Full paper

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

OPEN ACCESS

Citation: Lin, G., Van Eck, N.J., Hou, H., & Hu, Z. (2022). The changing role of cited papers over time: An analysis of highly cited papers based on a large full text dataset. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22204). https://doi.org/10.5281/zenodo.6948268

Collection: https://zenodo.org/communities/sti2022grx/

The changing role of cited papers over time: An analysis of highly cited papers based on a large full text dataset[1]

Gege Lin[*,**], Nees Jan van Eck[**], Haiyan Hou[*], and Zhigang Hu[*]

[*] *lingegenl@mail.dlut.edu.cn; houhaiyan@dlut.edu.cn; huzhigang@dlut.edu.cn*
WISE Lab, Institute of Science of Science and S&T Management, Dalian University of Technology, No.2 Linggong Road, Dalian, 116024 (China)

** *g.lin@cwts.leidenuniv.nl; ecknjpvan@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Kolffpad 1, Leiden, 2333 BN (The Netherlands)

Introduction

Citation analysis is used in research evaluations around the world and has impact on the spending of billions of dollars on research and the work and lives of millions of researchers. It is therefore crucial to address any associated issues or limitations (Zhao & Strotmann, 2020). Central to critiques of citation analysis practices has long been that it treats all citations equally, whether they are crucial to the citing paper or perfunctory (Aljuaid, Iftikhar, Ahmad, Asif, & Tanvir Afzal, 2021; Zhu, Turney, Lemire, & Vellino, 2015). Cited references are also located in different parts of the citing paper (Thelwall, 2019a), mentioned only once or multiple times (Ding, Liu, Guo, & Cronin, 2013; Hu, Lin, Sun, & Hou, 2017), and citations include one or more references (Lin, Hou, & Hu, 2019; Lyu, Ruan, Xie, & Cheng, 2021). In addition, cited references will be cited with different sentiments and fulfill different functions. Citation counts reveal little about the reasons why a paper has been cited (Lyu et al., 2021; Vyas et al., 2020).

Furthermore, the publication year of the citing paper is an important factor to the cited paper. For example, in library and information science, two year old publications are most frequently cited, which indicates that a short citation interval is the most common practice (Yan & Ding, 2010). Other scholars have found that the larger the citation interval, the higher the probability that a reference is cited in a method section (Otto, Ghavimi, Mayr, Piryani, & Singh, 2019). This indicates that treating all citations equally has consequences in detecting and assessing research impact.

Although scholars have reported findings on citation location, citation sentiment, and citation type, previous research has rarely studied the evolution of the role of citations based on a large full-text corpus covering a long time span. If we could also consider the citation interval when

[1] This work was supported by the National Natural Science Foundation of China (Grant No. 71974030) and research results of Liaoning Province economic and social development research project in 2022, Liaoning Federation of Social Science (2022lslybkt-034).

performing full-text citation analysis, the contribution of the cited paper to the citing paper could be evaluated in a more effective and reasonable way. Therefore, this study will choose a relatively large full-text dataset to analyze the way in which papers are cited in the full text and how this changes over time. The following exploratory research questions drive this study:

- RQ1: When a paper is cited, how does the citation location change over time? Will it be cited more in the beginning, middle, or end of the full text?
- RQ2: When a paper is cited, how does the cited reference type change over time? In other words, will the cited reference be mentioned once or multiple times in the full text?
- RQ3: When a paper is cited, how does the in-text citation type change over time? Will the paper be cited together with other references or cited alone?
- RQ4: When a paper is cited, how does the citation sentiment change over time? This study considers three types of sentiment, namely positive, negative, and neutral.

In this study, we analyze the evolutionary characteristics of the role of citations based on a large and disciplinarily broad full-text collection of scientific publications. The considered aspects include citation location, cited reference type, in-text citation type, and citation sentiment. The remainder of the paper is structured as follows. The data and methods section describes the performed data acquisition, data processing, and data analysis. The key observations are reported in the results and discussion section. The final section contains conclusions, limitations and suggestions for future work.

