Journal of Informetrics 7 (2013) 746–755

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Quantitative evaluation of alternative field normalization procedures

Yunrong Li[a], Filippo Radicchi[b], Claudio Castellano[c], Javier Ruiz-Castillo[a,*]

[a] Universidad Carlos III de Madrid, Departamento de Economía, Spain
[b] Universitat Rovira i Virgili, Department d'Enginyeria Quimica, Spain
[c] Istituto dei Sistemi Complessi (ISC-CNR), and Dipartimento di Fisica, "Sapienza" Universita' di Roma, Italy

ARTICLE INFO

Article history:
Received 18 April 2013
Received in revised form 7 June 2013
Accepted 11 June 2013
Available online 12 July 2013

Keywords:
Citation analysis
Citation practices
Normalization procedures
Citation inequality

ABSTRACT

Wide differences in publication and citation practices make impossible the direct comparison of raw citation counts across scientific disciplines. Recent research has studied new and traditional normalization procedures aimed at suppressing as much as possible these disproportions in citation numbers among scientific domains. Using the recently introduced *IDCP* (*I*nequality due to *D*ifferences in *C*itation *P*ractices) method, this paper rigorously tests the performance of six cited-side normalization procedures based on the Thomson Reuters classification system consisting of 172 sub-fields. We use six yearly datasets from 1980 to 2004, with widely varying citation windows from the publication year to May 2011. The main findings are the following three. Firstly, as observed in previous research, within each year the shapes of sub-field citation distributions are strikingly similar. This paves the way for several normalization procedures to perform reasonably well in reducing the effect on citation inequality of differences in citation practices. Secondly, independently of the year of publication and the length of the citation window, the effect of such differences represents about 13% of total citation inequality. Thirdly, a recently introduced two-parameter normalization scheme outperforms the other normalization procedures over the entire period, reducing citation disproportions to a level very close to the minimum achievable given the data and the classification system. However, the traditional procedure of using sub-field mean citations as normalization factors yields also good results.

1. Introduction

The number of citations that a scientific paper has accumulated is often interpreted as a proxy of its influence within the scientific community. Although the relation between citations and effective scientific influence is still under active debate (Adler, Ewing, & Taylor, 2009; MacRoberts & MacRoberts, 1989, 1996), citation numbers are often used in assessment exercises, and their practical role in modern science is becoming more and more central. However, because the significance of citations is content- and discipline-dependent (Bornmann & Daniel, 2008), the direct comparison of raw citation numbers received by articles belonging to different fields is plagued with difficulties. A paper in Biochemistry typically accumulates more citations than a paper in Mathematics but this does not necessarily imply that the former is more influential than the latter. Different scientific disciplines strongly differ in citation practices, and as a consequence the typical number of citations that a paper in a given field receives may strongly differ from the number of citations typical of another field.

Abbreviation: IDCP, citation Inequality due to Differences in Citation Practices.

* Corresponding author. Tel.: +34 91 624 95 88.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.06.001

Many bibliometric indicators have been developed with the aim of assessing the relevance of scientific research activities at different levels: journals (Garfield, 2006), scientists (Egghe, 2006; Hirsch, 2005), departments (Davis & Papanek, 1984), institutions (Kinney, 2007), *etc.* These indicators, however, are often based on raw citation numbers, and thus have several limitations when used to perform comparisons across different fields of research.

This paper is concerned with situations in which the actual number of citation counts of individual publications – and not only their location in a percentile distribution (or a percentile class) – is needed. A good example is provided by citation impact indicators with the property that receiving one more citation increases the citation impact level. Monotonic indicators with this property constitute an important class including, for example, frequently used average-based indicators. To understand the fundamental nature of the problem posed in this case by the inherent disproportion in citation numbers among scientific disciplines, assume for simplicity that there are only two levels of aggregation consisting, say, of a number of sub-fields and the all-sciences case.[1] Citations received by publications within a given sub-field are assumed to be comparable, while citations received by publications in two different sub-fields are not. To analyze the citation impact of research units within or across heterogeneous sub-fields one could use scale- and size-independent indicators, such as relative indicators that consider the ratio of the research unit mean citation to the sub-field mean citation, or the class of Foster, Greer, and Thorbecke low- and high-impact indicators introduced into scientometrics by Albarrán, Ortuño, and Ruiz-Castillo (2011a); Albarrán, Ortuño, and Ruiz-Castillo (2011b). However, to analyze the citation performance of research units in the all-sciences case, one must first confront the non-comparability of citation counts between publications in different sub-fields, an unavoidable problem at the highest aggregation level.

The different normalization procedures of raw citation counts at the level of the individual publication that have been proposed, can be classified in two conceptually different classes:

1. Target (or cited-side) procedures, in which citation weights or normalization factors are functions of the cited papers. This class includes many different types of normalization techniques such as: (i) field averages (see *inter alia* Braun, Glänzel, & Schubert, 1985; Moed, Burger, Frankfort, & van Raan, 1985; Moed, De Bruin, & van Leeuwen, 1995; Moed & van Raan, 1988; Schubert, Glänzel, & Braun, 1983; Schubert, Glänzel, & Braun, 1988; Radicchi, Fortunato, & Castellano, 2008; Schubert & Braun, 1986, 1996; Vinkler, 1986, 2003); (ii) average-based scalar difference from the mean (Glänzel, 2011); (iii) two-parameter reverse engineering (Radicchi & Castellano, 2012a), and (iv) exchange rates (Crespo, Li, & Ruiz-Castillo, 2013a; Crespo, Li, Herranz, & Ruiz-Castillo, 2013b).
2. Source (or citing-side) procedures, in which citation weights are functions of the citing papers, studied by, *inter alia*, Glänzel, Schubert, Thijs, and Debackere (2011); Leydesdorff and Opthof (2010); Moed (2010); Waltman and Van Eck (2012); and Zitt and Small (2008).

