Scientometrics (2014) 98:1601–1615
DOI 10.1007/s11192-013-1154-x

What a difference a colon makes: how superficial factors influence subsequent citation

Maarten van Wesel · Sally Wyatt · Jeroen ten Haaf

Received: 21 May 2013 / Published online: 18 October 2013

Abstract Getting cited is important for scholars and for the institutions in which they work. Whether because of the influence on scientific progress or because of the reputation of scholars and their institutions, understanding why some articles are cited more often than others can help scholars write more highly cited articles. This article builds upon earlier literature which identifies seemingly superficial factors that influence the citation rate of articles. Three Journal Citation Report subject categories are analyzed to identify these effects. From a set of 2,016 articles in Sociology, 6,957 articles in General & Internal Medicine, and 23,676 articles in Applied Physics, metadata from the Web of Knowledge was downloaded in addition to PDFs of the full articles. In this article number of words in title, number of pages, number of references, sentences in the abstract, sentences in the paper, number of authors and readability were identified as factors for analysis.

Keywords Citations · Readability · References · Sociology · Applied Physics · General & Internal Medicine

MSC 62-07

JEL Classification Z00

M. van Wesel (✉)
Department of Family Medicine, Faculty of Health, Medicine and Life Sciences, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands
e-mail: M.vanWesel@Maastrichtuniversity.nl

S. Wyatt
Department of Technology and Society Studies, Faculty of Arts and Social Sciences, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

J. ten Haaf
Department of Education and Research Services, University Library, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

Introduction

Writing highly cited articles is an important goal for scholars. It gives prestige to the authors and the institutions with which they are associated. Measurements of citations are used to rank and evaluate universities, departments and individual scholars, as well as the countries in which they are located (Haslam et al. 2008; Ball et al. 2009). More importantly, whether or not claims in an article become facts depends on if and how later papers refer to them (Latour 1987). Scientific facts are settled by broad agreement (Collins 1990). "Scientific activity is not 'about nature,' it is a fierce fight to *construct* reality." (Latour and Woolgar 1986).

A scholar who is able to align large numbers of other scholars (Latour 1987; Latour and Woolgar 1986; Collins 1990) has a greater impact on what becomes a fact and what not in his or her field, than a scholar who is unable to align other scholars. Aligning these scholars and obtaining their agreement "may involve funding, status, or persuasive ability" (Martin and Groth 1991). Being cited by others is a signal of this influence on scientific progress.

Are claims and content all that matters? What if seemingly superficial factors influence the number of times an article is cited? The many guidebooks on how to write research papers suggest that there are tricks to writing better papers. In addition to this literature, often based on authors' own experiences or methods, there is also evidence that there are factors that influence the frequency with which articles are cited. For example, it has long been established that scholars of higher rank are more promptly and widely cited (Merton 1968) than less well-known scholars. Having an established name on a paper might ensure that a paper is not ignored, the worst fate to befall a scientific paper (Latour 1987).

In the remainder of this introduction, we summarize earlier research about non-content related factors that affect subsequent citation, including length of titles and abstracts, numbers of pages, authors and cited references, and readability. We then outline our own methodology for selecting articles for analysis and for operationalizing our selected variables. We discuss the results for each of the three subject categories we analyzed, separately and in comparison, before making some recommendations about how to write highly cited articles in Sociology, General & Internal Medicine, and Applied Physics.

In research guidebooks it is recommended to use keyword and title search, preferably in indexes and/or bibliographies, and to base selection of articles to read on their abstracts (Booth et al. 2003; Neuman 1991). This indicates the importance of a catchy title, a good selection of keywords and an attractive abstract. In addition to the content, the readability of an abstract might contribute to its attractiveness. While Haslam et al. (2008) assumed informative and attention-capturing titles might improve impact, they found no association between the catchiness of a title and the impact of an article in the field of Social and Personality Psychology. Furthermore, in a regression of, what they refer to as, organization characteristics of an article they did find that title length had a small negative effect and the presence of a colon in the title had a positive effect on the impact. A possible explanation for this is that a colon may indicate scholarly complexity and distinction (Haslam et al. 2008). Stremersch et al. (2007) hypothesized that title length would have an impact on the number of citations an article in marketing would receive, but could not confirm this with their data. Jacques and Sebire (2010), in comparing highly and lowly cited articles in three medical journals, found a positive correlation between the number of citations received and the length of the title, the presence of a colon and the presence of an acronym. Jamali and Nikzad (2011), however, found a negative correlation between the number of citations and the title length and the presence of a colon in a set of six PLoS journals.

