Journal of Informetrics 6 (2012) 169–176

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

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

Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor

Rüdiger Mutz*, Hans-Dieter Daniel

ETH Zurich Professorship for Social Psychology and Research on Higher Education, Mühlegasse 21, 8001 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 26 August 2011
Received in revised form
20 December 2011
Accepted 21 December 2011

Keywords:
Journal impact factor
Normalization
McCall's area transformation
Rubin Causal Model

ABSTRACT

The journal impact factor (JIF) proposed by Garfield in the year 1955 is one of the most prominent and common measures of the prestige, position, and importance of a scientific journal. The JIF may profit from its comprehensibility, robustness, methodological reproducibility, simplicity, and rapid availability, but it is at the expense of serious technical and methodological flaws. The paper discusses two core problems with the JIF: first, citations of documents are generally not normally distributed, and, furthermore, the distribution is affected by outliers, which has serious consequences for the use of the mean value in the JIF calculation. Second, the JIF is affected by bias factors that have nothing to do with the prestige or quality of a journal (e.g., document type). For solving these two problems, we suggest using McCall's area transformation and the Rubin Causal Model. Citation data for documents of all journals in the ISI Subject Category "Psychology, Mathematical" (Journal Citation Report) are used to illustrate the proposal.

1. Introduction

Scientific journals differ with respect to their position and prestige within the scientific community. One of the most commonly used and prominent indicators of a journal's position and prestige is the journal impact factor (JIF), which was introduced in 1955 by Garfield (1999):

> A journal's impact factor is based on 2 elements: the numerator, which is the number of citations in the current year to any items published in a journal in the previous 2 years, and the denominator, which is the number of substantive articles (source items) published in the same 2 years. (p. 979)

At the very beginning the JIF was to aid selection of highly cited and large journals for the Science Citation Index (Garfield, 1955, 2006). Nowadays, the JIF is used to generate rankings of journals to help scientists find important journals with potential excellent (in the sense of highly cited) contributions (Todorov & Glänzel, 1988). The JIF profits much from the fact that this measure can be easily reproduced from data provided by Thomson Reuters ISI, for instance, and it is available fast in connection with other journal impact measures (e.g., immediacy index) in the Journal Citation Reports (JCR). Additionally, the JIF is comprehensible, simple, and clearly defined, and comparable over time (Glänzel & Moed, 2002). However, there are some clear flaws, which have led to controversial discussions about the correctness of using the JIF to compare and evaluate journals (e.g., Boor, 1982; Leydesdorff & Bornmann, 2011a; Moed, Van Leeuwen, & Reeduk, 1999). In their state-of-the-art report, Glänzel and Moed (2002) listed several serious flaws of the JIF: among others, normalization for reference practices in different disciplines is missing; the merits of the citing journals are not taken into consideration; the peak in citations is not

* Corresponding author.
E-mail address: mutz@gess.ethz.ch (R. Mutz).

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2011.12.006

always 2 years; the citation frequency is affected by an age bias; one single measure might not represent the prestige and the position of a scientific journal (p. 174). For instance, Neuhaus, Marx, and Daniel (2009) found in their comparative analysis (Thomson Reuters Scientific versus Chemical Abstracts Service) for wide-scope journals (*Angewandte Chemie*, *Journal of the American Chemical Society*) that the literature databases offer only a rather unreliable indicator of the document type. Further, their findings showed that the composition of the journals in terms of length of the citation windows and thematic focus of the journals have a considerable impact on the overall JIF.

In this contribution we focus on two core problems with the JIF: first, citations of articles are generally not normally distributed, and what is more, the distribution of citations is affected by extreme values or outliers (Bornmann & Mutz, 2011; Bornmann, Mutz, Neuhaus, & Daniel, 2008). This fact has serious consequences for the use of the mean value in the calculation of the JIF, because mean values react very sensitively to outliers in general. A few extremely highly cited papers suffice to result in a strongly positive bias of the JIF. The mean value is of great importance in statistics, because it is arithmetically defined. The quadratic differences between each data point $\Sigma(x-\alpha)^2$ from a parameter α is a minimum, if α is the mean value. In other words, if one does not know anything about a single data point of a distribution (e.g., number of citations), the arithmetic mean is the best and most informative value with the smallest average (quadratic) residual to the real data point, the so-called expected value. However, due to the quadratic term the mean value is also strongly influenced by skewed distributions and outliers. Alternatively, the median (50% of the data are below the median) can be used, but it is not arithmetically defined. Additionally, robust statistic (Huber, 1981) offers several methods to investigate the stability of statistical procedures if the assumptions of statistical tests are violated. Instead, we favor in our contribution an approach (McCall‘s area transformation), which not only keeps in line with the original JIF definition based on mean values, but also considers current discussion on this problem (e.g., Bornmann & Mutz, 2011; Leydesdorff & Bornmann, 2011b).