While the development of cross-disciplinary citation indicators dates back to the 1980s, only recently have scholars started to apply them to large sets of empirical data, and test their performances statistically. Three methods have been proposed to assess the performance of a generic normalization procedure quantitatively: (i) between-group variance (Leydesdorff & Bornmann, 2012); (ii) a fairness test based on ranking (Radicchi & Castellano, 2012a, 2012b), and (iii) the Inequality due to Differences in Citation Practices method (*IDCP* hereafter) (Crespo et al., 2013a, 2013b).

Between-group variance is the simplest of the three tests, but, by construction, it vanishes for indicators normalized by field averages. This makes its applicability very limited. Although based on different principles, both the fairness and the *IDCP* tests leverage on strict statistical formalisms that do not require any strong assumption (*i.e.*, they are distribution free statistical tests). The fairness test has already been applied in three instances: to test the performance of indicators based on the two-parameter reverse engineering (Radicchi & Castellano, 2012a); to compare field averages and one version of "fractional citation counting", which is part of source normalization procedures (Radicchi & Castellano, 2012b), and to test the performances of normalized Impact Factors of journals (Leydesdorff, Radicchi, Bornmann, Castellano, & de Nooye, 2012). The *IDCP* method has been used for field averages, exchange rates, and Glänzel type normalizations (Crespo et al., 2013a, 2013b), as well as the comparison of target and source normalization procedures in Waltman & van Eck (2013).

In this paper, we perform an extensive analysis of six normalized indicators of the target or cited-side variety, and assess their performance in the citation distribution for all articles in all fields – the *all-fields* case – using the *IDCP* method. The dataset consists of publications in 172 sub-fields (or Web of Science subject-categories) indexed by Thomson Reuters. The publications appeared in six different years, spanning a period of more than two decades from 1980 to 2004. This feature allows us to analyze temporal trends in citation practices, as well as the sensitivity of our results to a large range of citation windows.

Among the main results, we find that, relative to overall citation inequality, the effect of differences in citation practices across sub-fields has a similar importance independently of the year of publication and the length of the common citation window. Similarly, the ranking of normalization procedures in terms of their ability to reduce such an effect is essentially the same over the entire period. Our findings in this last respect are in line with previous results. Firstly, the similarity of citation

[1] A similar problem arises when articles belonging to a number of closely related but heterogeneous sub-fields need to be aggregated into a single intermediate category, such as the aggregation of Cardiac & Cardiovascular Systems, Hematology, Oncology, and other sub-fields into the discipline "Internal Clinical Medicine" (for a recent example of the difficulties raised even at the sub-field level, see Van Eck, Waltman, Van Raan, Klautz, & Peul, 2012).

distributions within the 172 sub-fields classification system explains why several normalization procedures work reasonably well in diminishing the impact on citation inequality of differences in citation practices across sub-fields. Secondly, the reverse engineering procedure out-performs other normalization methods (see Radicchi & Castellano, 2012a, 2012b). Thirdly, normalization by field averages yields also good results (see Crespo et al., 2013a, 2013b; Radicchi & Castellano, 2012a, 2012b; Radicchi et al., 2008).

The rest of the paper is organized into four sections. Section 2 briefly discusses the evaluation methods, and introduces the six normalization procedures to be evaluated. Section 3 presents the data, and some descriptive statistics. Section 4 contains the empirical results, while Section 5 offers some concluding comments.

2. Methods

2.1. The measuring framework

Given a classification system into a number of scientific fields, Crespo et al. (2013a) introduces a simple model in which the number of citations received by an article is a function of two variables: the article's underlying scientific influence, and the field to which it belongs to. Consequently, the citation inequality in the all-fields case is the result of two factors: differences in scientific influence within homogeneous fields, and differences in citation practices across fields.

Suppose we have an initial citation distribution $\boldsymbol{Q} = \{c_l\}$ consisting of N distinct articles, indexed by $l = 1,\ldots, N$, where c_l is the number of citations received by article l. Assume that here are S sub-fields, indexed by $s = 1,\ldots, S$. In the multiplicative approach each article is wholly counted as many times as necessary in the several sub-fields to which it is assigned. In this way, the space of articles is expanded as much as necessary beyond the initial size in what we call distribution $\boldsymbol{C}$. If we let Ns be the number of distinct articles in sub-field s, the corresponding ordered citation distribution can be described by $\boldsymbol{c_s} = (c_{s1},\ldots, c_{si},\ldots, c_{sNs})$ with $c_{s1} \leq c_{s2} \leq \ldots \leq c_{sNs}$, where c_{si} is the number of citations of article i in sub-field s, and $c_{si} = c_l$ for some article l in the initial distribution $\boldsymbol{Q}$. Of course, $\boldsymbol{C} = \cup_s \boldsymbol{c_s}$, and the total number of articles in the sub-field extended count is $M = \Sigma_s\, Ns > N$. The above assignment of articles in $\boldsymbol{C}$ into the S sub-fields forms what we call classification system K.

As in Crespo et al. (2013a), partition each citation distribution $\boldsymbol{c_s}$ into Π quantiles of equal size, $\boldsymbol{c_s^\pi}$, and citation mean $\boldsymbol{\mu_s^\pi}$, for $\pi = 1,\ldots, \Pi$. For each π, define the citation distribution $\boldsymbol{c^\pi} = (\boldsymbol{c_1^\pi}, \ldots, \boldsymbol{c_s^\pi}, \ldots, \boldsymbol{c_S^\pi})$. Clearly, the number of articles in $\boldsymbol{c^\pi}$ is $\Sigma s\, Ns/\Pi = M/\Pi$, and the set of vectors $(\boldsymbol{c^1},\ldots, \boldsymbol{c^\pi},\ldots, \boldsymbol{c^\Pi})$ form a partition of distribution $\boldsymbol{c}$. For each π, let $\boldsymbol{m^\pi} = (\boldsymbol{\mu_1^\pi}, \ldots, \boldsymbol{\mu_s^\pi}, \ldots, \boldsymbol{\mu_S^\pi})$ be the distribution in which each publication in quantile $\boldsymbol{c_s^\pi}$ is assigned the mean citation in that quantile, $\boldsymbol{\mu_s^\pi}$. Under the assumptions of the model, the citation means $\boldsymbol{\mu_s^\pi}$ holding π constant are directly comparable across sub-fields. Therefore, for each π, the citation inequality of $\boldsymbol{m^\pi}$, abbreviated as $I(\pi)$,

$$I(\pi) = I(\boldsymbol{m^\pi}) = I(\boldsymbol{\mu_1^\pi}, \ldots, \boldsymbol{\mu_s^\pi}, \ldots \boldsymbol{\mu_S^\pi}), \tag{1}$$

captures the citation inequality attributable to differences in citation practices across sub-fields at quantile π (see Crespo et al., 2013a, 2013b for details).

