Journal of Informetrics 9 (2015) 419–429

Contents lists available at ScienceDirect

Journal of Informetrics

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

Does quality and content matter for citedness? A comparison with para-textual factors and over time

Lutz Bornmann[a,*], Loet Leydesdorff[b]

[a] Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany
[b] Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, 1001 NG Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 10 March 2015
Accepted 10 March 2015
Available online 1 April 2015

Keywords:
Citation impact
F1000
Web of Science
Prediction

ABSTRACT

Using (binomial) regression analysis, we run models using citation windows of one to ten years with both annual citation and cumulative citations as dependent variables, and with both bibliometric and quality indicators (judgments of peers) as independent variables. The bibliometric variables are the Journal Impact Factor (JIF) of the publication medium, the numbers of authors and pages, and the statistical citedness of the references used within the paper. We find that the JIF has a larger influence on the citation impact of a publication than the quality (measured by judgments of peers). However, the number of pages and the quality of the references are less influential. The influence of JIF peaks after three years and then declines (in most regression analyses), but remains higher than the influence of quality judgments even after ten years. These results call into question a discrepancy between the algorithmically based indicators and the qualitative judgments by experts. The latter seems less predictive for future citation than a combination of algorithmic constructs. The results of this study can contribute to the empirical specification of the relevance of a normative versus a constructivist theory of citation.

1. Introduction

During the last decade, citation analysis in evaluative bibliometrics has invested in refining the methods of measuring and comparing among different document sets in terms of their "citedness" and therefore perhaps "impact" after appropriate normalization for differences among fields of science, document types, and over time. Factors which influence the citedness of papers, such as the publication venue, the number of co-authors, the length of a paper, the quality of its references, have been studied using different databases: Web-of-Science (WoS), Scopus, and Google Scholar. In an overview of multivariate analyses of predictors of citations Onodera and Yoshikane (2014) summarized these independent variables as possibly strong predictors of citedness: the Journal Impact Factor (JIF) appeared as a strong predictor in 12 out of the 13 studies analyzed; but the number of references and other features of references were equally strong predictors. The number of authors was a strong predictor in only five out of 13 studies, and the length of the papers only in four.

From the perspective of a normative theory of citation (Small, 2004), one would expect that the intellectual quality of a paper becomes increasingly important in determining the respective citation rates the longer the citation window is

* Corresponding author. Tel.: +49 89 2108 1265.
E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), loet@leydesdorff.net (L. Leydesdorff).

http://dx.doi.org/10.1016/j.joi.2015.03.001

(see also Bornmann, Leydesdorff, & Wang, 2014). But how would one measure intellectual quality and its influence on citation? Referee reports preceding publication are confidential and cannot easily be compared across journals and fields. However, a quality assessment system based on peer review is available in the bio-medical sciences: F1000 (Du, Tang, & Wu, 2015; Waltman & Costas, 2014; Wouters & Costas, 2012). F1000 provides a post-publication peer review system of the biomedical literature in terms of papers from medical and biological journals. This service is part of the Science Navigation Group, a group of independent companies that publish and develop information services for the professional biomedical community and the consumer market.

F1000 Biology was launched in 2002 and F1000 Medicine followed in 2006. The two services were merged in 2009 and today constitute the F1000 database. Submissions to F1000 are selected by peer-nominated global "Faculty" of leading scientists and clinicians who then rate the papers and explain their importance (F1000, 2012). This means that only a relatively small set of papers from the medical and biological journals is reviewed, while most of this literature is actually not rated (Kreiman & Maunsell, 2011; Wouters & Costas, 2012).

The Faculty of F1000 nowadays consists of more than 5000 experts worldwide, assisted by approximately 5000 associates, who are organized into more than 40 subjects (which are further subdivided into more than 300 sections). On average, 1500 new recommendations are contributed by the Faculty each month (F1000, 2012). Faculty members can choose and evaluate any paper that interests them; however, "the great majority pick papers published within the past month, including advance online papers, meaning that users can be made aware of important papers rapidly" (Wets, Weedon, & Velterop, 2003, p. 254). Although many papers published in popular and high-profile journals (e.g. *Nature*, *New England Journal of Medicine*, *Science*) are evaluated, 85% of the papers selected come from specialized or less known journals (Wouters & Costas, 2012).

The papers selected for F1000 are rated by the expert-members as "good," "very good," or "exceptional," which is equivalent to the recommendation scores (RSs) of 1, 2, or 3, respectively. In many cases, a paper is not rated only by one single member, but by several.

In this study, we use the RSs as an independent variable in the regression analysis on a par with the other independent variables (JIF, number of authors, number of pages, and quality of the references) for the prediction of citations over time. Since the independent variables are on different measurement scales, the coefficients of the binomial regression analysis cannot be directly compared. We use (i) the marginal effects of changing each independent variable holding all other variables constant, and (ii) the increments in the variance explained (R^2) by adding independent variables as measures for the contribution to the prediction.

Both citation rates and accumulated citations during a ten-year period have been tested in a series of models, but in this study we are able to address the question of how the relative influences of independent variables change over time. The main findings are that the quality of the *journal* (measured as JIF), in which the paper was published, significantly contributes to the citation more than the quality of the paper as rated by F1000. The numbers of pages and the normalized citedness (mean normalized citation score, MNCS) of the references in each paper have less effect on citation than RS. In most of the regression analyses, the effects of the JIF and the numbers of authors diminish with time beyond two or three years.

2. Methodology

2.1. Data

In January 2014, F1000 provided one of us with data on all recommendations (and classifications) made, and the bibliographic information for the corresponding papers in their system ($n = 149{,}227$ records). This dataset contained a total of 104,633 different digital object identifiers (DOI), among which all are individual papers (with incidental exceptions). The approximately 30% reduction of the dataset by the identification of unique DOIs can be attributed to the fact that many papers received more than a single recommendation from Faculty members, and therefore appeared more than once in the dataset.

