Scientometrics (2012) 92:377–390
DOI 10.1007/s11192-012-0670-4

The generalized propensity score methodology for estimating unbiased journal impact factors

Rüdiger Mutz · Hans-Dieter Daniel

Received: 31 January 2012 / Published online: 17 February 2012

Abstract The journal impact factor (JIF) proposed by Garfield in the year 1955 is one of the most commonly used and prominent citation-based indicators of the performance and significance of a scientific journal. The JIF is simple, reasonable, clearly defined, and comparable over time and, what is more, can be easily calculated from data provided by Thomson Reuters, but at the expense of serious technical and methodological flaws. The paper discusses one of the core problems: The JIF is affected by bias factors (e.g., document type) that have nothing to do with the prestige or quality of a journal. For solving this problem, we suggest using the generalized propensity score methodology based on the Rubin Causal Model. Citation data for papers of all journals in the ISI subject category "Microscopy" (Journal Citation Report) are used to illustrate the proposal.

Keywords Journal impact factor · Causal inference · Generalized propensity score · Rubin Causal Model

Introduction

One of the most commonly used and prominent citation-based indicators of the performance and significance of a scientific journal 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).

R. Mutz (✉) · H.-D. Daniel
Social Psychology and Research on Higher Education, ETH Zurich, Muehlegasse 21, 8001 Zurich, Switzerland
e-mail: mutz@gess.ethz.ch

H.-D. Daniel
Evaluation Office, University of Zurich, Muehlegasse 21, 8001 Zurich, Switzerland

In the early days, the JIF provided for the selection of large and highly cited journals for the Science Citation Index (Garfield 1955, 2006). Nowadays, the JIF is used to find important journals with excellent in the sense of highly cited contributions (Todorov and Glänzel 1988). The JIF is simple, reasonable, clearly defined, and comparable over time and, what is more, can be easily calculated from data provided by Thomson Reuters (Glänzel and Moed 2002). However, the JIF reveals some substantial flaws, which have provoked severe discussions about the use of JIF to evaluate and compare journals (e.g., Moed et al. 1999; Leydesdorff and Bornmann 2011; Neuhaus et al. 2009). Glänzel and Moed (2002) gave a comprehensive overview of the several flaws of the JIF (e.g., normalization for reference practices in different disciplines, the two-year interval) in a state-of-the-art report. Recently, Vanclay (2012) stresses some technical limitations of the JIF such as the fallacy of overprecision (display with three decimals), missing confidence intervals, and the lack of matches between citing and cited papers.

In this contribution we focus on the following core problem with the JIF. "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 and Moed 2002, p. 173). Glänzel and Moed (2002, p. 178) mentioned five core factors that may influence the JIF: document type, the paper's age, the author's social status (due to the author's institution, for instance), subject matter, and the time interval of observation (i.e., the citation window). In a previous publication (Mutz and Daniel 2012) we recommended stratification on a single covariate based on Rubin Causal Model (Rubin 1974, 2004) as a statistical tool to correct the JIF for bias factors. However, this approach is restricted to a small set of bias factors (1–2 covariates). In this contribution we propose a more general approach to correct and adjust the JIF for an arbitrary number of bias factors, the so-called "generalized propensity score" methodology (Imai and van Dyk 2004; Imbens 2000; Zanutto et al. 2005).

In the following, the statistical background of generalized propensity scores are outlined. Next, the proposal will be illustrated using citation data for the subject category "Microscopy" of the ISI Journal Citation Report (JCR).

The generalized propensity score methodology

The JIF is strongly affected by factors that have nothing to do with the significance and performance of a scientific journal (Glänzel and Moed 2002, p. 173). One of the most important factors influencing the JIF is the document type of citable information (e.g., articles, reviews, letters). There is much empirical evidence that the JIF is positively biased in favor of reviews, because on average reviews are more cited than, for example, articles (Braun et al. 1989; Glänzel and Moed 2002; Moed and van Leeuwen 1995). However, if scientific journals in a certain journal set vary in their proportions of reviews in an certain observational interval, then any comparisons of journals based on JIF rankings are strongly biased and, therefore, unfair.