Second, “there is a wide spread belief that the ISI Impact Factor is affected or ‘disturbed’ by factors that have nothing to do with (journal) impact” (Glänzel & Moed, 2002, p. 173). Glänzel and Moed (2002, p. 178) named the following five factors that may influence and bias the JIF: document type, subject matter, the paper’s age, the paper’s social status (due to the author’s institution, for instance), and the observation period (i.e., the citation window).

In the following we propose one solution for each of the two core problems and illustrate our proposal using journal citation data for the ISI Subject Category “Psychology, Mathematical” of the JCR.

2. Skewed citation distribution: McCall’s area transformation

As mentioned above, the distributions of citations are skewed, and the mean value-based JIF is strongly affected by extreme values or outliers. As a solution to this problem we suggest *McCall’s area transformation procedure* (Krus & Kennedy, 1977; McCall, 1922), which is quite close but not redundant to the percentile approach suggested by Bornmann and Mutz (2011), Bornmann, Mutz, Marx, Schier, and Daniel, (2011), and Leydesdorff and Bornmann (2011b) with similar objectives. In contrast to an ordinary linear transformation of a scale (e.g., $5 \times X + 10$), McCall’s (nonlinear) area transformation not only transforms the skewed citation distribution into a standard normal distribution (z-distribution) but also standardizes its third moment, that is, the skewness of the distribution (Dekking, Kraaikamp, Lopuhaö, & Meester, 2005).

Similar to Leydesdorff and Bornmann (2011b) percentiles are used in the first step. Given a distribution of rank ordered citations, for each citation category a percentile rank $(100 \times p)$% can be calculated, which is the percent proportion that $(100 \times p)$% of the citations fall below this value. For instance, a percentile rank of $p = 0.20$ for articles with 50 citations means, that 20% of all articles have less than 50 citations. Whereas Leydesdorff and Bornmann (2011b) stop here and base their journal impact approach on these percentiles, we go one step further. The percentile ranks or cumulative frequencies of citations scores are transformed into z-values of the standard normal distribution. For each percentile rank a certain z-value can be accurately assigned. For instance, a percentile rank or proportion of 0.5 corresponds to a z-value of zero, a proportion of 0.975 to a z-value of 1.96. This procedure has some advantages over the pure percentile approach, suggested by Leydesdorff and Bornmann (2011b). The standard normal distribution is defined precisely by its bell curve as shape, by its area under the curve of 1.0, by its mean value of 0 and its standard deviation of 1.0. z-values can be used as an ubiquitous currency, they can be added and averaged, they can be linearly transformed $(a \times z + \mathrm{b})$ into any other scale. Outliers of citations are also considered, but due to their low proportion their z-value is shrunken towards the mean. Whereas percentiles are uniformly distributed, z-values are per se normally distributed as presumed by most statistical procedures. In the end a new scale for citations is generated with all necessary properties (e.g., normal distribution) not only for calculating the JIF, but also for any further statistical analysis.

In detail, regarding citation data the JIF calculation procedure consists of five steps: first, the citation data for each journal of a certain ISI-subject category or field are collected according to Garfield’s definition of JIF and then pooled. Second, absolute frequencies of each citation category (0, 1, 2,. . .) are calculated for each journal.

For example, journal “A” has 30 articles with no citations, 20 articles with one citation, and 10 articles with two citations. Additionally, within each citation category the journals are ranked in ascending order of the number of articles (see Table 1, column 1). Journals with more papers in a certain citation category (e.g., 40 articles with 1 citations) are ranked higher than journals with fewer papers in the same category (e.g., 20 articles with 1 citation). Second, the number of articles in a citation category per journal is converted to proportions (column 4 in Table 1). Third, the cumulative proportions (column 5 in Table 1) are transformed into z-values of a standard normal distribution using z-Tables, which are included in most introductory statistic books. If two or more journals have the same citation frequencies or rank ties (e.g., 10 articles with

Table 1
Calculation of z-values (fictitious data for two journals).