An effective way to boost impact might be sought in working together with others. There are various reasons why collaboration might positively influence the number of times an article is cited. Some argue that it positively affects the quality of a paper (for instance Haslam et al. 2008) as there will be a more extensive internal review process. Collaboration also increases the opportunities for self-citation (for instance Smart and Bayer 1986) and increases the network of scholars into which a paper can easily be introduced (for instance Frenken et al. 2005). Conclusions about whether or not collaboration indeed has a positive impact on citations vary. In an analysis of 270 articles in three applied fields (Clinical Psychology, Educational Measurement, and Management Science), Smart and Bayer (1986) conclude that "collaboration generally has little effect on aggregate quality, regardless of field, as measured by citation indices". Furthermore, their conclusion holds irrespective of whether or not self-citations are included (Smart and Bayer 1986). More recently, Haslam et al. (2008) found first author eminence and total author eminence influenced impact in the field of Social and Personality Psychology, although the number of authors had no significant influence. Webster et al. (2009) uncovered a significant positive relation between the number of authors and the number of citations a paper receives in Evolutionary Psychology. Similar relationships were found in the fields of Biology & Biochemistry, Chemistry, Mathematics and Physics (Vieira and Gomes 2010). Using raw data from the Web of Science over a ten-year period, Glänzel and Thijs (2004) were able to conclude that "multi-authorship increases above all the probability to be cited by others". Multi-authored papers are cited more, but the increase in self-citation rates is weaker than the increase in foreign citations (Glänzel and Thijs 2004). Important outliers in their set are single-authored papers, which have a very low share of self-citations. Furthermore Franceschet and Costantini (2010), in their study of 18,500 Italian research outputs, conclude that collaboration has a positive influence on the impact of papers. Important exceptions being hyperauthored papers, as is common in Physics, which receive fewer citations than papers with a smaller group of authors. Frenken et al. (2005) found that the number of authors (and the number of organizations) had a positive impact in the field of Biotechnology and Applied Microbiology. Within the field of Information Science and Technology, collaboration has a significant positive influence on citation rates (Levitt and Thelwall 2009).

Another important factor is the number of references a paper contains. In the past, this was stable at ten references per paper (Price 1963), but it is widely assumed this number has since increased. Larivière et al. (2008) have shown that while the growth of publications in medical fields and in natural sciences & engineering is progressively slowing down since 1980 the number of references has not leveled off, which would indicate a growth in the number of references per paper. A paper that itself contains many references to previous work is likely to develop a stronger standing than a paper with no or few references (Latour 1987; Latour and Woolgar 1986). References are used to increase a paper's power of persuasion (Gilbert 1977). Webster et al. (2009) suggest that, among other reasons, a form of reciprocal altruism ("I cite you, you cite me") could cause a paper with many references to be cited more often. They found a linear relation between a log transformation of the number of citations and the number of references (Webster et al. 2009), however they also indicate that there could be untested and unknown other variables influencing this relationship. Similar results were found by Vieira and Gomes (2010).

Several scholars have found a positive relationship between article length and the number of citations an article receives (Haslam et al. 2008; Wang et al. 2012; Vieira and Gomes 2010; Hudson 2007), simply because longer articles more often contain more findings.

Hartley et al. (1988), in a short literature review about the relation between readability and prestige, found indications that readability can have both a positive and negative effect on prestige, and thus concluded that superior measuring instruments were needed. For journals in the field of Marketing an increase in readability might negatively influence credibility (Stremersch et al. 2007). An article which is very readable might be thought of as simplistic, whereas an article that is difficult to read "presents us with a choice of whether to judge the author inept for not being clear, or ourselves stupid for not grasping what is going on." (Botton 2001), suggesting there is an optimum somewhere between 'too easy' and 'too hard' to read.

Different techniques have been developed to measure readability, including the Flesch Reading Ease Score (Flesch 1948) and Flesch-Kincaid Grade Level (Kincaid et al. 1975). These types of measurements have been criticized for only looking at surface level linguistic information (Crossley et al. 2008; Lin et al. 2009). Nonetheless the Flesch Reading Ease Score correlates quite well with comprehension (Fry 1968), and is widely used in readability research (see for instance Hayden 2008; Weeks and Wallace 2002; Wager and Middleton 2002; Roberts et al. 1994; Friedman et al. 2004; Villere and Stearns 1976; Hartley et al. 1988).

The Flesch-Kincaid Grade Level (FKGL) expresses the US school grade level or the years of education the reader should have completed in order to understand the text, while the Flesch Reading Ease Score (FRES) expresses readability on a scale, that for practical considerations can be thought of as ranging from 0 to 100, where a higher score indicates easier readability (e.g. 0–30 very difficult, 90–100 very easy). Whilst both FKGL and FRES are used in research, FRES appears to be the most used, even in more recent studies, and will be used in the rest of this study. Both formulas are included in Microsoft Word, and implemented as follows (Microsoft 2003).

$$\begin{aligned} \text{Flesch Reading Ease Score} &= 206.835 - (1.015 * \text{Total Words/Total Sentences}) \\ &\quad - (84.6 * \text{Total Syllables/Total Words}) \\ \text{Flesch-Kincaid Grade Level} &= (0.39 * \text{Total Words/Total Sentences}) \\ &\quad + (11.8 * \text{Total Syllables/Total Words}) - 15.59 \end{aligned} \tag{3}$$

Using these internal functions of Microsoft Word, the readability of a text can be calculated. For instance, the readability of this introduction, from head to tails, has a FRES of 31.3 and a FKGL of 14.

