, multidimensional distinctions also are possible.

In terms of the statistics, our main message is to keep significance in differences among citation distributions analytically separate from impact, which we defined—in analogy to the (vector-)summation of momenta in physics—as summations of products. Thus defined, the percentile rank approach


DOI: 10.1002/asi

of the *I3* enables us to take both the size and the shape of the distribution into account, and impacts among different units of analysis (e.g., journals, nations, universities, institutions, individuals) can be compared (i.e., added and subtracted) as percentages. Whether these units of analysis (e.g., individuals, research groups, small countries such as Monaco) are large enough for the comparison can be decided on statistical (instead of normative or moral) grounds. Differences which are not significant can be discouraged for usage in policymaking and research management.

## Policy Implications

The common assumption in citation impact analysis hitherto has been normalization to the mean. In our opinion, the results then are necessarily flawed because the citation distributions are often highly skewed. Highly productive units then can be disadvantaged because they publish often in addition to higher cited papers also a number of less-cited ones which depress their average performance. We became aware of this when we tried to reproduce the performance of seven PIs of the Academic Medical Center of the University of Amsterdam who had been evaluated by the CWTS. The first and sixth positions among these seven were swapped when we used mean percentile ranks (Leydesdorff et al., 2011). Thus, the effect of the proposed change in the paradigm of impact assessment can be highly significant, in terms of both the statistics and policy implications. In this case, for example, the ranking in terms of impact was used as input to the funding scheme of these PIs; the research group of the sixth PI thus suffered a loss in funding because of this group's productivity other than in the top-1% (T. Opthof, personal communication, 26 October 2010).

In this study, we generalized the nonparametric approach. Long-standing assumptions such as *Nature* and *Science* that would have higher citation impact (given higher IFs) when compared with *PNAS* were shown earlier as erroneous consequences of parametric assumptions. *Current Science* and the *Chinese Science Bulletin* were ranked at the 22nd and 20th place, respectively, in the set of multidisciplinary journals, with IFs of 0.782 and 0.898, respectively, but were rated as the fifth- and sixth-largest impact journals, respectively, using *I3*. Thus, this analysis in terms of percentiles shows the increased importance of these two multidisiciplinary journals while the IFs do not.

Narin's (1976) original scheme of cross-tabling journals and nations in evaluative bibliometrics can be considered as two dimensions for the aggregation into subsets of papers. *I3* is additive and therefore allows for the comparison of subsets which may differ along both axes. Actually, one also can compute a contribution of *I3* for a set that differs in terms of both addresses and journals (or other parameters), given that the reference sets for the percentile ranks are properly set as the relevant total sum sets. Whereas the h-index and its derived statistics seem to allow for such comparisons, *I3* does not simplify the computation by discarding the bulk of the references in the tails of the distributions. Note that papers with zero citations do not contributed to *I3* because *I3* is an impact indicator that takes both publication quantities and (normalized) citation impact into account. Using *I3* for the evaluation, one is no longer "punished" for one's productivity.

## Acknowledgments

We are grateful to Rüdiger Mutz for comments on previous drafts.

## References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirement for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. Journal of the American Society for Information Science and Technology, 54(6), 550–560.

Bensman, S.J. (2007). Garfield and the impact factor. Annual Review of Information Science and Technology, 41(1), 93–155.

Bensman, S.J., & Wilder, S.J. (1998). Scientific and technical serials holdings optimization in an inefficient market: A LSU Serials Redesign Project Eexercise. Library Resources and Technical Services, 42(3), 147–242.

Bornmann, L. (2010). Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. Journal of Informetrics, 4(3), 441–443.

Bornmann, L., & 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), 228–230.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics in Science and Environmental Politics, 8(1), 93–102.

Bornmann, L., Leydesdorff, L., & Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. Journal of Informetrics, 4(3), 211–220.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2011). Is interactive open access publishing able to identify high-impact submissions? A study on the predictive validity of atmospheric chemistry and physics by using percentile rank classes. Journal of the American Society for Information Science and Technology, 62(1), 61–71.

