Characterizing the highly cited articles: a large-scale bibliometric analysis of the top 1% most cited research

Pablo Dorta-González a,*, Yolanda Santana-Jiménez b

a Universidad de Las Palmas de Gran Canaria, TiDES Research Institute, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain. E-mail: pablo.dorta@ulpgc.es
b Universidad de Las Palmas de Gran Canaria, Departamento de Métodos Cuantitativos en Economía y Gestión, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain. E-mail: yolanda.santana@ulpgc.es

* Corresponding author

Abstract

We conducted a large-scale analysis of around 10,000 scientific articles, from the period 2007-2016, to study the bibliometric or formal aspects influencing citations. A transversal analysis was conducted disaggregating the articles into more than one hundred scientific areas and two groups, one experimental and one control, each with a random sample of around five thousand documents. The experimental group comprised a random sample of the top 1% most cited articles in each field and year of publication (highly cited articles), and the control group a random sample of the remaining articles in the Journal Citation Reports (science and social science citation indexes in the Web of Science database).

As the main result, highly cited articles differ from non-highly cited articles in most of the bibliometric aspects considered. There are significant differences, below the 0.01 level, between the groups of articles in many variables and areas. The highly cited articles are published in journals of higher impact factor (33 percentile points above) and have 25% higher co-authorship. The highly cited articles are also longer in terms of number of pages (10% higher) and bibliographical references (35% more). Finally, highly cited articles have slightly shorter titles (3% lower) but, contrastingly, longer abstracts (10% higher).
**Introduction**

It is well known that about 20% of papers obtain more than 80% of citations, while other papers are either not cited at all or are infrequently cited (Garfield 2006). Based on this, when a particular paper is cited more frequently than others of a similar topic and age, it is usually concluded that it has a higher quality compared to the other papers (Bornmann et al. 2012). However, there are other reasons why researchers cite papers: to support their own claims, methodology or findings (supportive citations); to present different points of view; and even to criticize the cited paper (Aksnes 2003; Harwood 2008).

Among the factors influencing the number of citations, Tahamtan, Askar Safipour and Ahamdzadeh (2016) identified three general dimensions: (i) **Paper related aspects**: quality, novelty, interest, field and topic, typology, study design, methodology, results and discussion, figures and appendices, titles and abstracts, references, length, age, early citation, and accessibility; (ii) **Journal related aspects**: journal impact factor, language, scope, and visibility; and (iii) **Author(s) related aspects**: number of authors, reputation, academic category, self-citations, international and national collaboration, country, gender and age, productivity, and funding.

Some authors have studied the scientific aspects influencing citations (Buela-Casal and Zych 2010; Callaham et al. 2002; Patterson and Harris 2009; Stremersch et al. 2007). In addition to the quality of the paper, the methodology (Bhandari et al. 2007) together with the novelty of the subject and the popularity of the topic (Chen 2012; Peng and Zhu 2012) seem to be the main scientific aspects influencing citations.

However, in this paper we study the bibliometric or formal aspects influencing citations. We restrict the analysis to research articles in order to avoid the typology bias. It is well known that some types of documents receive more citation than others. Generally, review articles receive more citations than research articles (Biscaro and Giupponi 2014; Fu and Aliferis 2010).
Our study differs from others in the literature in various ways. We conducted a large-scale transversal analysis of around 10,000 papers disaggregated into more than one hundred scientific areas. We used a novel methodology comparing two different groups of papers, those highly cited and those not highly cited. The population analysed are the research articles published in the period 2007-2016 in journals of the Journal Citation Reports (science and social science citation indexes in the Web of Science database). Our aim was to verify the existence or otherwise of bibliometric aspects which displayed significant differences between the group of top 1% most cited articles according to their field and year of publication (highly cited articles) and the remaining articles. In each of these two groups of articles, we considered a random sample of around five thousand documents.

**Bibliometric or formal aspects influencing citations: state of the art**

There is no strong evidence in the literature in favour of the thesis that some formal aspects contribute to a paper achieving more citations. The results mainly depend on the methodology employed and there is no consensus with respect to some of these aspects about the real effect on citations. A brief revision of the most relevant bibliometric factors considered in the literature is described below.

**Field and age**

The number of citations varies according to the characteristics of the disciplines and topics (Costas et al. 2009; Dorta-González et al. 2014). Hot topics usually attract more attention and receive more citations (Fu and Aliferis 2010), but the size of the literature (number of papers published in the field) also contributes to the number of citations a paper receives (Biscaro and Giupponi 2014).