Given a citation distribution $\boldsymbol{C}$, in the implementation of this model one uses an additively decomposable citation inequality index, I, defined as

$$I(\boldsymbol{C}) = \left(\frac{1}{N}\right) \sum_l \left(\frac{c_l}{\mu}\right) \log\left(\frac{c_l}{\mu}\right) \tag{2}$$

where μ is the mean of distribution $\boldsymbol{C}$. Applying the decomposability property of citation inequality index I first to the partition $\boldsymbol{c} = (\boldsymbol{c^1},\ldots, \boldsymbol{c^\pi},\ldots, \boldsymbol{c^\Pi})$, and then to the partition $\boldsymbol{c^\pi} = (\boldsymbol{c_1^\pi}, \ldots, \boldsymbol{c_s^\pi}, \ldots, \boldsymbol{c_S^\pi})$ for each π, it can be shown that the total citation inequality in the all-sciences case, $I(\boldsymbol{C})$, can be decomposed into the sum of three terms, one of them being the *IDCP* (*I*nequality due to *D*ifferent *C*itation *P*ractices) term under classification system K:

$$IDCP(K) = \sum_\pi \upsilon^\pi\, I(\pi) \tag{3}$$

where v_s^π is the share of total citations in quantile π of sub-field s, and $v^\pi = \Sigma_s\, v_s^\pi$. Therefore, $IDCP(K)$ is a weighted average of the key expressions in (1), with weights v^π so that $\Sigma_\pi\, v^\pi = 1$. Note that, due to the skewness of science, the weights v^π will rapidly increase with π.[2]

The impact of any normalization procedure can be evaluated by the reduction in the $IDCP(K)$ term after normalization. However, in order to assess the soundness of this methodology we first ask: what is the lowest possible value for the *IDCP* term? In the case of infinite, real valued and identically distributed data, the *ICDP* term would be equal to zero. However, real citation numbers do not satisfy the former requirement. Consequently, we search for the lowest value of the *IDCP* term that

[2] As far as the two remaining terms in the decomposition are concerned, one refers to the citation inequality that takes place within the $\boldsymbol{c_s^\pi}$ quantiles, while the other measures the citation inequality in the distribution where each article in any field is assigned the mean citation of the quantile to which it belongs. For high Π, the first term is expected to be small, while the second – capturing the skewness of science in the all-sciences case – is expected to be large. For details, see Crespo et al. (2013a).

is achievable given the data and the classification system K, and use it as a reference for assessing the ability of the different normalization procedures to effectively remove the effect on total citation inequality of differences in citation practices across sub-fields in K.

In order to reach this goal, we adopt the rank percentile approach (see Bornmann & Marx, 2013, for a very able summary of this approach and the recent literature related to it). Given a sub-field s, we assign to each paper with c_{si} citations a score r_{si} equal to the fraction of papers within the same sub-field that have accumulated a number of citations lower than or equal to c_{si}. Thus, a value of 0.9 means that the paper in question is among the 10% most cited publications, while a value of 0.5 indicates that the paper has received the median citation rate in this sub-field. According to this rule, scores have values in the range 0–1, preserve the natural order (including ties) of the original citation sequence, and in each sub-field have exactly the same distribution: the uniform one. Therefore, given a classification system, this way of assigning scores to papers represents a sort of "perfect normalization" scheme for which the *IDCP* term provides the best performance that can be achieved for a given data set.

2.2. Normalization procedures

The following six normalization procedures will be empirically investigated in Section 4.

1. *Normalization by sub-field average*. Each paper i in sub-field s receiving c_{si} citations, is assigned a score equal to $c'_{si} = c_{si}/\mu_{s1}$, where μ_{s1} is the average number of citations received by papers in sub-field s. Thus, c'_{si} represents the relative impact, in terms of citations, of paper i within sub-field s.
2. We additionally consider a slight variation of the former normalization scheme, where μ_{s1} is calculated excluding uncited publications. This different approach has been used by Radicchi et al. (2008), and has been also suggested by Waltman, van Eck, and Van Raan (2011) and Abramo, Cicero, and D'Angelo (2012), because it is supposed to lead to higher levels of reduction of citation disproportions among scientific fields.
3. *Normalization by median value*. This represents a simple modification of the previous indicator, where the only difference is that the number of citations c_{si} received by a paper is divided by the sub-field s's median value, m_s (instead of by the average value μ_{s1}). Since in practice there are several sub-fields for which $m_s = 0$, we calculate the median citation number of each category by excluding uncited publications.
4. *Normalization by two-parameter reverse engineering*. Radicchi and Castellano (2012a) introduce a normalization scheme based on the use of two parameters empirically estimated from the data. For each s, these parameters are the best estimates of the prefactor a_s and the α_s exponent of a power-law transformation able to make different citation distributions collapse on top of each other. This means that if the score of a paper is computed as $c'_{si} = (c_{si}/a_s)^{1/\alpha s}$, then the distribution of c'_{si} values is no longer dependent on the specific sub-field considered. In particular, when two distributions have the same exponent, the transformation necessary for their collapse is linear, and the method reduces to normalization by field average. Radicchi and Castellano (2012a) demonstrate that, for the vast majority of sub-fields, the values of α_s are very similar, and the citation distributions are nearly the same when plotted as a function of the normalized values c'_{si}. However, a limited number of sub-fields are characterized by widely changing values of the transformation parameters, so that the distribution of their c'_{si} values does not follow a universal law.
5. *Glanzel's normalization*. This normalization involves the transformation of the raw data of any sub-field citation distribution with N papers, $\boldsymbol{c_s} = (c'_{s1}, \ldots, c'_{sN})$, by the formula $c'_{si} = c_{si}/(\mu_{s2} - \mu_{s1})$, where μ_{s1} is the average citation of sub-field s, and μ_{s2} is the average citation defined over the publications receiving a number of citations equal to or greater than μ_{s1}.
6. *Exchange rates normalization*. Crespo et al. (2011a,b) find that the similarity of the shape of citation distributions over 22 broad fields or 219 sub-fields allows the effect of idiosyncratic citation practices to be rather well estimated over a wide range of intermediate quantiles where citation distributions behave as if they essentially differ by a scale factor. Consequently, a set of average-based measures, called exchange rates, can be estimated profitably over that interval.