Methodology

In order to analyze how the factors identified above affect citations, we selected three different subject categories, namely Sociology, General & Internal Medicine, and Applied Physics. Journal names for the 10 journals with the highest impact factor were extracted from the Journal Citation Report for 2005. Using the Web of Knowledge (WoK) advanced search function, information for the document type 'Article' was collected over the period 1996 to 2005, creating a corpus of these three categories. Records of these papers were downloaded between 3 and 11 February 2011, and stored in a database for further analysis. As most articles are cited within 5 years of publication, it was important to choose an early cut-off date.

For the papers identified, we attempted to collect full-text PDFs via the publishers. We searched on the issue and volume, the journal and the article title. Not all journals or journal issues fall within the scope of the library subscription of Maastricht University. Other

journals only contained reviews. Not all articles were found, sometimes due to misspellings in WoK where some titles seem to have been read using object character recognition (OCR) which can lead to mistaken characters, for instance 'rn' is read as 'm', and vice verse. Some text could not be extracted for analyses, as articles were sometimes locked for text extraction, which might be unintended. Articles with more than 100 words in the full text and five words in the abstract were included. Only journals for which at least 50 % of the articles were found, extracted and contained at least 100 words are included in the analysis (see Table 1 for an overview of the number of journals and articles included in the analysis compared against the number of articles in the Web of Knowledge, per included category).

Text was extracted from the downloaded PDF files for analysis. For each paper the following information was recorded:

- Number of pages, cited reference count, times cited count: all directly from the WoK record
- Length of the title and the number of authors: based on or inferred from the WoK record
- Number of sentences in the abstract and FRES of the abstract: based on the WoK abstract, and analyzed by Microsoft Word
- Number of sentences in the full text and FRES of the full text: based on the downloaded paper and analyzed by Microsoft Word.

As the citation and reference counts are expected to be positively skewed (Wang et al. 2012; Webster et al. 2009), a log transformation [Log Times Cited = Log10(Times Cited + 1) and Log Reference count = Log_{10}(Reference count + 1)] was applied as exemplified by Webster et al. (2009). Since the author count is also highly positively skewed, a log transformation [Log Author Count = Log_{10}(Author Count)] was also applied. Following the suggestion by Haslam et al. (2008) the presence of a (semi-) colon in the title was indicated by a binary variable.

The relation between the independent variables and the number of citations was analyzed using bivariate correlation for each category. Since there are indications that both readability and its opposite might have a negative impact on the number of citations, the relationship between the number of citations a paper receives and its readability might be parabolic. As bivariate correlations and linear regressions are linear, the square root of the readability scores is also be included in the analysis, as this makes the relationship behave in a more linear fashion.

A more advanced statistical analysis is required since the journal in which an article is published might have characteristics independent of the individual article that influence the number of citations received and the elapsed time since publication also influences the number of citation. A linear regression for each category using dummy variables for the journal and publication year was created. However, since the journal itself changes over time, for instance when a new editor takes over, time and journal cannot be seen as independent of each other. Therefore these dummies are combined in one dummy representing a journal in a year. To look at the effects of some factors independent of the journal in which the papers were published, a model without the journal/year dummy was first created.

Results

From Tables 2, 3, and 4 it is immediately clear that there is a high correlation between all the factors which could lead to multicollinearity in the regression model. Whilst this could have a negative impact on the reliability of the coefficient estimates in the regression

Table 1 Overview of the number of journals and articles included in the analysis compared to the number of articles in the Web of Knowledge, per category

	Journals in analysis	Articles in analysis	Articles in WoK
Sociology	9[a]	2,016	2,443
General & Internal Medicine	5[b]	6,957	11,444
Applied Physics	5[c]	23,676	31,498

[a] "Annual Review of Sociology", "American Journal of Sociology", "American Sociological Review", "Social Networks", "Sociology of Health & Illness", "British Journal of Sociology", "Social Problems", "Journal of Marriage and the Family", and "Law & Society Review"

[b] "New England Journal of Medicine", "Lancet", "Plos Medicine", "Canadian Medical Association Journal", and "Medicine"

[c] "Nature Materials", "Advanced Functional Materials", "Progress in Photovoltaics", "Plasma Processes and Polymers", and "Applied Physics Letters"

model, the predictive power of the model remains intact. Before making inferences about the coefficient estimates we tested for multicollinearity.

The number of Words in Title is significantly correlated ($p < .01$) with the Log Times Cited in all three subject categories (see Tables 2, 3, 4). This correlation is negative in both Sociology and Applied Physics (articles with shorter titles received more citations), but positive in Internal & General Medicine (the longer the title, the more citations received), confirming a hypothesis put forth by Stremersch et al. (2007) and results obtained by Jacques and Sebire (2010).

