Principles for comparing sets of documents in citation analysis: From independent samples to comparing sub-samples in terms of percentile ranks

Lutz Bornmann,[1] Loet Leydesdorff,[2] Rüdiger Mutz,[3] and Tobias Opthof[4]

[1] *bornmann@gv.mpg.de*
Max Planck Society, Hofgartenstrasse 8, D-80539 Munich (Germany)

[2] *loet@leydesdorff.net*
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, NL-1012 CX Amsterdam (The Netherlands)

[3] *mutz@gess.ethz.ch*
ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Zähringerstrasse 24, CH-8092 Zurich (Switzerland)

[4] *t.opthof@inter.nl.net*
Experimental Cardiology Group, Center for Heart Failure Research, Academic Medical Center K2-105, 1105 AZ Amsterdam (The Netherlands); Department of Medical Physiology, University Medical Center Utrecht, 3508 CM Utrecht (The Netherlands)

Abstract

Using citation analysis, sets of documents can be compared as independent samples; for example, in terms of average citation counts using potentially different reference sets. From this perspective, the size of samples matters only for the statistical significance testing of differences and the error estimation. Using the percentile rank approach, differences among citation distributions can be studied in a single scheme. The comparison among the sets reveals that different sizes of the samples affect the weighing of the probabilities and therefore the rankings. We distinguish among (1) the normalization of papers against external reference sets, (2) the normalization in terms of frequencies relative to the margin-totals of independent versus dependent samples, and (3) the potentially normative definition of percentile rank classes for the evaluation (e.g., top-1% most highly cited; median, etc.).

Introduction

Last year (2010), a short controversy flourished about the normalization of citation indicators. The controversy (Bornmann, 2010; Leydesdorff & Opthof, 2010 and 2011; Van Raan *et al*., 2010a; Waltman *et al*., 2011) led CWTS (the Center for Science and Technology Studies) shortly thereafter to propose a new crown indicator "*MNCS*" (the Mean Normalized Citation Scores) and several derivatives of this indicator (such as *MNCS1* and *MNCS2*; Van Raan *et al*., 2010b; Waltman *et al*., in press). Using *MNCS*, the "rate of averages" (*CPP/FCSm*: that is, the average citation per publication divided by the mean citation rate in the corresponding fields) thus was changed into the "averaging of rates"—as Gingras & Larivière (2011) summarized the core issue of the debate. The "new crown indicator" does not suffer from the shortcomings of the old one (Waltman *et al*., 2011). In our opinion, however, several issues which can be raised with respect to the normalization of citation scores have not yet sufficiently been discussed. Closure of the debate by establishing a new "crown indicator" could from this perspective be premature.

First, the new crown indicator (as the old one) is based on using (arithmetic) averages of – as a rule – highly skewed citation distributions. Both Bornmann & Mutz (2011) and Leydesdorff & Opthof (2011) raised the issue that it might be better to use the median and non-parametric statistics instead. More specifically, Bornmann & Mutz (2011) proposed an elaboration into a scheme which they called the "percentile rank approach" and which is already in use as the evaluation scheme in the *Science & Engineering Indicators* of the National Science

Foundation of the USA (NSB, 2010), prepared biannually by the American corporation ipIQ. In this scheme the focus is not only on (relative) citation rates, but also on the top-cited papers (Bornmann *et al*., 2010a).

Bornmann *et al*. (2008) raised the issue of using journals or groups of journals (aggregated into so-called Subject Categories by the Institute of Scientific Information (ISI) of Thomson Reuters) as systems of reference for the normalization. Rafols & Leydesdorff (2009) argued that these ISI Subject Categories are provided for reasons other than bibliometric measurement and had been shown as faulty in more than 40% of individual cases (see here also Boyack *et al*., 2005; Garfield & Pudovkin, 2002, at p. 1113n.). The use of these categories for journal classification as in the field-normalization of many of the existing indicators (including *MNCS*) therefore would be unfortunate. The ECOOM center in Leuven (Belgium) developed its own classification scheme (Glänzel & Schubert, 2003), but Rafols & Leydesdorff (2009) showed that this classification of journals does not improve on the ISI Subject Categories; the latter are finer grained and therefore less error-prone than the newly proposed ones (Leydesdorff & Rafols, 2009).