In the following, statistical concepts for the analysis of experimental designs (e.g., randomized controlled experiments, observational studies) are adopted to solve this problem. Experimental designs are very common in medicine and social sciences to test a causal hypothesis of certain treatments (e.g., impact of drugs) in comparison to an untreated control. As matter of fact, journals might be like treatments in an experimental design that vary in their impact (i.e. amount of citations) on the papers (treatment units) that are published in them. One prominent statistical methodology to analyze experimental design is the so-called Rubin Causal Model (Rubin 1974, 2004) and its "potential

outcome" concept, respectively. Causal effects are defined as the differences between potential outcomes that where measured under different exposures of the same units (i.e. articles) to treatments (i.e. journals). For instance, the causal effect of publishing a single paper i in *Science* involves comparison of the outcome (e.g., number of citations) two years later with the outcome a paper would have received, had it been published in a journal other than *Science*. If $Y_i(1)$ is defined as the outcome with publishing in *Science* ($t = 1$) and $Y_i(2)$ is defined as outcome publishing in another journal ($t = 2$), then the difference $\text{ICE}_i = Y_i(1) - Y_i(2)$ is the individual causal effect of publication in *Science* on the respective outcome as, for instance, number of citations. Let S_i the indicator (random variable) of publishing a paper i in the journal t (Imbens 2000, p. 707):

$$S_i(t) = \begin{cases} 1 & \text{if } T_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

However, for a single paper only one potential outcome ($Y_i(1)$ or $Y_i(2)$) can be observed at the same time, one potential outcome is always missing. This is the fundamental problem of causal inference (Holland 1986). For paper i, the realized outcome Y_i in contrast to the potential outcome $Y_i(\text{t})$ can be expressed as

$$Y_i = \sum_{t=1}^{m} Y_i(t) \cdot S(T_i = t), \tag{2}$$

where $m = 2$ treatments. Note, that for a paper i only one potential outcome is observed (either for journal 1 or for journal 2).

Instead, we can estimate the average causal effect (PFE) for journal 1 versus journal 2 in a population of papers, which is defined as $\text{ACE}_i = E(Y_i(1)) - E(Y_i(2))$ with $E(Y_i(1))$ is nothing but the observed JIF of *Science* and $E(Y_i(2))$ is the observed JIF of the comparison journal. However, the observed mean differences or prima facie effect between the two journals equals the average causal effect ACE only, if the papers are fully randomly assigned to the journals. In this case each paper has the same probability to be assigned to *Science*. Through randomization the journals no longer differ in any factors (e.g., proportions of document types). As a consequence, first, the prima facie effects are the true estimates of the average causal effect. Second, the JIF as the expected value across the observed potential outcomes is an unbiased estimate of the true impact of a journal. However, in bibliometrics it is illusory to assume a random assignment mechanism for papers to journals. The idea of the RCM is to introduce pre-submission covariates $\mathbf{X}$ (e.g., number of references, type of document, number of pages), in order to identify the assignment mechanism. The ideal covariate concurrently correlates with the journal assignment and influences the mean citations, but is not itself affected by the respective journal. For example, journals vary with respect to the proportion of various document types. Various document types, differ with respect to mean citations. Reviews attract on average more citations than articles or letters do. The findings of the study of Neuhaus et al. (2009) regarding the journal "Angewandte Chemie" and "The Journal of the American Chemical Society" showed that information available in the Science Citation Index is a rather unreliable indication of the document type and is, therefore, inadequate for comparative analysis. Therefore, the number of references should be used as covariate or proxy of the real document type.

The average causal effects are unbiased or unconfounded, if the following so-called *strong ignorability condition* can be held (Rubin 2007):

$$Y(t) \perp S(t) | \boldsymbol{X} \tag{3}$$

The potential outcomes $Y(t)$ are independent of the journal assignment $S(t)$ given the covariates $\mathbf{X}$. In the case of one covariate the strong ignorability condition can be easily satisfied by comparing the JIFs for each subclass of the covariate (e.g., document types) or by summing these JIFs across the subclasses weighted by the marginal frequencies (e.g., frequencies of the document types across the two journals) to yield an overall unconfounded JIF for each journal (Cochran 1968; Mutz and Daniel 2012; Rubin 1977, 2006).

In a fictitious example (see Table 1) journal 1 is positively biased in favor for reviews ($N = 80$ reviews, $N = 20$ articles). The overall prima facie effect (PFA) between the two journals amounts to $62-38 = 24$ citations. However, the mean differences between the two journals with respect to the different document types, as well as the overall mean difference (i.e. ACE) weighted by the marginal frequency 0.50, are zero ($0.50 \times [70 - 70] + 0.50 \times [30 - 30] = 0$). In sum, the unconfounded average causal effect is zero and the unconfounded, i.e. unbiased JIFs amount both to 50.