For all three categories the number of pages correlates significantly ($p < .01$) and positively with the Log of the number of times an article is cited, in concurrence with earlier literature. Another measurement for article length, the number of sentences in an article also correlates significantly ($p < .01$) and positively with the Log times an article is cited in all three categories (see Tables 2, 3, 4).

The length of the abstract, in terms of numbers of sentences, correlates positively and significantly ($p < .01$) with the Log times cited in both General & Internal Medicine and Applied Physics, but not in Sociology (see Tables 2, 3, 4).

The Log of the number of references an article contains correlates positively and significantly ($p < .01$) with the Log of the number of references the article received in all three categories (see Tables 2, 3, 4).

For all three categories a positive, significant correlation ($p < .01$) between Log of the Author Count and Log times cited has been found (see Tables 2, 3, 4).

The square root of the readability of the abstract as measured by the Flesch Reading Ease Score has a negative, significant, correlation ($p < .01$) with the Log times cited in all fields. The square root of the FRES of the whole text correlates, negatively and significantly ($p < .01$) with the Log times cited in Sociology and General & Internal Medicine, but not in Applied Physics (see Tables 2, 3, 4).

Regression

All variables were entered in a regression model as predictor variables for each of the subject categories (see Table 5). The journal/year dummies were entered in model two to further explain the variance of the Log Times Cited (see Table 6).

Table 2 Correlations between the different variables included in the Sociology category

	Log Times Cited	Words in Title	Number of Pages	Log Reference Count	Sentences in Abstract	Sentences in Full Text	Log Author Count	SQRT Abstract FRES
Words in Title	−.046[a]							
Number of Pages	.122[a]	.042						
Log Reference Count	.232[a]	.057[b]	.444[a]					
Sentences in Abstract	.010	.022	.046[b]	.060[a]				
Sentences in Full Text	.265[a]	.026	.716[a]	.617[a]	.105[a]			
Log Author Count	.191[a]	.075[a]	−.123[a]	−.060[a]	.032	−.027		
SQRT Abstract FRES	−.093[a]	−.041	−.057[b]	−.194[a]	.223[a]	−.098[a]	.006	
SQRT Full Text FRES	−.058[a]	−.057[b]	−.037	−.222[b]	.067[a]	.098[a]	.082[a]	.304[a]

[a] Correlation is significant at the .01 level (2-tailed)

[b] Correlation is significant at the .05 level (2-tailed)

Compared to Sociology and Applied Physics the variance in the Log Times Cited General & Internal Medicine is explained to a high degree (31.7 %) by these seemingly superficial factors (see Table 5).

For all three fields, the variance explained increases after adding the journal/year dummies and these increases are significant ($p < .01$), the variance in the Log Times Cited for General & Internal Medicine is explained beyond the 50 % (see Table 6).

The (unstandardized) Beta values of the significant predictors from Tables 7, 8 and 9 are combined in Table 10, that standardized Beta values are also given. From these standardized Betas we can see that a standard deviation change in the number of sentences in the full text has the largest impact on the log of the times an article is cited (and thus on the number of times an article is cited) in Sociology in the first model. In the second model an increase in the number of pages results in the largest change in the log of the times an article is cited. Likewise, we can see that in General & Internal Medicine the largest change in the log of the number of times an article is cited is caused by a change in the log of the author count (first model) and the number of pages (second model). In Applied Physics an increase in the number of sentences in the Full Text (first model) and log of the author count (second model) results in the largest increase in the log of the times an article is cited.

In the first model, for all three fields there were no parameters with a VIF (Variance Inflation Factor) greater than five, which would indicate multicollinearity. After adding the journal/year dummies in the second model for all three fields, the number of pages had a VIF greater than five (Sociology: 8.347; General & Internal Medicine: 5.643; Applied Physics: 6.986). In Sociology the number of sentences in the full text was also above the five threshold (7.202), in General & Internal Medicine and Applied Physics the VIF was close too, but did not break the threshold (4.837 and 4.920 respectively).

Table 3 Correlations between the different variables included in the General & Internal Medicine category

	Log Times Cited	Words in Title	Number of Pages	Log Reference Count	Sentences in Abstract	Sentences in Full Text	Log Author Count	SQRT Abstract FRES
Words in Title	.166[a]							
Number of Pages	.435[a]	.216[a]						
Log Reference Count	.413[a]	.166[a]	.688[a]					
Sentences in Abstract	.314[a]	.303[a]	.505[a]	.553[a]				
Sentences in Full Text	.394[a]	.187[a]	.834[a]	.756[a]	.523[a]			
Log Author Count	.417[a]	.267[a]	.344[a]	.164[a]	.277[a]	.235[a]		
SQRT Abstract FRES	−.108[a]	−.074[a]	−.086[a]	−.063[a]	.125[a]	−.047[a]	−.104[a]	
SQRT Full Text FRES	−.165[a]	.064[a]	−.248[a]	−.247[a]	−.037[a]	−.155[a]	−.085[a]	.514[a]