For bibliometric analysis in the current study, citation counts (between the date of publication and the end of 2013) and other bibliometric variables (such as the JIF) were matched at the paper level using an in-house database of the Max Planck Society (MPG) based on the WoS and administered by the Max Planck Digital Library (MPDL). In order to create a link between the individual papers and the bibliometric data, two procedures were used: (1) A total of 90,436 papers in the dataset could be matched with a single paper in the in-house database using the DOI as a key; (2) in the case of 4205 papers of the 14,197 remaining papers (in which no match could be achieved using the DOI), the name of the first author, the journal, the volume and issue numbers could be matched. Thus, bibliometric data is available for 94,641 papers of the 104,633 total (91%). This percentage approximately agrees with the value of 93% found by Waltman and Costas (2014), who used a similar procedure to match data from F1000 with the bibliometric data using the in-house database of the Centre for Science and Technology Studies (CWTS) in Leiden.

In order to obtain a ten-year citation window, we include in this study only papers which were published between 2000 and 2004. This reduces the data set to $n = 9898$ papers; none of the variables used in this study are missing for any of these documents. This step thus ensures that annual citations over ten years are available for all the papers. Here the publication year of the paper is taken as the first year, and citations are available for papers up to the year 2013 in the in-house database of MPDL mentioned above.

2.2. Statistical procedures and software

The statistical software package Stata 13.1 (http://www.stata.com/) is used for the analysis; in particular, we use the Stata commands nbreg, regress, fitstat, and mchange (Long & Freese, 2014) for estimating a series of regression models. The outcome variables (number of citations) in the models are count variables. They indicate "how many times something has happened" (Long & Freese, 2014, p. 481). The Poisson distribution is often used to model information on the basis of counts. However, this distribution rarely fits in a statistical analysis of bibliometric data, due to overdispersion (Allison, 1980; Chen & Leydesdorff, 2013). "That is, the [Poisson] model underfits the amount of dispersion in the outcome" (Long & Freese, 2014, p. 507).

Since the standard model to account for overdispersion is the negative binomial (Hausman, Hall, & Griliches, 1984), negative binomial regression models are calculated in the present study (Hilbe, 2007). In the case of citation counts, one can also consider zero-inflated negative binomial regression models (ZINBRM), since one can expect many papers with zero citations. However, these models assume that there are two different mechanisms for citations: a first mechanism would lead to the fact that a paper is cited at all, and a second mechanism that allows other citations to follow the first. Since this assumption is not specified for citations (there is no known difference in the mechanism for the first citation and the later citations), we did not use ZINBRM in this study. Following the recommendations of Thelwall and Wilson (2014), however, we additionally calculated ordinary least squares regression models based on logarithmized citation data. The comparison of the results of these models with those of the negative binomial regression models will prove the stability of the results.

The regression models are calculated with robust standard errors, since it cannot be assumed that the papers and their attributions are independent of one another for every case in the data set. Thus, in these models the traditional standard errors are replaced by robust standard errors. These standard errors "are considered robust in the sense that they are correct in the presence of some types of violations of the assumptions of the model" (Long & Freese, 2014, p. 103). For example, some papers have the same authors or were evaluated by the same Faculty members. Hilbe (2014) even gives the following general advice for the modeling of count data: "Unless your Poisson or negative binomial model is well fitted and meets its respective distributional assumptions, use robust or empirical standard errors as a default" (p. 133).

The regression models are used in this study to test the influence of the various factors named above (such as the number of authors) on the citation counts of a paper over a citation period of ten years. Whereas in one part of the regression models the annual citation counts of the papers between the publication year (c0) and the ninth year (c9) after publication constitute the dependent variable, for the other part of the models the cumulative citation counts between the publication year (cc0 = c0) and the ninth year (cc9) constitute the dependent variable. The cumulative citations refer to the counts in the years between publication and the year in question (for example, cc2 includes the first two years after publication—including the publication year).

3. Results

3.1. Descriptive statistics

Table 1 lists the dependent and independent variables included in the regression models. Whereas the cumulative citations in Table 1 show a definite upward trend over the years (from 3.3 citations in cc0 up to 140.97 citations in cc9), the annual citations reach a peak at about three years after publication (17.18 citations in model c3), and then the citations decline slightly in the following years. In model 10 (c9), the papers are still cited 13.42 times on average in the tenth year.

The lower part of Table 1 shows the independent variables which are used in all these models. The first variable in the table—the *mean RSs*—reflects the quality attributed by Faculty to the F1000 papers. Since some papers received more than a single recommendation from different Faculty members, an average score is calculated for each paper. The *JIF* in Table 1 refers to the publication year of the papers under study: if a paper was published in the year 2004, the JIF for this same year is used in the analysis. As the numbers in the table show, the JIFs for the papers vary widely: they range from 0.53 to 35.04.

In addition to the *number of authors* and the *size of the paper* (number of pages) the model also includes the mean impact of the cited references as an independent variable. Since we can assume that the F1000 papers in the data set have cited papers from very different publication years and subject categories, the mean normalized citation scores (MNCS) of the cited references was used instead of the raw citation counts of the papers cited. Using this indicator allows for the comparison of the impact of papers cited across various time periods and subject categories (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). For each paper, an average value is obtained from the NCSs of the cited references of a paper in the F1000 data set.

Since NCSs for the papers in the MPDL in-house database are only available for papers published after 1979, only the impact of the papers published since then can be used in this study (in each case the citation impact of the reference between its publication year and the end of 2013). Furthermore, only the citation impacts of papers can be used which are in the source code of WoS (i.e., are linked to a paper indexed in WoS) (Marx & Bornmann, 2015). However, since all papers in the F1000 data set are from the biomedical area, a large majority of the cited references were also linked in WoS.