Data and Methods

To analyze the changing role of cited papers over time, we focus on highly cited papers (HCPs) that are published long enough ago. In this way, we have the availability of a sufficient number of citations to the same cited papers and the possibility to systematically analyze any differences over time. We use the Web of Science (WoS) database, one of the largest comprehensive academic information resource covering most disciplines (Pranckutė, 2021), to select HCPs. In WoS, we search for all papers that (1) are cited at least 1000 times, (2) are published in 2000, (3) are written in English, and (4) are classified as article. In this way, we retrieved 883 HCPs. In total, the retrieved HCPs are cited by 1,542,690 unique papers.

To study in more detail how the HCPs are cited, we collect the full text of the citing papers. We sourced full-text data from the Elsevier ScienceDirect corpus that was also used in previous studies (Boyack, van Eck, Colavizza, & Waltman, 2018; Lamers et al., 2021) and that is hosted at the Centre for Science and Technology Studies (CWTS) at Leiden University. This corpus contains the full texts of nearly five million English-language research articles, short communications, and review articles published in Elsevier journals between 1998 and 2016. The corpus comprises articles from nearly 3,000 Elsevier journals. We collected the full text of 220,335 citing papers, accounting for 14.3% of total number of citing papers. The full text of the citing papers was parsed and references, citation sentences, in-text citations, and reference mentions were identified. We then determined different aspect of the citations to the HCPs, such as citation location, citation frequency, and citation sentiment. Three HCPs were ignored because we could not successfully find them in the full-text data. In the end, our case dataset therefore includes 880 HCPs.

Some information on the 220,335 papers citing our HCPs and for which we have the full text available is given in the Table 1. Because HCPs can be cited together in the same citing paper, we found 251,645 references that point to HCPs in the 220,335 citing papers. HCPs can also be

cited in the same in-text citation, so the number of in-text citations is slightly less than the number of reference mentions.

Table 1. The basic full-text and citation information of the citing papers.

	Referring to all cited works	Referring to the HCPs only
References	12,259,405	251,645
Reference mentions	18,006,353	360,419
In-text citations	11,742,513	360,366
Citation sentences	9,638,642	357,099

The terminology used in previous studies on in-text citations is not fully consistent. To avoid confusion, we define our terminology here. The first two concepts are related to how often a reference is mentioned in the full-text:

- Multiple Mentioned Reference (MMR): a MMR is a reference that is mentioned more than once in the body of the citing paper (Hu et al., 2017). It is also called "CountX mentions" in (Ding et al., 2013) and "Multi-citations" in (Zhao, Cappello, & Johnston, 2017).
- Single Mentioned Reference (SMR): a SMR is a reference that is mentioned only once in the body of the citing paper. It is also called "CountOne mentions" in (Ding et al., 2013) and "Uni -citations" in (Zhao et al., 2017).

The other two concepts are related to the number of references in an in-text citation:

- Multi-Reference Citation (MRC): a MRC is an in-text citation that includes more than one reference (American Psychological Association, 2019; Lin et al., 2019). It is also called "citation string" in (Lyu et al., 2021; Petrić, 2007).
- Single-Reference Citation (SRC): a SRC is an in-text citation that only includes one reference.

Results and Discussion

Citation location in different citing years

We first analyzed the location of citations to the HCPs in the full text of the citing papers. Text progression was used to represent the citation location. The average of the citation location per citing year was calculated using the arithmetic mean. We also divided the citation location based on the text progression into three equal parts: begin part (0%-33%), middle part (33%-66%), end part (66%-100%).

Figure 1 shows that the average in-text location of citations to HCPs decreases over the citing years. This means that, on average, HCPs are cited earlier in the text as the difference in publication year between the citing papers and the HCPs increases. Figure 2 shows that there is little change in the percentage of HCP in-text citations located in middle part over the citing years (it fluctuates around 25%), while the percentage located in the begin part is increasing and in the end part is decreasing. It suggests that the citations to HCPs in the first citing years are about equally common in the opening sections (36%), such as introduction or background, and the final sections (37%), such as discussion or conclusion. In later citation years, this changes and citations increasingly occur in the opening sections instead of the final sections (56% vs 19%).

Figure 1. Average location of citations to HCPs in the full text of papers in different citing years.