[a] Correlation is significant at the .01 level (2-tailed)

Discussion

Our analysis shows that some of the variance in the number of citations an article receives can be explained by seemingly superficial factors that have nothing to do with the content of the article. In the Sociology articles, 13.3 % of the variance in the log times cited can be explained by such factors. Changes in the log of the number of references, the log of the number of authors, and the number of sentences in the full text and the number of pages have the most influence. Adding the journal and year of publication to the model explains 31.7 % of the variance in the log times cited. The variables with the most influence are the log of the numbers of references, log of the number of authors, the number of words in the title, the presence of a colon in the title and the number of pages.

In General & Internal Medicine articles, 31.7 % of the variance in the log of the number of times an article is cited can be explained by superficial factors such as the log of the number of authors, the log of the number of references, the presence of a colon in the title and the number of pages. When journal and publication year dummies are added, the model can explain 50.9 % of the variance. Relevant factors are the log of the number of references and number of authors, number of pages and the square root of the Flesch Reading Ease Score (FRES) of the abstract.

In Applied Physics, only 6.7 % of the variance in the log of the number of times an article is cited can be explained by factors such as the log of the numbers of references and authors, number of pages and the square root of the full text FRES. In the second model, this rises to 12.2 % of the variance. The log of numbers of references and authors plus the numbers of pages and title words are significant factors.

While the influence of these superficial factors varies between fields, it is clear that such factors are not trivial as they can influence the number of citations an article obtains. Adding the journal and year dummies has an effect on the influence of some of the more superficial variables on the variance in the frequency with which an article is cited. When we look at the two Sociology models, adding journal/year dummies changes the influence of a standard deviation change to the number of sentences in the full text from positive to

Table 4 Correlations between the different variables included in the Applied Physics category

	Log Times Cited	Words in Title	Number of Pages	Log reference Count	Sentences in Abstract	Sentences in Full Text	Log Author Count	SQRT Abstract FRES
Words in Title	−.089[a]							
Number of Pages	.033[a]	.004						
Log Reference count	.172[a]	−.011	.287[a]					
Sentences in Abstract	.049[a]	.021[a]	.215[a]	.113[a]				
Sentences in Full Text	.138[a]	−.015[b]	.761[a]	.531[a]	.315[a]			
Log Author Count	.140[a]	.053[a]	−.053[a]	.043[a]	.033[a]	.023[a]		
SQRT Abstract FRES	−.031[a]	.029[a]	−.087[a]	−.082[a]	.282[a]	−.044[a]	−.005	
SQRT Full Text FRES	−.007	.022[a]	−.138[a]	.056[a]	.018[a]	.073[a]	.054[a]	.445[a]

[a] Correlation is significant at the .01 level (2-tailed)

[b] Correlation is significant at the .05 level (2-tailed)

Table 5 Summary of the first model regression

	Adjusted R^2	F	p
Sociology	.133	$F_{9,2006} = 35.421$	.000
General & Internal Medicine	.317	$F_{9,6947} = 360.375$	.000
Applied Physics	.067	$F_{9,23666} = 188.034$	.000

Table 6 Summary of the second model regression

	Adjusted R^2	F	p
Sociology	.317	$F_{97,1918} = 10.620$	.000
General & Internal Medicine	.509	$F_{43,6913} = 168.894$	.000
Applied Physics	.122	$F_{36,23639} = 92.344$	.000

negative. This suggests that there is a difference between the distributions of these variables between journals. Also there are differences between the categories, for instance, in Sociology, the square root of the full-text FRES is not an important explanatory variable, though it is in General & Internal Medicine, and for the abstracts of Applied Physics and General & Internal Medicine articles.

Why this difference in distribution between journals exists is not clear from this research. Possibly some factors influence acceptance rate of papers in some journals or some factors are influenced in the editing process. Another suggestion might be that it depends on the specific subfield in which a journal operates; this could be especially true in Sociology which seems a broader field than General & Internal Medicine and Applied Physics. Also, currently, we have no explanation for the between-field variation of the

Table 7 Descriptive statistics, unstandardized Beta and *p* values for variables in the regression models for the Sociology category