In addition to classifications of journals grouping potentially heterogeneous sets, journals themselves can be heterogeneous in terms of document types, citation half-lives, cognitive substance, etc. (Leydesdorff, 2008; Moed, 2010). Bornmann *et al*. (2008) therefore proposed to use classification schemes at the level of individual papers such as the Medical Subject Headings (MeSH) of Medline, that is, the publicly available database of the National Institute of Health of the USA. Bornmann *et al.* (2011; in press), for example, applied the percentile-rank approach using the classifications of *Chemical Abstracts*. Leydesdorff & Opthof (2010) proposed to appreciate differences among individual papers by using fractional counting of the citations in terms of the number of references in the *citing* papers with the argument that differences in so-called "citation potentials" (Garfield, 1979, at p. 365) are generated on this side of the citation process. Similar proposals had been done by Moed (2010), Zitt (2010), and Zitt & Small (2008) for the normalization of journal impacts using citing-side normalizations (Leydesdorff & Bornmann, 2011).

In this study, we focus on *cited-side* normalizations and try to take the discussion about this subject one step further by raising in addition to the problem of normalization, the problem of evaluation (for example, in terms of 1% most highly cited papers), and the issue of how to appreciate differences in productivity (publication rates) using citation analysis. With the exception of the *h*-index—which is also defined in terms of numbers of publications that meet a specified criterion (Hirsch, 2005)—citation indicators hitherto have abstracted from publication behavior and differences in productivity rates. However, in evaluation practices one often is confronted with questions such as how to weigh one paper in the top-1% (of the most cited papers) range against five or ten papers in the top-5% range, etc. The current schemes do not allow for quantitative assessments of such comparisons.

In summary, we distinguish first a number of analytical questions and then elaborate on the percentile rank approach for developing a set of criteria that can be met with this new indicator for citation analysis. Let us list these criteria:

1. A citation-based indicator were to be defined so that the choice of the reference set(s) can be varied by the analyst independently of the question of the evaluation scheme. In other words, these two dimensions of the problem have to be disentangled;
2. The citation indicator should leave room for different evaluation schemes, for example, by funding agencies. Some agencies may be interested in the top-1% (e.g., the National Science Board; NSB, 2010) while others may be interested in answers to questions of whether papers funded by the agency perform significantly better than comparable non-funded ones (e.g., Bornmann *et al*., 2010b);

3. The indicator should preferentially allow for taking productivity into account. Thus, one should, for example, be able to compare two papers in the 39th percentile with a single paper in the 78th percentile (with or without weighting the differences in ranks in an evaluative scheme as specified under 2.);
4. The indicator should provide the user, among other things, with a relatively straightforward criterion for the ranking (for example, a percentage of a maximum) that can additionally be tested for its statistical significance in relation to comparable (sets of) papers;
5. It should be possible to indicate statistical error of the measurement.

In this study, we try to make important steps in relation to these stated objectives. For this purpose, we joined forces between our two teams which have previously been involved independently in the noted controversy. First, we replicated the measurements of CWTS (2008) and Opthof & Leydesdorff (2010) for the purpose of establishing the percentile ranks of citations of the papers under study in their respective reference sets, and secondly, we elaborate on the ideas of Bornmann & Mutz (2011) for developing percentile ranks as schemes which allow us to compare across sets using non-parametric statistics. Using the percentile rank values allows us to express differences in terms of numbers which can be considered as percentages and we will specify how differences among these numbers can be tested for their significance. Using the six-rank scheme of the National Science Foundation (Bornmann *et al*., 2010a; NSB, 2010), for example, we show the effect of the non-linear transformation implied when using such an evaluation scheme.

Methods and materials

Because as academics, we do not have the possibilities for manipulating yearly volumes of the *Science Citation Index* similarly to the quasi-industrial centers which license the database for evaluation purposes, we used the Web-of-Science interface at the Internet and confined the normalization to seven sets and the comparable documents (in terms of document types) in the same journals and publication years. Although the choice of the normalization baseline matters (Colliander & Ahlgren, 2011), this normalization is not fundamental to the analytical approach, but serves us as an example. In our scheme, one needs one normalization or another against a reference set for each paper (Radicchi *et al*., 2008). One could, for example, consider the un-normalized citation rates as a zero-normalization because all reference sets are then set equal to unity.

Because the ISI split the category of "articles" into articles and proceedings papers in the period under study (in October 2008), we will consider "articles OR proceedings papers" as our reference sets in the publishing journals in the specific years of publication of 241 source documents. These source documents were published by seven Principal Investigators (PIs) in the Academic Medical Center (AMC) of the University of Amsterdam. The PIs belong to a group of 232 scientists evaluated by CWTS (2008 and 2010). Opthof & Leydesdorff (2010) provide reasons for selecting these seven scientists in terms of the distributions of citations as a representative sample given the range in the larger group. The seven authors published 23, 37, 22, 32, 37, 65, and 32 papers, respectively, during this period. The seven document sets overlap in seven coauthored papers. Thus, 248 – 7 = 241 papers could be attributed to seven document sets. The seven sets constitute our units of evaluation.