With a value of 10.95, the MNCS for the cited references is large (a value of 1 is the worldwide average). On the one hand, the large value is the result of a few extremely large NCSs in the dataset. On the other hand, the highly cited F1000 papers

Table 1
Dependent and independent variables included in the negative binomial regression models based on (1) annual and (2) cumulative citation counts (n = 9898 papers published between 2000 and 2004).

Variable	Mean	Standard deviation	Minimum	Maximum
(1) Annual citation counts (dependent variable)				
c0 (model 1)	3.30	6.84	0	169
c1 (model 2)	14.27	19.55	0	493
c2 (model 3)	17.18	23.18	0	553
c3 (model 4)	16.87	23.76	0	705
c4 (model 5)	16.39	24.04	0	642
c5 (model 6)	15.74	23.75	0	561
c6 (model 7)	15.12	23.38	0	452
c7 (model 8)	14.57	23.30	0	467
c8 (model 9)	14.11	23.50	0	447
c9 (model 10)	13.42	23.32	0	544
(2) Cumulative citation counts (dependent variable)				
c0 (model 1)	3.30	6.84	0	169
cc1 (model 2)	17.56	25.16	0	591
cc2 (model 3)	34.75	47.06	0	1082
cc3 (model 4)	51.61	69.46	0	1596
cc4 (model 5)	68.00	92.11	0	2238
cc5 (model 6)	83.74	114.36	0	2799
cc6 (model 7)	98.87	136.07	0	3251
cc7 (model 8)	113.44	157.53	0	3673
cc8 (model 9)	127.55	179.08	0	4046
cc9 (model 10)	140.97	200.44	1	4395
(3) Independent variables (for all models)				
Mean recommendation score	1.43	0.55	1	3
Journal Impact Factor (JIF)	13.78	9.27	0.53	35.04
Number of authors	6.00	4.82	1	86
Number of pages	8.02	3.65	1	44
Mean impact of references	10.95	14.16	0.53	245.98

(see the results of Bornmann, 2014) seem to lean on the shoulders of other highly-cited papers (Bornmann, de Moya-Anegón, & Leydesdorff, 2010; Merton, 1965).

3.2. Results of the regression analysis

Table 2 shows the results of the 20 negative binomial regression models based on (1) annual and (2) cumulative citation counts. As these results show, most of the coefficients are statistically significant. This was to be expected, since with the large sample used in this study, even small effects lead to statistically significant results (Kline, 2004). Note that the coefficients of the regression on each independent variable cannot be compared with one another because the coefficients are not standardized.

In order to express the size of the effects of the independent on the dependent variables, two further evaluations are undertaken as a follow up to the regression analyses. The results are shown in Fig. 1:

(1) Marginal effects were calculated, which "indicate the change in the rate [here: citation rates] for a given change in one independent variable, holding all other variables constant" (Long & Freese, 2014, p. 493). With the help of the marginal effects one can visualize how the citation counts in one year would change in dependency on an independent variable while all other independent variable in the model were held constant. Since the independent variables in this study are of very differently scales, Fig. 1 shows the effect of one standard deviation (SD) change of an independent variable on the dependent variable. For example, a standard deviation change in the RS (a value of 0.55, see Table 1) on average increases a paper's expected citation rate for the first year after publication (c1) by 1.8, holding other variables at their observed values.

(2) In addition to the effect of a marginal change of one standard deviation, Fig. 1 shows the increments in R^2 in the right column. This is another approach to comparing the weight of variables in a regression model (Acock, 2014). Since negative binomial regression analyses do not rely on ordinary least squares, various pseudo-R^2 coefficients are output as an additional result of the regression analysis. Among these coefficients—any of which might provide information about the quality of a model—we chose to use Cragg–Uhler's (Nagelkerke) coefficients in this study (Cragg & Uhler, 1970): This coefficient is independent of sample size, it explains the variability of the dependent variable explained by the model, it lies between 0 and 1, and it agrees with the coefficient R^2, if both coefficients (R^2 and pseudo-R^2) can be calculated for a model. A value of zero means that the model says nothing about the variability of the dependent variable; a value of one means that the variability in the dependent variable is completely explained. For each independent variable in

Table 2
Results of 20 negative binomial regression models based on (1) annual and (2) cumulative citation counts (n = 9898 papers from 2000 to 2004).