With respect to the age of the article, in general the number of citations increases in the first years after publication before reaching a peak and then gradually decreasing over time (Dorta-González and Dorta-González 2013b). One possible reason is that the information becomes increasingly outdated and obsolete (Barnett and Fink 2008).
**Co-authorship and visibility**

The number of authors indicates the extent of the scientific collaboration. Papers with more authors are more likely to obtain a higher number of self-citations, external citations and visibility (Biscaro and Giupponi 2014; Peng and Zhu 2012).

To increase this visibility, researchers also try to publish their papers in high impact journals (to reach more readers and become more frequently cited). The *journal impact factor* (JIF) can be considered a proxy of visibility and prestige, which are of high importance for a document to be cited (Dorta-González et al. 2017; Dorta-González and Santana-Jiménez 2018). Publishing papers in high impact journals would result in more citations than publishing in low impact ones (Aksnes 2003; Callaham et al. 2002; Fu and Aliferis 2010; Garner et al. 2014; Peng and Zhu 2012). However, the impact factor is a consequence of citations, and is often considered a cause of citations. Thus, considering this aspect in predicting citations is controversial (Bhandari et al. 2007).

**Length of the document and references**

The length of a paper is among the factors increasing the number of citations (Falagas et al. 2013; Peng and Zhu 2012; van Wesel et al. 2014). This might be due to the fact that longer papers contain more information. However, some other studies show there is no relationship between the length of a paper and the number of citations (Royle et al. 2013; Walters 2006).

The list of bibliographical references in a paper can be considered the knowledge of the author(s) about the literature. Thus, the number of references, their prestige as well as the variety of the references in a paper increase the frequency of citation (Biscaro and Giupponi 2014; Chen 2012; Falagas et al. 2013; van Wesel et al. 2014).

**Title and abstract**

The characteristics of the title and abstract are not identified as determinant for citations in the literature. The title affects the number of downloads more than the number of citations. Papers with titles in question form are downloaded more than those with
descriptive titles, but they are less frequently cited (Jamali and Nikzad 2011). Furthermore, titles with two components separated by a colon increases the number of citations (Jacques and Sebire 2010). Other punctuation marks such as a comma also increase citations (Buter and van Raan 2011).

Although some authors did not find a significant correlation between title length and citations (Jamali and Nikzad 2011; Rostami et al. 2014), others maintain that the title length negatively affects citations, in that longer titles receive less citations than shorter one (Stremersch et al. 2015; Subotic and Mukherjee 2014).

Finally, papers with longer abstracts receive more citations (van Wesel et al. 2014).

**Methodology**

The on-line search application of the *Web of Science* database, currently managed by Clarivate Analytics and available at apps.webofknowledge.com, was used for the data search.

Two citation indexes in the Journal Citation Reports were selected (Science Citation Index Expanded and Social Sciences Citation Index). In addition, the basic search option was employed with the following search criteria: Document Types = (Article) AND Year Published = (2007-2016). This basic search resulted in a total population of around ten millions research articles published between 2007 and 2016.

Regarding the design of the research, two groups of research articles were considered, an experimental and a control group. The experimental group was filtered using the Essential Science Indicators (ESI) Top Papers criterion, refining to the highly cited papers category which uses the ESI to locate the top 1% most cited documents according to their field and year of publication. This search resulted in a total of 99,479 *Highly Cited Articles* (HCA).

Then, 5,000 of these HCA articles were randomly selected. After discarding some anomalous documents with empty relevant data, a total of 4,956 articles remained. This sample represents 4.98% of the total population in the HCA group.
For the control group, the HCA were first excluded and then a simple random sample of 5,000 articles was made, resulting in 4,998 Non-Highly Cited Articles (NHCA) after discarding an anomalous pair of data.

Therefore, the total sample size was \( n = 9,954 \) research articles published between 2007 and 2016. For this random sample, the following variables were exported directly from the database: Author, Year, Title, Abstract, Source, Page Count, Times Cited, Cited Reference Count, and Research Areas.

In the disaggregated analysis, of the 137 research areas in the Journal Citation Reports, those in which the number of articles in either of the two groups (HCA and NHCA) was less than five cases were discarded. This is because we consider there are not enough data to draw conclusions. After this discarding process, the final number of research areas in the disaggregated analysis was 107.