For these 241 documents and their corresponding reference sets in the journals published in the same years, we determined citation rates in early November 2010. For each paper thus a number of citations per paper ("CPP" in the terminology of CWTS) and "journal citation score" ("JCS") can be computed and thus for each set a so-called *CPP/JCSm* (mean citation rate divided by mean journal citation score) can be calculated both in terms of a "rate of

averages" or as an "average of rates" (that is, an *MNCS* but then defined at the journal level; Van Raan *et al*., 2010b, at p. 291).

In order to move to the percentile rank approach, the citation of each paper is rated in terms of its percentile. In each set, the number of papers with citations smaller than the citation of a paper *i* is expressed as a percentage. In other words: if 65.4 % of the papers were below that of the i^{th} paper with a certain citation, then the percentile score of this paper would be rounded to and classified into the 65^{th} percentile class. Thus, for each set a column vector with 100 values (from the 0^{th} to the 99^{th} percentile, but with ranks 1 to 100) is created. Note that the seven column vectors—representing the seven sets—are now equal in size and thus comparable.

From this matrix (7 columns with each 100 rows), the six percentile impact classes used by the NSF (NSB, 2010; cf. Bornmann *et al*., 2010a) for the evaluation were aggregated as follows:

(1) bottom 50% (papers with a percentile less than the 50^{th} percentile),
(2) 50^{th} – 75^{th} (papers within the [50^{th}; 75^{th}[percentile interval),
(3) 75^{th} – 90^{th} (papers within the [75^{th}; 90^{th}[percentile interval),
(4) 90^{th} – 95^{th} (papers within the [90^{th}; 95^{th}[percentile interval),
(5) 95^{th} – 99^{th} (within the [95^{th}; 99^{th}[percentile interval),
(6) top 1% (papers with a percentile equal to or greater than the 99^{th} percentile).

Thus, a matrix of seven cases (rows) and six variables (columns) is generated. Note that the scores in this matrix are non-parametric (ordinal-scaled) while the previous ones of 100 percentile classes were interval-scaled. In other words, this transformation by aggregation into six classes is non-linear. Thus, the percentile scores are transformed for the purpose of a normative evaluation. We use this evaluative scheme of the NSF in this study as an example of such an evaluation scheme.

The mean percentile rank scores are calculated by weighting the relative frequencies $p(x)$ in the k sets with their rank x, as follows:

$$R(x) = \sum_{k=1}^{k} x \cdot p(x) \qquad (1)$$

Thus, one paper in the 78^{th} percentile weights twice as much as a paper in the 39^{th} percentile in the case of the hundred percentile ranks while in the case of six ranks the paper in the 78^{th} percentile would count three times as much as a paper in the 39^{th} percentile. The maximum weight in the case of 100 classes [R(100)] is consequently 100, while this maximum is six in the case of six classes [$R(6)$]; namely, for the case that all papers are placed in the highest class, respectively. The minimum is always 1, that is, when all papers are to be placed in the first (and lowest) category.

Above or below medium performance can be tested by testing the 100 percentile ranks against a median value of 50 using, for example, Wilcoxon's signed-rank test (which is available under the non-parametric tests in SPSS) and against a similar reference value for $R(6)$. This latter reference value can be obtained by the sum of the products of proportions of the percentile classes (50:25:15:5:4:1) multiplied with the rank numbers of the classes: for example, each count in the bottom-50% class counts as one, and a count in the top-1% as six. One thus obtains an expected value of $R(6)$ for the case of random attribution, as follows: 0.50*1 + 0.25*2 + 0.15*3 + 0.05*4 + 0.04*5 + 0.01*6 = 1.91. The observed distributions can be tested against this expected value using, for example, chi-square statistics.

In the case of both $R(6)$ or $R(100)$—or any other scheme for the evaluation—the sets can be tested against each other for statistical significance of the differences using Dunn's test or Mann-Whitney's U test. First, one should test whether differences among the scientists under study are significant using Kruskal-Wallis (rank variance analysis). If the null hypothesis is

not rejected (that is, no significant differences among the sets are found) then the analysis should be finished here. In the other case, one can further test the differences using the Mann-Whitney U test on each two samples or Dunn's test including an *ex-post* Bonferroni correction for multiple comparisons (in a single pass). The Mann-Whitney test is more conservative—that is, less inclined to flag differences as significant—than Dunn's test, that is, the non-parametric version of the ANOVA-based Bonferroni correction (Levine, 1991, pp. 68 ff.) Since Opthof & Leydesdorff (2010) used the latter test, we staid with this choice.