Annual citation counts	Model 1 c0	Model 2 c1	Model 3 c2	Model 4 c3	Model 5 cc	Model 6 c5	Model 7 c6	Model 8 c7	Model 9 c8	Model 10 c9
Mean recommendation score	0.22***	0.22***	0.23***	0.23***	0.24***	0.24***	0.26***	0.27***	0.26***	0.28***
	(7.46)	(12.33)	(12.73)	(11.72)	(11.29)	(10.54)	(10.95)	(10.42)	(9.25)	(9.39)
Journal Impact Factor	0.06***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***
	(29.65)	(46.42)	(42.69)	(40.19)	(36.51)	(35.28)	(33.95)	(31.41)	(29.24)	(28.09)
Number of authors	0.04***	0.03***	0.04***	0.04***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***
	(10.16)	(15.11)	(15.69)	(14.76)	(13.54)	(12.61)	(11.67)	(11.00)	(10.46)	(9.56)
Number of pages	−0.01**	0.01*	0.01***	0.01***	0.01***	0.01***	0.01***	0.02***	0.02***	0.02***
	(−2.84)	(2.15)	(4.12)	(4.80)	(4.45)	(4.54)	(4.39)	(4.36)	(4.12)	(4.19)
Mean citation impact of linked references	0.00***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
	(4.62)	(9.22)	(8.34)	(7.94)	(7.12)	(6.64)	(5.84)	(6.18)	(5.55)	(5.41)
Constant	−0.31***	1.09***	1.28***	1.23***	1.19***	1.14***	1.09***	1.02***	1.02***	0.92***
	(−4.58)	(27.61)	(33.46)	(30.42)	(27.13)	(24.90)	(22.57)	(19.57)	(18.63)	(15.59)
Cumulative citation counts	**Model 1 c0**	**Model 2 cc1**	**Model 3 cc2**	**Model 4 cc3**	**Model 5 cc4**	**Model 6 cc5**	**Model 7 cc6**	**Model 8 cc7**	**Model 9 cc8**	**Model 10 cc9**
Mean recommendation score	0.22***	0.22***	0.22***	0.22***	0.23***	0.23***	0.24***	0.24***	0.24***	0.25***
	(7.46)	(11.88)	(12.96)	(12.87)	(12.72)	(12.47)	(12.40)	(12.26)	(11.98)	(11.78)
Journal Impact Factor	0.06***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***	0.05***
	(29.65)	(44.84)	(45.87)	(45.13)	(43.70)	(42.64)	(41.75)	(40.76)	(39.67)	(38.70)
Number of authors	0.04***	0.03***	0.04***	0.04***	0.04***	0.04***	0.04***	0.03***	0.03***	0.03***
	(10.16)	(14.57)	(15.69)	(15.75)	(15.44)	(15.07)	(14.68)	(14.29)	(13.94)	(13.54)
Number of pages	−0.01**	0.00	0.01**	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
	(−2.84)	(0.81)	(2.58)	(3.52)	(3.91)	(4.18)	(4.33)	(4.44)	(4.47)	(4.50)
Mean citation impact of linked references	0.00***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
	(4.62)	(8.67)	(8.93)	(8.78)	(8.44)	(8.11)	(7.74)	(7.54)	(7.27)	(7.05)
Constant	−0.31***	1.30***	1.98***	2.36***	2.63***	2.83***	2.99***	3.11***	3.23***	3.32***
	(−4.58)	(31.59)	(52.58)	(63.49)	(69.62)	(73.61)	(76.35)	(77.69)	(78.52)	(78.67)

Notes. t statistics in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

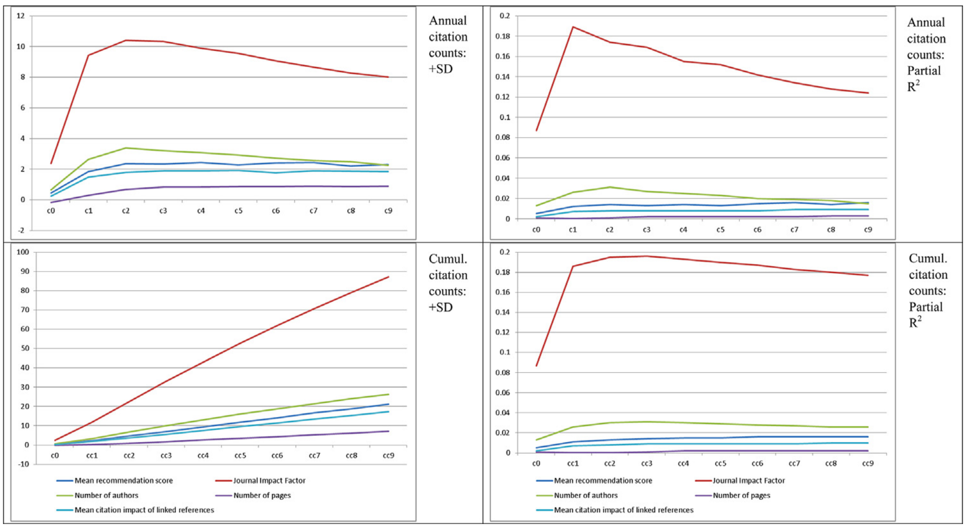

Fig. 1. Marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) citation counts.

the model, the increase of the coefficient when that variable is entered, is provided after all the other variables already entered the model (Acock, 2014).

Fig. 1 shows the marginal effects (left graphics) and partial pseudo R^2s (right graphics) of the independent variables which were included in the regression models based on annual (c0 to c9, upper graphics) and cumulative (cc0 to cc9, lower graphics) citation counts. The results shown in the various graphics look very similar—with the exception of the one in the bottom-left quadrant. Over the entire period of ten years, the correlations between JIF and the citation counts are significantly higher than for the other independent variables. This result suggests strongly that an important role is played for predicting the later citation of a paper (during a period of ten years) by the journal in which the paper is published. However, the graphics also show that this correlation becomes weaker over the years. The correlation rises at first until the second or third year after a paper appears, and thereafter tends to fall off.

Since JIF itself can be considered as a two-year moving average, this highest correlation after two or three years is no surprise. However, this decreasing correlation over the years does not appear in the visualization of the marginal effects based on the cumulative citations in the graph at the bottom-left. In this case, a continuously increasing effect can be seen over the years for the JIF and the other independent variables. Since for the cumulative citations—unlike the annual citations—every year provides more citations (as the dependent variable), an increase of one standard deviation in an independent variable is associated with more citations for analytical reasons. Note that this continuous increase, however, is not visible with the partial pseudo R^2s, which are based on the cumulative citations. Using this measure, it becomes equally apparent that after a peak in the third or fourth year, JIF explains a decreasing part of the variability in the cumulative citation counts. Although the decline of the partial pseudo R^2s is less than in the case of the annual citation counts, the result still indicates a decreasing effect of JIF even on cumulative citations in the course of time.

The number of authors of these F1000 papers is the independent variable which, after the JIF, exercises the strongest effect on the citation counts (albeit at a significantly lower level). This independent variable also shows an initially increasing correlation with the citation impact over the years, which then (after about the second or third year) decreases. The number of authors is perhaps a multiplication effect of specific audiences for each author involved (Narin, Stevens, & Whitlow, 1991); this social (audience) effect can be expected to lose significance after a number of years.

