Use of Bibliometrics Data to Understand the Citation Advantages of Different Open Access Categories in Covid-19 Related Studies

Xiaoju Chen  
Carnegie Mellon University, USA  
xiaojuc@andrew.cmu.edu

Neelam Bharti  
Carnegie Mellon University, USA  
nbharti@andrew.cmu.edu

Matthew R. Marsteller  
Carnegie Mellon University, USA  
matthewm@andrew.cmu.edu

ABSTRACT
The number of Open Access (OA) research articles is trending upward. This research aims to understand the correlations between different OA types and the impact of OA research articles evaluated based on the citation numbers. To avoid bias caused by the publication year, we chose to use COVID-19 studies in different fields to take advantage of this topic’s quick turnaround. We analyzed the bibliometrics data and citation numbers (excluding self-citations) of around 42,000 English language articles published in 2020 related to COVID-19. We evaluated different types of OA categories such as Gold, Bronze, and Hybrid articles separately. Results show that amongst all OA categories, Hybrid/Green and Bronze/Green OA articles had significant citation advantages. Green OA articles returned more citations than articles with the other OA status. Gold OA articles have no citation advantages compared to non-OA articles. Gold/Green OA articles had the highest self-citation rates, followed by Non-OA articles. The results of the study can be used in understanding different OA categories and the reasons for OA choices. Certain strategies can be made accordingly to improve the awareness of OA in different fields and help OA publishers to improve the OA services.

KEYWORDS
Open Access; Bibliometrics analysis; Research impact; Self-citation; Covid-19.

INTRODUCTION
Open Access (OA) grants free access to academic information such as research publications and data. OA models for scholarly communication make research available to readers without any financial, legal, or technical barriers, unlike the subscription models, where the reader pays to access. According to several recent trends, OA papers can be classified as Gold, Green, Hybrid, and Bronze (Martín-Martín et al., 2018). This classification is also adapted by databases like Scopus and Web of Science. Gold OA refers to Open Access journals that provide immediate access to the articles published; Green OA are articles that are deposited in a repository and made publicly available by the authors; Hybrid OA articles are articles in subscription-based journals with an Article Processing Charge (APC); Bronze OA articles are articles that are made free-to-read on the publisher website, without an explicit open license. Most studies that investigated the citation advantages of OA articles focused on Green and Gold OA articles (Zhang & Watson, 2017; Dorta-González & Santana-Jiménez, 2017; Young & Brandes, 2020). Most recently, researchers started to pay attention to different types of OA articles and compare their citation advantages. For example, Zhang & Watson (2017) found that for the same citation period considered, both Non-OA article and Green OA articles have more citation numbers than Gold OA articles in the physical science field; hybrid OA articles had the highest citation rate compared to Gold OA, Green OA, and Non-OA article. Wang et al. (2015) discovered that within the same journal (Nature Communications), hybrid articles have a significant citation advantage over Non-OA articles for all time periods studied. Piwowar et al. (2018) found that Green OA articles had slightly more citations, followed closely by Hybrid OA. In all of these studies, major OA categories were introduced and compared; however, the more detailed OA data were not separately examined. For instance, a Gold OA article can be archived by the authors and made Green OA; these Gold/Green OA articles should be separately investigated to understand OA’s citation advantage. In addition, the existing research often ignored the impacts of self-citation. A few studies had separately investigated articles with or without self-citation impacts (Norris et al., 2008; Gargouri et al., 2010; Clements, 2017; Basson, 2019). However, in these studies, self-citation was excluded to avoid the impact of self-citation to understand OA’s real advantages; the correlation between self-citations and different types of OA was not studied.

In this study, we aimed to understand the citation advantages of different types of OA articles as well as the correlation between self-citation and different OA categories. We chose the articles related to the novel coronavirus, COVID-19 studies. The novel coronavirus caused an ongoing global pandemic of coronavirus illness. The virus was first identified in Wuhan (China) in December 2019. Because the COVID-19 was discovered in December 2019, the research related to this topic was published in January 2020 onward. We chose this topic to take advantage of its

84th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 3, 2021 | Salt Lake City, UT. Author(s) retain copyright, but ASIS&T receives an exclusive publication license.
short turnaround time period. It eliminates the citation advantages caused by various citation periods and allows the understanding of the citation advantage caused merely by different OA types.