Citations	Journal	Number of articles	Proportion	Cumulative proportions	z-Value
0	A	30	0.23	0.23	−0.74
0	B	20	0.15	0.38	−0.29
1	B	40	0.31	0.69	0.50
1	A	20	0.15	0.84	1.02
2	A	10	0.08	0.92	1.43
2	B	8	0.06	0.98	2.16
5	B	2	0.02	1.00	3.72
5	A	0	0.00	–	–

Note: The cumulative proportion of 0.9999 was used instead of 1.00 to calculate the z-value.

4 citations each), these journals get the same z-value for a specific citation category (or rank position), i.e. the average z-value of these journals. The last cumulative proportion of 100% is set to 99.99% in order to avoid the problem that in the asymptotic defined z-distribution the proportion of 0 and 100% are not defined. After mean centering, the generated z-values are more or less normally distributed. It must be noted that due to possible ties the z-values may not be precisely normally distributed. Fourth, the JIF$_z$ for each journal is calculated by averaging the z-values of each journal. Unfortunately, there is no software to calculate McCall's area transformation, however, the procedure can be easily programmed using common statistical software as, for instance, SAS or STATA.

McCall's area transformation is nonlinear in the sense that the rank position of all citation categories of each journal are maintained in the z-scale (ordinal scale) but not the differences between two citation categories (interval scale). That means that the difference between two z-values (e.g., $\Delta z = 1.5–1.0 = 0.5$) does not represent any linear transformed differences between two citation categories (e.g., $\Delta\text{cit} = 2 - 1 = 1$). However, the impact of outliers and extreme values is drastically reduced according to their frequencies. A further strength of this journal impact measure lies in its clear interpretation: each JIF$_z$ represents a certain percentile rank proportion according to the z-distribution: For instance, a journal with a JIF of 1.96 has a mean percentile rank position of 97.5%.

Beirlant, Glänzel, Carbonez, and Leemans (2007) proposed a quantile plotting approach that bears a certain similarity to our approach but uses ISI JIFs for the transformation already published in JCR.

3. Bias factors: covariate adjustment

The JIF is heavily influenced by factors that have nothing to do with the prestige or quality of a journal (Glänzel & Moed, 2002, p. 173). One of the most important bias factors is the document type of citable information. There is much empirical evidence that the JIF is positively biased in favor of reviews and negatively biased to the detriment of research letters (Braun, Glänzel, & Schubert, 1989; Glänzel & Moed, 2002; Moed & van Leeuwen, 1995). If scientific journals in a certain journal set differ in their proportions of letters or reviews, then any comparisons of the journals based on their JIFs are strongly biased and unfair.

In the following, we borrow statistical concepts from randomized controlled experiments, which are the most rigorous kind of comparison between groups, to generate a solution for this bias problem. Journals might be like treatments in an experiment or an observational study that have more or less (citation) impact on the articles (units of treatments) that are published in them. In the Rubin Causal Model (Rubin, 1974, 2004) – a statistical framework that is well accepted in many fields – the causal effects of treatments are the differences between the potential outcomes that were observed under different exposures of units to treatments. For example, the causal effect of publishing in *Nature* involves comparison of the citations 2 years later with the number of citations a paper would have had, had it been published in a journal other than *Nature*. $Y(0)$ is the outcome (i.e., number of citations) without publishing in *Nature*, and $Y(1)$ is the outcome with publishing in *Nature*, and the difference $Y(1) - Y(0)$ is the individual causal effect of publication in *Nature* on the number of citations. The fundamental problem of causal inference is that only one potential outcome can be observed at the same time (Holland, 1986). In other words, one of the potential outcomes is always missing.