In the case of a set of covariates propensity scores are more appealing (e.g., Guo and Fraser 2010). The propensity $r(t = 1, \mathbf{X})$ is the probability that a paper is assigned to journal $t = 1$ given a set of covariates. The propensity scores can be estimated by an ordinary logistic regression ($\log(r/(1 - r) = \mathbf{X}\boldsymbol{\beta}$), and are not only obtained for papers published in journal $t = 1$, but also for papers published in journal $t = 2$. Identical propensity scores in both journals reflect a balance in the distribution of the corresponding set of covariates $\mathbf{X}$ in journal 1 and journal 2. Thus, the strong ignorability condition (Eq. 3) can be transformed to

$$\begin{gathered} Y(t) \perp S(t) | \, r(t, \boldsymbol{X}) \\ 0 < r(t, \boldsymbol{X}) < \boldsymbol{1}, \end{gathered} \tag{4}$$

in the sense, that the potential outcomes Y(t) are independent from the journal assignment S(t) given the propensity scores $r(t, \mathbf{X})$. Please note, that the propensity scores vary within the interval [0, 1]. Therefore, covariates must be excluded from data analysis, which definitely predict the assignment to a journal ($r(t, \mathbf{X}) = 1$) or not ($r(t, \mathbf{X}) = 0$). For instance, a covariate "document type—review or not" must be excluded, if there are any journals in the journal set, which only publish reviews.

To satisfy the strong ignorability condition, the stratification is on the propensity scores instead of stratification on a single covariate as shown in Table 1. 5 strata (quintiles) might be sufficient to remove about 90% of initial biases in the journals (Rosenbaum and Rubin 1984, p. 521). In order to apply the propensity scores for balancing the two groups, it must

Table 1 Number of papers and mean number of citations for two journals and two document types (fictitious data)

Document type	Journal 1		Journal 2		Total N_{pap}
	N_{pap}	$Mean_{cit}$	N_{pap}	$Mean_{cit}$	
Review	80	70	20	70	100 (50%)
Research article	20	30	80	30	100 (50%)
Total	100	62	100	38	200 (100%)

N_{pap} number of papers, $Mean_{cit}$ mean number of citations

be guaranteed that there is an overlapping between the distributions of propensity of the two journals. A further assumption is the *Stable Unit Treatment Value Assumption* (SUTVA), which claims, that the potential outcomes of one paper are not affected by the journal assignment of any other paper (Rubin 2005, p. 323). This assumption is violated, for example, if similar papers are published in two journals. Then it might be the case, that the publication of the paper in journal B has an impact on the citations of the similar paper in journal A. However, such rare cases can be neglected.

The classical propensity score methodology is restricted to binary cases (e.g., treatment versus control, journal A versus journal B). However, there are several attempts to embed this concept in a larger class of a so-called *generalized propensity score* methodology for multi-valued treatments (Imai and van Dyk 2004; Imbens 2000; Feng et al. 2011; Kluve et al. 2012; Lu et al. 2011; Rosenbaum 2010; Spreeuwenberg et al. 2010; Wang et al. 2001; Zanutto et al. 2005). The treatments may be categorical with more than two groups, ordinal with ranked treatments (e.g., dose–response relationships), or continuous. If the strong ignorability assumption (Eq. 2) is held, then the entire set of m potential outcomes of a paper must be independent from the m journals in a SCR journal set given the covariates. In order to avoid this rather strong assumption, Imbens (2000, p. 707) introduced the concept of *weak unconfoundedness*. The assignment of a journal (treatment) T is weakly unconfounded given pre-submission (pre-treatment) covariates $\mathbf{X}$ (or the generalized propensity scores r), if Eq. 3 (or Eq. 4) is held, where $S_i(t)$ equals 1, if paper i is published in journal t, and 0, if paper i is published in any other journal of the journal set. In other words, the binary comparison of two journals is replaced by the binary comparison of a special journal with the entire journal set with the respective journal removed from the set.