	N	Mean	SD	Model 1		Model 2	
				B	*p*	*B*	*p*
Log Times Cited	2,016	1.686 (30.214)	.212 (47.139)				
(Constant)				.857	.000	.590	.000
Colon in Title	2,016			−.018	.382	−.008	.666
Words in Title	2,016	11.050	3.902	−.008	.004	−.009	.000
Number of Pages	2,016	21.419	9.312	−.005	.000	.016	.000
Log Reference Count	2,016	1.686 (52.876)	.212 (22.929)	.187	.002	.368	.000
Sentences in Abstract	2,016	5.968	2.228	−.003	.504	.009	.044
Sentences in Full Text	2,016	490.426	182.064	.001	.000	.000	.000
Log Author Count	2,016	.225 (1.932)	.221 (1.208)	.412	.000	.263	.000
SQRT Abstract FRES	2,016	3.578 (16.435)	1.907 (12.160)	−.007	.221	−.006	.240
SQRT Full Text FRES	2,016	4.877 (25.956)	1.475 (11.310)	−.026	.000	.014	.153

Table 8 Descriptive statistics, unstandardized Beta and *p* values for variables in the regression models for the General & Internal Medicine category

	N	Mean	SD	Model 1		Model 2	
				B	*p*	*B*	*p*
Log Times Cited	6,957	1.345 (193.440)	.322 (315.510)				
(Constant)				.730	.000	.815	.000
Colon in Title	6,957			−.147	.000	−.022	.115
Words in Title	6,957	12.059	4.521	.005	.001	.008	.000
Number of Pages	6,957	5.921	2.693	.026	.000	.050	.000
Log Reference Count	6,957	1.345 (26.338)	.322 (15.315)	.466	.000	.362	.000
Sentences in Abstract	6,957	10.043	4.037	.005	.004	.012	.000
Sentences in Full Text	6,957	215.178	93,751	.000	.974	.000	.000
Log Author Count	6,957	.804 (10.613)	.348 (31.951)	.507	.000	.308	.000
SQRT Abstract FRES	6,957	3.412 (15.111)	1.863 (11.392)	−.017	.000	−.026	.000
SQRT Full Text FRES	6,957	4.962 (25.202)	.7623 (7.249)	.003	.775	.020	.023

influence of the factors studied. We can only speculate as to whether this has to do with different citation practices, or with the training, position and time allocated to research by the people writing it up (full-time scholars vs. doctors who do some research along side their clinical practice).

While these results are based on statistical analysis, they could be used to help people to prepare articles that might become more highly cited. For instance, we notice, when we look at the first and second models, that the number of references and the number of authors explain some of the variance in the number of citations articles received in all three of the fields. This does not mean that one should artificially inflate the number of references (for instance by coping references from other articles, as discussed in Ramos et al. 2012) and the number of authors. The positive effect of an increase in the number of

Table 9 Descriptive statistics, unstandardized Beta and *p* values for variables in the regression models for the Applied Physics category

	N	Mean	SD	Model 1		Model 2	
				B	*p*	*B*	*p*
Log Times Cited	23,676	1.254 (31.395)	.475 (51.792)				
(Constant)				.935	.000	.540	.000
Colon in Title	23,676			.034	.006	.042	.001
Words in Title	23,676	10.719	3.583	−.012	.000	−.012	.000
Number of Pages	23,676	3.275	1.414	−.047	.000	−.016	.003
Log Reference Count	23,676	1.218 (16.783)	.166 (7.525)	.315	.000	.351	.000
Sentences in Abstract	23,676	4.617	1.533	.003	.183	.010	.000
Sentences in Full Text	23,676	121.932	36.828	.002	.000	.001	.000
Log Author Count	23,676	.644 (4.959)	.219 (2.397)	.284	.000	.287	.000
SQRT Abstract FRES	23,676	3.971 (19.081)	1.820 (12.367)	−.001	.597	−.004	.032
SQRT Full Text FRES	23,676	6.151 (38.142)	.551 (6.422)	−.042	.000	−.011	.106

Table 10 Unstandardized and standardized (between brackets) Beta values of significant predictors in first and second regression models for all three categories

	Model 1			Model 2		
	Sociology	General & Internal Medicine	Applied Physics	Sociology	General & Internal Medicine	Applied Physics
Colon in Title		−.147	.340			.042
		(−.111)	(.018)			(.021)
Words in Title	−.008	.005	−.012	−.009	.008	−.012
	(−.066)	(.039)	(−.093)	(−.075)	(.061)	(−.090)
Number of Pages	−.005	.026	−.047	.016	.05	−.016
	(−.108)	(.123)	(−.139)	(.329)	(.234)	(−.047)
Log Reference Count	.187	.466	.315	.368	.362	.351
	(.088)	(.261)	(.110)	(.172)	(.203)	(.123)
Sentences in Abstract		.005		.009	.012	.010
		(.038)		(.042)	(.083)	(.031)
Sentences in Full Text	.001		.002	0	0	.001
	(.304)		(.180)	(−.174)	(−.068)	(.085)
Log Author Count	.412	.507	.284	.263	.308	.287
	(.202)	(.307)	(.131)	(.129)	(.186)	(.132)
SQRT Abstract FRES		−.017			−.026	−.004
		(.−.056)			(−.085)	(−.016)
SQRT Full Text FRES	−.026		−.042		.020	
	(−.084)		(−.048)		(.026)	

references should be understood in the context of the persuasion factor of papers that build on previous literature, as well as some reciprocal altruism. Also the positive impact of an increase in the number authors should be understood as arising from the extension of the

network of scholars into which work can easily be introduced, as well as a possible increase in the quality of a paper resulting from rigorous internal review.