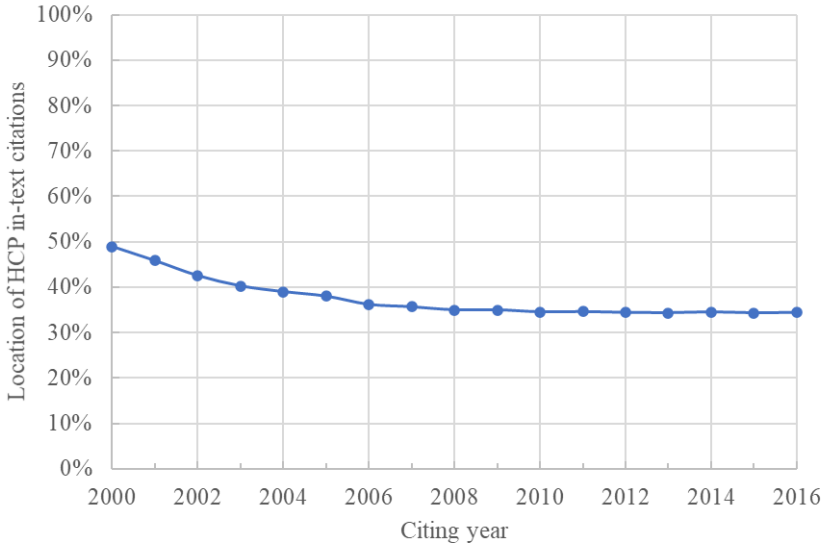
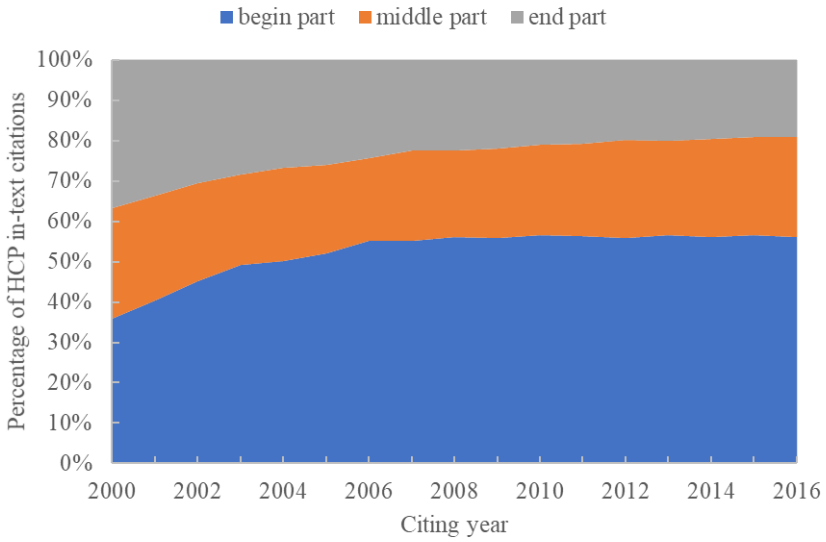


Figure 2. Percentage of citations to HCPs located in the begin part, middle part, and end part of the full text of papers in different citing years.



Cited reference type in different citing years

References can be mentioned more than once in the full text. More and more scholars argue that these MMRs are more essential to the citing work than SMRs (Ding et al., 2013; Hu et al., 2017; Zhao & Strotmann, 2020). In this study, we analyzed the number of times references to HCPs are mentioned in the full text and how the percentage of SMRs and MMRs changes over the citing years.

Figure 3. Percentage of reference to HCPs that are mentioned only once (SMR) or multiple times (MMR) in the full text of papers in different citing years.