Contrary to expectations, the mean RSs play less of a role in the later citation impact than the JIF and the number of authors. Since the RSs entered the analysis as proxies for the quality of the F1000 papers, one would have expected a significantly higher correlation with the citation counts. However, in the case of the RSs—in contrast to JIF and the number of authors—it appears that the effect does not decline after the second or third year after publication. The effect continues at a level that may be low, but is still constant. There are two characteristics of the RS variable which may partly explain why its effect in the models is relatively small. (1) The RS variable is frequently based on the opinion of just one Faculty member. The results of Bornmann (in press) show that F1000 papers received between one and 20 recommendations from different Faculty members. Most of the papers (around 94%) have a single recommendation (around 81%) or two recommendations (around 13%). (2) The reliability analysis of the F1000Prime peer review system by Bornmann (in press) shows a rather low level of agreement between Faculty members. This result is in agreement with most other studies which have been published on journal peer review (Bornmann, 2011). Both characteristics of the RS variable can lead to an un-reliable quality assessment of F1000 papers and thus, a small effect in the models of this study.

The lowest correlation with the citation counts appears in Fig. 1 for the independent variables "number of pages" and "mean citation impact of the linked references". Since current research always builds on the results of earlier research, one could have imagined a stronger effect of the average impact of the references (Bornmann et al., 2010; Merton, 1965). Research which continues from important earlier research might also be important itself, and lead to citations accordingly. However, the effect of these two independent variables remains constant after two or three years.

In order to test the stability of the reported empirical results, we re-calculated the regression models based on three different model specifications. In the following, only the visualizations of the regression models' results are shown. The coefficients and *t*-statistics are not shown to reduce the size of this paper.

(1) Squared terms: Following the arguments of Bornmann and Williams (2013) one could claim that the results of the regression models shown in Table 2 and Fig. 1 are based on questionable assumptions. For example: can one assume that the more pages a paper has, the higher the impact will be in terms of citations? It is probably more reasonable to assume that, after a certain point, additional pages generate decreasing benefit or even diminish the likelihood of a paper to be cited. Similarly, one might expect diminishing returns for higher JIFs, i.e., it is better to be published in a more influential journal, but above a certain JIF value the benefits may decrease. To address such questions, an additional 20 regression models were calculated, in which squared terms for JIF, number of authors, and paper length were added.

Squared terms allow for the possibility that the variables involved eventually have diminishing benefits or even a negative effect on the dependent variable (Berry & Feldman, 1985). For example, while a single-page paper may be too short to have much impact, a paper that gets too long may be less likely to be read and cited. In most of our regression models, the squared terms are negative, highly significant, and theoretically plausible. Fig. 2 shows the marginal effects (+SD) and partial pseudo R^2s of the independent variables which were calculated subsequent to the regression models. However, as the results in the figure show, they hardly differ from those resulting from the models without squared terms (see Fig. 1). Therefore, we can let this assumption rest.

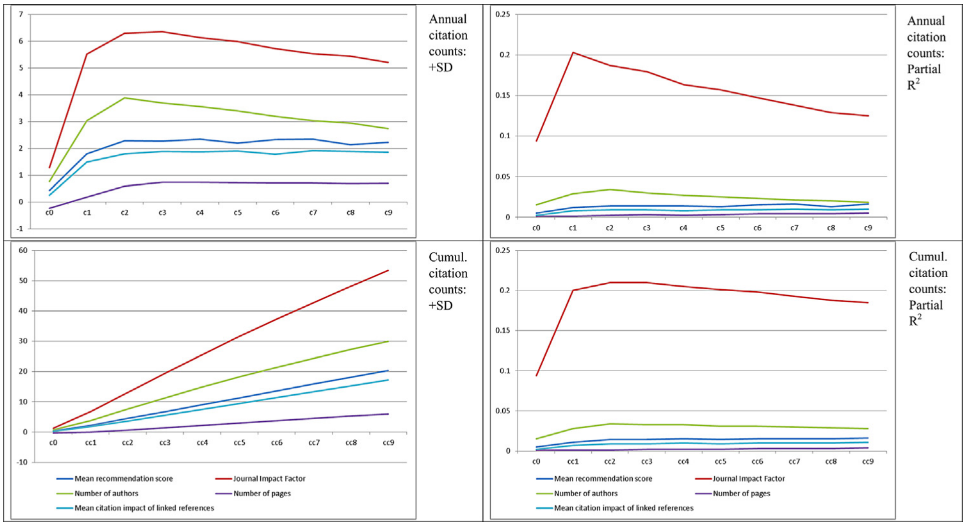

Fig. 2. Marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) citation counts (regression models with squared terms for JIF, number of authors, and number of pages included).

(2) Total RSs: In the case of publications with multiple RSs in F1000, we took the average score as the quality variable (see above). One could argue that one should take the total score instead. The total scores are also visible for every paper in the F1000Prime system, which are used to rank the papers in each discipline. A publication with one recommendation with a RS of 3 can be expected to be of lower quality than another publication with two recommendations, one with a RS of 3 and one with a RS of, say, 2. The first publication has been recommended by only one Faculty member. The second publication has been recommended by two members. One Faculty member has given the highest RS to this publication, just like in the case of the first publication, but in addition there is a second member who has also given a RS, albeit not the highest one. Thus, we tested whether the total rather than the average RSs of a publication as independent variable changes the results.

Fig. 3 shows the marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) citation counts. As the comparison of the graphics with those in Fig. 1 shows, the results are similar. We see diminishing correlations for the JIF and the number of authors and more or less constant correlations for the RSs (sum scores), the impact of the references and the number of pages. The most visible difference between Figs. 1 and 3 is that the RS has a higher correlation with citations (higher than that with the number of authors). The total score seems to reflect the quality of a paper better than the average score.