We also download the JIF from the Journal Citation Reports. Web of Science uses a journal classification system where each journal is assigned to one or several subject categories. According to the JIF, each journal is placed in a percentile within each category. In this paper, we used the best percentile for each journal, that is, the highest of them all. This is the reason why the median is above the 50th percentile, even in the NHCA group.

Finally, we linked the two datasets by the journal. Both the search for the data and its export to the dataset were done during the first week of September 2017.

As the main statistical tools, we used the median and a non-parametric median test. The median is the value separating the higher half from the lower half of a data sample. That is, the middle value of a data set. The basic advantage of the median in describing data compared to the mean is that it is not skewed so much by extremely large or small values, and so it may give a better idea of a typical value. Because of this, the median is of central importance in robust statistics.

Finally, a non-parametric median test was chosen to compare the HCA and NHCA groups since the variables considered in the study do not follow a normal distribution. The non-parametric median test is a statistical tool that tests the null hypothesis that the medians of the populations from which two or more samples are drawn are identical.
Results and discussion

Data distribution and linear correlation between variables

The hypothesis of normality was rejected for all the analysed variables. Normal contrasts were performed and the frequency histograms corroborated asymmetry and distributions very far from the normal. For this reason, it was decided to use the median in this paper as a measure of central tendency, which is quite common in bibliometric studies (Dorta-González and Dorta-González 2013a).

In relation to independence between the variables analysed, within the 107 areas linear correlations higher than 0.5 were found only between number of references and number of pages, and usually the HCA group had a higher coefficient. Pearson’s correlation coefficients for the 10 research areas with the largest sample sizes are shown in Table 1. As can be observed, in the HCA group these coefficients are above 0.58 in most of the areas (7 out of 10), but in only two 2 areas in the NHCA group. Therefore, in the highly cited group, the longest papers are supported by a greater number of references.

[Table 1 about here]

However, no linear correlations were found between any other pair of variables analysed. Interestingly, although the number of authors might be expected to have an impact on both article length and number of citations, no correlation was found in this study between number of authors and number of pages, nor between number of authors and JIF.

Medians by groups of papers, and equality of median tests in aggregated areas

A non-parametric median test was chosen to compare the HCA and NHCA groups as the variables in this study do not follow a normal distribution (Table 2). For all the variables analysed there are significant differences, below the 0.01 level, between the groups of articles (HCA and NHCA). The highly cited articles are published in higher impact factor journals (33 percentile points higher), have more authors (25% higher) and are longer in terms of number of pages (10% higher) and bibliographical references.
(35% higher). In addition, highly cited articles have slightly shorter titles (3% lower) but, contrastingly, longer abstracts (10% higher).

[Table 2 about here]

Medians by groups of papers, and equality-of-medians tests in disaggregated areas

In order to reduce the field effect, Table 3 analyses the previous aspects but disaggregating for each of the 107 research areas. This information is also summarized in Figure 1. Note that the medians are higher for the JIF percentile in the HCA group in all research areas. Moreover, in the HCA group the medians are clearly higher in most of the research areas for the rest of the aspects except for length of title. That is, the highly cited articles are, in general, more extensive in number of pages, which is the result of the work of a greater number of authors who reference a greater number of documents. In addition, highly cited articles have slightly shorter titles but, contrastingly, longer abstracts.

[Table 3 and Figure 1 about here]

A summary of Table 3 in relation to the significance level is presented in Table 4. Note that, in general, the percentage of research areas with significant differences between the two groups of papers increases when only considering the 30 research areas (of the total of 107) which had more than 50 papers in both the HCA and NHCA.

[Table 4 about here]

From highest to lowest significance in the results, the following comments can be made about Table 4.

Journal Impact Factor percentile

Of the 107 areas considered, 79 display differences between the median percentiles corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of areas showing differences between medians rises to 86. When large samples are considered, the equality-of-medians hypothesis is rejected in all research areas with more than 50 observations in each group. In all cases, the percentile is higher in the HCA group. This result is strong
evidence in favour of the hypothesis that publishing in journals with a high impact factor contributes to achieving more citations for a paper.

**Number of authors**

In 36 of 107 areas, there exist differences between the median number of authors corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of areas showing differences between rises to 51. When large samples are considered, the equality-of-medians hypothesis is rejected in 80% of the research areas with more than 50 observations in each group. In all cases, the median number of authors is higher in the HCA group. This result is strong evidence in favour of the hypothesis that collaborations contribute to a paper achieving more citations.