The overall so-called family-wise alpha error (Type I) across all possible pairwise comparisons increases with the number of these pairwise comparisons *c*. For each pairwise comparison an adjusted alpha error of 0.05 divided by the number of all possible pairwise comparisons was used instead of an alpha error of 0.05. For n=7 there are $c=n*(n-1)/2=7*6/2=21$ comparisons, and the adjusted alpha-level therefore amounts to 0.05/21=0.0023. In general, this alpha-level of 0.05/c can be used as the significance level with the *ex-post* Bonferroni correction in the non-parametric case (Levine, 1991, pp. 68 ff.).

Another issue is the normalization of the relative frequencies $p(x)$ in terms of the respective margin totals ($p_i = f_i / n_i$; $n_i = \Sigma_i f_i$). If a scientist with 10 publications would have one publication in the 99th profile, this publication contributes for 1/10th times 100—the rank number—and therefore 10 percent points to his/her citation rank profile $R(100)$ given the function in Equation 1 (above). However, a scientist with 100 publications of which one in the 99th profile, would add only a single percent point to this score. (An analogous reasoning can be elaborated for $R(6)$.) Since publications in the highest ranks are scarce—given the well-known skewness in empirical citation distributions—this system thus can be expected to disadvantage productive scientists. For reasons of space constraints, we do not provide these more advanced statistics in this abbreviated version of the envisaged full paper.

This effect disappears when the frequencies are not calculated relative to each subset (e.g., the *œuvre* of each scientist or group), but to the total set under study. The weighting is then similar for each scientist in the aggregated set. In order to make the resulting ranks comparable with the individually weighted (for example, as percentages), the results have to be multiplied again with k (in our case, $k = 7$). We distinguish between the two normalization by writing below $R(6)$ and $R(100)$ when normalizing over the six categories of 100 percentile classes for each set (vector) as independent samples, and $R(6,k)$ and $R(100,k)$ when normalizing also over the second dimension of the k subsets of a single sample.

Using $R(100,k)$, the effects of different publication rates are completely taken out of the equation: each publication in our case of the 248 documents under study has a weight of (1/248) in its percentile rank. The resulting percentile ranks [$R(100)$] thus can directly be compared with one another across the sets: two publications in the 39th percentile in one set now weigh equally as one publication in the 78th percentile in another in $R(100,k)$. However, using $R(6,k)$ a non-linear transformation is involved. Furthermore, papers in the same rank are equally appreciated using both $R(6,k)$ and $R(100,k)$. We will discuss the differences of and similarities between the two normalizations (in terms of relative frequencies) in the next section in empirical terms.

In summary, we have thus constructed an indicator in which the different criteria specified above are analytically distinguished. The reference sets can be determined for each individual paper; for example, as the set of all papers in the same journal in the same year and of the same document type. Each set can be attributed a comparable score between 1 and 100 in the case of $R(100)$ and $R(100,k)$ or between 1 and 6 in the case of $R(6)$ and $R(6,k)$. (The latter score can also be expressed as a percentage of six.) The six classes or any other normative scheme for the assessment can directly be derived from the matrix of the percentile values because the evaluative scale is based on specific aggregation rules which can be chosen

differently depending on the purposes of the evaluation. The scores can be read as percentages of the maximum possible score for each document set under study.

A disadvantage of our scores might seem to be that the idea of a "world average" provided by the old "crown indicator" as a baseline (*CPP/FCSm* = 1) has to be abandoned. In our opinion, an average is always sample-dependent unless one knows the population. The sample of documents can be as large as all documents contained in the *Science Citation Index*, but also the latter remains a sample which is based, for example, on Garfield's (1971) Law of Concentration. More importantly, the concept of a "world average" as a standard confounds the analytically different questions of external normalization against a reference set for each paper and internal normalization as a relative frequency with that of evaluation standards. By this integration into a single number, one looses the possibility of using statistics and the indication of error. Instead, CWTS and ECOOM used "rules of thumb" for indicating significance in the deviation from the world standard as 0.5 (Van Raan, 2005) or 0.2 (CWTS, 2008, at p. 7; cf. Schubert & Glänzel, 1983; Glänzel, 1992 and 2010).[5]

The statistics implied in our procedures may seem sophisticated and at first sight complex because they involve non-parametric routines. When fully elaborated, these statistics can be automated in SPSS as a batch job. In this study, however, we guide the reader step-by-step through the procedures using relatively small sets as an example in order to enable users to reproduce the percentile-rank evaluation using their own datasets, their reference sets, and potentially different evaluation schemes.

Results

The distributions of the 100%-percentiles for each paper of the seven scientists under study are shown in Figure 1, both as scatter plots and box plots. The black dots in the boxes represent the arithmetic mean, the stars the minima and maxima, respectively. The borders of the boxes indicate the 25%, 50%, and 75% quantiles of individual distributions. Obviously, all scientists score across the whole variance, but in some cases (e.g., Scientists 2 and 3) the concentration in the top half is larger than at the bottom. Scientists 5 and 6 have publications in the 0^{th} percentile.