Further lessons can be gleaned for sociologists: do not use titles that are too long. An article with a title less than the mean will, all other things being equal, receive more citations than an article with a title longer then the mean. The articles themselves ought to be longer than the mean (as measured by the numbers of pages, at least) and the number of sentences in the abstract should also be greater than the mean in sociology. In contrast, a longer title does help in General & Internal Medicine, as do more pages and more sentences in the abstract. Here there is also the somewhat paradoxical result that the abstract should be less readable than the mean abstracts but the article itself should be more readable, both as measured by the square root of the Fresh Reading Ease Score. In Applied Physics, not only the title should be shorter than the mean but also the article itself. Both the abstract and the full text should contain more sentences than the mean in and the abstract should not be less easy to read. Short articles with many sentences could indicate short sentences should be used, but could also indicate one should avoid too many figures and tables in an article, which would inflate article length. Future research could shine some light on this matter.

For those variables that do not surface as significant, we cannot claim they do not contribute to the number of citations an article receives. It could well be that the sample lacks the ability to discriminate between highly and lowly cited articles for these variables.

There are limitations to this research: *Applied Physics Letters* accounts for 93.9 % of the sample, overshadowing all other journals in the Applied Physics category. In a future research project this could be circumvented by another way of selecting journals and articles to create a more homogenous set of articles. Whilst General & Internal Medicine and Applied Physics are subfields of the broader fields of Medicine and Physics, respectively, Sociology itself is a broad field, making it a more diverse category compared to the other two categories. There are also limitations to the text extraction method which are summarized in "Appendix".

Some research questions remain;

- Are there differences between journals in the same category?
- Would a more homogenous set of articles produce the same results?
- Do these factors already play a role in the selection of papers, or are they introduced during the editing process (as suggested by Roberts et al. 1994; Wager and Middleton 2002 with respect to readability)
- Does readability surface as an influential factor when using more advanced techniques, such as the soft fuzzy rough set model (Wang et al. 2012)
- Why does the influence of these factors vary so much between the three fields?

If scholars or their institutions want to contribute to scientific literature, and to be seen to contribute, and if they wish promote their individual and collective reputations in rankings and evaluations, they need to be aware of how the invisible hand in science works, and how it can be influenced. Form and style also influence how well individual scholars and their institutions fare in the global competition that scientific publication has become.

Acknowledgments The authors would like to thank Loet Leydesdorff for his helpful comments. Furthermore we believe that additional comments from the two anonymous reviewers have increased the quality of this article, for which we are grateful.

Appendix: limitations to the automatic processing of article text

The automatic processing of article text offers great advantages in terms of speed thus increasing the sheer number of articles that can be processed, but there are disadvantages in terms of accuracy. Some of the causes will be briefly discussed below. Even though the problems discussed below have an impact on the readability score of individual articles, our assessment is that, as all articles in a journal/year set seem to suffer the same problems, it reduces the inter-article differences per subset. Therefore we do not expect the problems to reduce the reliability of our analysis.

Continuous print

Many journals offer articles as a single unit in a PDF, but some journals print articles continuously, i.e. not always starting an article on a new page. Thus the PDF of an article may also contain pages of another article, most likely its reference list or first page. These pages are also included in the analysis of the target article.

OCR mistakes

As mentioned in the main text regarding titles of articles in the Web of Science, articles themselves also suffer from OCR problems. Unfortunately not all PDFs are created from their original source. One of the most common problems this respect is mistaking the letter "m" with the combination "rn", and possibly also vice verse.

Footer/header

Depending on the way an article is created and presented, sentences across pages are read continuously (as a human reader would do) or are read as continuing in the footer and header, thus including items such as the page number, article or journal name.

Affiliations, addresses, and references

All text in the articles is extracted, icluding the affiliations and addresses. As these do not follow normal language conventions they will have an impact on readability.

References

Ball, R., Mittermaier, B., & Tunger, D. (2009). Creation of journal-based publication profiles of scientific institutions—A methodology for the interdisciplinary comparison of scientific research based on the J-factor. *Scientometrics, 81*(2), 381–392. doi:10.1007/s11192-009-2120-5.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2003). *The craft of research (2nd ed., Chicago guides to writing, editing, and publishing)*. Chicago: University of Chicago press.

Botton, A. D. (2001). *The consolations of philosophy*. Baltimore, MD: Penguin Books Ltd.

Collins, H. M. (1990). *Artificial experts: Social knowledge and intelligent machines (inside technology)*. Cambridge, MA: MIT Press.

Crossley, S., Greenfield, J., & McNamara, D. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly, 42*(3), 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. doi:10.1037/h0057532.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*(4), 540–553. doi:http://dx.doi.org/10.1016/j.joi.2010.06.003.