**Number of characters in abstract and titles**

Most people find publications nowadays via Google Scholar or other online sources. The search algorithms used by Google and other search engines assign more importance to words appearing in a title compared with an abstract or the body text of a paper. If the article title includes keywords that other researchers in the field are likely to search for, then the document is much more likely to show up on the search returns. From a bibliometric perspective, the length of titles and abstract is therefore of interest.

In our results, in 19 of the 107 areas considered there are differences between the median length of the abstract corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of research areas showing differences between medians rises to 39. When large samples are considered, the equality-of-medians hypothesis is rejected at the 0.05 level in 57% of the research areas with more than 50 observations in each group. In most cases, the abstracts are longer in the HCA group.

However, in only 9 of the 107 areas are there differences between the median length of the title corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of areas showing differences between medians rises to 19. When large samples are considered, the equality-of-medians hypothesis is rejected at the 0.05 level in 33% of the research areas with more
than 50 observations in each group. In most cases, the length of the title is smaller in the HCA group.

These results provide empirical evidence in favour of the thesis that longer abstracts and shorter titles contribute to a paper achieving more citations.

**Number of pages and references**

In 56 of the 107 areas, there exist differences between the median number of pages corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of research areas showing differences between medians rises to 62. When large samples are considered, the equality-of-medians hypothesis is rejected at the 0.01 level in 60% of research areas with more than 50 observations in each group (67% of areas at the 0.05 level). In most cases, the median number of pages is higher in the HCA group.

In 38 of the 107 areas, there are differences between the median number of references corresponding to the HCA and NHCA groups at a significance level of 1%. If a significance level of 5% is considered, the number of research areas showing differences between medians rises to 51. When large samples are considered, the equality-of-medians hypothesis is rejected at the 0.01 level in 67% of research areas with more than 50 observations (80% of areas at the 0.05 level). In all cases, the median number of references is higher in the HCA group.

These results are evidence that a higher number of pages and longer list of references contribute to a paper achieving more citations.

**Question form in title and abstract**

The title is very important for the visibility of a paper. However, the abstract is the key to persuading potential readers to finally read the paper. The frequency of the most common punctuation marks in the title and abstract are shown in Table 5. Question form titles appear in 2% of the HCA group and slightly less in the NHCA group (1.86%). However, question marks in the abstract are rare and only appear in approximately 1% of cases.
The colon in titles is quite frequent (23% in the HCA group and 18% in the NHCA group). This use is related to the size of the title because authors frequently use the colon to link sentences instead of other longer rhetorical figures. Therefore, the greater presence of this punctuation mark in the HCA group may explain the fact that titles are slightly shorter within this group.

A descriptive title maximizes the possibilities that readers correctly remember the arguments to rediscover what they are looking for. However, some authors adopt question form titles in the belief that they will be more attractive and increase the number of readers and citations. In an attempt to resolve this issue and determine whether question form titles influence the number of citations per year, a median test for this variable was performed, distinguishing between the HCA and NHCA groups (Table 6).

The results show differences in the HCA group in the median number of citations per year between papers titled in descriptive and question forms at a significance level of 1%. According to the results, in the HCA group the median number of citations per year is higher for papers with a descriptive title. No significant differences were found in the NHCA group.

Citations per year in the highly cited article group

It has been seen that there are significant differences in several characteristics between the two groups of articles, the HCA and the NHCA. It was also decided to determine whether there existed significant differences within the HCA group. For the purpose of brevity, we will limit ourselves to describing some of the variables graphically.

Scatterplots between citations per year and the variables that most affect the impact are shown in Figure 2. As can be observed, in the case of JIF percentile, the vast majority of the highly cited articles are published in journals of the first quartile (percentile above 75%). Furthermore, an exponential relation between the two variables can be clearly seen. Within the select group of articles in the first quartile, there are numerous cases
with more than a hundred citations per year in the analysed period. In five cases the number of citations per year is more than five hundred. However, in the group of articles published in journals of the second quartile (percentile 50 to 75) only a small number of papers with more than a hundred citations are observed.

[Figure 2 about here]

For number of authors there are two different trends depending on a specific threshold. Up to about 10 authors, there is a positive effect on citations. However, above that level there is no clear effect on citations. Surprisingly, there are about twenty papers with more than a hundred authors and one of them close to one thousand authors.

Something similar happens with number of pages. There are two different trends. Up to around 15 pages, there is a positive effect on citations. However, above that level there is again no clear effect on citations. Perhaps surprisingly, there are many cases of articles more than fifty pages long.