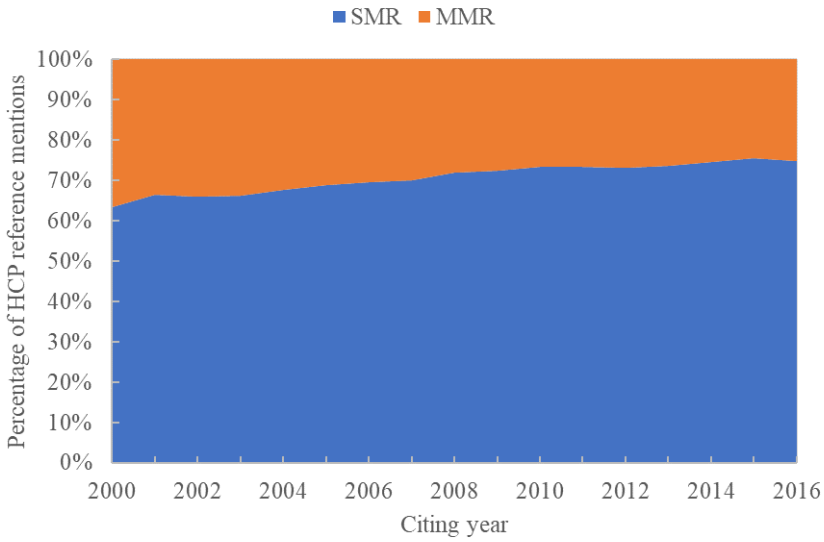


Figure 4. Average number of mentions of references to HCPs in the full text of papers in different citing years.

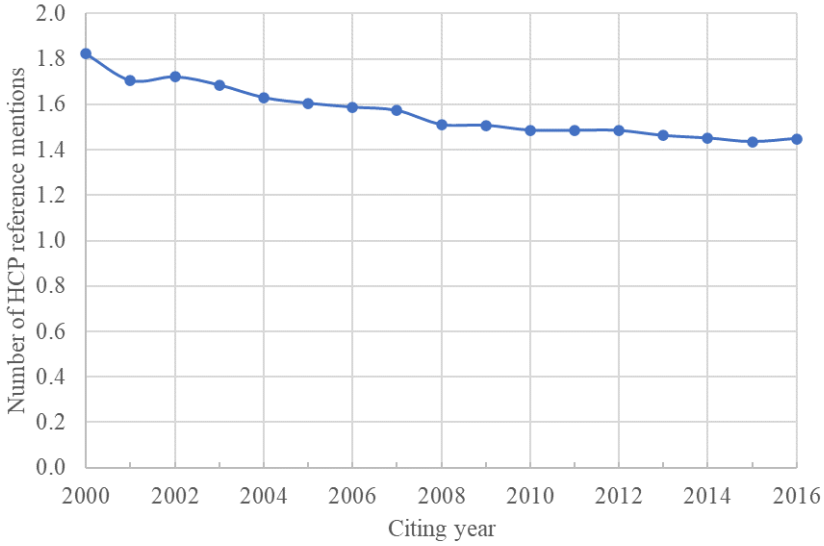


Figure 3 shows that in the first few citing years many cited articles are cited multiple times in the citing articles, while that percentage decreases as the citation interval increases. The percentage MMRs reduced from 37% to 25%. The exact opposite trend is visible for SMRs that are mentioned only once in the citing paper. It is also useful to know the average number of times that each reference is mentioned. Figure 4 shows that the average number of times the HCPs are mentioned in the full text of the citing papers is decreasing from 1.8 in 2000 to 1.4 in 2016. This result is in line with the earlier finding that references mentioned only once are typically more highly cited than those that are mentioned multiple times (Boyack et al., 2018).

In-text citation type in different citing years

In this study, we also analyzed the in-text citations in which the HCPs are referenced. We analyzed the number of references that these in-text citations contain and how the percentage

of SRCs and MRCs changes over the citing years. Figure 5 shows that the percentage of MRC increases over the citing years, from 32.7% in 2000 to 45.7% in 2016. It means that HCPs are more likely to be cited along with other references in the same in-text citation in later citing years. This could indicate that as HCPs get older, they tend to serve more and more as general references and become less essential to the papers in which they are cited. Previous research has therefore advised research evaluators to avoid the use of citation impact formulas that overvalue HCPs by treating their citations as equally as the citations of less-cited papers (Thelwall, 2019b).

The average number of references in HCP in-text citations is shown in Figure 6. The number increases quickly in the first few citing years and reaches a peak in 2007, followed by a slight decline until 2016. In other words, HCPs are cited together with more other references as they age, but there is a limit to this growth.

