Received 30 May 2020
Revised 9 January 2021
6 March 2021
Accepted 17 March 2021

# Analysis of the scientific literature's abstract writing style and citations

Haotian Hu
*School of Information Management, Nanjing University, Nanjing, China and Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing, China*

Dongbo Wang
*School of Information Management, Nanjing Agricultural University, Nanjing, China, and*

Sanhong Deng
*School of Information Management, Nanjing University, Nanjing, China and Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing, China*

## Abstract

**Purpose** – The citation counts are an important indicator of scholarly impact. The purpose of this paper is to explore the correlation between citations of scientific articles and writing styles of abstracts in papers and capture the characteristics of highly cited papers' abstracts.

**Design/methodology/approach** – This research selected 10,000 highly cited papers and 10,000 zero-cited papers from the WOS (2008-2017) database. The Coh-Metrix 3.0 textual cohesion analysis tool was used to quantify the 108 language features of highly cited and zero-cited paper abstracts. The differences of the indicators with significant differences were analyzed from four aspects: vocabulary, sentence, syntax and readability.

**Findings** – The abstracts of highly cited papers contain more complex and professional words, more adjectives, adverbs, conjunctions and personal pronouns, but fewer nouns and verbs. The sentences in the abstracts of highly cited papers are more complex and the sentence length is relatively longer. The syntactic structure in abstracts of highly cited papers is relatively more complex and syntactic similarities between sentences are fewer. Highly cited papers' abstracts are less readable than zero-cited papers' abstracts.

**Originality/value** – This study analyses the differences between the abstracts of highly cited and those of zero-cited papers, reveals the common external and deep semantic features of highly cited papers in abstract writing styles, provide suggestions for researchers on abstract writing. These findings can help increase the scientific impact of articles and improve the review efficiency as well as the researchers' abstract writing skills.

**Keywords** Citations, Highly cited, Abstract writing styles, Coh-Metrix, Linguistic features

**Paper type** Research paper

## 1. Introduction

The citation number of academic papers not only reflects the degree of recognition of the researcher, but also becomes an important indicator of a scholar's research capabilities (e.g. h-index) and journal scholarly impact (e.g. Impact Factor, IF). Therefore, factors that contribute to an increase in the number of citations have aroused research interest among researchers.

This research is supported by the National Natural Science Foundation of China (Grant No. 71774084, 71503124 and 71673143) and National Social Science Fund of China (Grant No. 19BTQ062). The authors express their sincere gratitude towards Institute for Intelligent Systems, the University of Memphis, provider of Coh-Metrix tool, for their help and guidance. The authors thank two anonymous reviewers for helpful observations and suggestions.

Online Information Review
Vol. 45 No. 7, 2021
pp. 1290-1305
© Emerald Publishing Limited
1468-4527
DOI 10.1108/OIR-05-2020-0188

Most of the existing researches have investigated into factors such as the title of the paper (Guo *et al.*, 2018), its authors (Potthoff and Zimmermann, 2017) and the references (Gong *et al.*, 2019). These types of text have been extensively researched because they are readily accessible and would not require too much computational resources to process. Convenient as it is, this approach brings certain drawbacks. Take the title data of the paper as an example. The average number of words in the title of the papers collected by WOS (Web of Science) is 9.53 words (Guo *et al.*, 2018). That is to say, the number of citations of papers can be judged only by the essays with less than 10 words, which obviously leaves out a lot of information.

However, the abstract text contains richer information regarding the content of the article. An abstract is a highly concise version of the paper. Researchers often read the abstract to grasp the main idea of a paper (Dowling *et al.*, 2018) and decide whether they would like to continue reading or cite the paper. Therefore, we believe that there is a certain correlation between abstracts and citation numbers. In addition, compared with full-text data, abstracts are easier to obtain. At present, studies on the relationship between the abstract and the number of citations are rare, while the available literature is mainly based on features like the length (Didegah and Thelwall, 2013; Wesel *et al.*, 2014; Sohrabi and Iraj, 2017; Xie *et al.*, 2019) and readability (Gazni, 2011; Didegah and Thelwall, 2013; Wesel *et al.*, 2014; Lei and Yan, 2016; Didegah *et al.*, 2018; Fages, 2020) of abstracts. The research gap is that none of these studies have analyzed and extracted the syntactic structure, lexical complexity and deep semantic information of the abstract texts. We think to some extent the potential influencing factors might be omitted.

Our study, a comprehensive one, has taken the surface features and deep features of the abstract into consideration and explored the relationships between the writing styles of abstracts and the number of citations of the paper from the perspectives of linguistic coherence and consistency. With the help of the Coh-Metrix 3.0 tool (Graesser *et al.*, 2004), we calculated 108 language coherence indicators for top 10,000 highly cited papers and 10,000 zero-cited paper abstracts in the WOS (2008–2017) database. The differences between the abstracts of highly cited and those of zero-cited papers were analyzed from four aspects, namely lexical selection, sentence characteristics, syntactic use and overall readability, and the linguistic features of abstracts of highly cited papers were summarized as well. Our study can help researchers write targeted standard abstracts in a more comprehensive and accurate way. In addition, the findings can help predict the citations to some extent.

## 2. Literature review

Title is one of the most accessible, processed and analyzed texts in the paper. Previous studies have explored the relationship between the number of citations and the length of the title (Guo *et al.*, 2018), the type of sentence involved (Jamali and Nikzad, 2011) and the characteristics of the title (Gnewuch and Wohlrabe, 2017; Rostami *et al.*, 2014; Nair and Gibbert, 2016). The findings indicate that there is a certain degree of correlation between the title and the number of citations. The above studies have two issues to be resolved. The first issue is that title-based research usually only analyses the external characteristics of the title, such as the length of the title, the sentence structure, the number of keywords included. However, it is rare to find study mined deeply into the intrinsic features of titles. The second issue is that the title is a very short text, which leads to the inevitable defect that it contains only a very small amount of useful information. Therefore, it is difficult to predict the number of citations only by the title.

As a result, some researchers set their sights on the author data of the paper. Ibáñez *et al.* (2013) revealed that the citation number of papers composed by two or less than two authors outnumbered those with more than two authors, and collaborative papers written by authors from different countries were more cited than the ones written by authors from the same country in computer science field. Potthoff and Zimmermann (2017) found that scholars preferred to cite papers whose authors were of the same gender. These studies revealed that there is a certain bias in the process of citing papers. However, although the author's academic influence, gender and collaborators will significantly affect the number of citations, these tend to be the author's own attributes, rather than the attributes of the text writing style. Therefore, for general researchers, it is difficult to expect to improve the number of citations by adjusting the author factor.