Finally, as can be deduced from Figure 2, all of the ten most cited papers (with more than four hundred citations per year) were published in journals above the percentile 90, and only one had more than a hundred authors. In addition, three are over twenty-five pages long. The most cited document is a major collaboration of around five hundred authors. The second and third most cited articles are two very extensive documents of more than two hundred and fifty pages.

Conclusions

There is no strong evidence in the literature in favour of the thesis that some formal aspects contribute to achieving more citations for a paper. The results mainly depend on the methodology employed and there is no consensus in some aspects about the real effect on citations.

This large-scale study, both in terms of sample size and the number of areas considered, analyses metadata associated with the publications and concludes that some of them have a significant influence in explaining the impact of documents.
Although the number of authors might be expected to have an impact on both article length and number of citations, no correlation was found in this study between number of authors and number of pages, nor between number of authors and JIF.

Highly cited articles differ from the other articles in most bibliometric aspects. There are significant differences, below the 0.01 level, between highly cited and non-highly cited articles in many variables and areas. Highly cited articles are published in journals with a higher impact factor (33 percentile points above) and have more authors (25% more). Highly cited articles are also longer in terms of number of pages (10% higher) and bibliographical references (35% more). Finally, highly cited articles have slightly shorter titles (3% lower) but, contrastingly, longer abstracts (10% higher).

The practical implications of these results are related mainly to the impact of the publication journal and the impact of the article. The publication journal is very important in relation to the impact of the research because the journal impact factor percentile is usually a good measure of visibility and readership. From the point of view of research impact, it is preferable for titles to be descriptive and short, and the abstract to be the part that extensively describes the conclusions and methodological aspects.
References

Aksnes, D. W. (2003). Characteristics of highly cited papers. Research Evaluation, 12(3), 159–170.

Barnett, G. A., & Fink, E. L. (2008). Impact of the internet and scholar age distribution on academic citation age. Journal of the American Society for Information Science and Technology, 59(4), 526–534.

Bhandari, M., Busse, J., Devereaux, P. J., Montori, V. M., Swiontkowski, M., Tornetta III, P., & Schemitsch, E. H. (2007). Factors associated with citation rates in the orthopedic literature. Canadian Journal of Surgery, 50(2), 119–123.

Biscaro, C., & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. PLoS One, 9(6), e99502.

Bornmann, L., Schier, H., Marx, W., & Daniel, H. D. (2012). What factors determine citation counts of publications in chemistry besides their quality? Journal of Informetrics, 6(1), 11–18.

Buela-Casal, G., & Zych, I. (2010). Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. Psicothema, 22(2), 270–276.

Buter, R. K., & van Raan, A. F. J. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. Journal of Informetrics, 5(4), 608–617.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. Journal of the American Medical Association, 287(21), 2847–2850.

Chen, C. M. (2012). Predictive effects of structural variation on citation counts. Journal of the American Society for Information Science and Technology, 63(3), 431–449.

Costas, R., Bordons, M., Van Leeuwen, T. N., & Van Raan, A. F. J. (2009). Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. Journal of the American Society for Information Science and Technology, 60(4), 740–753.

Dorta-González, P., & Dorta-González, M. I. (2013a). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. Scientometrics, 95(2), 645–672.

Dorta-González, P., & Dorta-González, M. I. (2013b). Impact maturity times and citation time windows: The 2-year maximum journal impact factor. Journal of Informetrics, 7(3), 593–602.
Dorta-González, P., Dorta-González, M. I., Santos-Peñate, D. R., & Suárez-Vega, R. (2014). Journal topic citation potential and between-field comparisons: The topic normalized impact factor. *Journal of Informetrics, 8*(2), 406–418.

Dorta-González, P., González-Betancor, S. M., & Dorta-González, M. I. (2017). Reconsidering the gold open access citation advantage postulate in a multidisciplinary context: an analysis of the subject categories in the Web of Science database 2009-2014. *Scientometrics, 112*(2), 877–901.

Dorta-González, P., & Santana-Jiménez, Y. (2018). Prevalence and citation advantage of gold open access in the subject areas of the Scopus database. *Research Evaluation, 27*(1), 1–15.

Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. *PLoS One, 8*(2), e49476.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association, 295*, 90–93.

Garner, J., Porter, A. L., & Newman, N. C. (2014). Distance and velocity measures: Using citations to determine breadth and speed of research impact. *Scientometrics, 100*(3), 687–703.