Brumback, R. A. (2008a). Response to correspondence, "Worshiping false idols: The impact factor dilemma": Correcting the record. Journal of Child Neurology, 23(9), 1092.

Brumback, R. A. (2008b). Worshiping false idols: The impact factor dilemma. Journal of Child Neurology, 23(4), 365.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178(4060), 471–479.

Garfield, E., & Sher, I.H. (1963). New factors in the evaluation of scientific literature through citation indexing. American Documentation, 14, 195–201.

Gingras, Y., & Larivière, V. (2011). There are neither "king" nor "crown" in scientometrics: Comments on a supposed "alternative" method of normalization. Journal of Informetrics, 5(1), 226–227.

Glänzel, W. (2010). On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper. Journal of Informetrics, 4(3), 313–319.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, USA, 102(46), 16569–16572.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), 7–15.

King, D.A. (2004, July 15). The scientific impact of nations. Nature, 430, 311–316.

Levine, G. (1991). A guide to SPSS for analysis of variance. Hillsdale, NJ: Erlbaum.


DOI: 10.1002/asi

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluation. Journal of the American Society for Information Science and Technology, 59(2), 278–287.

Leydesdorff, L. (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), 1327–1336.

Leydesdorff, L., Bornmann, L., Mutz, R., & 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), 1370–1381.

Leydesdorff, L., & Opthof, T. (2011). Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS. Journal of Informetrics, 5(1), 224–225.

Leydesdorff, L., & Rafols, I., (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. Journal of Informetrics, 5(1), 87–100.

Lundberg, J. (2007). Lifting the crown—Citation z-score. Journal of Informetrics, 1(2), 145–154.

Moed, H.F., & van Leeuwen, T.N. (1996). Impact factors can mislead. Nature, 381(6579), 186.

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. Washington, DC: National Science Foundation.

National Science Board. (2010). Science and Engineering Indicators. Washington, DC: National Science Foundation. Available at: http://www.nsf.gov/statistics/seind10/

Nisonger, T.E. (1999). JASIS and library and information science journal rankings: A review and analysis of the last half century. Journal of the American Society for Information Science, 50(11), 1004–1019.

Nisonger, T.E., & Davis, C.H. (2005). The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl–Davis study. College and Research Libraries, 66(4), 341–377.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. Journal of Informetrics, 4(3), 423–430.

Pringle, J. (2008). "Worshiping false idols: The impact factor dilemma": Correcting the record. Journal of Child Neurology, 23(9), 1092.

Pudovkin, A.I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology, 53(13), 1113–1119.

Pudovkin, A.I., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. CollNet Journal of Scientometrics and Information Management, 3(2), 3–10.

Rafols, I., & 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), 1823–1835.

Rossner, M., Epps, H.V., & Hills, E. (2007). Show me the data. Journal of Cell Biology, 179(6), 1091–1092.

Rossner, M., Van Epps, H., & Hill, E. (2008). Irreproducible results—A response to Thomson Scientific. The Journal of General Physiology, 131(2), 183–184.

Rousseau, R., & Leydesdorff, L. (2011). Simple arithmetic versus intuitive understanding: The case of the impact factor. ISSI Newsletter, 7(1), 10–14.

Seglen, P. O. (1992). The skewness of science. Journal of the American Society for Information Science, 43(9), 628–638.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. Journal of Information Science, 11(4), 147–159.

Van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review, and project selection: A case study in the social and behavioral sciences. Research Evaluation, 18(4), 273–288.

van Raan, A.F.J., van Leeuwen, T.N., Visser, M.S., van Eck, N.J., & Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. Journal of Informetrics, 4(3), 431–435.

van Raan, A.F.J., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & Waltman, L. (2010, September). The new set of bibliometric indicators of CWTS. Paper presented at the 11th International Conference on Science and Technology Indicators (pp. 291–293), Leiden, The Netherlands.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. Journal of the American Society for Information Science and Technology, 59(13), 2070–2086.