Figure 5. Percentage of HCP in-text citations that contain only one reference (SRC) or muliple references (MRC) in different citing years.

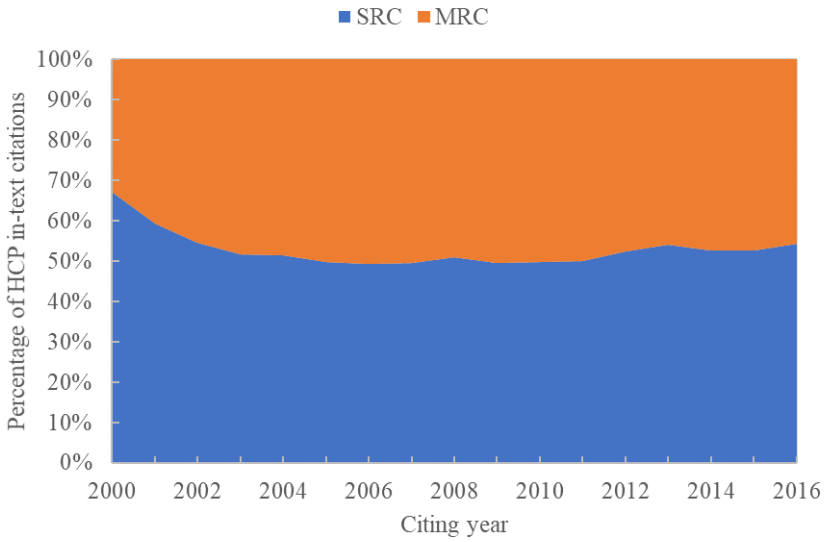
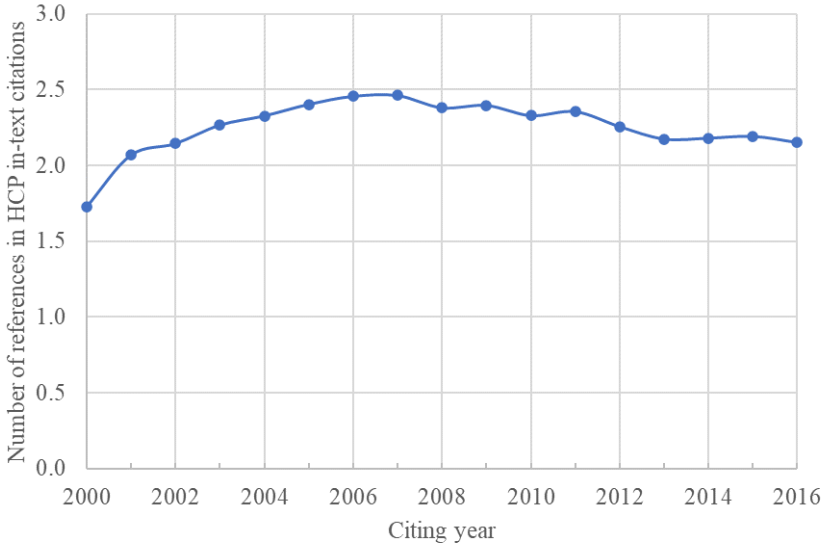


Figure 6. Average number of references in HCP in-text citations in different citing years.



Citation sentiment in different citing years

Citation sentiment is one of the essential aspects in citation content analysis, which is useful for analyzing citation influence and the recommendation of scientific publications (Chen and Nguyen 2020). Citation sentiment indicates the attitude and opinion polarity of the author(s) of the citing paper in relation to the cited paper (Vyas et al., 2020). VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that determines for each given text the proportion of the text that is associated with a positive, negative, or neutral sentiment (Hutto & Gilbert, 2014). VADER is available in the NLTK package for Python and we used this to analyze how the sentiment of sentences that contain citations to our HCPs change over time.

Figure 7 shows that the largest proportion of citation sentence text is associated with a neutral sentiment, higher than 90% in all citing years. Although the proportion associated with a positive sentiment has increased slightly over the citing years, it still takes up a small percentage (4.7%-6.1%). The proportion associated to a negative sentiment shows a weak decline from 3.8% to 3.5%.