References can best represent the citation of literature. Therefore, researchers investigated into the rules of citation by means of references. Wesel *et al.* (2014) found that the number of references is significantly related to citations in disciplines like sociology, general and internal medicine and applied physics. Gong *et al.* (2019), on the basis of CSSCI data sources in China, found that for all disciplines, the citation number of papers containing foreign references was significantly higher than that without foreign references. Stevens and Duque (2019) found that when the references were ordered alphabetically, papers rank in the front had a higher possibility of being cited. Reference factor is more feasible as a means to increase the number of citations. However, there is still not much the author can do about this.

The above-mentioned information categories such as title, author, reference and other texts are easier to process. However, compared with the abstract and full text, the information contained is quite limited. Sohrabi and Iraj (2017) explored the abstract text further and found that the ratio of the occurrence of keywords in the abstract to the length of the abstract could be used to predict the number of citations. In addition to studying the correlation between the length of abstract and the number of citations (Didegah and Thelwall, 2013; Wesel *et al.*, 2014; Didegah *et al.*, 2018), the readability of abstract has also been explored. Vergoulis *et al.* (2019) used four readability metrics and three scientific impact measures to calculate the readability and scientific impact of all abstracts in the Open Citations COCI dataset. The results show that the overall readability of scientific publications is declining, and the readability of abstracts is not significantly related to scientific impact. Lei and Yan (2016) selected four core journals in the field of information science, used FRE and SMOG readability indicators to explore the relationship between the number of citations and the readability of abstracts, also found that there was no significant correlation. Gazni (2011) explored the correlation between abstracts produced by five top institutions in the world and reading difficulty and found that abstracts of highly cited papers are often difficult to read. Didegah *et al.* (2018) analyzed abstracts of papers written by Finnish authors in WOS and found that the number of citations with easier abstracts to read was significantly reduced compared with those with more difficult abstracts. Didegah and Thelwall (2013) found that the readability of abstract was negatively correlated with the number of citations in Biology and Biochemistry. For Economics (Didegah *et al.*, 2018; Fages, 2020), the abstracts of papers that are easy to read have a positive correlation with the citations and popularity. In addition, Hafeez *et al.* (2019) found that in Psychiatry, the use of structured abstracts can enhance the academic impact of journals and papers. The abstract is the data type with the longest text length in the metadata of the paper. It can present the main frame and experimental flow of the paper comprehensively. By analyzing the abstract, more factors affecting the number of citations can be discovered. This is one of the considerations why we chose abstract text in WOS as the experimental data to explore the correlation between abstract and citations. We believe that the readability of the abstract should be related to the number of citations (Gazni, 2011; Didegah *et al.*, 2018; Dowling *et al.*, 2018). The reason is that an easy-to-understand writing method is certainly more conducive to readers' understanding than an obscure writing method.

Compared with studies only focused on a few factors, some researchers have explored dozens of factors that influence the citations at the same time, which is more likely to draw more comprehensive and reliable conclusions. Tahamtan *et al.* (2016) identified 28 relevant factors that affected the number of cited papers and summarized the correlation between each influencing factor and the number of citations. Xie *et al.* (2019) explored 66 influencing factors that may relate to the citation numbers of papers and identified 46 influencing factors that were significantly related to the number of cited papers in the field of library and information science. Such studies are sufficient in breadth but insufficient in depth. Metadata is still rarely analyzed from the semantic level. The conclusions obtained by previous studies were mostly statistical results based on external characteristics. The research of Xie *et al.* (2019) provided ideas for the selection of indicators for our experiment: extensively identify several potential factors related to the number of citations, filter out significantly related factors and analyze these factors to find their correlation with the number of citations.

We found that most of the previous researches were based on short text level, when exploring the issue of “what factors are related to the number of citations?”. However, the information that can be expressed in such texts is very limited and cannot fully reflect the true writing style of the paper, making it difficult to find out the significant relationship between them and the number of citations. In addition, when conducting scientific studies, researchers rarely decide whether to cite a paper just by reading the title and author information of the paper. Therefore, previous researches often lack a reasonable explanation of causality; the robustness and universality of conclusions need to be further tested.

Abstract is the summary of the paper, which contains far more information than the title and other metadata. Theoretically, it is more closely related to the number of citations and more worthy of investigation. However, most of the previous studies looked at only a few indicators, such as abstract length and readability, and the extraction of abstract semantic information was not deep enough. To solve the problems, we used the 108 indicators provided by the Coh-Metrix tool to conduct a comprehensive and in-depth analysis of the abstracts of WOS papers. Explored deeper text features from vocabulary, syntax, semantics and cohesiveness, so as to help us better understand the relationship between the abstract language style and the number of citations. In addition, we provided suggestions for researchers to write high-level abstracts and to maximize the impact of their papers.

## 3. Method

### *3.1 Source of data*

The SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH indexed documents in the WOS (Web of Science) core collection database were selected as the data source. The time range was limited to 2008–2017. Papers after 2017 were not selected because they were published late and have not accumulated enough citations (Lei and Yan, 2016). We manually downloaded the records in the above time range from WOS and stored them in the SQL Server database. The downloaded WOS data included multiple document formats such as article, proceedings paper, retracted publication, editorial material, etc. To ensure that the articles are all research papers (Didegah and Thelwall, 2013), we limited the document types to article and proceedings paper. The number of citations in different disciplines is not the same (Slyder *et al.*, 2011; Marx and Bornmann, 2015; Tahamtan *et al.*, 2016), but because our research wants to explore the universal difference between abstracts with high citation and those with zero citation in the whole field of scientific research, so when selecting highly cited papers, we valued the number of citations more than the subject category. In addition, we have limited the publication type to journal and the language to English. Finally, 10,000 papers with the highest number of citations (between 26,150 and 372) in this decade were selected. The distribution of the number of highly cited papers over time is shown in Figure 1.

As can be seen from Figure 1, the number of papers gradually decreased with the increase of years. In 2008, the number of highly cited papers was the largest, with 2,432 papers, and in 2017, the number of highly cited papers was the least, with only 2 papers. This is because on