3. Data

3.1. The dataset

We use the dataset already analyzed in Radicchi and Castellano (2012a). It consists of six subsets, each including all publications in 8304 scientific journals in the following years: 1980, 1985, 1990, 1995, 1999, and 2004. Journal titles are collected from the Journal Citation Reports database (http://science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D). We restrict our attention only to documents written in "English", and classified as "Article", "Letter", "Note" or "Proceedings Paper" for a total of 2,906,615 publications. We retrieve from the Web of Science database (WoS, isiknowledge.com, field "times cited") the number of citations each document has accumulated from its publication year up to the week of May 23–31, 2011. Note that the citation windows vary across the yearly subsets, ranging from seven years (and five months) for the papers published in 2004, to 31 years (and five months) for the 1980 subset.

In what follows, the S sub-fields in the classification system K introduced in Section 2.1 are identified with 172 subject-categories distinguished in the Web of Science by Thomson Reuters. As already emphasized by the inventors themselves

Table 1
The skewness of science. Averages (and standard deviations) over 172 sub-field citation distributions in 1980–2004 *versus* previous results for articles published in 1998–2002 with a five-year citation window classified in 219 sub-fields.

	Percentage of articles in category			Percentage of total citations accounted for by category		
	1	2	3	1	2	3
Results from our dataset, selected years:						
1980	73.2 (4.3)	19.0 (2.6)	7.7 (2.1)	21.1 (4.6)	32.1 (2.3)	46.9 (5.5)
1985	73.1 (4.3)	19.1 (2.5)	7.8 (2.2)	21.7 (5.0)	31.9 (2.4)	46.4 (5.6)
1990	72.0 (3.7)	19.8 (2.2)	8.2 (1.8)	21.8 (4.3)	32.4 (1.7)	45.8 (5.0)
1995	71.1 (3.6)	20.3 (2.1)	8.6 (1.6)	22.5 (4.0)	32.7 (1.4)	44.8 (4.5)
1999	70.2 (3.6)	20.8 (2.1)	9.0 (1.8)	23.3 (4.0)	33.0 (1.5)	43.7 (4.1)
2004	68.6 (3.5)	21.7 (2.0)	9.7 (1.7)	24.3 (3.6)	33.4 (1.4)	42.3 (3.5)
Previous results over 219 sub-fields for articles published in 1998–2002 with a five-year citation window. Table 1, p. 391 in Albarrán et al. (2011c):						
	68.6 (3.7)	–	10.0 (1.7)	29.1 (1.6)	–	44.9 (4.6)

Category 1 = articles with a low number of citations, below μ_1.
Category 2 = articles with a fair number of citations, above μ_1 and below μ_2.
Category 3 = articles with a remarkable or outstanding number of citations, above μ_2.
where μ_1 = mean citation of each citation distribution, μ_2 = mean citation of articles with a number of citations above μ_1.

(Pudovkin & Garfield, 2002), this classification is known to have several weak points. One of them is that publications in the periodical literature are assigned to sub-fields *via* the journal in which they have been published. Many journals are assigned to a single sub-field, but many others are assigned to two, three, or even more sub-fields. For example, the percentage of single-category papers in the datasets used in this paper tends to diminish with time: it is 67% in 1980, but only 56% in 2004. Therefore, between one third and 44% of all papers in our dataset are assigned to several sub-fields.

To tackle this problem, two paths can be followed. The first is a fractional strategy, according to which each publication is fractioned into as many equal pieces as necessary, with each piece assigned to its corresponding sub-field. The second follows a multiplicative strategy in which each paper is counted as many times as necessary in the several sub-fields to which it is assigned. In this paper we adopt the multiplicative approach. This leads to a substantial increase in the total number of "papers": 42% in 1980, 45% in 1985, 48% in 1990, 56% in 1995, 58% in 1999 and 61% in 2004. However, judging from the findings obtained by Crespo et al. (2013b) using a similar dataset, we expect the results presented in this paper to be comparable to those that could have been obtained with a fractional strategy.

3.2. Descriptive statistics within each year

For each year, Table A in the Appendix of the Working Paper version of this article, Li et al. (2013), presents the number of documents by size, while Table B includes sub-field mean citations and standard deviations.[3] Within each year, it comes as no surprise that sub-fields are very different in two key respects: Firstly, they are of very different sizes. In 1980 and 1995, for example, mean sub-field sizes are 2103 and 2753, with standard deviations 2341 and 2930. The range of variation in 1995, for example, is illustrated by Andrology and Engineering, Marine with 201 and 306 documents, and Engineering, Electrical & Electromagnetic and Biochemistry & Molecular Biology with 19,938 and 34,475 documents. Secondly, reflecting the wide differences in citation practices that motivate this paper, sub-fields have very different mean citations. In 1990 and 2004, for example, mean sub-field citations are 8.4 and 8.6, while standard deviations are 7.9 and 4.4. The range of variation in 2004 is illustrated, for example, by Cell Biology and Engineering, Marine with 17.8 and 0.6 mean citations, respectively.