(3) Ordinary least squares regression: The results of Thelwall and Wilson (2014) show that negative binomial regression seems to be unsafe for citation data, because it can give false positives at a much higher rate than indicated by the confidence

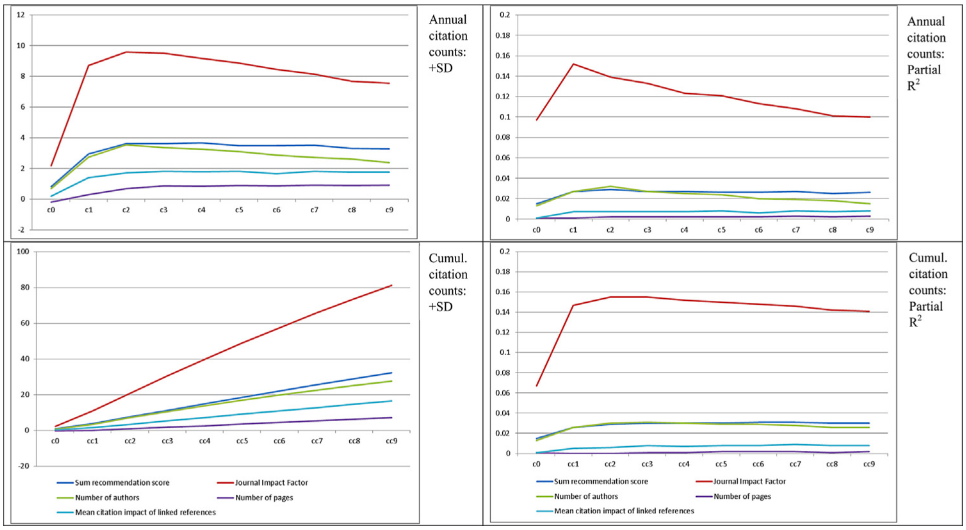

Fig. 3. Marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) citation counts. The models are based on the sum of the recommendation scores instead of the mean.

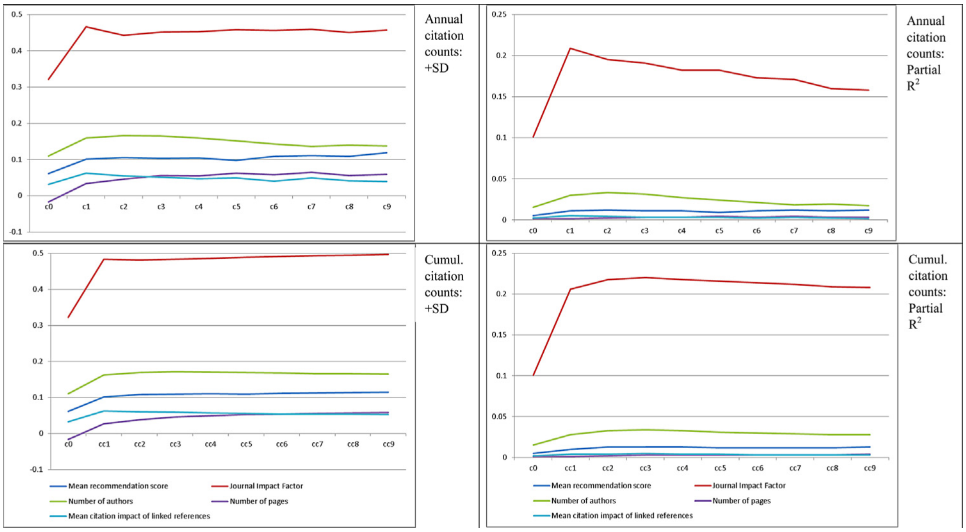

Fig. 4. Marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) citation counts (logarithmized citations + 1). The results are based on ordinary least squares regressions instead of negative binomial regressions (Thelwall & Wilson, 2014).

level used. As limitations of their study, the authors point out that "they only deal with the simplest case of a single factor but it seems likely that the same conclusions would be drawn for more complex factors. Moreover, the unreliability of negative binomial regression for citations is probably overestimated by the simulation approach used here since real citation data will be bounded" (p. 967). Despite these limitations of their results, we re-calculated the models by adding 1 to the citation data, took the logarithm, and used ordinary least squares regression.

The results are shown in Fig. 4. The figure shows the marginal effects (+SD) and partial pseudo R^2s of the independent variables included in the regression models based on annual (c0 to c9) and cumulative (c0 to cc9) (logarithmized) citation counts. According to the size of the correlations between independent variables and citations, the results are in agreement with those visualized in Fig. 1. The effect of the JIF on citations is the highest, followed by the number of authors. The difference of Fig. 4 to Fig. 1 is that the diminishing effects of the independent variables (JIF and number of authors) in Fig. 4 are not as clearly visible as in Fig. 1.

4. Discussion and conclusions

Since factors that influence citation counts (such as publication year and field of science) are used to normalize citation counts in advanced bibliometrics, the role of such factors has been an important topic in bibliometric research. Were it to be established that factors other than publication year and subject category exercise a systematic influence on citations (namely factors which have hardly anything to do with the quality of a paper), bibliometricians would have to consider whether these other factors should also be taken into account when normalizing citations. Bibliometrics is thus concerned with the various factors influencing citations in order to approach closer to the measurement of the quality construct with the help of citation statistics.

Unfortunately, there has not yet been a standard for the validation of citation counts in terms of their correlation with quality: "There is no standard for the validation of counts of papers and citations as they relate to quality. For example, much can be written about a mistake" (Panel for Review of Best Practices in Assessment of Research, 2012, p. 34). As the results of this study show, one can assume the existence of other factors which may exercise a (significant) effect on citation rates.

Our results are in agreement with most other studies published to date. The journal in which a paper was published, emerges in this study as the most important factor (which is also in agreement with most other studies (Didegah & Thelwall, 2013; Onodera & Yoshikane, 2014)). But it is also apparent from regression models of the current study that the influence of some factors changes with time (Bornmann et al., 2014). This means we found diminishing effects for both independent variables in some, but not in all regression analyses. In these models, it can be seen that the correlation with citation counts weakens both for JIF and for the number of authors—after a peak in the second or third year after the publication of a paper. As the citation counts for a paper can be expected to decrease beyond three or more years, the effects of these two independent variables are also reduced. We mentioned that JIF can be considered as a two-year moving average, so that it best indicates short-term impact (Bensman, 2007; de Solla Price, 1970). The number of authors was suggested to be associated with an audience effect that equally can be expected to fade away after a few years.