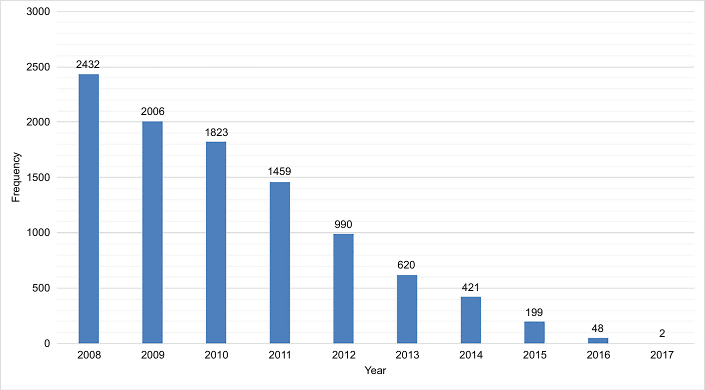


**Figure 1.**
Highly cited papers distributed over time

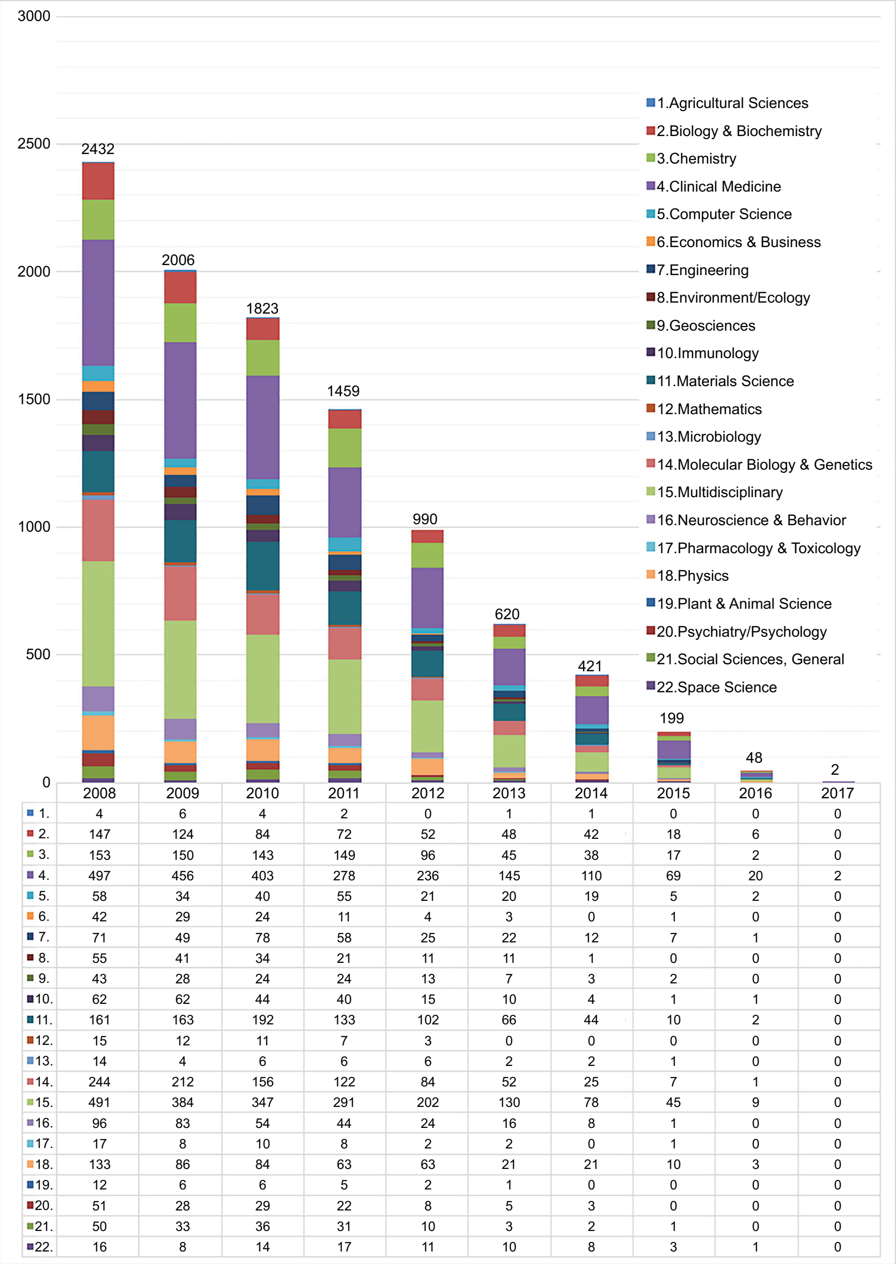
the one hand, the earlier a paper is published, the larger the time window it can be retrieved, and on the premise of the same quality of the paper, the more likely it is cited. On the other hand, since the selected top 10,000 cited papers are based on the number of cited papers in all WOS databases from 2008 to 2017, it is understandable that those published earlier enjoy a better chance of being cited, hence a higher possibility of being ranked among the top 10,000 cited papers.

The data extraction process of zero-cited papers is more complicated. The choice of database, time limit, document type, etc. was consistent with the highly cited papers. In order to ensure the consistency of the discipline distribution of the zero-cited papers and the highly cited papers dataset, we classified the selected highly cited papers into 22 disciplines according to the ESI (Essential Science Indicators) discipline category (Didegah and Thelwall, 2013) provided by Clarivate Analytics. Figure 2 shows the distribution of the number of papers from various disciplines over time. We selected the same number of zero-cited papers according to the number of highly cited papers from each discipline in each year. However, since the number of zero-cited papers in "Multidisciplinary" in 2008 and 2009 was relatively small, we selected a total of 219 zero-cited papers from the same discipline in 2010 and 2011, which were close in time, as supplements. Through the above restrictions, we finally ensured that the distribution of highly cited papers and zero-cited papers in terms of quantity, time, subject and document type are the same.

The above two sets of data constituted the highly cited group and zero-cited group respectively and were used for subsequent comparisons.

### *3.2 Coh-Metrix tool*

Coh-Metrix is a text automatic analysis tool developed by the University of Memphis (Graesser *et al.*, 2004; McNamara *et al.*, 2013, 2014), which can comprehensively analyze the coherence and consistency of input text by calculating surface and deep language features. The latest version is Coh-Metrix 3.0. This tool is widely used for evaluation of writing quality (Zedelius *et al.*, 2019; Macarthur *et al.*, 2019), writing ability (Perin and Lauterbach, 2018) and cognitive understanding (Wiley *et al.*, 2017). Some researchers have also proposed new text analysis methods based on Coh-Metrix tools (Wolfe *et al.*, 2019) or developed new application platforms (Shi *et al.*, 2018; McNamara *et al.*, 2013). In addition, relevant researchers have carried out some research on the measurement of text readability and complexity (Tortorelli, 2020; Spencer *et al.*, 2019), which also provides some valuable guidance and reference for the selection of method in this paper.



| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|
| ■ 1. | 4 | 6 | 4 | 2 | 0 | 1 | 1 | 0 | 0 | 0 |
| ■ 2. | 147 | 124 | 84 | 72 | 52 | 48 | 42 | 18 | 6 | 0 |
| ■ 3. | 153 | 150 | 143 | 149 | 96 | 45 | 38 | 17 | 2 | 0 |
| ■ 4. | 497 | 456 | 403 | 278 | 236 | 145 | 110 | 69 | 20 | 2 |
| ■ 5. | 58 | 34 | 40 | 55 | 21 | 20 | 19 | 5 | 2 | 0 |
| ■ 6. | 42 | 29 | 24 | 11 | 4 | 3 | 0 | 1 | 0 | 0 |
| ■ 7. | 71 | 49 | 78 | 58 | 25 | 22 | 12 | 7 | 1 | 0 |
| ■ 8. | 55 | 41 | 34 | 21 | 11 | 11 | 1 | 0 | 0 | 0 |
| ■ 9. | 43 | 28 | 24 | 24 | 13 | 7 | 3 | 2 | 0 | 0 |
| ■ 10. | 62 | 62 | 44 | 40 | 15 | 10 | 4 | 1 | 1 | 0 |
| ■ 11. | 161 | 163 | 192 | 133 | 102 | 66 | 44 | 10 | 2 | 0 |
| ■ 12. | 15 | 12 | 11 | 7 | 3 | 0 | 0 | 0 | 0 | 0 |
| ■ 13. | 14 | 4 | 6 | 6 | 6 | 2 | 2 | 1 | 0 | 0 |
| ■ 14. | 244 | 212 | 156 | 122 | 84 | 52 | 25 | 7 | 1 | 0 |
| ■ 15. | 491 | 384 | 347 | 291 | 202 | 130 | 78 | 45 | 9 | 0 |
| ■ 16. | 96 | 83 | 54 | 44 | 24 | 16 | 8 | 1 | 0 | 0 |
| ■ 17. | 17 | 8 | 10 | 8 | 2 | 2 | 0 | 1 | 0 | 0 |
| ■ 18. | 133 | 86 | 84 | 63 | 63 | 21 | 21 | 10 | 3 | 0 |
| ■ 19. | 12 | 6 | 6 | 5 | 2 | 1 | 0 | 0 | 0 | 0 |
| ■ 20. | 51 | 28 | 29 | 22 | 8 | 5 | 3 | 0 | 0 | 0 |
| ■ 21. | 50 | 33 | 36 | 31 | 10 | 3 | 2 | 1 | 0 | 0 |
| ■ 22. | 16 | 8 | 14 | 17 | 11 | 10 | 8 | 3 | 1 | 0 |

**Figure 2.**
Papers from 22 disciplines distributed over time

The Coh-Metrix software can calculate 108 indicators in 11 groups of input text, namely, descriptive, text easability principal component scores, referential cohesion, latent semantic analysis (LSA), lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information and readability.

### 3.3 Data analysis method

We used the Coh-Metrix tool to calculate a total of 108 indicators such as the cohesiveness and legibility of the highly cited and zero-cited paper abstracts. Using Python 3.5.2, we combined the papers' citation number and 108 Coh-Metrix indexes and stored them in Microsoft Office Excel 2016 workbook. The independent sample *t*-test was carried out by Python's Scipy library to test the difference between the quantified output indicators of the highly cited group and the zero-cited group.

## 4. Result

Independent sample *t*-test shows that 91 of the 108 Coh-Metrix output indicators have significant differences. The Independent sample *t*-test uses two-tailed test. In order to reduce the probability of false positives, we used Bonferroni Correction to correct the *t*-test results. Finally, 85 indicators have differences at the significance level $p = 0.05/108$. Because of limited space, we only selected indicators that have significant differences among the following seven easy to interpret and understand categories of indicators: descriptive, word information, connectives, readability, referential cohesion, syntactic complexity and syntactic pattern density. Please note that in Table 1–Table 7, the highly cited group in the "Group" field is represented by "1", and the zero-cited group is represented by "0".

### 4.1 Descriptive

Descriptive indicators include statistical data such as word length, sentence length and paragraph length. The surface characteristics of the abstract can be analyzed from a macro perspective. Table 1 shows the indicators, scores and significance test results related to descriptiveness.

The number of sentences and words, length of paragraphs and sentences of the highly cited group abstracts are significantly higher than those in the zero-cited group. This indicates that the length of abstracts of highly cited paper is usually longer and more informative. This finding is consistent with previous research (Sohrabi and Iraj, 2017; Xie *et al.*, 2019). A longer abstract can provide a more comprehensive and detailed introduction to the overall content of the article, so that readers can understand the core research results. Shorter abstracts, instead, are more likely to suffer from problems such as incomplete introduction to the content of the article. It is safe to conclude that a paper with a longer

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Number of sentences | 1 | 8.280 | 4.095 | 20.061 | 0.000 |
| | 0 | 7.138 | 3.954 | | |
| Number of words | 1 | 196.626 | 89.937 | 32.235 | 0.000 |
| | 0 | 157.164 | 83.052 | | |
| Length of paragraphs | 1 | 8.280 | 4.095 | 20.061 | 0.000 |
| | 0 | 7.138 | 3.954 | | |
| Length of sentences | 1 | 26.086 | 12.681 | 8.428 | 0.000 |
| | 0 | 24.520 | 13.586 | | |
| Length of words (syllables) | 1 | 1.973 | 0.161 | 25.552 | 0.000 |
| | 0 | 1.910 | 0.188 | | |
| Length of words (letters) | 1 | 5.768 | 0.446 | 29.993 | 0.000 |
| | 0 | 5.564 | 0.515 | | |

**Note(s):** Length of paragraphs: average number of sentences in each paragraph; Length of sentences: average number of words in each sentence; Length of words: average number of syllables or letters in all of the words

**Table 1.**
Descriptive results and significance test

**Table 2.** Word information results and significance test

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Incidence score of nouns | 1 | 338.212 | 40.784 | −6.183 | 0.000 |
|  | 0 | 342.149 | 48.885 |  |  |
| Incidence score of verbs | 1 | 95.398 | 23.429 | −9.000 | 0.000 |
|  | 0 | 98.594 | 26.682 |  |  |
| Incidence score of adjectives | 1 | 146.533 | 37.688 | 23.956 | 0.000 |
|  | 0 | 132.412 | 45.320 |  |  |
| Incidence score of adverbs | 1 | 31.178 | 17.125 | 22.035 | 0.000 |
|  | 0 | 25.675 | 18.174 |  |  |
| Incidence score of all personal pronouns | 1 | 14.648 | 12.033 | 11.472 | 0.000 |
|  | 0 | 12.551 | 13.770 |  |  |
| Incidence score of first person pronoun singular | 1 | 0.312 | 2.174 | −6.129 | 0.000 |
|  | 0 | 0.565 | 3.511 |  |  |
| Incidence score of first person pronoun plural | 1 | 9.379 | 9.326 | 31.485 | 0.000 |
|  | 0 | 5.253 | 9.205 |  |  |
| Incidence score of third person pronoun singular | 1 | 0.088 | 1.396 | −12.652 | 0.000 |
|  | 0 | 0.665 | 4.348 |  |  |
| Word acquisition age score | 1 | 404.260 | 33.096 | 8.316 | 0.000 |
|  | 0 | 399.731 | 43.256 |  |  |
| Word familiarity rating | 1 | 550.568 | 12.205 | −6.156 | 0.000 |
|  | 0 | 552.080 | 21.307 |  |  |
| Word concreteness rating | 1 | 381.162 | 30.101 | −7.727 | 0.000 |
|  | 0 | 384.728 | 34.975 |  |  |
| Word imagery rating | 1 | 402.162 | 25.204 | −7.495 | 0.000 |
|  | 0 | 405.163 | 31.095 |  |  |