VADER also provides a compound score that is a 'normalized, weighted composite score' providing a single unidimensional measure of sentiment for a given text. The compound score ranges from -1 (most extreme negative) to +1 (most extreme positive). Figure 8 shows the average sentiment compound score of sentences that contain citations to our HCPs. This result also shows a slight rise in sentiment as HCPs age (from 0.04 to 0.11).

Figure 7. Proportion of the text of sentences containing citations to HCPs associated with a positive, negative, or neutral sentiment over the citing years.

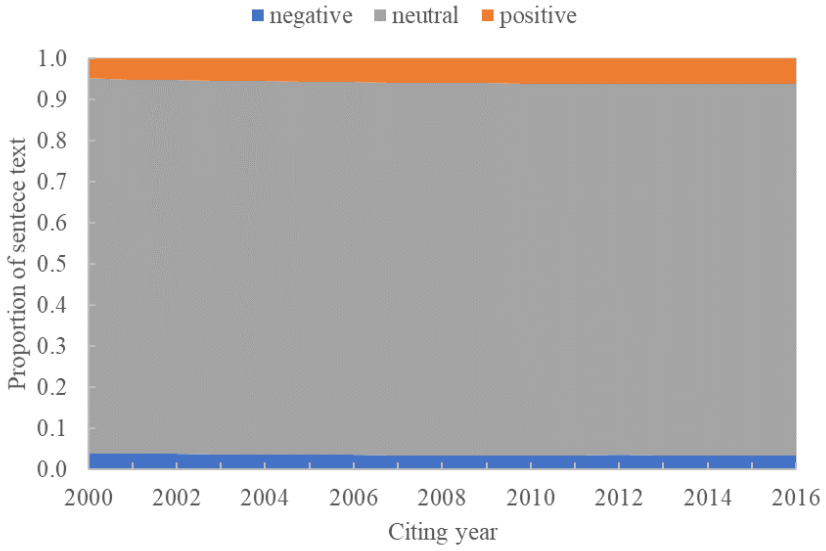
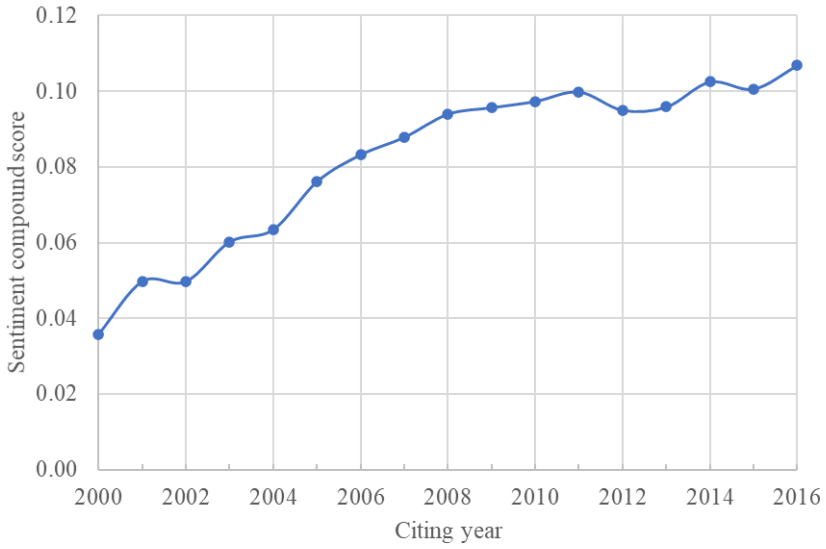


Figure 8. Average sentiment compound score of the text of sentences containing citations to HCPs over the citing years.



Conclusions

In this paper, we have analyzed the changing role of cited papers over time. This was done by taking 880 HCPs as a case dataset and by collecting the full text of more than 220 thousand papers that are citing those HCPs. Based on the collected full text data, we have analyzed how different aspects of the citations to the HCPs change over time. The aspects included in our analysis are the location of the citation in the full text, cited reference type, in-text citation type, and citation sentiment. The conclusion of our analysis can be summarized as follows.