In addition to JIF and the number of authors, the current study also used the mean RSs (peer judgments), the number of pages, and the MNCSs of the linked references into account. Unlike JIF and the number of authors, these three factors showed no decreasing effects over the years. After a rise in the first few years, their effect hardly changes during the following years. Here, the results of all calculated regression models are in agreement.

In summary, the results of the current study indicate that the citation impact of a paper is influenced by various factors which—seen over time—can have a different effect on the impact. The JIF and the number of authors show—compared with other factors—a (strong) correlation with the citation count of a paper: The higher the reputation of a journal (measured by JIF) and the more authors a paper has, the stronger the expected boost on the impact of the paper, particularly two to three years after its appearance.

Furthermore, the results about the three other independent variables suggest that the factors which indicate the quality or the contents of a paper have a relatively constant effect on the number of citations over the years. Whereas the number of pages of a paper can perhaps indicate the number of research results reported in a paper, the mean citation impact of the linked references can be expected to reflect the quality of the research on which the authors of a paper have based themselves in their research. The mean RSs is based on an assessment of the quality of a paper by experts in the field.

The results of this study can contribute to the empirical specification of the relevance of a normative versus a constructivist theory of citation. From the former perspective—usually attributed to Merton (1988)—citation tends to be considered as an atom of peer recognition (see also Kaplan, 1965), whereas the constructivist perspective emphasizes the rhetorical function of citation in the weaving of texts (Cozzens, 1989; Gilbert, 1977; Luukkonen, 1997). Our results suggest that the quality or the content of a paper can be expected to lead to a constant citation impact over the years—if one would have measured only these two aspects as independent generators of citation counts. However, the typical development of the citation curves over the years (as can also be seen in Table 1) seems mainly to be attributable to the effect of the JIF of the journal (!) and—perhaps—the number of authors. In our opinion, the impact of a paper can thus be analytically divided into two parts: first, the impact of a paper is related to its cognitive content, but, secondly, a larger part of the impact can be attributed to the publication medium of a paper and other para-textual characteristics (this differentiation was proposed by Randić, 2009).

Even if citations thus do not mainly measure the quality of the research (Macilwain, 2013), it nevertheless seems possible to extract a partial correlation for the quality aspect of the citations. However, this quality aspect is hidden behind a range of extrinsic factors which—alongside quality—also exercise an influence on the citation rates. As Colliander (2014) formulated: "Consider a case where two articles have the same utility and influence with respect to the judgment of the peers they are addressing. If the two sets of peers are not identical then the citation counts of the two articles are partly dependent on factors such as the citation behavior in the two sets (propensity to cite)" (p. 1). In a similar vein, Leydesdorff and Amsterdamska (1990) formulated: "(. . .) whether or not a scientific contribution will be cited seems to depend in the first instance on whether *citing* authors can *use* the reference in their texts. Whether they in turn will be able to do so depends on the current development of the field; and hence, citations are not a valid indicator of the quality of cited papers at the moment they are published" (p. 326, Cozzens, 1989).

Textual dynamics seem to exercise such a strong effect on the citation impact, that the citation impact of a paper in the first years after its publication can even be predicted from these factors. This conclusion was drawn, for example, by Yu, Yu, Li, and Wang (2014), who collected a large set of para-textual features of a paper for the explanation of citation impacts, namely features of authors, features of the journal of publication, and features of citations involved in constructing a paper's feature space.

In the interpretation of the results of this study (and also of other similar studies) one should keep in mind that not only citation impact, but also the idea of quality is a multidimensional phenomenon. Not only citations display many dimensions, but also quality, as the following quotation makes clear: "Research quality is a complex, multidimensional attribute that takes into account various factors such as originality, rigor, and scientific impact (but does not include consideration of broader socio-economic impacts of research)" (Council of Canadian Academies, 2012, p. 11). Martin and Irvine (1983) first distinguished between the "impact" as an aspect of scientific quality (the 'impact' of a publication describes its *actual* influence on surrounding research activities at a given time) and "importance" (the influence on the advance of scientific knowledge) or "quality" (how well the research has been done). These authors considered "impact" as the most important indication for scientific progress. Since both citation impact and quality are multidimensional phenomena, however, and the various dimensions are not identical, the interpretation of results which are derived from the correlation between the two phenomena remains a difficult undertaking.

5. Recommendations for future studies

The current study was restricted to papers from the biomedical area and their citations. Since citation is a relatively common practice in these sciences, the evaluation of the annual citation rates can as a rule be based on large numbers. Leydesdorff and Amsterdamska (1990) also used biomedical data. The biomedical fields are special in having often very fast research fronts (de Solla Price, 1970). Therefore, it would be desirable to repeat this study with data from other fields of science, in order to check the generality of the results; but in these cases, attention should be paid to the requirement of sufficiently large samples.

While most of the F1000 papers can be attributed to the biomedical area in some broad sense, the results of Bornmann and Marx (2015) show that these papers are quite subject heterogeneous and that they are distributed over a large number of WoS subject categories. Since we controlled for the JIF in the regression models when looking at the independent effect of the other independent variables, one could argue that we are addressing this problem. However, it would be desirable that some kind of field-normalization of citation data should be used in future studies (especially if data from very heterogeneous fields are used). Here, it should be noted that the generation of field-normalized citation data for single citing years is a sophisticated process.