**Note(s):** The incidence scores are occurrence per 1,000 words (McNamara *et al.*, 2014). Word acquisition age score: the higher the score, the later the word is learned by children (e.g. word *cortex* scores higher than *milk*); Word familiarity rating: sentence with more familiar word to adults gets higher score (e.g. word *water* scores higher than *calix*); Word concreteness rating: word with higher score is more concrete and less abstract (e.g. word *box* scores higher than *protocol*); Word imagery rating: low imagery word is difficult to construct a mental image and has lower score (e.g. word *hammer* scores higher than *overtone*)

**Table 3.** Connectives results and significance test

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| All connectives incidence | 1 | 78.930 | 22.916 | 13.030 | 0.000 |
|  | 0 | 74.475 | 25.373 |  |  |
| Logical connectives incidence (e.g., and, or) | 1 | 28.190 | 14.576 | 17.122 | 0.000 |
|  | 0 | 24.465 | 16.152 |  |  |
| Adversative and contrastive connectives incidence (e.g., | 1 | 9.511 | 8.792 | 19.954 | 0.000 |
| although, whereas) | 0 | 7.028 | 8.800 |  |  |
| Expanded temporal connectives incidence (e.g., first, until) | 1 | 15.522 | 12.567 | 11.519 | 0.000 |
|  | 0 | 13.434 | 13.066 |  |  |
| Additive connectives incidence (e.g., and, moreover) | 1 | 46.835 | 17.690 | 17.497 | 0.000 |
|  | 0 | 42.278 | 19.115 |  |  |
| Positive connectives incidence (e.g., also, moreover) | 1 | 72.674 | 22.745 | 5.070 | 0.000 |
|  | 0 | 70.961 | 24.988 |  |  |
| Negative connectives incidence (e.g., however, but) | 1 | 7.736 | 7.732 | 21.982 | 0.000 |
|  | 0 | 5.384 | 7.397 |  |  |

**Note(s):** The incidence scores are occurrence per 1,000 words

**Table 4.**
Readability results and significance test

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Flesch Reading Ease Score | 1 | 16.141 | 12.210 | −34.501 | 0.000 |
| | 0 | 22.662 | 14.428 | | |
| Flesch-Kincaid Grade Level | 1 | 17.454 | 5.053 | 16.537 | 0.000 |
| | 0 | 16.222 | 5.477 | | |
| Second Language Readability Score | 1 | 6.301 | 5.172 | −15.659 | 0.000 |
| | 0 | 7.602 | 6.506 | | |

**Note(s):** The scores of Flesch Reading Ease Score (Score = 206.835 - (1.015 × Average sentence length) – (84.6 × Average number of syllables per word)) and Second Language Readability Score are positively correlated with readability; Text with higher Flesch-Kincaid Grade Level (Score = (0.39 × Average sentence length) + (11.8 × Average number of syllables per word) – 15.59) is harder to read

**Table 5.**
Referential Cohesion results and significance test

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Local noun overlap | 1 | 0.635 | 0.253 | 13.357 | 0.000 |
| | 0 | 0.583 | 0.298 | | |
| Local argument overlap | 1 | 0.697 | 0.243 | 15.473 | 0.000 |
| | 0 | 0.638 | 0.291 | | |
| Local stem overlap | 1 | 0.761 | 0.225 | 19.691 | 0.000 |
| | 0 | 0.690 | 0.283 | | |
| Global noun overlap | 1 | 0.572 | 0.233 | 11.474 | 0.000 |
| | 0 | 0.531 | 0.274 | | |
| Global argument overlap | 1 | 0.631 | 0.228 | 13.360 | 0.000 |
| | 0 | 0.584 | 0.273 | | |
| Global stem overlap | 1 | 0.703 | 0.220 | 18.135 | 0.000 |
| | 0 | 0.640 | 0.269 | | |
| Local anaphor overlap | 1 | 0.108 | 0.169 | 4.918 | 0.000 |
| | 0 | 0.096 | 0.183 | | |
| Global content word overlap | 1 | 0.106 | 0.050 | −3.573 | 0.000 |
| | 0 | 0.109 | 0.070 | | |

**Note(s):** All of the indices are binary, mean. Noun overlap: measures of local and global overlap between sentences in terms of nouns; Argument overlap: the overlap between sentences in terms of nouns and pronouns; Stem overlap: the overlap that a noun in one sentence is matched with a content word in a previous sentence that shares a common lemma; Anaphor overlap: the overlap that the later sentence contains a pronoun which refers to a pronoun or noun in the earlier sentence; content word overlap: content words that overlap between sentence pairs

abstract will be more likely to be cited. In addition, readers are prone to such mentality that the papers with long abstracts will be more detailed and serious than those with short abstract papers, and thus the research results will be more valuable.

The average word syllables and average word letters of the highly cited group abstracts are also significantly higher than those in the zero-cited group, which indicates that highly cited papers often contain longer, less-frequently used words that are unfamiliar to the readers. Words with the above-mentioned characteristics are often professional terms in a specific scientific field. This shows that the use of terms in the abstracts makes them more professional and might increase the citation number.

### *4.2 Word information*

Words are the basic units of the abstract. The part of speech and the legibility of the words will affect the readability of the abstract. Table 2 shows the indicators, scores and significance test results related to word information.

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Number of words before main verb | 1 | 5.849 | 2.801 | −13.306 | 0.000 |
| | 0 | 6.494 | 3.958 | | |
| Number of modifiers per noun phrase | 1 | 1.308 | 0.272 | −10.629 | 0.000 |
| | 0 | 1.352 | 0.307 | | |
| Minimum editorial distance (POS tags) | 1 | 0.596 | 0.074 | 9.795 | 0.000 |
| | 0 | 0.582 | 0.116 | | |
| Minimum editorial distance (words) | 1 | 0.871 | 0.081 | 17.265 | 0.000 |
| | 0 | 0.843 | 0.139 | | |
| Minimum editorial distance (lemmas) | 1 | 0.853 | 0.080 | 19.840 | 0.000 |
| | 0 | 0.822 | 0.137 | | |
| Syntactic structure similarity adjacent | 1 | 0.089 | 0.035 | −19.446 | 0.000 |
| | 0 | 0.101 | 0.053 | | |
| Syntactic structure similarity global | 1 | 0.085 | 0.029 | −21.547 | 0.000 |
| | 0 | 0.097 | 0.046 | | |

**Note(s):** Minimum editorial distance refers to the minimum editing times needed to convert one sentence into another, scores between adjacent sentences are computed from part of speech tags, words and lemmas respectively; Syntactic structure similarity is the proportion of intersection tree nodes between the syntactic trees of all adjacent sentences or across paragraphs