To make correct causal inferences (unbiased comparisons), we must have a model for how units get assigned to treatments. One of the most important assignment mechanisms is randomization: all articles in a certain field are randomly assigned to the journals in a special journal set, and, therefore, each article has the same probability to be assigned to *Nature*, for instance. In the end, the journals no longer differ in any factors (e.g., proportions of document types), which guarantees the JIF as a mean value across the observed potential outcomes to be an unbiased measure of journal impact. However, in bibliometrics it is totally unrealistic to randomly assign papers to journals. In this case covariates can help to identify the assignment mechanism (Cochran, 1968; Rubin, 1977) and afterwards correct the mean values (i.e., JIF) for the covariates. Covariates are properties of documents that have an impact on the number of citations but are not themselves causally affected by the journals that are under study. The covariate as bias factor not only shapes or affects the number of citations but also differs in its frequency distribution between different journals. The rationale of the procedure will be illustrated in the following example of two journals (see Table 2)

Table 2
Number of documents and mean number of citations for two journals and two document types (fictitious data for two journals).

Document type	Journal A		Journal B		Total N_{Art}
	N_{Doc}	Mean_{cit}	N_{Doc}	Mean_{cit}	
Review	16	80	8	60	24 (60%)
Research Article	4	10	12	40	16 (40%)
Total	20	66	20	48	40 (100%)

N_{Doc}, number of documents; Mean_{cit}, mean number of citations.

Journal A publishes mostly reviews (16 of 20 papers), whereas journal B publishes predominately research articles (12 of 20 papers). Due to this bivariate frequency distribution, the assignment of the documents to journal A or B depends strongly on the document type (Cramer's $V = 0.41$, $\chi^2(1) = 6.7$, $p < .05$). Cramer's V is a measure of the relation between two categorical variables, and vary between −1.0 (perfect negative relationship) and 1.0 (perfect positive relationship). In the result, there is an obvious mean difference or prima facie effect between the two journals of 66 − 48 = 18 citations.

However, the observed prima facie effect is confounded due to a strong bias in favor of journal A. To remove this bias, the conditional means (review, research articles) are weighted by the marginal distribution of the covariate: that is, by 60% or .60 for the reviews, and by 40% or .40 for research articles, respectively. In the end, this kind of adjustment procedure removes entirely the mean differences between the two journals in the number of citations: the mean number of citations amounts, finally, to $0.6 \times 80 + 0.40 \times 10 = 52$ for journal A, and $0.6 \times 60 + 0.40 \times 40 = 52$ for journal B, respectively. As an alternative, it is also possible to calculate *conditional causal effect* for each level of the covariate and combine them afterwards using the marginal frequencies to calculate the average causal effect: that is $(80 - 60) \times 0.60 + (10 - 40) \times 0.40 = 0$. In the case of two or three categorical covariates, the combinations of the two or three covariates can be used. In the case of more than three covariates or continuous variables, propensity score analysis offers a statistically well-founded option for adjustment (e.g., Guo & Fraser, 2010; Rubin, 2004, 2006). It must be noted that this procedure is not identical to the usual covariance adjustment in analysis of variance (ANOVA).

To make causal inferences in general it must be guaranteed that the treatment assignment does not depend on the potential outcomes $Y(0)$ and $Y(1)$ given the covariates, which is called the "strong ignorability" condition. For a detailed introduction to the statistical background of causal inference and further assumptions such as SUTVA ("stable unit treatment value assumption"), see, Rubin (2004, 2006), for example.

4. Application: ISI Subject Category "Psychology, Mathematical"

Following Garfield's (1955, 2006) definition, we retrieve the JIF data for the year 2010 of all journals of the SCI section "Psychology, Mathematical." The data comprise the total number of citations in the year 2010 of all citable documents (articles, letter, reviews, etc.) that were published in the years 2008 and 2009 in the 11 scientific journals of the Thomson ISI Subject Categories, as mentioned above. There are no special reasons for choosing this journal set, except that the process of data generation for the chosen ISI Subject Category should not be too costly. However, it cannot be denied that there is a critical discussion in the field of psychology about the use and misuse of JIFs for comparing journals (Anseel, Duyck, De Baene, & Brysbaert, 2004; Boor, 1982; Egloff, 2006; Rotton, Levitt, & Foos, 1993; Schweizer, 2010). Unfortunately, as Glänzel and Moed (2002) point out, the journal impact factors included in the Science Citation Index or the Social Sciences Citation Index are inaccurate:

> In particular, ISI classifies documents into types. In calculating the numerator of the IF, ISI counts citations to all types of documents, whereas as citable documents in the denominator ISI includes as a standard only research articles, notes, and reviews. (p. 181)

Editorials, letters to the editor, and other types of documents, when they are cited, are not included in the denominator of the ISI JIFs. Thus, the JIFs published in JCR (not yet published at the time of submitting this paper) may not fully agree with the JIFs calculated in our study.