However, once we use scale- and size-invariant statistical techniques that allow us to focus on the shape of citation distributions, we discover that – within each year – citation distributions are extremely similar. To show this, we use the Characteristic Scores and Scales (CSS hereafter) technique, introduced by Schubert, Glänzel, and Braun (1987) in the analysis of citation distributions. For each sub-field s in a given year, we compute the characteristic scores μ_{s1} and μ_{s2} already introduced in Section 2.2. Consider the partition of sub-field citation distributions into three broad classes: documents with none or few citations below μ_{s1}; fairly cited articles, with citations above μ_{s1} and below μ_{s2}; and articles with a remarkable or outstanding number of citations above μ_{s2}. Table 1 presents the average and standard deviation over the 172 sub-fields of the percentage of articles in the three classes in every year, as well as the corresponding statistics for the percentages of the total number of citations accounted by each class.

Two points should be emphasized, one referring to the situation within each year, and another to the evolution over time. Firstly, within each year, the small standard deviations in Table 1 indicate that sub-field citation distributions are very similar. Specifically, they are highly skewed in the sense that a large proportion of articles get none or few citations while a small percentage of them account for a disproportionate amount of all citations. The evidence for more than two decades can be summarized with a single picture illustrating the partition of documents into the three classes, as well as the percentages of total citations accounted by each class (Fig. 1). As can be seen in Table 1, the situation closely resembles the one described in Albarrán, Crespo, Ortuño, and Ruiz-Castillo (2011) for articles with a common, five-year citation window

[3] Nanoscience & Technology and Robotics are missing in 1980, and Cell & Tissue Engineering in the 1980–1990 period.

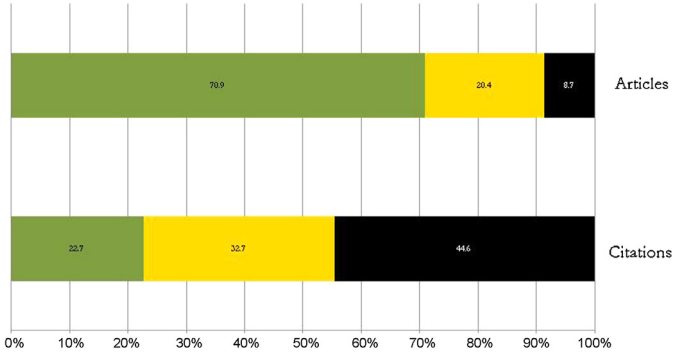

Fig. 1. Percentage of articles in three broad classes, and percentage of citations accounted for by each class (Characteristic Scores and Scales technique).

Table 2
The evolution of total citation inequality and the *IDCP* term.

Year	(1) Total citation inequality	(2) *IDCP*	(3) = (2)/(1), in %
1980	1.058	0.124	11.7
1985	1.088	0.143	13.1
1990	1.030	0.139	13.5
1995	0.966	0.137	14.2
1999	0.890	0.120	13.4
2004	0.790	0.099	12.5
Average			13.1
Std. dev.			0.9

published in 1998–2003 in a wide array of 219 sub-fields. It is important to note that, as has been emphasized in the recent literature on normalization (Crespo et al., 2013a, 2013b; Radicchi & Castellano, 2012a, 2012b; Radicchi et al., 2008), and as we will presently see in the next section, this similarity between citation distributions within each year paves the way for meaningful comparisons of citation counts across our 172 sub-fields.

Secondly, in spite of the summary illustrated in Fig. 1, it should be noted that the publication and citation percentages presented in Table 1 evolve smoothly during the 1980–2004 period. As the citation window increases from seven years for documents published in 2004 up to 31 years for documents published in 1980, sub-field citation distributions become somewhat more skewed. Specifically, for the two polar years in question, on average between 69% and 73% of all articles receive citations below the mean and only account for, approximately, between 24% and 21% of all citations, while articles with a remarkable or outstanding number of citations represent about 10% or 8% of the total, and account for, approximately, between 42% and 47% of all citations. As we will presently see in the next section, these small differences over time play a crucial role in the robustness of our results to differences in publication dates and citation window lengths.

4. Empirical results

4.1. The importance of the IDCP term over the 1980–2004 period

With the exceptions cited in note 1, the six yearly datasets are characterized by the same classification system into 172 sub-fields. However, apart from publication dates and lengths of citation windows, many features of science change over the 1980–2004 period. Specifically, the distribution of documents by sub-field and sub-field citation means vary considerably over the six years we study (see Tables A and B in the Appendix in Li et al., 2013). Therefore, the first question that should be analyzed is how total citation inequality $I(\boldsymbol{C})$ and the *IDCP* term for the raw citations depend on time.

We have just seen in Table 1 that, on average, sub-field citation distributions become less skewed as we consider closer publication dates and smaller citation windows. Correspondingly, as can be seen in column 1 in Table 2, except for a slight increase between 1980 and 1985, $I(\boldsymbol{C})$ decreases when we move in that direction.

The evolution of the *IDCP* term is more complex. To begin with, Fig. 2 illustrates what happens to the expression $I(\pi)$ as a function of π (Eq. (1) in Section 2.1) for different yearly datasets when the number of quantiles is equal to 100 (since $I(\pi)$ is too high for low percentiles, for clarity only the curves for $\pi > 50$ are included in Fig. 2).[4] It is convenient to distinguish between two regimes. Firstly, as we move from 1999 to 2004, the change in publication dates and the shortening of citation windows have a systematic effect: yearly curves move downward, and present a U shape with an intermediate percentile interval in which $I(\pi)$ remains essentially constant. Since the *IDCP* term for each year is simply a weighted average of $I(\pi)$

[4] As in Crespo et al. (2013a), all results in this paper are robust to the number of quantiles.