Let T be multinomially distributed, then the generalized propensity scores could be estimated by a multinomial logistic regression (Imai and van Dyk 2004, p. 856). For m journals, there are m sets of generalized propensity scores for paper i (Feng et al. 2011):

$$\begin{aligned} r(1, \boldsymbol{X_i}) &= p(T = 1|\boldsymbol{X_i}) \\ r(2, \boldsymbol{X_i}) &= p(T = 2|\boldsymbol{X_i}) \\ &\cdots \\ r(m, \boldsymbol{X_i}) &= p(T = m|\boldsymbol{X_i}) \\ \text{where } r(1, \boldsymbol{X_i}) + r(2, \boldsymbol{X_i}) + \cdots + r(m, \boldsymbol{X_i}) &= 1.0 \end{aligned}$$

Let $\beta(t,\ r)$ denote the expected outcome (mean citation) of a paper in journal t given generalized propensity score $r(t, \mathbf{X}) = r$. If the journal assignment is weakly unconfounded given the covariates $\mathbf{X}$, for all journals ($t = 1$ to T) the following results are obtained (Feng et al. 2011; Imbens 2000, p. 708):

$$\beta(t,\ r) = E\{Y(t)|r(t,\ \boldsymbol{X}) = r\} = E\{Y|T\ = t,\ r(T,\ \boldsymbol{X}) = \boldsymbol{r}\} \tag{5}$$

The expected value of a paper in journal t with respect to the distribution of $r(t, \mathbf{X})$ is (Imbens 2000, p. 708)

$$E\{Y(t)\} = E\{\beta(t,\ r(t,\ \boldsymbol{X}))\}, \tag{6}$$

where $E\{Y(t)\}$ is the unconfounded JIF of a journal.

Following Imbens (2000, p. 708), Feng et al. (2011) offer a *propensity score weighting formula* to finally estimate the unconfounded expected value for a journal, and the unconfounded JIFs, respectively:

$$\mathrm{JIF}_t = \hat{E}\{Y(t)\} = \left[\sum_{i=1}^{n} \frac{Y_i \cdot S(T_i = t)}{r(t, \mathbf{X}_i)}\right] \cdot \left[\sum_{i=1}^{n} \frac{S(T_i = t)}{r(t, \mathbf{X}_i)}\right]^{-1} \quad (7)$$

where the generalized propensity scores $r(t, \mathbf{X})$ are normalized (summed to 1.0 in each journal t). Alternatively, the stratification method, mentioned above, can be applied using the m propensity scores for each paper.

Materials and methods

According to Garfield's (1955, 2006) definition, we retrieve the JIF data for the year 2010 of all journals of the SCI section "Microscopy". The data encompass the total number of citations in the year 2010 of all citable papers (articles, letter, reviews, etc.) that were published in the years 2008 and 2009 in the 9 scientific journals of the Thomson ISI SCR section, as mentioned above ($N = 2{,}138$ papers). There are no special reasons for choosing this journal set, except that the process of data generation for the chosen SCR subject category should not be too costly.

Regretfully, as Glänzel and Moed (2002) point out, the JIFs included in the Science Citation Index or the Social Sciences Citation Index is inaccurate:

> In particular, ISI classifies papers into types. In calculating the numerator of the IF, ISI counts citations to all types of papers, whereas as citable papers 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 papers, when they are cited, are not included in the denominator of the ISI JIFs. Thus, the JIFs published in JCR may not fully agree with the JIFs calculated in our study.

For balancing purposes the following four characteristics of the papers are included in the generalized propensity score estimation (Table 2): the publication year ($0 = 2008$, $1 = 2009$), the number of authors, the number of pages, and the number of references. The paper type was not included in the analysis for two reasons: First, one journal of the journal set, "Micron—The International Research and Review Journal for Microscopy", publishes only papers, which are coded as reviews in ISI Web of Science.

Second, the ISI-document type might not be an accurate indicator of the true document type. For instance, the journal Micron does not only aim at publication of reviews (www.journals.elsevier.com/micron/). Instead, we include the number of references as an indicator of document type, which also might not be affected by the journal assignment. The higher the number of references, the more the paper reviews literature. Rubin and Thomas (1996) recommended to include in the propensity score estimation even unimportant covariates at the expense of efficiency loss. It is better to include unimportant covariates than to enhance bias by leaving out important covariates.

In the first step, the generalized propensity scores were estimated using multinomial logistic regression. In the second step the overlapping of the propensity distributions between a journal and all other journals was inspected. In the third step, following Hirano and Imbens (2004, p. 81) the balancing of the covariates are tested using t tests. In the last step, the unconfounded or unbiased JIFs are estimated by the propensity score weighting method (Imbens 2000, p. 708; Feng et al. 2011). To estimate standard errors of the JIFs a resampling method (bootstrapping) was used (Fan et al. 2001). The covariates were mean centered.