Harwood, N. (2008). Publication outlets and their effect on academic writers’ citations. *Scientometrics, 77*(2), 253–265.

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), 2.

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88*(2), 653–661.

Patterson, M. S., & Harris, S. (2009). The relationship between reviewers’ quality-scores and number of citations for papers published in the journal physics in medicine and biology from 2003-2005. *Scientometrics, 80*(2), 343–349.

Peng, T. Q., & Zhu, J. J. H. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology, 63*(9), 1789–1803.

Rostami, F., Mohammadpoorasl, A., & Hajizadeh, M. (2014). The effect of characteristics of title on citation rates of articles. *Scientometrics, 98*(3), 2007–2010.
Royle, P., Kandala, N. B., Barnard, K., & Waugh, N. (2013). Bibliometrics of systematic reviews: Analysis of citation rates and journal impact factors. *Systematic Reviews, 2, 74.*

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

Subotic, S., & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science, 40*(1), 115–124.

Tahamtan, I., Askar Safipour, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics, 98*(3), 1601–1615.
Table 1: Pearson correlation coefficient by groups of articles between the variables N References and N Pages for the 10 research areas of largest sample size

| Area                                | Sample size | Pearson correlation coefficient |
|-------------------------------------|-------------|--------------------------------|
|                                     | HCA   | NHCA  | HCA   | NHCA  |
| Biochemistry & Molecular Biology    | 197   | 180   | 0.62  | 0.60  |
| Business & Economics                | 171   | 101   | 0.39  | 0.22  |
| Chemistry                           | 719   | 492   | 0.73  | 0.41  |
| Computer Science                    | 149   | 186   | 0.58  | 0.39  |
| Engineering                         | 466   | 597   | 0.59  | 0.34  |
| Environmental Sciences & Ecology    | 265   | 192   | 0.63  | 0.53  |
| General & Internal Medicine         | 285   | 116   | 0.74  | 0.72  |
| Materials Science                   | 453   | 288   | 0.67  | 0.32  |
| Mathematics                         | 280   | 269   | 0.42  | 0.36  |
| Physics                             | 550   | 636   | 0.34  | 0.51  |

Table 2: Medians by groups of papers, and equality-of-medians tests in 6 bibliometric aspects for aggregated research areas

|                      | Highly Cited Articles (n=4,956) | Non-Highly Cited Articles (n=4,998) | Non-parametric Test |
|----------------------|---------------------------------|----------------------------------|---------------------|
|                      | Median                          | Median                           | p-value             |
| N Authors            | 5                               | 4                                | .00                 |
| N Title characters   | 90                              | 93                               | .00                 |
| N Abstract characters| 1,278                           | 1,160                            | .00                 |
| N References         | 41                              | 30                               | .00                 |
| N Pages              | 10                              | 9                                | .00                 |
| JIF Percentile       | 91                              | 58                               | .00                 |

[Table 3 is at the end of the document]
Table 4: Number of areas, from all 107 research areas and the 30 areas with more than 50 papers in both groups, where there are significant differences between groups of articles

| Areas | p-value | N Authors | N Title characters | N Abstract characters | N References | N Pages | JIF Percentile |
|-------|---------|-----------|--------------------|-----------------------|--------------|---------|---------------|
|       |         | HCA > NHCA | 36 | 1 | 15 | 38 | 43 | 0.01 |
|       |         | HCA < NHCA | 0 | 8 | 4 | 0 | 8 | 0.01 |
| Total (of 107) | | 36 | 9 | 19 | 38 | 56 | 79 | 0.01 |
|       | p ≤ .05 | HCA > NHCA | 51 | 6 | 30 | 51 | 49 | 86 |
|       |         | HCA < NHCA | 0 | 13 | 9 | 0 | 8 | 0.05 |
| Total (of 107) | | 51 | 19 | 39 | 51 | 62 | 86 | 0.05 |
|       | p ≤ .01 | HCA > NHCA | 21 | 0 | 5 | 20 | 14 | 30 |
|       |         | HCA < NHCA | 0 | 5 | 3 | 0 | 4 | 0.01 |
| Total (of 30) | | 21 | 5 | 8 | 20 | 18 | 30 | 0.01 |
|       | p ≤ .05 | HCA > NHCA | 24 | 2 | 10 | 24 | 15 | 30 |
|       |         | HCA < NHCA | 0 | 8 | 7 | 0 | 4 | 0.05 |
| Total (of 30) | | 24 | 10 | 17 | 24 | 20 | 30 | 0.05 |