The ordering of the scientists from one to seven was based by Opthof & Leydesdorff (2010) on the ranking of these scientists in the original report of CWTS (2008). These rankings were based on *CPP/JCSm* in the CWTS terminology. Table 1 provides first the replication of this *CPP/JCSm* on the basis of our downloads of data (in November 2010)[6] and then in the third column the ranking based on the alternative indicator proposed by Opthof & Leydesdorff (2010) and ever since elaborated by CWTS into the new crown indicator *MNCS* (or more precisely, the journal-equivalent of this indicator *MNCS/MNJS*; cf. Van Raan *et al*., 2010, at p. 291). The values for the four indicators proposed above [$R(6)$, $R(100)$, $R(6,k)$, and $R(100,k)$] follow in the next columns. The consequent ranks are added in each column between brackets, providing a one to the highest value and a seven to the lowest. Table 2 provides both the Pearson correlations (lower triangle) and rank-order correlations (Kendall's tau-b) between these indicators.

Notes

[5] Schubert & Glänzel (1983) based their reasoning on normal distributions (Glänzel, 2010). The reasoning can be used to estimate error in large sets (Glänzel, *personal communication*, 16 November 2009), but, in our opinion, this estimator is insufficiently precise for evaluations of smaller sets.

[6] Our values deviate from *CPP/JCSm* in that we did not correct for self-citations (CWTS, 2008; Opthof & Leydesdorff, 2010).

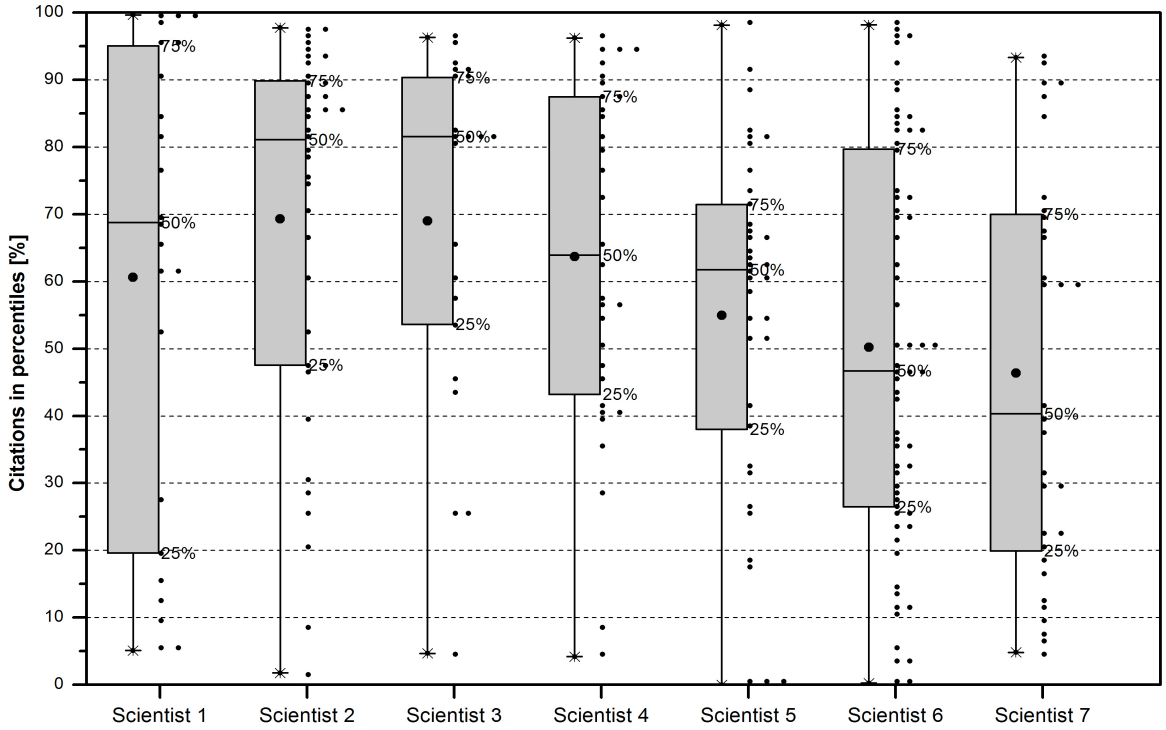

Figure 1. Boxplots for citations of each paper in percentiles separated for seven scientists