All data analysis was performed with SAS (Allison 1999; SAS Institute Inc. 2009).

Table 2 Characteristics for all papers in the journal set "Microscopy", published 2008–2009

Covariates	Histochemistry/ Cell biology	Journal of Electron Microscopy	Journal of Microscopy Oxford	Micron	Microscopy–Microanalysis	Microscopy Research Technique	Scanning	Ultramicroscopy	Ultrastructural Pathology
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Publication year	0.45 (0.50)	0.60 (0.49)	0.35 (0.48)	0.44 (0.50)	0.34 (0.48)	0.50 (0.50)	0.39 (0.49)	0.45 (0.50)	0.53 (0.50)
Number of authors	5.24 (2.72)	4.56 (2.74)	4.28 (2.22)	4.29 (2.14)	4.39 (2.41)	4.77 (2.55)	3.80 (1.98)	4.41 (2.13)	4.72 (2.76)
Number of pages	10.32 (4.54)	6.29 (3.00)	7.56 (3.75)	6.39 (3.48)	6.04 (4.17)	7.03 (3.05)	7.36 (4.85)	6.02 (3.03)	5.97 (2.52)
Number of references	57.76 (46.93)	24.14 (13.50)	27.90 (22.02)	35.54 (46.07)	22.49 (22.20)	38.45 (31.40)	26.25 (20.50)	24.30 (12.79)	26.74 (23.67)
Number of papers	300	77	427	311	178	231	87	449	78

Results

Generalized propensity score estimation

In the first step we were applying multinomial logistic regression with journal assignment T as the dependent variable, in order to estimate the generalized propensity scores. For 9 journals overall 8 intercept parameters and 8 regression parameters for each covariate were estimated. In order to save space, we only report the test statistics instead of the parameters for two models, one model with the main effects (Model 1), and one model with the main and two-way interaction effects (Model 2).

Overall, model 2 outperforms model 1 (Table 3): The Akaike's Information Criterion (AIC) decreases from 8151.25 (Model 1) to 8029.52 (Model 2), the Pseudo-R^2 as a measure of the amount of explained variance increases from 0.24 to 0.31, as well as the overall Wald χ^2, which tests, whether all parameters are zero.

According to Feng et al. (2011) and Rubin and Thomas (1996) overparametrization is not crucial. As mentioned above, it is better to include unimportant covariates to yield the best estimates of generalized propensity scores than to enhance bias by leaving out important covariates. In sum, number of references, number of pages and their interaction are the most important predictors of the journal assignment with the highest Wald χ^2.

Check on overlapping of the propensity score distributions

As mentioned above, corresponding papers of a journal and any another journal with the same propensity score values are balanced with respect to the distribution of covariates. Therefore, it must be guaranteed that there is some overlapping between the generalized propensity scores of the two groups.

Table 3 Test statistics for the multinomial logistic regression model used to estimate generalized propensity scores ($N = 2{,}138$ papers)

Covariate	Model 1 Wald χ^2 ($df = 8$)	Model 2 Wald χ^2 ($df = 8$)
Publication year	36.59*	40.86*
Number of authors	59.27*	22.35*
Number of pages	147.52*	88.57*
Number of references	156.17*	107.35*
Number of authors × number of pages		28.60*
Number of authors × number of references		25.20*
Number of pages × number of references		71.17*
Year × number of authors		17.60*
Year × number of pages		36.05*
Year × number of references		19.29*
Overall Wald Test ($df = 80$)	422.87*	549.46*
AIC	8151.25	8029.52
Pseudo-R^2	0.24	0.31

df degrees of freedom, *AIC* Akaike's Information Criterion

* $p < 0.05$

Figure 1 contains box plots of the estimated generalized propensity score for each journal and each in comparison to all other journals ("other"). Box plots depict the distribution of data (Bornmann et al. 2008). The lower boundary of the box marks the first quartile (25th percentile), a line within the box indicates the median, and the upper boundary of the box marks the third quartile (75th percentile). Error bars (whiskers) above and below the box indicate the 90th and 10th percentiles.

Except for the journal "Histochemistry/Cellbiology" there is always some overlapping between the corresponding distributions (journal vs. "others"). But even for the journal "Histochemistry/Cellbiology" in the first quartile (25%) there are still 28 papers in any other journal ("other") with the same generalized propensity scores. Therefore, the assumption of overlapping propensity score distributions can be hold.