In the first step, we applied McCall's area transformation to transform the citation data for all articles of each journal to z-values (see Fig. 1). For each citation number and journal, a z-value of the cumulated probability of articles was calculated. As mentioned above, the citation data of articles are strongly skewed and heavily influenced by outliers or extreme values (Fig. 1a). However, McCall's area transformation performs almost perfectly: the histogram of the z-values approximates a normal distribution with mean value of zero. The null hypothesis that the data are normally distributed cannot be rejected (Shapiro–Wilk $W = 0.98$, $p \geq 0.05$).

In the second step, the different kinds of JIFs were calculated, separately for each journal in the ISI Subject Category (Table 3) for the year 2010. Although the journals vary strongly both in the number of publications and the total number of citations, the journal impact factors in terms of SCR (SCR JIF) differ only slightly, from 0.30 (*Applied Measurement in Education*) to 2.25 (*Behavior Research Methods*), or 2 citations per document on the average. These small differences are maintained in the z-values, which vary only within about one standard deviation [−1.05, 0.68].

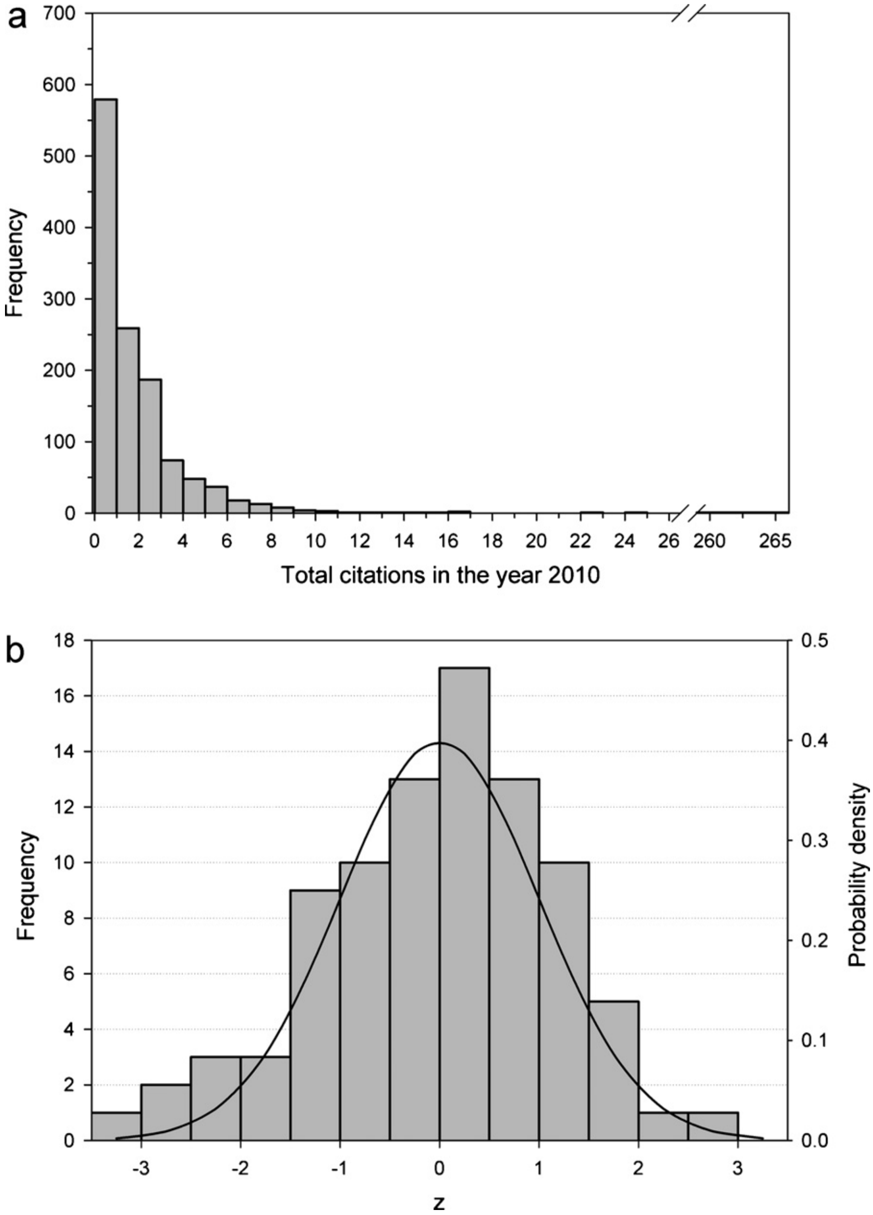

Fig. 1. Total citations in the year 2010 of all articles (ISI Subject Category "Psychology, Mathematical") published in the year 2008 and 2009: (a) raw data. (b) McCall area-transformed z-data (with the normal distribution as curve).

For instance, the journal *Behavior Research Methods* achieves the highest percentile rank of 0.75. All articles published in this journal in the years 2008 and 2009 achieve on the average 75% of the citation distribution of all documents in all journals in the ISI Subject Category "Psychology, Mathematical." Journals with z-values of zero or below, such as *Journal of Mathematical Psychology* or Applied *Measurement in Education*, perform on the average 50% or well below the average with respect to the distribution of the total citations across all citation categories in all journals.