**Table 6.** Syntactic complexity results and significance test

| Indice | Group | Mean | SD | *t* | Sig |
|---|---|---|---|---|---|
| Noun phrase density | 1 | 387.899 | 47.129 | 5.089 | 0.000 |
| | 0 | 384.295 | 52.862 | | |
| Verb phrase density | 1 | 138.497 | 34.356 | −11.426 | 0.000 |
| | 0 | 144.360 | 38.119 | | |
| Adverbial phrase density | 1 | 18.643 | 12.055 | 21.339 | 0.000 |
| | 0 | 14.934 | 12.521 | | |
| Preposition phrase density | 1 | 134.894 | 25.477 | −9.611 | 0.000 |
| | 0 | 138.831 | 32.077 | | |
| Agentless passive voice form density | 1 | 12.515 | 10.375 | −40.799 | 0.000 |
| | 0 | 19.751 | 14.386 | | |
| Negation density | 1 | 2.454 | 4.079 | −5.292 | 0.000 |
| | 0 | 2.815 | 5.453 | | |
| Gerund density | 1 | 18.216 | 13.130 | 9.292 | 0.000 |
| | 0 | 16.420 | 14.183 | | |

**Note(s):** The density scores are occurrence per 1,000 words

**Table 7.** Syntactic Pattern Density results and significance test

The noun and verb usage of highly cited group are significantly smaller than the zero-cited group, and the adjective and adverb usage rate are significantly higher than the zero-cited group. This implies that if a paper's abstract contains fewer nouns and verbs while contains more adjectives and adverbs, it will be more likely to be cited.

With regard to the pronoun usage of the highly cited group, it tends to use more personal pronouns, especially the first person pronoun plural. However, the usage rate of the first person pronoun singular and the third person pronoun singular is lower than that of zero-cited group. Although there is no significant difference in the incidence score of third person pronoun plural ($p = 0.265$), the highly cited group scores 0.08 higher than the zero-cited group. This implies that the abstracts of highly cited papers reveal a tendency of having group personal pronouns such as "we/our". Yet first-person singular pronouns such as "I/me" are used less frequently. This is because scientific research is usually carried out by research teams. Therefore, when introducing research in the abstract of the paper, "our research" is often used instead of the phrase "my research". Meanwhile, "their research" is used instead of "his or her research".

The highly cited group acquisition age score is higher than the one of zero-cited group, and the word familiarity rating is lower than the one of zero-cited group. These two indicators show that words in the highly cited papers' abstracts belong to those that are acquired by the children in the advanced level and are not familiar to the adults. The following are two sentences from the abstracts with a word acquisition age score of 546 and 309 respectively. (1)"Furthermore, inhibition of autophagy delays the senescence phenotype, including senescence-associated secretion." (2)"The approach is based on measuring the distances between the faces in an embedding of a planar graph." The concreteness rating and the imaginable rating of the highly cited group are lower than those of the zero-cited group, which also indicates that papers with higher citations pose higher difficulty for readers to understand. Combined with the two indicators of the length of word in Table 1, we can conclude that in the abstracts of highly cited papers, it is more likely to encounter less frequently used, complex words or terms that are difficult to understand. They are generally relating to abstract concepts and methods.

### *4.3 Connectives*

Connectives, as an important part of articles, function as connecting contextual contents. Table 3 shows the indicators, scores and significance test results associated with the connectives.

Except for causal connectives ($p = 0.322$), the usage rate of all other types of connectives of the highly cited group is higher than that of the zero-citation group. The results show that compared with papers with zero citations, abstracts with more citations usually have more connectives to enhance cohesion between words and sentences, which unite abstracts as a whole. The use of connectives makes the logical expression of sentences more rigorous, makes the expression of research process and conclusion more abundant and organized, but it may increase the difficulty of comprehension to some extent.

### *4.4 Readability*

Coh-Metrix provides three common readability calculation methods: the Flesch Reading Ease Score, the Flesch Kincaid Grade Level and the Second Language Readability Score. Table 4 shows the scores and significance test results of these three readability indicators.

Flesch Reading Ease Score (Vergoulis *et al.*, 2019) and the Second Language Readability Score (Crossley *et al.*, 2008) of the highly cited group are both lower than the ones of the zero-cited group, and Flesch Kincaid Grade Level (Fages, 2020) is higher than that of the zero-cited group. All three indexes indicate that the readability of abstracts with more citations in WOS is lower than that of papers with zero citation. This result suggests that abstracts of highly cited papers are not always easy to read. This finding also verifies the previous studies (Gazni, 2011; Didegah *et al.*, 2018; Didegah and Thelwall, 2013). The reason lies in the characteristics of the papers in the highly cited group. They involve innovation and professionalism, which means new concepts, methods, models, etc. In order to ensure the conciseness of the abstracts, information is condensed and details omitted, thereby raising the difficulty of reading and understanding of the abstracts. This result also confirms the conclusion drawn in Section 4.1 and Section 4.2, i.e. highly cited abstracts often reflect strong professionalism. In order to present the difference in readability more intuitively, we selected the following two sentences from the abstracts with Flesch Reading Ease Score of 1 and 57 respectively. (1) "Superlenses have great potential in applications such as biomedical imaging, optical lithography and data storage." (2) "It was observed that a total of 33 water harvesting structures were constructed between 2004 and 2007."

### *4.5 Referential cohesion*

Referential cohesion refers to the overlap or common reference of content words (i.e. noun, verb, adjective and adverb) between sentences. Referential cohesion is divided into local cohesion (overlap between adjacent sentences) and global cohesion (overlap between all sentences in the abstract). In particular, referential cohesion is subdivided into more specific categories for analysis. Table 5 shows the indicators, scores and significance test results of the referential cohesion module.

The local and global noun overlap, argument overlap and stem overlap in the highly cited group are all higher than that in the zero-cited group. That is, there is a greater probability of the same noun, pronoun and cognate in the two preceding and following sentences of the highly cited paper abstracts, whether between adjacent sentences or among all sentences in the text. The local anaphor overlap in the highly cited group is also higher than that in the zero-cited group, which indicates that the abstracts of the highly cited papers tend to refer to the noun or pronoun in the preceding sentence by pronouns, thus connecting the text content before and after the pronouns. Global content word overlap is the only indicator in which the highly cited group scores lower than the zero-cited group. For noun overlap, argument overlap, stem overlap and content word overlap, the overlapping words in the two sentences are either the same or stem from the same lemma. This ensures that the referential relations between the two sentences are one-to-one correspondence, and that the two words refer to the same thing. In anaphor overlap, the pronoun included in the latter sentence can refer to both the pronoun in the preceding sentence and the noun in the preceding sentence, which may cause unclear referential relation to some extent, resulting in difficulty of understanding.

### *4.6 Syntactic complexity*

Simple syntactic structure is easy to understand. On the contrary, complex syntactic structure contains various modifiers, subject-subordinate sentences, inverted sentences and other special sentence patterns, which increases the difficulty of understanding and memorizing. Table 6 shows the indicators, scores and significance test results of the syntactic complexity module.