On average, HCPs are cited earlier in the text as they get older. There is little change in the percentage of citations to HCPs that are located in middle part of the full text of the citing papers over the citing years, while the percentage located in the begin part is increasing and in the end part is decreasing. Second, many HCPs are mentioned multiple times in the full text of the citing papers in the first few citing years, while that percentage decreases as the HCPs get older. In addition, the average numbers of time the HCPs are mentioned in the full text of the citing papers is decreasing from 1.8 in 2000 to 1.4 in 2016. Third, HCPs are more likely to be cited along with other references in the same in-text citation in later citing years. This could indicate that as HCPs get older, they tend to serve more and more as general references and become less essential to the papers in which they are cited. Also, HCPs are cited together with more other references as they age, but there is a limit to this growth. Our last finding is that there is only a very weak increase in citation sentiment over the citing years. The largest proportion of the text of sentences in which HCPs are cited is associated with the neutral sentiment, followed by the positive and the negative sentiment.

There are several limitations to this study that should be noted. First, although the number of full texts of the citing papers that we have collected is large, it still accounts for only a relatively modest share of the total number of citing papers (14.3%). Additional full text data from other sources may yield different results. Second, fields of research were not considered, which may hide some evolutionary features that may be present at a more granular field or discipline level. Third, we limited ourselves in the number of citation characteristics studied. For example, we did not attempt to analyze the change in the relatedness between citing papers and cited papers. Despite these limitations, our results can be regarded as weak evidence that the reason why

papers are cited may change over time. In the future, we look forward to additional studies examining the evolutionary characteristics of citations, at more granular levels, using full-text data from multiple sources, considering different research areas, and using semantic analyses. Such studies have the potential to influence our understanding of citation theory and behavior, and to have practical impact on applications such as information search and retrieval and the accurate modeling of the structure and dynamics of science.

References

Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Tanvir Afzal, M. (2021). Important citation identification using sentiment analysis of in-text citations. *Telematics and Informatics*, *56*, 101492.

American Psychological Association. (2019). *Publication Manual of the American Psychological Association, (2020)*. American Psychological Association.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*(3), 583–592. Elsevier.

Hu, Z., Lin, G., Sun, T., & Hou, H. (2017). Understanding multiply mentioned references. *Journal of Informetrics*, *11*(4), 948–958. Elsevier.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225.

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Investigating disagreement in the scientific literature. (P. Rodgers, Ed.)*ELife*, *10*, e72737. eLife Sciences Publications, Ltd.

Lin, G., Hou, H., & Hu, Z. (2019). Understanding Multiple References Citation. (pp. 2347–2357). Presented at the ISSI.

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: A meta-synthesis. *Scientometrics*, *126*(4), 3243–3264. Springer.

Otto, W., Ghavimi, B., Mayr, P., Piryani, R., & Singh, V. K. (2019). Highly cited references in PLOS ONE and their in-text usage over time. *ArXiv:1903.11693 [cs]*. Retrieved April 29, 2022, from http://arxiv.org/abs/1903.11693

Petrić, B. (2007). Rhetorical functions of citations in high-and low-rated master's theses. *Journal of English for Academic Purposes*, *6*(3), 238–253. Elsevier.

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, *9*(1), 12. Multidisciplinary Digital Publishing Institute.

Thelwall, M. (2019a). Should citations be counted separately from each originating section? *Journal of Informetrics*, *13*(2), 658–678. Elsevier.

Thelwall, M. (2019b). Are classic references cited first? An analysis of citation order within article sections. *Scientometrics*, *120*(2), 723–731. Springer.

Vyas, V., Ravi, K., Ravi, V., Uma, V., Setlur, S., & Govindaraju, V. (2020). Article citation study: Context enhanced citation sentiment detection. *ArXiv:2005.04534 [cs]*. Retrieved April 28, 2022, from http://arxiv.org/abs/2005.04534

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, *61*(8), 1635–1643.

Zhao, D., Cappello, A., & Johnston, L. (2017). Functions of uni-and multi-citations: Implications for weighted citation analysis. *Journal of Data and Information Science*, *2*(1), 51–69.

Zhao, D., & Strotmann, A. (2020). Deep and narrow impact: Introducing location filtered citation counting. *Scientometrics*, *122*(1), 503–517. Springer.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427.