In the third step, the journal impact factor is corrected for the document type. Due to the low frequencies, only two categories of document types were chosen: research articles and letters or other documents (reviews, editorials, etc.). The two conditions for covariates were somewhat fulfilled: first, the journals differ statistically significantly in their frequencies of the two document types ($\chi^2(10) = 37.9$, $p < .05$, Cramer's $V = 0.17$). Second, an analysis of variance (ANOVA) that bases on the McCall transformed z-values of citation categories across journals and document types shows a significant main effect between journals ($F(10, 96) = 2.72$, $p < .05$), document types ($F(1, 96) = 21.67$, $p < .05$) but no significant interaction ($F(10, 96) = 1.01$, $p \geq .05$). However, ANOVA does not use the statistically correct terms for causal inference. Next, following the procedure outlined above, the unconfounded journal impact factors (cJIF$_z$) are calculated using the relative frequency of the document types (87.6% are research articles, 12.4% are other documents) as marginal frequency and weight in the

Table 3
JIF values for the ISI Subject Category (SC) "Psychology, Mathematical" for the year 2010, sorted by the SCR JIF values.

Journal	No. of documents	No. of citations	SCR JIF	JIF_z	r_{JIFz}	pr_{JIFz}	$cJIF_z$	r_{cJIFz}	pr_{cJIFz}
Behavior Research Methods	275	619	2.25	0.68	1	0.75	0.80	2	0.79
Psychonomic Bulletin	329	712	2.16	0.63	2	0.74	0.82	1	0.79
Journal of Educational and Behavioral Statistics	48	69	1.44	−0.37	7	0.36	−0.33	7	0.37
Psychometrika	97	128	1.32	0.25	3	0.60	0.19	3	0.58
British Journal of Mathematical and Statistical Psychology	64	78	1.22	−0.18	5	0.43	−0.12	6	0.45
Journal of Mathematical Psychology	94	111	1.18	−0.09	4	0.46	0.00	4	0.50
Applied Psychological Measurement	82	78	0.95	−0.52	9	0.30	−0.37	8	0.35
Journal of Educational Measurement	52	41	0.79	−0.85	10	0.19	−0.61	10	0.27
Journal of Classification	37	29	0.78	−0.41	8	0.34	−0.39	9	0.35
Educational and Psychological Measurement	118	89	0.75	−0.25	6	0.40	−0.10	5	0.46
Applied Measurement in Education	44	13	0.30	−1.05	11	0.15	−1.00	11	0.15

No., number; SCR JIF, JIF calculated according to the SCR definition; JIF_z, JIF calculated based on area-transformed *z*-data; r_{JIFz}, rank position of a journal regarding JIF_z; pr_{JIFz}, percentile rank of JIF_z; $cJIF_z$, JIF_z corrected for document type; r_{cJIFz}, rank position of the journal regarding $cJIF_z$; p_{cJIFz}, percentile rank of $cJIF_z$.

Table 4
Rank correlation (Kendall's tau) between the different JIFs and other variables (N = 11 journals).