Words before the main verb of the main clause and modifiers per noun phrase in the highly cited group are all fewer than those in the zero-cited group, indicating that the abstracts of the papers with more citations are less likely to use adverbs, adjectives, etc. to modify the main actions and the nouns in noun phrases. The minimum editing distance is one of the common indexes to measure text similarity. The calculation results of the three minimum editing distances based on part of speech tags, words and lemmas show that the score of highly cited group are all higher than that of the zero-cited group, indicating that the similarities between adjacent sentences of the highly cited paper abstract are small and the content differences of the statements are large. Syntactic structure similarity adjacent and global scores of highly cited group are all relatively low, which not only in accord with the finding of minimum editing distance in adjacent sentences, but also demonstrates that the syntactic structure similarity of any two sentences in the highly cited paper abstract is relatively low compared with that of the zero-cited paper. A good abstract must take into account syntactic brevity and content comprehensiveness. Therefore, there will be deletions of unnecessary modifiers and reductions in semantic and syntactic duplication between sentences.

### 4.7 Syntactic pattern density

Syntactic pattern density is used to illustrate the statistical incidence score of various types of words and phrases. Table 7 shows various indicators, scores and significance test results of the syntactic pattern density module.

Compared with the zero-cited group, the highly cited group has a higher incidence of noun phrases, adverbial phrases and gerunds, indicating that the abstracts of the highly cited papers have relatively complex syntax and contain relatively dense information (Graesser *et al.*, 2004). The incidence of verb phrases, preposition phrase, agentless passive voice forms and negation expressions are lower in the highly cited group. The reason may be the abstract of a highly cited paper often takes gerunds in the beginning of the sentence. Noun phrases and adverbial phrases, instead, are more often used for richer and more precise explanation of the research in the abstract.

## 5. Discussion

In addition to the above findings, regarding Coh-Metrix indicators with significant differences, the standard deviation (SD) of means of the highly cited group are almost all less than that of the zero-cited group. This indicates that the overall data fluctuation range of papers with more citations is smaller, and the shared characteristics reflected in these papers are consistent and universal, while papers with zero citation are exactly the opposite.

Our study reveals the differences in abstract writing style between highly cited papers and zero-cited papers in WOS database.

First is the difference in lexical selection: (1) Abstracts of highly cited papers tend to use more complex, difficult and professional words. Compared with zero-cited papers, words from highly cited abstract contain more syllables and letters; they require older acquisition age; they have lower familiarity degree of adults; they are involved with concepts and methods that are more abstract and professional. (2) Abstract of highly cited papers tends to use fewer nouns and verbs but more adjectives, adverbs and personal pronouns. As regards personal pronouns, the plural form of first person is more frequently used, but first and third person pronouns singular are less commonly used. (3) Abstracts of highly cited papers contain more conjunctions, but there was no significant difference in the use of causal conjunctions.

In terms of sentence characteristics, (1) the abstract length of highly cited papers is relatively longer. The average number of sentences, the average number of words and the average length of sentences are all larger than those of zero-cited paper. A longer abstract will allow for a more detailed and clearer introduction of the research content. (2) Because of more conjunctions and complex terms involved, sentences in the abstracts of highly cited papers may be harder to understand. (3) The overlap of nouns, pronouns and content words makes the connections between sentences in abstracts of highly cited papers stronger and more accurate.

In terms of syntactic use, the syntactic structure in abstracts of highly cited papers is relatively more complex. The abstracts of highly cited paper use fewer modifiers in noun phases and fewer words before main verbs, but more noun phrases, adverbial phrases and gerunds. There are comparatively fewer similarities between adjacent sentences in both content and syntactic structure.

In terms of overall readability, highly cited paper abstracts are less readable than zero-cited paper abstracts, which are mainly reflected in their lower overall syntactic simplicity and more complex concepts. As for referential cohesion, the antecedent reference relationship of highly cited paper abstracts is rather clear and not easy to result in misreference phenomenon, although the anaphor overlap in the abstract is relatively high.

Considering the above differences, we provide the following suggestions for researchers to write better abstracts. (1) Increase the number of words in the abstract and use terms in the abstracts to present the new concepts, methods and models involved. Personal pronouns should be plural in the first person, not in the first and third person pronoun singular. Use more conjunctions between sentences except causal conjunctions. (2) Use complex syntactic structures, more noun phrases, adverbial phrases, gerunds, but fewer verb phrases, prepositional phrases and negative sentences. Reduce the number of modifiers in noun phrases and control the number of nouns and pronouns to ensure one-to-one correspondence between references and avoid the problem of referential ambiguity. (3) Reduce both the adjacent and global similarity of content and syntactic structure between sentences and use referential cohesion strategies to enhance the relevance between adjacent sentences. (4) Although papers with less readable abstracts are more likely to be cited, under the premise of a comprehensive introduction to the research content, the authors should ensure the readability of the abstract as much as possible.

## 6. Conclusion

Our study adopted Coh-Metrix analysis tool to calculate the characteristics of 11 categories of 108 indexes for highly cited abstracts and zero-cited abstracts in the WOS (2008–2017) database, analyzed and interpreted the indexes that have significant difference between highly cited and zero-cited groups. The characteristics of highly cited paper abstracts were summarized from four aspects: lexical selection, sentence characteristics, syntactic use and overall readability. Based on these characteristics, we provided some suggestions for researchers on writing high quality abstracts.

The number of citations is crucial for both researchers and research institutions. Our findings can guide researchers to write abstracts that have higher possibility to be cited and improve researchers' abstract writing skills. In addition, these findings can be used to predict the scientific impact of the paper in the future through the abstract of the paper, which can effectively shorten the review time of journal reviewers and improve review efficiency. Admittedly, there are still some shortcomings in our research. Due to the large amount of papers contained in WOS database, only 10,000 papers with high citation rates and zero citation rates are selected as samples for analysis, which may not be enough to reflect the real situation of the overall data. Therefore, in the follow-up studies, we will increase the sample size and conduct in-depth research on the characteristics of most-cited abstract and zero-cited abstract according to different time periods.

## References

Crossley, S.A., Salsbury, T., McCarthy, P.M. and McNamara, D.S. (2008), in Sloutsky, V., Love, B. and McRae, K. (Eds), "LSA as a measure of coherence in second language natural discourse", *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Washington, DC, Cognitive Science Society, pp. 1906-1911.

Didegah, F. and Thelwall, M. (2013), "Which factors help authors produce the highest impact research? Collaboration, journal and document properties", *Journal of Informetrics*, Vol. 7 No. 4, pp. 861-873.

Didegah, F., Bowman, T.D. and Holmberg, K. (2018), "On the differences between citations and altmetrics: an investigation of factors driving altmetrics versus citations for Finnish articles", *Journal of the Association for information Science and Technology*, Vol. 69 No. 6, pp. 832-843.

Dowling, M., Hammami, H. and Zreik, O. (2018), "Easy to read, easy to cite?", *Economics Letters*, Vol. 173, pp. 100-103.

Fages, D.M. (2020), "Write better, publish better", *Scientometrics*, Vol. 122 No. 3, pp. 1671-1681.