METHOD
For this study, we aimed to find a good number of articles that represent all types of OA status to allow comparisons. We did not attempt to include as many publications as possible; instead, we believe that a good sample of articles would be sufficient. Therefore, the bibliometrics and citation data were acquired from the Scopus database, which includes most of the important publications. We defined the search terms were defined as “COVID-19 OR coronavirus OR sars-cov-2” to avoid exclusion of earlier publications before the name COVID-19 was defined by the WHO (World Health Organization). The search was also limited to English language publications only, “final” as the publication stage, and articles only (exclude review, letter, etc.). To avoid the gap due to download time and analysis, only the publications from 2020 were considered. With these search terms and criteria, 42,696 articles were found in 28 different subject areas. We then downloaded the bibliometrics data and citation data for all these articles from Scopus. Because Scopus only allowed to download the information for a maximum of 2,000 articles, these 42,696 articles were downloaded in 23 separate files. We also downloaded the full citation records and citation records excluding self-citation from Scopus in 23 separate files.

All the data were cleaned in Python after download. After removing duplication, 41,243 unique records based on title, author, and journal name were identified. Then, these records were merged with the citation files based on the article titles. Here, to avoid errors due to formatting, natural language processing packages were used to remove punctuations and tokenize the text. After merging, 41,897 records were matched with citation and self-citation excluded data. Resulted records were grouped by their open access categories shown in the bibliometrics data. Eight OA categories were identified: Gold; Gold and Green; Bronze; Bronze and Green; Hybrid; Hybrid and Green; Non-OA; Green. The citation data for each category were analyzed separately. Articles with the same number of citations were grouped together. The citation numbers with or without self-citations were analyzed separately to understand the impact of self-citations.

RESULTS AND DISCUSSION
Citation advantages of OA articles
Figure 1 shows the overall citation situation from eight OA or Non-OA categories. The values on the x-axis indicate the citation times, and the y-axis shows the ratio of OA articles to the total number of articles with the number of citations on the x-axis for each category. Note that the y-axis is in a logarithmic scale to make the data more disguisable.

The results showed that all green OA articles had significantly more citations compared to articles with the other OA status. For example, Hybrid/Green OA articles generally had the highest citation rates, especially for articles with more than 100 citation times; however, Hybrid OA articles (without Green OA status) had much lower citation rates. Overall, Hybrid, Bronze, and Non-OA with additional Green OA status had higher citation rates. Regardless of having Green OA status, Gold OA articles had lower citation rates; this might be related to the OA journals' quality. In the subject fields studied, Gold OA journals tended to have lower impact factors, this might contribute to the lower citation numbers. Gold OA articles without Green OA status had similar citation rates with Non-OA articles. Results in Figure 2 also supported these conclusions. For all the articles that had been cited at least once, Gold OA and Non-OA articles had the lowest values. Gold/Green OA articles had a slight improvement than Gold OA only articles; however, the self-citation rate was comparatively high for this category.

For articles with a few citations, there were no significant differences between Hybrid/Green, Bronze/Green, and Non-OA/Green articles; however, Hybrid/Green and Bronze/Green OA articles were leading for articles with more than ten citations. This citation advantage was more significant for articles with more than 100 citations, for OA categories with Green OA or without Green OA status. A possible explanation for this is that authors might have made the final manuscript freely available through a repository with a link to the journal article. For the topic of COVID-19, these important works could quickly attract attention and result in more citation numbers. Similarly, many journal publishers made the COVID-19 related article freely available for everyone as their social responsibility through their website disregarding the availability of the final manuscript leading to more citations. It would be interesting to study the social impact of this research with the citation impact.

Previous research has discussed the reasons for a higher citation rate for Green OA articles in general. Authors might self-archive their better-quality work, and articles made available as preprints have an extended period to be cited (Kurtz et al., 2005; Davis & Fromerth, 2007). These arguments were supported by the results in our study: Green OA articles had notably more citations than their counterparts in the same OA category. In this study, because of the topic we chose and the restricted citation period only to 2020, the differences could be more significant. Articles that were not Green OA might not have been available on time to be cited.
Figure 1. Overall citation numbers (without self-citation), in logarithmic scale

Figure 2. Articles that have been cited at least once

Self-citation
Table 1 shows the number of articles with self-citations. Each row shows the citation times; each row’s entry represents the absolute difference between articles with self-citations and without self-citations. For example, the first entry for Gold/Green OA column shows that of the 9873 Gold/Green OA articles, 10.1% of these articles had zero citations if self-citation was not included. Similarly, the second row for the same column shows that 5.3% of the Gold/Green OA articles with 1 to 4 citations would not have as many citations if self-citations were excluded. The results from Table 1 show that Gold/Green OA articles had the highest rate of self-citations (the difference was approximately 10%). In contrast, other OA or Non-OA categories had similar self-citation rates (from 2% to 4%). Surprisingly, Gold OA without Green status had the lowest self-citation rate. The analysis indicates that OA categories were not a factor affecting self-citation.