Table 5: Frequency of most common punctuation marks in title and abstract

| Character | Highly Cited Articles (n=4,956) | Non-Highly Cited Articles (n=4,998) |
|-----------|---------------------------------|-------------------------------------|
|           | Frequency | Percentage | Frequency | Percentage |
| Title     | ?         | 99 | 2.00% | 93 | 1.86% |
|           | :         | 1,153 | 23.26% | 916 | 18.33% |
|           | .         | 24 | 0.48% | 48 | 0.96% |
| Abstract  | ?         | 48 | 0.97% | 46 | 0.92% |

Table 6: Median of citations per year and median test by groups for titles in question and descriptive forms

|               | Question form | Descriptive | Median Test p-value |
|---------------|---------------|-------------|---------------------|
| HCA Cites per year | 19             | 25.5         | 0.00                |
| Frequency     | 99             | 4856        |                     |
| NHCA Cites per year | 0.75           | 1            | 0.17                |
| Frequency     | 93             | 4905        |                     |
Figure 1: Medians by groups of articles in 107 research areas
Figure 2: Scatterplots for the highly cited articles group
Table 3: Medians by groups and equality-of-medians tests (green colour below 0.05 level) for 6 bibliometric aspects in 107 research areas (Source: Web of Science)

| Area                                      | n | N Authors | N Title characters | N Abstract characters | N References | N Pages | JIF Percentile |
|-------------------------------------------|---|-----------|--------------------|-----------------------|--------------|---------|----------------|
| 1 Acoustics                               | 5 | 15        | 3                  | 2                    | 0.176        | 95      | 82             |
| 2 Agriculture                             | 56| 127       | 4                  | 4                    | 0.751        | 100     | 107            |
| 3 Allergy                                 | 5 | 4         | 7                  | 3                    | 0.294        | 91      | 105            |
| 4 Anesthesiology                          | 8 | 12        | 6                  | 5                    | 0.528        | 144     | 90.5           |
| 5 Anthropology                            | 8 | 9         | 4                  | 2                    | 0.09         | 121     | 71             |
| 6 Automation & Control Systems            | 63| 19        | 3                  | 2                    | 0.07         | 93      | 86             |
| 7 Behavioral Sciences                     | 8 | 7         | 2                  | 5                    | 0.833        | 63      | 99             |
| 8 Biochemistry & Molecular Biology        | 197| 180      | 7                  | 5                    | 0.001        | 90      | 104.5          |
| 9 Biodiversity & Conservation             | 12| 9         | 4.5                | 3                    | 0.445        | 96.5    | 105            |
| 10 Biomedical Social Sciences             | 12| 4         | 4                  | 1.5                  | 0.042        | 93.5    | 84             |
| 11 Biophysics                             | 14| 40        | 7                  | 5                    | 0.214        | 91.5    | 106.2          |
| 12 Biotechnology & Applied Microbiology   | 66| 89        | 5.5                | 4                    | 0.108        | 90      | 115            |
| 13 Business & Economics                   | 171| 101      | 2                  | 2                    | 0.194        | 74      | 80             |
| 14 Cardiovascular System & Cardiology     | 146| 78        | 12                 | 7                    | 0            | 119     | 110.5          |
| 15 Cell Biology                           | 104| 67       | 11                 | 6                    | 0            | 98      | 103            |
| 16 Chemistry                              | 719| 492      | 5                  | 4                    | 0            | 100     | 103            |
| 17 Communication                          | 14| 7         | 3                  | 2                    | 0.186        | 85      | 102            |
| 18 Computer Science                       | 149| 186      | 4                  | 3                    | 0            | 77      | 72             |
| 19 Construction & Building Technology     | 8  | 30        | 4.5                | 3                    | 0.078        | 95.5    | 83.5           |
| 20 Crystalligraphy                        | 5  | 8         | 2                  | 4.5                  | 0.279        | 112     | 90.5           |
| 21 Dermatology                            | 12| 24        | 6.5                | 4                    | 0.024        | 88.5    | 75.2           |
| 22 Developmental Biology                  | 5  | 9         | 7                  | 5                    | 0.057        | 99      | 129            |
| 23 Education & Educational Research       | 30| 38        | 3.2                | 3                    | 0.218        | 74.5    | 97             |
| 24 Electrochemistry                       | 65| 60        | 5                  | 4                    | 0.001        | 107     | 122            |
| 25 Endocrinology & Metabolism             | 58| 70        | 7                  | 5                    | 0            | 108.5   | 115.5          |
| 26 Energy & Fuels                         | 318| 108       | 5                  | 3                    | 0            | 103     | 96.5           |
| 27 Engineering                            | 466| 597       | 4                  | 3                    | 0            | 94      | 90             |
| 28 Environmental Sciences & Ecology       | 265| 192       | 4                  | 3                    | 0.001        | 98      | 101.5          |
| 29 Evolutionary Biology                   | 33| 10        | 4                  | 3                    | 0.481        | 93      | 117.5          |
| 30 Fisheries                              | 5  | 12        | 4                  | 4.5                  | 0.707        | 104     | 118.5          |
| 31 Food Science & Technology              | 73| 82        | 5                  | 2                    | 0.023        | 109     | 108.5          |
| 32 Forestry                               | 13| 31        | 5                  | 3                    | 0.027        | 106     | 107            |
| 33 Gastroenterology & Hepatology          | 55| 55        | 11                 | 7                    | 0            | 98      | 111            |
| 34 General & Internal Medicine            | 285| 116       | 12                 | 5                    | 0            | 86      | 92.8           |