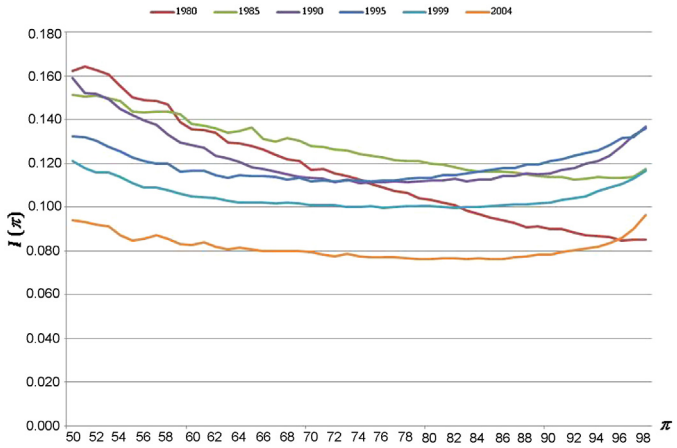

Fig. 2. Citation inequality attributable to differences in citation practices across sub-fields, $I(\pi)$ as a function of the percentile π with raw data. All yearly datasets.

Table 3
The reduction of the *IDCP* term in % as a consequence of applying the different normalization procedures.

	Median without 0s (1)	Glänzel (2)	Exchange mean rates (3)	Mean without 0s (4)	Mean (5)	Two parameters (6)	Perfect normalization (7)
1980	52.8	52.6	63.6	66.8	71.3	84.4	90.8
1985	64.3	63.3	70.9	75.0	78.0	82.4	95.1
1990	64.6	66.1	69.9	78.4	80.7	89.4	96.4
1995	62.2	70.7	67.8	81.7	83.3	94.0	96.4
1999	64.8	67.9	70.3	80.9	82.2	93.3	96.4
2004	63.9	63.5	71.5	79.4	80.8	92.5	96.5
Average	62.1	64.0	69.0	77.0	79.4	89.3	94.9
Std. dev.	4.6	6.2	2.9	5.5	4.3	5.0	2.2

expressions (see Eq. (3) in Section 2.1), we expect *IDCP* terms to decline as we move from 1999 to 2004, a fact observed in column 2 in Table 2. Secondly, changes in publication dates and/or changes in citation window lengths at the beginning of the period generate a change of regime: the curves $I(\pi)$ as a function of π for 1985 and 1980 have a negative slope and intersect some of the other curves. However, as shown in column 2 in Table 2, the *IDCP* term keeps increasing from 1999 to 1985, but slightly decreases in 1980.

Interestingly enough, the net result of these changes is that the ratio $IDCP/I(\mathbf{C})$ remains approximately constant (see column 3 in Table 2). This means that, in spite of the many differences between the six yearly datasets, the relative importance of the differences in citation practices across sub-fields is of a similar order of magnitude over the entire period, representing about 13% of total citation inequality.[5]

4.2. The relative performance of alternative normalization procedures

Before presenting the results for the six regular normalization procedures, it is convenient to assess how well the "perfect normalization" scheme does. The results of the reduction of the absolute value of the *IDCP* term are in column 7 in Table 3. On average, the best normalization procedure achievable with our data generates a 95% reduction of the *IDCP* term. Not surprisingly in view of the two regimes found in the previous section, the ability of this procedure to reduce the *IDCP* term in 1985, and above all, in 1980, is lower than in 1990–2004.

Recall that the evidence in Table 1 indicates that, within each year, sub-field citation distribution shapes are very similar. This should translate into generally good results for reasonable normalization procedures. In particular, it is instructive to begin with the $I(\pi)$ curves in Fig. 2. As pointed out already, these curves are relatively constant over a certain percentile interval during the 1990–2004 period. This is exactly what was found in Crespo et al. (2013a, 2013b, Fig. 1), indicating that, to a large extent, sub-field citation distributions behave as if they essentially differ by a scale factor of constant size over that interval. Therefore, we expect that a set of average-based exchange rates can be estimated with some precision over that interval.[6] On the other hand, using exchange rates and sub-field means as normalization factors should capture

[5] This percentage might be compared with what has been previously observed for other classification systems. See Li et al. (2013) for details.
[6] Results for exchange rates, standard deviations, and coefficients of variation are in columns 1–3 in Table C in the Appendix in Li et al. (2013).

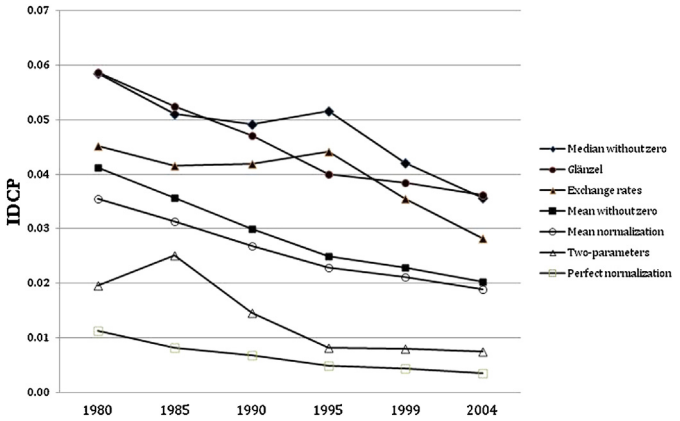

Fig. 3. A comparison of the *IDCP* term in absolute value after applying the different normalization procedures in the six yearly datasets.

Table 4
A comparison of the percentage that the *IDCP* term represents in % relative to total citation inequality after applying the different normalization procedures.