Non-OA articles had slightly more self-citation than other categories except for Gold/Green. It was possibly because Non-OA articles had a similar number of self-citations compared to all other categories. Still, because the overall citation numbers were smaller, the self-citation rates were more significant than usual. It is unclear why Gold/Green OA had more self-citations, and Gold OA had the lowest self-citation rate.
Table 1. Difference of citation rate without self-citations, represents the absolute difference between articles with self-citations and without self-citations

| Citation times | Gold/ Green | Hybrid/ Green | Bronze/ Green | Non OA | Non OA/ Green |
|----------------|-------------|---------------|---------------|--------|---------------|
| 0              | 2.0%        | 10.1%         | 2.7%          | 2.2%   | 2.0%          | 1.9% |
| 1-4            | 1.4%        | 5.3%          | 1.7%          | 0.6%   | 1.0%          | 0.3% |
| 5-10           | 0.3%        | 2.4%          | 0.4%          | 1.0%   | 0.6%          | 0.8% |
| 11-29          | 0.2%        | 1.5%          | 0.6%          | 0.2%   | 0.2%          | 0.5% |
| 30-100         | 0.1%        | 1.0%          | 0.0%          | 0.4%   | 0.2%          | 0.3% |
| >100           | 0.0%        | 0.3%          | 0.0%          | 0.1%   | 0.0%          | 0.1% |
| Total number   | 5058        | 9873          | 900           | 1788   | 6582          | 8142 |

CONCLUSION

In this study, we used bibliometric analysis to understand the citation advantages of different OA types and their self-citation rates. Results showed that all green OA articles had significantly more citations compared to articles with the other OA status; regardless of having Green OA status, Gold OA articles had lower citation rates. For articles with a few citations, there were no significant differences between Hybrid/Green, Bronze/Green, and Non-OA/Green articles; however, Hybrid/Green and Bronze/Green OA articles were leading for articles with more than ten citations. In addition, we found that Gold/Green OA articles had the highest rate of self-citations. In contrast, other OA or Non-OA categories had similar lower self-citation rates. Surprisingly, Gold OA without Green status had the lowest self-citation rate. The analysis indicates that OA categories were not a factor affecting self-citation.

The preliminary results from this study showed citation advantages of Green OA articles; however, some questions remain unclear. Future work will involve understanding the reason why Gold/Green have more self-citation and Gold had the lowest self-citation rate. Also, we will consider other factors such as subject, journal quality and country of origin, aiming to find the true OA citation advantages and incentives of OA publishing in different fields. In future studies, we will also use statistic tests to verify the significance of the differences. The results of the study can be used to raise awareness of OA in different subject fields and help OA publishers to improve the OA services.

REFERENCES

- Basson, I. (2019). An investigation of open access citation advantage through multiple measures and across subject areas for articles published from 2005 to 2014 [Thesis, Stellenbosch : Stellenbosch University]. https://scholar.sun.ac.za:443/handle/10019.1/105966
- Clements, J. C. (2017). Open access articles receive more citations in hybrid marine ecology journals. FACETS. https://doi.org/10.1139/facets-2016-0032
- Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 203–215. https://doi.org/10.1007/s11192-007-1661-8
- Dorta-González, P., & Santana-Jiménez, Y. (2017). Prevalence and citation advantage of gold open access in the subject areas of the Scopus database. ArXiv:1708.06242 [Cs]. http://arxiv.org/abs/1708.06242
- Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE, 5(10), e13636. https://doi.org/10.1371/journal.pone.0013636
- Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. Information Processing & Management, 41(6), 1395–1402. https://doi.org/10.1016/j.ipm.2005.03.010
- Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. Journal of Informetrics, 12(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012
- Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. Journal of the American Society for Information Science and Technology, 59(12), 1963–1972. https://doi.org/10.1002/asi.20898
- Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlanger, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ. http://dx.doi.org/10.7717/peerj.4375
- Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. Scientometrics, 103(2), 555–564. https://doi.org/10.1007/s11192-015-1547-0
Young, J. S., & Brandes, P. M. (2020). Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *The Journal of Academic Librarianship, 46*(2), 102105. https://doi.org/10.1016/j.acalib.2019.102105

Zhang, L., & Watson, E. M. (2017). Measuring the Impact of Gold and Green Open Access. *The Journal of Academic Librarianship, 43*(4), 337–345. https://doi.org/10.1016/j.acalib.2017.06.004