Based on our results, we analytically divided the impact of a paper into two parts: first, its cognitive content, and, secondly, the publication medium and other para-textual characteristics. In principle such an analytical distinction can be made, but our empirical analysis only partially reflects this analytical distinction. Our empirical analysis is unable to distinguish in an accurate way between the effect of quality or cognitive content on the one hand and the effect of para-textual characteristics on the other hand. Thus, we would like to encourage further scientometric research on citation counts over longer time periods using independent variables which differentiate between the cognitive content and para-textual characteristics in a better way than we did.

In this study, we have obtained quality judgments of papers from F1000. However, F1000 provides scores only for a small proportion of all available publications (Waltman & Costas, 2014). Essentially, these scores represent recommendations by Faculty members. The fact that a publication has a score in F1000 means that it has been recommended. Publications not included in F1000 do not have a recommendation. Obviously, none of the F1000 members considered these publications to be of sufficient value or importance to provide a recommendation. From this point of view, it can be argued that the fact that a publication is included in F1000 is in itself an indication of quality. Indeed, the results of Bornmann (in press) show that publications included in F1000 on average can be expected to be of higher impact than publications that are not included in F1000. Thus, by using the F1000 data set in this study, the quality scale of the papers is biased to high performers. Future research should try to avoid this bias by using a dataset covering the full quality range.

Acknowledgements

Lutz Bornmann would like to thank Adie Chan, Ros Dignon and Iain Hrynaszkiewicz from F1000 for providing him with the F1000Prime data set. The bibliometric data used in this paper are from an in-house database developed and maintained by the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Thomson Reuters (Scientific).

References

Acock, A. C. (2014). *A gentle introduction to Stata*. College Station, TX, USA: Stata Press.
Allison, P. D. (1980). Inequality and scientific productivity. *Social Studies of Science*, *10*(2), 163–179.
Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, *41*, 93–155.
Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. Thousand Oaks, CA, USA: Sage Publications.
Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199–245.
Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, *8*(4), 935–950.
Bornmann, L. (2015). Inter-rater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for Information Science and Technology* (in press)
Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS ONE*, *5*(10), e11344.
Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, *8*(1), 175–180.
Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418.
Bornmann, L., & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, *7*(2), 562–574. http://dx.doi.org/10.1016/j.joi.2013.02.005
Chen, C., & Leydesdorff, L. (2013). Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, http://dx.doi.org/10.1002/asi.22968
Colliander, C. (2014). A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23193
Council of Canadian Academies. (2012). *Informing research choices: Indicators and judgment: The expert panel on science performance and research funding*. Ottawa, Canada: Council of Canadian Academies.
Cozzens, S. E. (1989). What do citations count? The rhetoric-first model. *Scientometrics*, *15*(5–6), 437–447.
Cragg, J. G., & Uhler, R. (1970). The demand for automobiles. *Canadian Journal of Economics*, *3*, 386–406.
de Solla Price, D. J. (1970). Citation measures of hard sciences, soft science, technology and non-science. In C. E. Nelson, & D. K. Polloc (Eds.), *Communication among scientists and engineers* (pp. 3–22). Lexington, MA, USA: Heath.
Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064. http://dx.doi.org/10.1002/asi.22806
Du, J., Tang, X., & Wu, Y. (2015). *The effects of research level and article type on the differences between citation metrics and F1000 recommendations*. Retrieved from http://arxiv.org/abs/1501.01076
F1000. (2012). *What is F1000?* Retrieved from http://f1000.com/about/whatis
Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, *7*(1), 113–122.
Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents R and D relationship. *Econometrica*, *52*(4), 909–938.
Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge, UK: Cambridge University Press.
Hilbe, J. M. (2014). *Modelling count data*. New York, NY, USA: Cambridge University Press.
Kaplan, N. (1965). The norms of citation behavior – Prolegomena to the footnote. *American Documentation*, *16*(3), 179–184.
Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC, USA: American Psychological Association.
Kreiman, G., & Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience*, *5* http://dx.doi.org/10.3389/fncom.2011.00048

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of citation analysis. *Science Technology & Human Values*, *15*(3), 305–335. http://dx.doi.org/10.1177/016224399001500303
Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, TX, USA: Stata Press, Stata Corporation.
Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics*, *38*(1), 27–37. http://dx.doi.org/10.1007/Bf02461121
Macilwain, C. (2013). Halt the avalanche of performance metrics. *Nature*, *500*(7462), 255.
Martin, B. R., & Irvine, J. (1983). Assessing basic research – Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.
Marx, W., & Bornmann, L. (2015). On the causes of subject-specific citation rates in Web of Science. *Scientometrics*, *102*(2), 1823–1827.
Merton, R. K. (1965). *On the shoulders of giants*. New York, NY, USA: Free Press.
Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *ISIS*, *79*(4), 606–623.
Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific cooperation in Europe and the citation of multinationally authored papers. *Scientometrics*, *21*(3), 313–323.
Onodera, N., & Yoshikane, F. (2014). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23209
Panel for Review of Best Practices in Assessment of Research, Panel for Review of Best Practices in Assessment of Research, Development Organizations, Laboratory Assessments Board, Division on Engineering, Physical Sciences, & National Research Council. (2012). *Best practices in assessment of research and development organizations*. The National Academies Press.
Randić, M. (2009). Citations versus limitations of citations: Beyond Hirsch index. *Scientometrics*, *80*(3), 809–818.
Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics*, *60*(1), 71–79.
Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971. http://dx.doi.org/10.1016/j.joi.2014.09.011
Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433–445.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. http://dx.doi.org/10.1016/j.joi.2010.08.001
Wets, K., Weedon, D., & Velterop, J. (2003). Post-publication filtering and evaluation: Faculty of 1000. *Learned Publishing*, *16*(4), 249–258.
Wouters, P., & Costas, R. (2012). *Users, narcissism and control – Tracking the impact of scholarly publications in the 21st century*. Utrecht, The Netherlands: SURFfoundation.
Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252. http://dx.doi.org/10.1007/s11192-014-1279-6