	Median without 0s (1)	Glänzel (2)	Exchange rates (3)	Mean without 0s (4)	Mean (5)	Two parameters (6)
1980	5.6	6.2	4.4	4.3	3.6	1.7
1985	4.9	5.5	4.0	3.7	3.2	2.1
1990	5.0	5.3	4.3	3.3	2.9	1.3
1995	5.6	4.8	4.8	3.0	2.7	0.8
1999	4.9	4.9	4.2	2.9	2.7	0.8
2004	4.8	5.2	3.8	2.9	2.7	0.9
Average	5.1	5.3	4.3	3.3	2.9	1.3
Std. dev.	0.4	0.5	0.3	0.5	0.4	0.5

reasonably well these scale factors separating sub-field citation distributions. This is what we find in columns 3 and 5 in Table 3. Naturally, the reduction in the *IDCP* term in 1985, and above all, in 1980, is smaller than in 1990–2004.

Interestingly enough, using sub-field means without zeros as normalization factors (column 4 in Table 3) performs better than using exchange factors but worse than using sub-field means computed over all documents. Finally, the median and the Glänzel normalization procedures (columns 1 and 2 in Table 3) do worse than the exchange rates, while the two-parameter scheme (column 6 in Table 3) performs extremely well, achieving reductions in the *IDCP* terms which are very close to the "perfect normalization" system, particularly in the 1995–2004 period with citation windows of 17, 12, and eight years. As in the former cases, the performance is much worse in 1980 than for the rest of the period.

The results of the contest are clearly illustrated in Fig. 3: the ranking of the six procedures is essentially the same over the entire period. Finally, Table 4 contains the results concerning the relative importance of the *IDCP* term after applying the different normalization schemes (excluding the "perfect normalization" that, understandably, does not perform well in

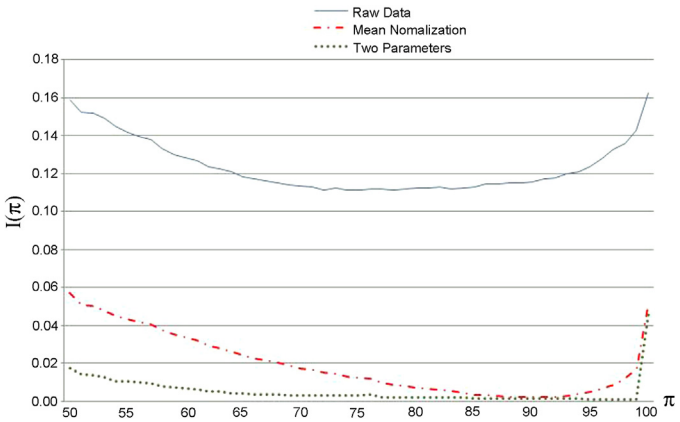

Fig. 4. Citation inequality attributable to differences in citation practices across sub-fields, $I(\pi)$ as a function of the percentile π before (raw data) and after applying the best normalization procedures to the 1990 Dataset.

relative terms because the total citation inequality in this case is extremely low). On average over the six years, the best two procedures – namely, the sub-field mean normalization and the two-parameter scheme – decrease the relative importance of the *IDCP* term from, approximately, 13%, to 2.9% and 1.3% after normalization – a remarkable improvement.

Further insight into the performance of the best normalization procedures is provided by Fig. 4, where $I(\pi)$ is plotted as a function of π both for the raw data and for the key normalization procedures for papers published in 1990 (to save space, similar results for other publication years are available on request). Note that since this quantity for the perfect normalization procedure is practically indistinguishable from the horizontal axis, it has not been included in Fig. 4. Also, because $I(\pi)$ is too large for many low percentiles, Fig. 4 only reports results for the interval (50, 100). The reduction of the $I(\pi)$ curve achieved by the best normalization procedures over the entire percentile range is clearly illustrated. At the very upper tail of citation distributions – namely, when it most matters – the performance of mean normalization clearly worsens (this is also the case with the alternative procedures not included in Fig. 4), although the two-parameter scheme keeps doing better than the rest in that interval. However, it should be emphasized that even after normalization differences in citation practices in the last percentile are very large indeed. At this level, the comparability of sub-field citation distributions becomes a much harder task.

5. Discussion and conclusions

While the use of citation numbers in research assessment exercises is becoming more and more relevant, there is still much room for the improvement of bibliometric indicators devoted to the quantification of research impact. In particular, there is a strong necessity to find proper ways of suppressing disproportions in raw bibliometric measures that are merely due to different citation practices in different fields. In this paper we have used a recently introduced measurement framework – the *IDCP* method – with two purposes: (i) to estimate the effect on citation inequality of differences in citation practices across sub-fields when using raw citation numbers, and (ii) to assess the effectiveness of six normalization procedures for reducing this effect.

We have used a dataset consisting of 2.9 million papers published in different years ranging from 1980 to 2004, and assigned to 172 distinct sub-fields according to the same classification system. Many aspects of scientific activity, including the citation process within this classification system, have changed considerably over this period. Nevertheless, this paper has unraveled a number of regularities that can be summarized as follows.