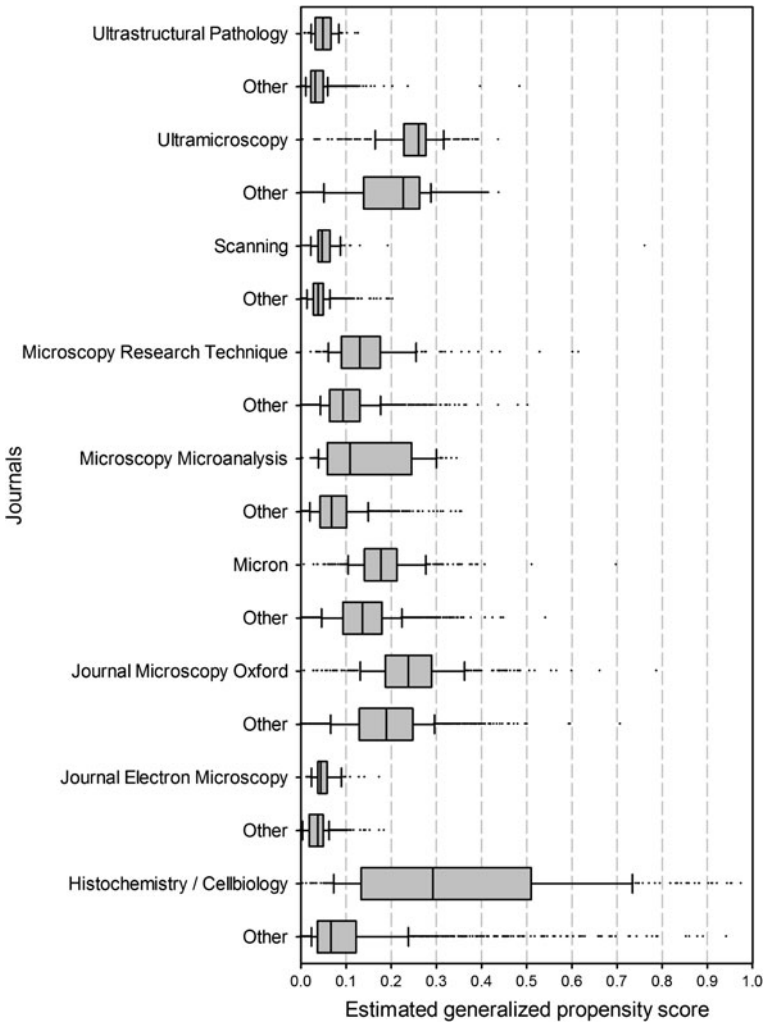

Fig. 1 Overlapping of the propensity score distribution between a journal and the other journals ("other")

Check on covariate balance

Once generalized propensity scores (GPS) are estimated, one must check on covariate balance. The balance is satisfactory, if the papers published in a special journal does not differ anymore on the average from the papers in other journals with respect to all covariates. Following Hirano and Imbens (2004, p. 81) for each journal t, we use five strata, defined by the quintiles of the propensity score for the respective journal t, i.e. $r(t, \mathbf{X}_i)$. For each strata, the mean difference (and its standard deviation) between journal t, and all other journals are calculated for each covariate and aggregated across the five strata using the marginal frequencies as a weight (see Table 1), in order to generate the overall mean difference and t test, respectively.

In Table 4 the t statistic for the mean differences between a journal and all other journals, separated for data, which are either adjusted ("Yes") or unadjusted ("No") for the generalized propensity score. As expected, in most cases the t statistic in the adjusted case is lower than the t statistic in the unadjusted case.

However, not all t values adjusted for GPS fall short of the significance level, some t value adjusted for GPS are even higher than the corresponding t value in the unadjusted case: For the covariate "number of references", for instance, the t statistic of 2.27 for the unadjusted data of the journal "Ultrastructural Pathology" is higher than the t statistic for the adjusted one (t value = 3.87). Out of 180 t statistics 37.8% are statistically significant before the generalized propensity score adjustment, but only 8.9% are statistically significant after the adjustment. Kluve et al. (2012, p. 14) pointed out, that t statistic might be prone to a "balance fallacy", i.e. for some covariates the t statistic drops which might be driven by increased variances instead of decreased mean differences. In our study we did not find any substantial differences in the standard errors between the unadjusted and the adjusted group.