	SCR JIF	JIF_z	$cJIF_z$	No. of publications	Total no. of citations
SCR JIF	1.00				
JIF_z	0.64*	1.00			
$cJIF_z$	0.60*	0.89*	1.00		
Number of publications	0.42	0.64*	0.75*	1.00	
Total number of citations	0.59*	0.81*	0.92*	0.84*	1.00

* $p < .05$.

calculation of the average values or average causal effects. In the result, the unconfounded journal impact factors ($cJIF_z$) do not differ much from the uncorrected ones. The *z*-values and percentile ranks of the best journals are somewhat higher for the corrected JIF_z than the corresponding values for the uncorrected ones (e.g., 0.80 instead of 0.68 for the journal *Behavior Research Methods*).

In the last step, rank correlations (Kendall's tau) between the three kinds of JIFs, the number of articles, and total number of citations across journals were calculated (Table 4). The usual journal impact factor calculated by Thomson ISI (SCR) correlates only moderately ($r_k \sim 0.6$) with the new journal impact factors (JIF_z, $cJIF_z$) – that is, the ranks of journals do not fully agree regarding the new and the old JIFs (Table 3). In other words, the proposal adds value to the traditional concept of the journal impact measure. Nevertheless, the corrected and the non-corrected JIF_z are highly correlated due to the fact that the impact of the document type as a bias factor is all in all rather low for the journal set "Psychology, Mathematical". The adjustment procedure itself, however, may not be to blame for these results. In contrast to the Thomson ISI journal impact factor, the new JIFs depend strongly on the total number of citations, indicated by high correlations between these variables and total number of citations ($r_k = .81$ and $r_k = .92$, respectively).

5. Conclusions

Without any doubt, the journal impact factor proposed by Garfield in the year 1955 (Garfield, 1955, 2006) is still one of the most prominent and common measures of the prestige, position, and importance of a scientific journal. The JIF may profit from its comprehensibility, robustness, methodological reproducibility, simplicity, and rapid availability but at the expense of serious technical and methodological flaws, as Glänzel and Moed (2002) outlined in their state-of-the-art report, for instance. Our contribution deliberately looked at only two problems, but they are core problems with the journal impact factor: the use of mean values in the face of skewed citation distributions and outliers, and the influence of factors that have nothing to do with the quality of a journal (called bias factors). As solutions to these problems, we discussed McCall's area transformation (McCall, 1922) and the Rubin Causal Model (RCM) (Rubin, 2004, 2006).

In a recent publication Glänzel, Schubert, Thijs, and Debackere (2011) differentiated between a priori and a posteriori normalization of citation indicators. A posteriori normalization comprises all mathematical transformations of standard journal impact factors, as they are available from JCR, for instance. The *quantile plotting* approach (Beirlant et al., 2007) and the characteristic scores and scales approach (Glänzel, 2011) are prominent examples of an a posteriori normalization procedure. In contrast, a priori normalization uses sources of individual documents in a "paper-to-paper" solution. Examples are the SNIP indicator proposed by Moed (2010), the fractional citation counting approach suggested by Leydesdorff and Opthof (2010) or the percentile approach proposed by Leydesdorff and Bornmann (2011b). In the light of our results, we argue strongly for a priori normalization procedures to consider outliers, and skewed citation distributions, respectively, and to adjust for bias factors. In our point of view, it makes no sense to normalize JCR journal impact factors afterwards that might in their very nature be affected by outliers and bias factors.

Additionally, the RCM offers a rigorous definition of bias factors in combination with methods to correct the journal impact for these influences: a bias factor is defined as a covariate that is, first, not influenced by the journal itself. Second, its frequency distribution differs between the journals of the respective journal set, and, third, the various levels of the covariate shows different effects on the total number of citations. This definition fully applies to the covariate "type of documents" (Braun et al., 1989). Scientific journals differ in the frequencies of certain documents, especially reviews or research articles. Additionally, different document types show different levels of total citations. For instance, reviews attract on average more citations than articles or letters do. It must be noted that the literature databases offer only a rather unreliable indicator of the document type (Neuhaus et al., 2009).