Table 1. Values and ranking (between brackets) using the various indicators

Scientist	*Avg(CPP)/ Avg(JCS) (= CPP/JCSm)*		*Avg(CPP/JCS)*		*R(100)*		*R(6)*		*R(100,k)*		*R(6,k)*	
1	1.99	(1)	2.04 (± 0.50)	(1)	61.17	(4)	2.83	(1)	39.71	(7)	1.83	(5)
2	1.42	(3)	1.56 (± 0.16)	(3)	69.81	(1)	2.68	(3)	72.91	(2)	2.79	(2)
3	1.45	(2)	1.60 (± 0.24)	(2)	69.55	(2)	2.77	(2)	43.19	(5)	1.72	(6)
4	1.17	(4)	1.32 (± 0.15)	(4)	64.34	(3)	2.34	(4)	58.12	(3)	2.12	(3)
5	1.03	(5)	1.04 (± 0.15)	(5)	55.49	(5)	2.00	(5)	57.95	(4)	2.09	(4)
6	0.86	(6)	1.04 (± 0.11)	(6)	50.69	(6)	1.91	(6)	93.00	(1)	3.50	(1)
7	0.71	(7)	0.87 (± 0.15)	(7)	46.88	(7)	1.72	(7)	42.34	(6)	1.55	(7)

Table 2. Rank-order correlations (Kendall's Tau-b; upper triangle) and Pearson correlations (lower triangle) between the various indicators.

	Avg(CPP)/ Avg(JCS)	*Avg(CPP/JCS)*	*R(100)*	*R(6)*	*R(100,k)*	*R(6,k)*
CPP/JCSm		0.98 **	0.62	1.00	-0.24	-0.05
Avg(CPP/JCS)	0.99 **		0.59	0.98 **	-0.20	0.00
R(100)	0.68	0.71		0.62	0.14	0.14
R(6)	0.93 *	0.95 **	0.89 **		-0.24	-0.05
R(100,k)	-0.38	-0.35	-0.13	-0.30		0.81 *
R(6,k)	-0.22	-0.18	-0.07	-0.16	0.98**	

Note: **. Correlation is significant at the 0.01 level (2-tailed); *. correlation is significant at the 0.05 level (2-tailed).

As was to be expected (Waltman *et al.*, in press), the first two indicators based on comparing citation scores versus journal citation scores parametrically are highly and significantly correlated ($r = 0.99$, $p < 0.01$; $\tau = 0.98$, $p < 0.01$). Using the six percentile classes [$R(6)$], the ranking is precisely the same as with these two average-based indicators. However, $R(100)$ deviates at one place from this shared pattern by replacing the number 1 to the fourth position behind the numbers 2, 3, and 4. This corresponds (not incidentally) with the visual impression obtained by inspecting Figure 1. $R(100)$ can be considered as a summary indicator of the patterns shown in Figure 1.

Table 3: Number of papers published by seven scientists categorized to six percentile impact classes

Percentile impact class	*Scientist 1*	*Scientist 2*	*Scientist 3*	*Scientist 4*	*Scientist 5*	*Scientist 6*	*Scientist 7*	*Total*
<50th (bottom 50%)	7	10	5	10	11	34	17	94
[50th; 75th[	6	5	4	8	18	14	9	64
[75th; 90th[	3	13	6	8	6	11	4	51
[90th; 95th[	1	5	5	5	1	1	2	20
[95th; 99th[	3	4	2	1	1	5	0	16
≥99th (top 1%)	3	0	0	0	0	0	0	3
Total	23	37	22	32	37	65	32	248

Table 3 informs us about the distributions of the seven sets across the six percentile impact classes. Scientist 2 has more papers than Scientist 1 in most categories, but not in the 99th percentile rank (class 6). Given the smaller size of the *œuvre* of Scientist 1 the three papers in this category weigh heavily, namely: each for (6 * [1/22] = 0.27). This leads to a contribution of 0.82 on a score of 2.83. More dramatically, however, is the difference between these two scientists when scoring in the 95th percentile. The three papers of Scientist 1 in this class contribute 5 * (1/22) * 3 = 0.68 to the score, while the four papers of Scientist 2 in this same class contribute only 5 * (1/37) * 4 = 0.54 to his/her score.

The example proves our point that citation scores that do not take publication rates into account "punish" productivity in terms of lower rankings. The six percentile classes make this quantitatively visible, but this same effect can also be expected using the average-based citations because they also operate on probability distributions while assuming independence among the samples. After such a normalization (e.g., using *z*-scores; cf. Radicchi *et al*., 2008) at the level of independent samples, the differences in size among the sets are manifested only in terms of the significance testing and error terms because in these computations the *n* of cases in each sample plays a role in the denominator (for example, as the square root of *n* in the case of computing the standard error of the measurement). However, citation analysts hitherto have paid insufficient attention to the question whether observed differences are also statistically significant; one rarely finds error estimates in the tables or error bars in the accompanying figures and graphs.