Frenken, K., Hölzl, W., & Vor, F. D. (2005). The citation impact of research collaborations: the case of European biotechnology and applied microbiology (1988–2002). *Journal of Engineering and Technology Management, 22*, 9–30. doi:10.1111/j.1435-5957.2010.00309.x.

Friedman, D. B., Hoffman-Goetz, L., & Arocha, J. F. (2004). Readability of cancer information on the internet. *Journal of Cancer Education, 19*(2), 117–122. doi:10.1207/s15430154jce1902_13.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 513–516, 575–578. doi:10.2307/40013635.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113–122. doi:10.2307/284636.

Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics, 61*(3), 395–404. doi:10.1023/B:SCIE.0000045117.13348.b1.

Hartley, J., Trueman, M., & Meadows, A. (1988). Readability and prestige in scientific journals. *Journal of Information Science, 14*(2), 69–75. doi:10.1177/016555158901500209.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., et al. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185. doi:10.1007/s11192-007-1892-8.

Hayden, J. D. (2008). Readability of the British Journal of Surgery. *British Journal of Surgery, 95*, 119–124. doi:10.1002/bjs.5994.

Hudson, J. (2007). Be known by the company you keep: Citations—quality or chance? *Scientometrics, 71*(2), 231–238. doi:10.1007/s11192-007-1671-6.

Jacques, T. S., & Sebire, N. J. (2010). The impact of article titles on citation hits: An analysis of general and specialist medical journals. *JRSM Short Reports*, *1*(1). doi:10.1258/shorts.2009.100020.

Jamali, H., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88*(2), 653–661. doi:10.1007/s11192-011-0412-z.

Kincaid, J. P., Fishburne, R. P, Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Springfield, VA: National Technical Information Service.

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology, 59*(2), 288–296. doi:10.1002/asi.20744.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Levitt, J. M., & Thelwall, M. (2009). Citation levels and collaboration within library and information science. *Journal of the American Society for Information Science and Technology, 60*(3), 434–442. doi:10.1002/asi.21000.

Lin, S.-Y., Su, C.-C., Lai, Y.-D., Yang, L.-C., & Hsieh, S.-K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet. *Computational Linguistics and Chinese Language Processing, 14*(1), 45–84.

Martin, B., & Groth, E. (1991). *Scientific knowledge in controversy: The social dynamics of the fluoridation debate (SUNY series in science, technology, and society)*. Albany, NY: State University of New York Press.

Merton, R. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. doi:10.1126/science.159.3810.56.

Microsoft. (2003). Readability scores. http://office.microsoft.com/en-us/word/HP051863181033.aspx. Accessed 23 March 2013.

Neuman, W. L. (1991). *Social research methods: Qualitative and quantitative approaches*. Boston: Allyn and Bacon.

Price, D. J. D. S. (1963). *Little science, big science* (George B. Pegram lectures, Vol. 1962). New York: Columbia University Press.

Ramos, M. A., Melo, J. G., & Albuquerque, U. P. (2012). Citation behavior in popular scientific papers: What is behind obscure citations? The case of ethnobotany. *Scientometrics, 92*, 711–719. doi:10.1007/s11192-012-0662-4.

Roberts, J. C., Fletcher, R. H., & Fletcher, S. W. (1994). Effects of peer review and editing on the readability of articles published in annals of internal medicine. *Journal of the American Medical Association, 272*, 119–121. doi:10.1001/jama.1994.03520020045012.

Smart, J. C., & Bayer, A. E. (1986). Author collaboration and impact: a note on citation rates of single and multiple authored articles. *Scientometrics, 10*(5–6), 297–305. doi:10.1007/BF02016776.

Stremersch, S., Verniers, I., & Verhoef, P. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing, 71*(3), 171–193. doi:10.1509/jmkg.71.3.171.

Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4*, 1–13. doi:10.1016/j.joi.2009.06.002.

Villere, M. F., & Stearns, G. K. (1976). The readability of organizational behavior textbooks. *The Academy of Management Journal, 19*(1), 132–137. doi:10.2307/255455.

Wager, E., & Middleton, P. (2002). Effects of technical editing in biomedical journals: A systematic review. *JAMA, 287*(21), 2821–2824. doi:10.1001/jama.287.21.2821.

Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, *93*, 635–644. doi:10.1007/s11192-012-0766-x.

Webster, G. D., Jonason, P. K., & Schember, T. O. (2009). Hot topics and popular papers in evolutionary psychology: Analyses of title words and citation counts in evolution and human behavior, 1979–2008. *Evolutionary Psychology, 7*(3), 348–362.

Weeks, W. B., & Wallace, A. E. (2002). Readability of British and American medical prose at the start of the 21st century. *British Medical Journal, 325*, 1451–1452. doi:10.1136/bmj.325. 7378.1451.