In conclusion, we would even go so far as to say that any kind of comparisons between journals or research groups or any comparison with reference values in bibliometrics in general must be based on such a causal framework to justify the fairness of the results. The suggested proposal might be very promising; however, it cannot be denied that there are some limitations:

- *Sampling dependency*: The empirical example presented can only illustrate the proposal. The empirical results depend strongly on the chosen data (i.e., year, JCR subject category).
- *Nonlinearity of the transformation*: McCall's area transformation is a nonlinear transformation that changes the status of the scale. Strictly speaking, in this case mean values may not be calculated. However, the concept of expected value proposed by Bornmann and Mutz (2011) still holds.
- *Assignment mechanism*: To adjust the journal impact factors correctly, the true assignment mechanism of the documents to the journals must be known. In bibliometrics we do not know exactly the actual assignment mechanism. In the propensity score analysis a whole set of many empirical covariates may help to describe the assignment mechanism and to adjust the citation indicators correctly.
- *Covariates*: The journal impact factors can only be corrected for covariates that differ in their frequencies between journals. For discipline, for instance, which might not vary within a specific journal set, the proposed adjustment procedure cannot be applied.

In spite of these limitations, we guess that the proposal offers some promising improvements of the journal impact factors, but further, detailed elaborations of the statistical background are needed.

References

Anseel, F., Duyck, W., De Baene, W., & Brysbaert, M. (2004). Journal impact factors and self-citations: Implications for psychology journals. *American Psychologist*, *59*, 49–51.

Beirlant, J., Glänzel, W., Carbonez, A., & Leemans, H. (2007). Scoring research output using statistical quantile plotting. *Journal of Infometrics*, *1*, 185–192.

Boor, M. (1982). The citation impact factor: Another dubious index of journal quality. *American Psychologist*, *37*(8), 975–977.

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 Infometrics*, *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*(Part 4), 857–879.

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

Braun, T., Glänzel, W., & Schubert, A. (1989). Some data on the distribution of journal publication types in the Science Citation Index Database. *Scientometrics*, *15*, 325–330.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313.

Dekking, F. M., Kraaikamp, C., Lopuhaö, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics: Understanding why and how*. New York, NY: Springer.

Egloff, B. (2006). Some remarks on impact factor. *Psychologische Rundschau*, *57*(2), 116–118.

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

Garfield, E. (1999). Journal impact factor: A brief review. *Journal of the Canadian Medical Association*, *161*(8), 979–980.

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *Journal of the American Medical Association*, *295*(1), 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*(1), 40–48.

Glänzel, W., & Moed, H. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics*, *87*, 415–424.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis—Statistical methods and applications*. London, UK: Sage.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley-Interscience.

Krus, D. J., & Kennedy, P. H. (1977). Lost: McCall's T scores: Why? *Educational and Psychological Measurement*, *37*, 257–261.

Leydesdorff, L., & Bornmann, L. (2011a). How fractional counting of citations 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*(2), 217–229.

Leydesdorff, L., & Bornmann, L. (2011b). Integrated indicators compared with impact factors: An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology*, *62*(11), 2133–2146.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, *61*(11), 2365–2369.

McCall, W. A. (1922). *How to measure in education*. New York, NY: Macmillan.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.

Moed, H. F., & van Leeuwen, T. N. (1995). Improving the accuracy of the Institute for Scientific Information's Journal Impact Factor. *Journal of the American Society of Information Science*, *46*, 461–467.

Moed, H. F., Van Leeuwen, T. N., & Reeduk, J. (1999). Towards appropriate indicators of journal impact. *Scientometrics*, *46*(3), 575–589.

Neuhaus, C., Marx, W., & Daniel, H.-D. (2009). The publication and citation impact profile of *Angewandte Chemie* and the *Journal of the American Chemical Society* based on the sections of *Chemical Abstracts*: A case study on the limitations of the Journal Impact Factor. *Journal of the American Society for Information Science and Technology*, *60*(1), 176–183.

Rotton, J., Levitt, M. J., & Foos, P. (1993). Citation impact, rejection rates, and journal values. *American Psychologist*, *48*, 911–912.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*(1), 1–26.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29*(3), 343–367.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, UK: Cambridge University Press.

Schweizer, K. (2010). Judging a journal by the impact factor: Is it appropriate and fair for assessment of journals? *European Journal of Psychological Assessment*, *26*(4), 235–237.

Todorov, R., & Glänzel, W. (1988). Journal citation measures: A concise review. *Journal of Information Science*, *14*, 47–56.