**Note:** Green colour indicates equality-of-medians test below 0.05 level.
| Subject                        | Papercount | Citations | Citations | Articlecount | Articlecount |
|--------------------------------|------------|-----------|-----------|--------------|--------------|
| Pediatrics                    | 16         | 64        | 7.5       | 4            | 0.057        |
| Pharmacology & Pharmacy       | 57         | 122       | 5         | 5            | 0.269        |
| Physical Geography            | 15         | 14        | 4         | 2            | 0.035        |
| Physics                       | 550        | 636       | 5         | 5            | 0.093        |
| Physiology                    | 6           | 37        | 6.5       | 4            | 0.068        |
| Plant Sciences                | 88         | 90        | 7         | 4            | 0           |
| Polymer Science               | 14         | 68        | 4         | 3.5          | 0.45         |
| Psychiatry                    | 64         | 41        | 7         | 5            | 0.002        |
| Psychology                    | 98         | 75        | 3         | 3            | 0.436        |
| Public Administration         | 18         | 23        | 3         | 2            | 0.026        |
| Public, Environmental & Occupational Health | 151 | 93 | 6 | 4 | 0.62 |
| Radiology, Nuclear Medicine & Medical Imaging | 34 | 55 | 9 | 5 | 0.001 |
| Rehabilitation                | 5           | 22        | 4         | 3            | 0.438        |
| Remote Sensing                | 12         | 16        | 4.5       | 3            | 0.172        |
| Research & Experimental Medicine | 69 | 55 | 10 | 5 | 0.017 |
| Respiratory System            | 33         | 23        | 8         | 6            | 0.017        |
| Rheumatology                  | 16         | 18        | 9.5       | 4.5          | 0.006        |
| Science & Technology - Other Topics | 819 | 190 | 7 | 4 | 0.016 |
| Social Issues                 | 5           | 5         | 3         | 2            | 0.058        |
| Social Sciences - Other Topics | 14         | 12        | 2.5       | 2.5          | 1            |
| Sociology                     | 8           | 16        | 2         | 1.5          | 0.24         |
| Sport Sciences                | 7           | 23        | 5         | 3            | 0.195        |
| Substance Abuse               | 10          | 6         | 4.5       | 2.5          | 0.182        |
| Surgery                       | 46          | 126       | 9         | 5            | 0           |
| Tele26s                       | 36          | 49        | 4         | 3            | 0.125        |
| Thermodynamics                | 48          | 52        | 4         | 3            | 0.029        |
| Toxicology                    | 15          | 37        | 6         | 4            | 0.202        |
| Transplantation               | 7           | 10        | 14        | 8            | 0.008        |
| Transportation                | 10          | 10        | 3         | 3            | 1           |
| Urban Studies                 | 8           | 13        | 3         | 2            | 0.248        |
| Urology & Nephrology          | 28          | 45        | 8         | 6            | 0.016        |
| Veterinary Sciences           | 12          | 41        | 7.5       | 3            | 0.007        |
| Virology                      | 21          | 30        | 11        | 4.5          | 0           |
| Water Resources               | 29          | 55        | 4         | 3            | 0.054        |
| Zoology                       | 8           | 26        | 3         | 3            | 0.702        |