Table 1, for example, contains the standard error of the measurement for the indicator proposed by Opthof & Leydesdorff (2010). The larger size of the error of the measurement for Scientist 1 (± 0.50) when compared with all others and the relatively low value of this parameter for Scientist 6 (± 0.11) could have flagged this spurious publication effect of the sample sizes (n_1 = 22 and n_6 = 65, respectively). However, the last two columns of Table 1 show the effects of this correction quantitatively: Scientist 1 becomes the seventh in rank using *R*(100,k) and fifth in rank using *R*(6). However, Scientist 6 is now the highest ranked one using both these indicators. In the case of Scientist 1, the higher appreciation of the top-percentile fully explains the difference of rating when using *R*(6,k) instead of *R*(100,k).

Using *R*(6,k) or *R*(100,k), the four papers of Scientist 2 in the 95th percentile rank (class 5) and the three papers of Scientist 1 in this same class are contributing proportionally, that is, 4:3, to their respective scores. As noted in the case of *R*(100,k) one weights two papers in the 39th percentile of one scientist equally to one paper in the 78th percentile of another. This is transformed in the case of using *R*(6,k) for normative reasons; for example, because one is more interested in most highly-cited papers when comparing nations or institutions.

The transparency of *R*(100,k) can be considered as an advantage, but a six-point scale such as *R*(6,100) may be felt as more functional to communications in the policy domain. Of course, the user (e.g., the policy maker) can suggest another scheme such as *R*(5,k) by specifying other classes. In some countries one is used to five point scales in the evaluation (e.g., the Netherlands), while in other countries six is in use as the highest score (e.g., Germany).

Conclusions and discussion

Our purpose in this study was to provide citation impact indicators which are no longer based on averages, but on percentile ranks. We specified a number of criteria for a more abstract scheme that can also be used to organize and schematize different citation impact indicators according to the three distinguished degrees of freedom: the selection of the reference sets, the evaluation criteria, and the choice for defining the samples as independent or not.

The proposed indicators [$R(6)$, $R(100)$, $R(6,k)$, $R(100,k)$] first improve on the averages-based indicators because one can abstract from the shape of the distribution of citations over papers. Secondly, the choice of the reference set for each paper is no longer related to the evaluation scheme. Both the reference sets can be chosen—for example, as individual journals, groups of journals (e.g., ISI Subject Categories), papers selected on specific criteria such as index terms or keywords, etc.—and the evaluation schemes can be specified; for example, in terms of six classes or differently. The latter choice is a normative one, while the former one needs analytical grounding.

The elaboration of the proposal of Bornmann & Mutz (2011) to use percentile ranks made us aware how sensitive citation-based indicators can be on sample sizes even after this first correction for differences in the shapes of the distributions. This result provided us with the major learning step of the study: one should compare "like with like" as Martin & Irvine (1983) once formulated in the early days of citation analysis, but one should not reduce this comparison to the specification of reference set(s) for each article. The document sets under study are to be compared among them *after* being normalized at the individual paper level against the reference sets. The normalization in terms of the external reference sets and thereafter the rewrite as percentiles was not yet sufficient since one needs the additional normalization as relative frequencies across the sets under study. By normalizing the relative frequencies in terms of the grand total of the combined sample, one eventually obtains percentile rank scores that account for both differences in size and shape of the citation distributions. These scores [$R(100,k)$ and $R(6,k)$, or more generally: $R(i,k)$] are comparable across sets.

Our data provided us with an opportunity to make a convincing case for this change in the framework of citation analysis—from considering sets as independent samples to subsamples of a single sample—by showing the dramatic mistake that one can make when one uses citation rates without taking sample sizes into account. Only because of the smaller sample size, Scientist 1 leaded the ranking: a paper in a similar rank contributed in this case $1/22 = 0.045$ to the total score while it would contribute only $1/37 = 0.027$ for Scientist 2. Scientist 6 with 65 papers thus was disadvantaged to the extent that s/he would lead the ranking when we corrected for this size effect. Since without this correction, the percentile ranks $R(6)$ and $R(100)$ correlated highly and significantly with the "Leiden" indicators (with or without the correction for the order of operations), the scores based on averages also suffer from this ignored size-effect.