In conclusion, the generalized propensity score adjustment provokes a high, but not perfect balance regarding all included covariates and their two-way interactions.

Estimation of the unconfounded JIFs

Finally, the unconfounded JIFs are calculated using the formula in Eq. 7 (Table 5). There are considerable differences between the GPS adjusted and unadjusted JIFs. For example, the raw JIF of the journal "Histochemistry/Cellbiology" shrinks about 26% from 4.68 to 3.44, the JIF of the journal "Micron" increases about 60% from 1.56 to 2.49.

Even the ranking of the journals changes slightly. The rank correlation amounts to 0.66 (Kendall's tau). The standard errors were calculated using bootstrapping: 1,000 samples with replacement were drawn from the original data for each journal with identical sample size as the original data (N = 2,138). Then, for each sample the generalized propensity scores were calculated and the GPS adjusted and unadjusted JIFs. The standard deviation of the estimated JIFs for each sample was used as an estimate of the standard errors of the JIF (SE) for a journal. The standard errors of the GPS adjusted JIFs are higher than the corresponding unadjusted one.

Conclusions

Without any doubt, the JIF proposed by Garfield in the year 1955 (Garfield 1955, 2006) is still one of the most commonly used and prominent citation-based indicators of the

Table 4 Check of balance given the generalized propensity score (GPS)—*t* statistics for mean equality regarding the set of covariates

Covariate	Adjusted for GPS?	Histochemistry/ Cell biology	Journal of Electron Microscopy	Journal of Microscopy Oxford	Micron	Microscopy–Microanalysis	Microscopy Research Technique	Scanning	Ultramicroscopy	Ultrastructural Pathology
Publication year	No	−0.72	−3.03*	3.69*	−0.35	2.54*	−2.17*	0.81	−1.04	−1.66
	Yes	−0.65	−1.94	0.57	2.12*	−0.36	−0.55	1.28	−0.37	−0.19
Number of authors	No	−5.14*	−0.17	2.79*	1.89	0.61	−1.62	3.41*	1.05	−0.80
	Yes	1.56	0.47	1.19	1.48	0.36	0.39	0.83	1.13	1.38
Number of pages	No	−13.32*	2.62*	−2.47*	4.30*	3.78*	0.68	−0.44	8.77*	4.19*
	Yes	−2.83*	3.37*	1.41	2.06*	−2.16*	1.89	0.94	3.74*	2.79*
Number of references	No	−10.27*	5.51*	4.70*	−1.15	6.36*	−2.85*	2.98*	11.02*	2.27*
	Yes	−3.49*	4.40*	2.98*	0.12	0.80	1.37	1.79	7.24*	3.87*
Number of authors × number of pages	No	−1.70	−0.37	−0.79	0.81	−0.97	0.44	1.08	1.43	2.03*
	Yes	−0.52	−0.52	−0.87	0.37	−0.84	0.14	0.56	0.37	0.92
Number of authors × number of references	No	−3.36*	4.59*	2.42*	−0.63	−0.07	1.91	3.15*	4.63*	2.76*
	Yes	−0.14	4.33*	2.22*	−0.45	1.39	2.50*	2.72*	4.75*	3.51*
Number of pages × number of references	No	1.32	−1.02	−1.27	0.46	−1.69	1.46	−1.71	−2.06*	1.53
	Yes	1.01	−2.17*	−0.61	−0.66	−0.84	0.03	−0.80	−2.37*	−0.61
Year × number of authors	No	−5.19*	−1.06	3.67*	0.78	1.48	0.13	2.51*	0.48	0.14
	Yes	1.40	0.17	2.83*	2.89*	0.26	0.61	0.73	1.09	0.48
Year × number of pages	No	−7.44*	0.25	−3.44*	5.98*	−1.31	−0.67	1.15	5.66*	2.18*
	Yes	−0.10	0.62	−0.40	3.86*	−0.88	1.92	1.64	1.93	0.79
Year × number of references	No	−7.04*	2.14*	0.74	0.23	2.24*	−3.40*	3.43*	6.34*	2.73*
	Yes	−0.88	0.34	1.74	0.29	1.84	−0.35	2.59*	1.96	1.45

* $p < 0.05$

Table 5 Generalized propensity score adjusted and unadjusted journal impact factors for the ISI SCR subject category "Microscopy" for the year 2010