Gazni, A. (2011), "Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world", *Journal of Information Science*, Vol. 37 No. 3, pp. 273-281.

Gnewuch, M. and Wohlrabe, K. (2017), "Title characteristics and citations in economics", *Scientometrics*, Vol. 110 No. 3, pp. 1573-1578.

Gong, K., Xie, J., Cheng, Y., Larivière, V. and Sugimoto, C.R. (2019), "The citation advantage of foreign language references for Chinese social science papers", *Scientometrics*, Vol. 120 No. 3, pp. 1439-1460.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. and Cai, Z. (2004), "Coh-Metrix: analysis of text on cohesion and language", *Behavior Research Methods, Instruments, andComputers*, Vol. 36 No. 2, pp. 193-202.

Guo, F., Ma, C., Shi, Q. and Zong, Q. (2018), "Succinct effect or informative effect: the relationship between title length and the number of citations", *Scientometrics*, Vol. 116 No. 3, pp. 1531-1539.

Hafeez, D.M., Jalal, S. and Khosa, F. (2019), "Bibliometric analysis of manuscript characteristics that influence citations: a comparison of six major psychiatry journals", *Journal of Psychiatric Research*, Vol. 108, pp. 90-94.

Ibáñez, A., Bielza, C. and Larranaga, P. (2013), "Relationship among research collaboration, number of documents and number of citations: a case study in Spanish computer science production in 2000–2009", *Scientometrics*, Vol. 95 No. 2, pp. 689-716.

Jamali, H.R. and Nikzad, M. (2011), "Article title type and its relation with the number of downloads and citations", *Scientometrics*, Vol. 88 No. 2, pp. 653-661.

Lei, L. and Yan, S. (2016), "Readability and citations in information science: evidence from abstracts and articles of four journals (2003–2012)", *Scientometrics*, Vol. 108 No. 3, pp. 1155-1169.

Macarthur, C.A., Jennings, A. and Philippakos, Z.A. (2019), "Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?", *Reading and Writing*, Vol. 32 No. 6, pp. 1553-1574.

Marx, W. and Bornmann, L. (2015), "On the causes of subject-specific citation rates in web of science", *Scientometrics*, Vol. 102 No. 2, pp. 1823-1827.

McNamara, D.S., Crossley, S.A. and Roscoe, R.D. (2013), "Natural language processing in an intelligent writing strategy tutoring system", *Behavior Research Methods*, Vol. 45 No. 2, pp. 499-515.

McNamara, D.S., Graesser, A.C., McCarthy, P.M. and Cai, Z. (2014), *Automated Evaluation of Text and Discourse with Coh-Metrix*, Cambridge University Press, Cambridge.

Nair, L.B. and Gibbert, M. (2016), "What makes a 'good' title and (how) does it matter for citations? A review and general model of article title attributes in management science", *Scientometrics*, Vol. 107 No. 3, pp. 1331-1359.

Perin, D. and Lauterbach, M. (2018), "Assessing text-based writing of low-skilled college students", *International Journal of Artificial Intelligence in Education*, Vol. 28 No. 1, pp. 56-78.

Potthoff, M. and Zimmermann, F. (2017), "Is there a gender-based fragmentation of communication science? An investigation of the reasons for the apparent gender homophily in citations", *Scientometrics*, Vol. 112 No. 2, pp. 1047-1063.

Rostami, F., Mohammadpoorasl, A. and Hajizadeh, M. (2014), "The effect of characteristics of title on citation rates of articles", *Scientometrics*, Vol. 98 No. 3, pp. 2007-2010.

Shi, G., Lippert, A., Shubeck, K.T., Fang, Y., Chen, S., Pavlik, P.I., Greenberg, D. and Graesser, A.C. (2018), "Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension", *Behaviormetrika*, Vol. 45 No. 2, pp. 615-633.

Slyder, J.B., Stein, B.R., Sams, B.S., Walker, D.M., Jacob Beale, B., Feldhaus, J.J. and Copenheaver, C.A. (2011), "Citation pattern and lifespan: a comparison of discipline, institution, and individual", *Scientometrics*, Vol. 89 No. 3, pp. 955-966.

Sohrabi, B. and Iraj, H. (2017), "The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts", *Scientometrics*, Vol. 110 No. 1, pp. 243-251.

Spencer, M., Gilmour, A.F., Miller, A.C., Emerson, A.M., Saha, N.M. and Cutting, L.E. (2019), "Understanding the influence of text complexity and question type on reading outcomes", *Reading and Writing*, Vol. 32 No. 3, pp. 603-637.

Stevens, J.R. and Duque, J.F. (2019), "Order matters: alphabetizing in-text citations biases citation rates", *Psychonomic Bulletin and Review*, Vol. 26 No. 3, pp. 1020-1026.

Tahamtan, I., Afshar, A.S. and Ahamdzadeh, K. (2016), "Factors affecting number of citations: a comprehensive review of the literature", *Scientometrics*, Vol. 107 No. 3, pp. 1195-1225.

Tortorelli, L.S. (2020), "Beyond first grade: examining word, sentence, and discourse text factors associated with oral reading rate in informational text in second grade", *Reading and Writing*, Vol. 33 No. 1, pp. 143-170.

Vergoulis, T., Kanellos, I., Tzerefos, A., Chatzopoulos, S., Dalamagas, T. and Skiadopoulos, S. (2019), "A study on the readability of scientific publications", in Doucet, A., Isaac, A., Golub, K., Aalberg, T. and Jatowt, A. (Eds), *Digital Libraries for Open Knowledge. TPDL 2019. Lecture Notes in Computer Science*, Vol. 11799, pp. 136-144.

Wesel, M., Wyatt, S. and Haaf, J. (2014), "What a difference a colon makes: how superficial factors influence subsequent citation", *Scientometrics*, Vol. 98 No. 3, pp. 1601-1615.

Wiley, J., Hastings, P., Blaum, D., Jaeger, A.J., Hughes, S., Wallace, P., Griffin, T.D. and Britt, M.A. (2017), "Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science", *International Journal of Artificial Intelligence in Education*, Vol. 27 No. 4, pp. 758-790.

Wolfe, C.R., Dandignac, M. and Reyna, V.F. (2019), "A theoretically motivated method for automatically evaluating texts for gist inferences", *Behavior Research Methods*, Vol. 51 No. 6, pp. 2419-2437.

Xie, J., Gong, K., Li, J., Ke, Q., Kang, H. and Cheng, Y. (2019), "A probe into 66 factors which are possibly associated with the number of citations an article received", *Scientometrics*, Vol. 119 No. 4, pp. 1429-1454.

Zedelius, C.M., Mills, C. and Schooler, J.W. (2019), "Beyond subjective judgments: predicting evaluations of creative writing from computational linguistic features", *Behavior Research Methods*, Vol. 51 No. 2, pp. 879-894.

**Corresponding author**
Dongbo Wang can be contacted at: db.wang@njau.edu.cn