In other words, the initial step of Lundberg (2007) and Opthof & Leydesdorff (2010) of introducing significance testing and error indication in the measurement of average citations (as it had already been done previously by other centers; cf. Gingras & Archambault (2011)) was not yet a sufficient step. By proceeding to the percentile rank approach of Bornmann & Mutz (2011), we could make the next step in this study that the assumption of operating on independent probability distributions when using the mean or the median (or any other statistics) requires a further reflection. In citation analysis, one compares samples which are no longer independent when compared. Without normalization across the samples, one changes the basis for the comparison when moving from one set to another.

In summary, using these indicators the citation analyst has three degrees of freedom: (1) one can choose external reference sets for each paper, (2) one can choose a normative evaluation

scheme, (3) one can choose sample definitions as independent or not. These three dimensions are analytically different, and the eventual scores will differ with these choices. *Vice versa*, there are no absolute citation impact scores which are independent of making these choices (such as a "world average"; Glänzel, 1992; Van Raan, 2005). As we argued, one can use zero-normalizations such as choosing *not* to normalize against reference sets (that is, assuming the reference values to be equal to unity). The citation impact enterprise is thoroughly probabilistic and precisely the probability distributions lead us to a strong preference for defining probabilities across sets such as when using $R(i,k)$ where i is the indicator for the (percentile rank) class and k the indicator of each subset. The new measure enables us to compare sets of different size among one another.

References

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

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*, 199-245.

Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2010a). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS ONE, 5*(10), e11344.

Bornmann, L., Leydesdorff, L., & Van den Besselaar, P. (2010b). A Meta-evaluation of Scientific Research Proposals: Different Ways of Comparing Rejected to Awarded Applications. *Journal of Informetrics, 4*(3), 211-220.

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., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high-profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society - Series A (Statistics in Society) 174*(4), 1-23.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of 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*, 93-102. doi: 10.3354/esep00084.

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 Physicsby Using Percentile Rank Classes, *Journal of the American Society for Information Science and Technology, 62*(1), 61-71.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the Backbone of Science. *Scientometrics,* 64(3), 351-374.

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: a case study of 20 natural science departments. *Journal of Informetrics, 5*(1), 101-113.

CWTS. (2008). AMC-specifieke CWTS-analyse 1997–2006 (access via AMC intranet; unpublished, confidential). Leiden, The Netherlands: CWTS.

Garfield, E. (1971). The mystery of the transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. *Current Contents, 3*(33), 5–6.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

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. (1992). Publication Dynamics and Citation Impact: A Multi-Dimensional Approach to Scientometric Research Evaluation. In P. Weingart, R. Sehringer & M. Winterhagen (Eds.), *Representations of Science and Technology. Proceedings of the International Conference on*

Science and Technology Indicators, Bielefeld, 10-12 June 1990 (pp. 209-224). Leiden: DSWO / Leiden University Press.
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.
Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics,* 56(3), 357-367.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572.
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., & Bornmann, L. (in press). How fractional counting affects the Impact Factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology* (doi: 10.1002/asi.21450).
Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: fractional counting of citations. *Journal of Informetrics, 4*(4), 644-646.
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. (2009). A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.
Lundberg, J. (2007). Lifting the crown - citation *z*-score. *Journal of Informetrics, 1*(2), 145-154.
Martin, B., & Irvine, J. (1983). Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio Astronomy. *Research Policy, 12*, 61-90.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265-277.
National Science Board (2010). Science and engineering indicators 2010, appendix tables. Arlington, VA, USA: National Science Foundation (NSB 10-01).
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.
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.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268-17272.
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.
Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics, 5*(1), 59-73.
Spaan, J. A. E. (2010). The danger of pseudoscience in informetrics. *Journal of Informetrics 4*(3), 439-440.
Van Raan, A. F. J. (2005). Measurement of central aspects of scientific research: performance, interdisciplinarity, structure. *Measurement, 3*(1), 1-19.
Van Raan, T. (2010). Bibliometrics: measure for measure. [10.1038/468763a]. *Nature, 468*(7325), 763-763.
Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J., & Waltman, L. (2010a). Rivals for the crown: reply to Opthof and Leydesdorff. *Journal of Informetrics, 4*(3), 431-435.
Van Raan, A. F. J., Eck, N. J. v., Leeuwen, T. N. v., Visser, M. S., & Waltman, L. (2010). *The new set of bibliometric indicators of CWTS.* Paper presented at the 11th International Conference on Science and Technology Indicators, Leiden, September 9-11, 2010; pp. 291-293.
Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2011). Towards a New Crown Indicator: Some Theoretical Considerations. *Journal of Informetrics, 5*(1), 37-47.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (in press). Towards a new crown indicator: An empirical analysis. *Scientometrics.* Arxiv preprint arXiv:1004.1632.

Zitt, M. (2010). Citing-side normalization of journal impact: A robust variant of the Audience Factor. *Journal of Informetrics, 4*(3), 392-406.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856-1860.