Journal	ISI-JIF	Unadjusted for GPS			Adjusted for GPS			$\frac{(JIF_p - JIF_r)}{JIF_r}$ %
		JIF_r	SE_r	$Rank_r$	JIF_p	SE_p	$Rank_p$	
Histochemistry/Cell biology	4.72	4.68	0.31	1	3.44	0.31	1	−26.49
Ultramicroscopy	2.06	2.00	0.12	2	2.30	1.02	3	15.22
Journal of Microscopy Oxford	1.87	1.77	0.10	3	2.10	0.27	4	18.65
Journal of Electron Microscopy	1.77	1.62	0.24	4	1.69	0.36	6	3.89
Microscopy Research Technique	1.72	1.58	0.18	5	1.63	0.33	7	3.26
Micron	1.65	1.56	0.11	6	2.49	0.65	2	60.07
Microscopy–Microanalysis	3.26	1.34	0.17	7	1.79	0.72	5	33.43
Scanning	1.33	1.14	0.19	8	1.28	0.33	8	11.36
Ultrastructural Pathology	0.73	0.72	0.11	9	0.61	0.12	9	−13.78

JIF journal impact factor, *ISI* ISI journal impact factor, published in SCR, *r* raw, *p* generalized propensity score (GPS) adjusted, *SE* boostraped standard error, *rank* rank of the journal according to the JIF

performance and significance of a scientific journal. The JIF may profit from its robustness, 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 or Vanclay (2012) mentioned in his recent review to the JIF. Our contribution deliberately looked at one core problem, the influence of factors that have nothing to do with the impact of a journal (called bias factors). As solution to this problem, we discussed the Rubin Causal Model (RCM) (Rubin 2004, 2006).

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, for instance, to the covariate "type of documents" (Braun et al. 1989). Scientific journals differ in the frequencies of certain papers, 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). The method of stratification on core covariates, as recently proposed by Mutz and Daniel (2012), is restricted to few covariates. However, the so-called assignment mechanism to journals is more complex, and requires many covariates to give a true picture of what is going on in journals. We suggest generalized propensity score as methodology to adjust JIF for an arbitrary number of covariates. A generalized propensity score is simply the probability, that a paper is published in a certain journal. Papers of different journals with the same propensity score are balanced according to all covariates included. As concluded in Mutz and Daniel (2012), 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).
- *Assignment mechanism*: To adjust the JIFs correctly, the true assignment mechanism of the papers to the journals must be known. In bibliometrics we do not know exactly the actual assignment mechanism. Additionally, it must be plausible, that journals do not influence the included covariates.
- *Covariates*: The JIFs 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 substantive improvements for the estimation of JIFs, as Vanclay (2012) calls for, but further empirical studies and theoretical statistical work are needed.

References

Allison, P. D. (1999). *Logistic regression using SAS: Theory and application*. Cary, NC: SAS Institute Inc.

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.

Fan, X., Felsöváli, Á., Sivo, S. A., & Keenan, S. C. (2001). *SAS for monte carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute Inc.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., & Li, X.-S. (2011). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*. doi: 10.1002/sim.4168 (published online February 24, 2011).

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

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

Hirano, K., & Imbens, G. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspective* (pp. 73–84). London: Wiley.

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

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(467), 854–866.

Imbens, G. (2000). The role of propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710.

Kluve, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society A, 175*(Part 2), 1–31.

Leydesdorff, L., & Bornmann, L. (2011). 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.

Lu, B., Greevey, R., Xu, X., & Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *American Statistician, 65*(1), 21–30.

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.

Mutz, R., & Daniel, H.-D. (2012, in press). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Infometrics*.

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.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

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. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

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

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20–36.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory in practice. *Biometrics, 52*, 249–264.

SAS Institute Inc. (2009). *SAS/STAT 9.2 user's guide*. Cary, NC: SAS Institute Inc.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Bussbach, J. J. V., Andrea, H., et al. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms. An introduction from a case study in mental health. *Medical Care, 48*(2), 166–174.

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

Vanclay, J. K. (2012). Impact factor: Outdated artifact or stepping-stone to journal certification? *Scientometrics* (accepted paper).

Wang, J., Donnan, P. T., Steinke, D., & MacDonald, T. M. (2001). The multiple propensity score for analysis of dose-response relationships in drug safety studies. *Pharmacoepidemiology and Drug Safety, 10*, 105–111.

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics, 30*(1), 59–73.