1. As observed in the past in other large datasets organized according to other classification systems, we find that, within each year, citation distribution shapes in our 172 sub-fields system are very similar to each other. Consequently, as also observed in the past, different normalization systems work reasonably well in the sense that the *IDCP* term – capturing the effect on citation inequality of differences in citation practices across sub-fields – is considerably reduced after normalization. In particular, on average over the entire period the three worst procedures reduce the *IDCP* term by 62–69%, while the three best reduce it by 77–89%. This is a remarkable result taking into account that the maximum reduction achievable with the data in the "perfect normalization" case is, on average, of 95% of the *IDCP* term.
2. As we go back in time and, consequently, citation windows become larger, two phenomena should be noted. Firstly, sub-field citation distributions become more skewed, and hence yearly overall citation inequality increases. Secondly, we must distinguish between two sub-periods: 2004–1990, in which the phenomena studied in this paper seem to evolve smoothly in time in a comparable fashion, and the 1985 and, above all, the 1980 datasets, in which the citation process seem to work quite differently. In the midst of a period of more than two decades with these features, the main results of the paper are the following two.
   (a) As we move toward earlier publication dates and greater citation windows, differences in citation practices within our classification system increase, causing the *IDCP* term to increase in absolute value in such a way that it represents, approximately, the same 13% of overall citation inequality over the entire period.
   (b) As we move back in time, normalization factors adjust to changes in overall citation inequality and to changes in the differences across sub-fields, with the following two consequences. Firstly, except for 1980, each of the six normalization procedures (as well as the "perfect" one) performs similarly over the 1985–2004 period. Secondly, the same ranking of procedures according to their ability to reduce the *IDCP* term is essentially maintained over the entire period.
3. The best normalization procedure, namely, the two-parameter reverse engineering scheme, performs very close to the "perfect normalization" over the entire support of citation distributions, except for the last 1% in which this – and, of course, the other five procedures – experience a dramatic worsening. Nevertheless, in 1995–2004 when citation windows vary between seven and sixteen years, this procedure reduces the size of the *IDCP* term by 92.5–94%, and the importance of the problem of idiosyncratic differences in citation practices from 13% to less than 1% of overall citation inequality.

As far as extensions are concerned, the following three important issues could be cited: (i) the study of normalization procedures using better classification systems; (ii) the comparison of normalization procedures based on different classification systems, or on no classification scheme at all, as in source or citing-side procedures; and (iii) the estimation of confidence intervals, so as to establish which of the differences found in this paper are statistically significant.

Acknowledgement

Ruiz-Castillo acknowledges financial help from the Spanish MEC through grant ECO2011-29762.

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012). How important is choice of the scaling factor in standardizing citations? *Journal of Informetrics*, *6*, 645–654.
Adler, R., Ewing, J., & Taylor, P. (2009). Citation statistics. *Statistical Science*, *24*, 1–14.
Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*, 385–397.
Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011a). The measurement of low- and high-impact in citation distributions: Technical results. *Journal of Informetrics*, *5*, 48–63.
Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011b). High- and low-impact citation measures: Empirical applications. *Journal of Informetrics*, *5*, 122–145.
Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*, 45–80.
Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO reports*, *14*, 226–230.
Braun, T., Glänzel, W., & Schubert, A. (1985). *Scientometrics indicators. A 32 country comparison of publication productivity and citation impact*. Singapore, Philadelphia: World Scientific Publishing Co. Pte. Ltd.
Crespo, J. A., Li, Y., Herranz, N., & Ruiz-Castillo, J. (2013). The effect on citation inequality of differences in citation practices at the web of science subject category level. *Journal of the American Society for Information Science and Technology* [in press]
Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). Differences in citation impact across scientific fields. *PLoS ONE*, *8*, e58727.
Davis, P., & Papanek, G. F. (1984). Faculty ratings of major economics departments by citations. *American Economic Review*, *74*, 225–230.
Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, *69*, 131–152.
Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association*, *295*, 90–93.
Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. *Journal of Information Science*, *37*, 40–48.
Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalization of citation indicators. The case of journal ranking. *Scientometrics*, *87*, 415–424.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of United States of America*, *102*, 16569–16572.
Kinney, A. L. (2007). National scientific facilities and their science impact on non-biomedical research. *Proceedings of the National Academy of Science of United States of America*, *104*, 17943–17947.
Leydesdorff, L., & Bornmann, T. (2012). 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*, *62*, 217–229.
Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. *Journal of Informetrics*, *4*, 644–646.
Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & de Nooye, W. (2012). Field-normalized impact factors: A comparison of rescaling versus fractionally counted ifs. *Journal of the American Society for Information Science and Technology* [in press].
Li, Y., Radicchi, F., Castellano, C., & Ruiz-Castillo, J. (2013). Quantitative Evaluation of Alternative Field Normalization Procedures, Working Paper 13–05, Universidad Carlos III (http://hdl.handle.net/10016/16741).
MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science and Technology*, *40*, 342–349.
MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*, 435–444.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*, 265–277.
Moed, H. F., & van Raan, A. F. J. (1988). Indicators of research performance. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 177–192). North Holland.
Moed, H. F., Burger, W. J., Frankfort, J. G., & van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy*, *14*, 131–149.
Moed, H. F., De Bruin, R. E., & van Leeuwen, Th. N. (1995). New bibliometrics tools for the assessment of national research performance: Database description, overview of indicators, and first applications. *Scientometrics*, *33*, 381–422.
Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*, 1113–1119.
Radicchi, F., & Castellano, C. (2012a). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.
Radicchi, F., & Castellano, C. (2012b). Testing the fairness of citation indicators for comparisons across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, *6*, 121–130.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Science of United States of America*, *105*, 17268–17272.
Schubert, A., & Braun, C. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, *9*, 281–291.
Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, *36*, 311–324.
Schubert, A., Glänzel, W., & Braun, T. (1983). Relative citation rate: A new indicator for measuring the impact of publications, in Tomov, D., & Dimitrova, L. (eds.), Proceedings of the first national conference with international participation in scientometrics and linguistics of scientific text, Varna.
Schubert, A., Glänzel, W., & Braun, T. (1987). A new methodology for ranking scientific institutions. *Scientometrics*, *12*, 267–292.
Schubert, A., Glänzel, W., & Braun, T. (1988). Against Absolute Methods: Relative Scientometric Indicators and Relational Charts as Evaluation Tools. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 137–176). North Holland.
Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2012). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, arXiv:1210.0442
Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*, 157–177.
Vinkler, P. (2003). Relations of relative scientometric indicators. *Scientometrics*, *58*, 687–694.
Waltman, L., & Van Eck, N. J. (2012). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, http://dx.doi.org/10.1007/s11192-012-0913-4 [in press]
Waltman, L. & van Eck, N.J. (2013) A systematic empirical comparison of different approaches for normalizing citation impact indicators, http://arxiv.org/abs/1301.4941
Waltman, L., van Eck, N. J., & Van Raan, A. F. J. (2011). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.
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*, 1856–1860.