# Are classic references cited first? An analysis of citation order within article sections[1]

Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK.

Early citations within an article section may have an agenda-setting role but contribute little to the new research. To investigate whether this practice may be common, this article assesses whether the average impact of cited references is influenced by the order in which they are cited within article sections. This is tested on 1,683,299,868 citations to 41,068,375 unique journal articles from 1,470,209 research articles in the PubMed Open Access collection, split into 22 fields. The results show that the first cited article in the Introduction and Background have much higher average citation impacts than later articles, and the same is true to a lesser extent for the Discussion and Conclusion in most fields, but not the Methods and Results. The findings do not prove that early citations are less central to the citing article but nevertheless add to previous evidence suggesting that this practice may be widespread. It may therefore be useful to distinguish between initial introductory citations when evaluating citation impact, or to use impact indicators that implicitly or explicitly give less weight to the citation counts of highly cited articles.
**Keywords**: citation order; in-text citations; highly cited articles

## Introduction

Academics writing up their research need to justify the importance and correctness of their work and may cite references to support these goals. Some sections may start with a general reference to introduce the topic to the reader or to summarise general knowledge in the field (Swales, 1990), whereas later references may give more direct support for the methods, theory or claims made in the new paper. Previous analyses of the purposes of individual citations has argued that some are perfunctory in the sense of contributing little to the new research (Moravcsik & Murugesan, 1975; Voos & Dagaev, 1976; Lin, 2018) and may not have been read by the authors (Klitzing, Hoekstra, & Strijbos, 2018). These general references may become highly cited for their simple peripheral role, or highly cited papers may be chosen for this role, as providers of implicitly validated evidence. Thus, it seems possible that highly cited papers attract new imitative citations because they are highly cited, especially if they are supporting a general claim or serve as concept markers for a topic (Case & Higgins, 2000; Shadish, Tolliver, Gray, & Sen Gupta, 1995). The existence of perfunctory citations is also supported by the hypothesis that citation practices are imitative. A degree of citation mimicking is a reasonable explanation for the highly skewed distribution of citation counts. This has been called the Matthew Effect (Merton, 1968) or the rich-get-richer (de Solla Price, 1976) law. There is no systematic information about the relationship between citation order and citation contribution, however. If early citations are routinely of a general kind then it might be possible to modify citation indicators to take this into account.

The Introduction is the most likely location for perfunctory citations (Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, 1998; Tang & Safer, 2008). Since an Introduction may be

[1] Thelwall, M. (in press). Are classic references cited first? An analysis of citation order within article sections. *Scientometrics*.

structured from general to specific, it seems likely that more perfunctory and more highly cited articles would be nearer the start. This is not supported by prior studies of the relationship between citation counts and position within the citing text. One study of Elsevier full text articles from 2015 in five categories found that for three fields (Biomedical and Health Sciences; Life and Earth Sciences; Physical Sciences and Engineering), the most cited articles occurred on average, a third of the way through the citing paper, whereas in two fields (Maths and Computer Science; Social Sciences and Humanities), the peak occurred two thirds through the citing paper (Figure 8 of: Boyack, van Eck, Colavizza, & Waltman, 2018). The same investigation found PubMed Central Open Access papers to have a similar shape to the first three fields. Studies of the positions of citations within the body of an article have found that there are relatively many at the start of an article, with a secondary peak near the end, although this varies by discipline (Boyack, van Eck, Colavizza, & Waltman, 2018; see also: Bertin, Atanassova, Gingras, & Larivière, 2016; Ding, Liu, Guo, & Cronin, 2013; Hu, Chen, & Liu, 2013). Older references have also been found to be nearer the start of articles or in the methods section in seven PLOS journals (Bertin, Atanassova, Gingras, & Larivière, 2016) or just before the middle of a much larger set (Boyack, van Eck, Colavizza, & Waltman, 2018).

This paper assesses the extent to which the average citation impact of an article varies with the order that it is cited within an article section, driven by the questions below. This differs from the most similar previous papers (Boyack, van Eck, Colavizza, & Waltman, 2018) by counting separately by article section, by using more recent articles, employing finer-grained categories, and by focusing on the order in which articles are cited rather than how far through the text each citation occurs. Focusing on sections may give finer-grained information because articles can be arranged in different orders, affecting where different types of citations are placed.

1. Does the average citation rate of articles cited in paper sections vary by the order in which they are cited?
2. Does the answer to the above vary by the discipline of the citing article?
3. Are there differences between sections in the relationship between reference order and citation impact?

# Methods

The PMC Open Access collection of full text documents in XML format (NCBI, 2015) was used for the raw data. This is a multidisciplinary collection with a strong biomedical focus but is apparently the largest free source of open access documents. This collection was downloaded in November 2017. Each article was parsed to extract the section names and references. Section names were extracted from section tags or from title tags immediately following section tags. Only main section names were used and subsections were given the same name as the hosting section. Section names were standardised to the main six by removing any initial numbers and identifying common variations (e.g., Literature Review for Background). References were identified by order within the text of each section, using the XREF tag. Instances where references were cited implicitly through ellipsis (e.g., “[1] – [5]”) were detected with heuristics and completed. Citations were only tracked for journal articles, to ensure accuracy, since other document types are harder to compare and merge.

The articles were split into the 22 main Science-Metrix fields (Archambault, Beauchesne, & Caruso, 2011) to categorise them approximately by broad field. This seems to more effective for field delineation than the Web of Science and Scopus classifications

(Klavans & Boyack, 2017). The Science-Metrix list was expanded by adding the 100 largest missing journals from the PMC collection. Fields with fewer than 30 articles in any authorship position (1-10 or 11+) and section were not reported due to the relatively unreliability of the averages for small sample sizes. The final dataset included 1,683,299,868 citations (counting multiple citations from different sections multiple times) to 44,479,287 journal articles (41,068,375 unique journal articles, combining articles cited in different sections) from 1,470,209 research articles in the PubMed Open Access collection.

The average citation count for articles in each position was calculated using geometric means. These are more appropriate than arithmetic means since citation data is highly skewed (Thelwall & Fairclough, 2015).

## Results

In all sections except the Methods, the first reference is the most cited, on average. This tendency is weak for the Results section and strongest for the Introduction and Background (Figure 1). References cited in the Methods section are most cited overall, irrespective of position in the reference list.

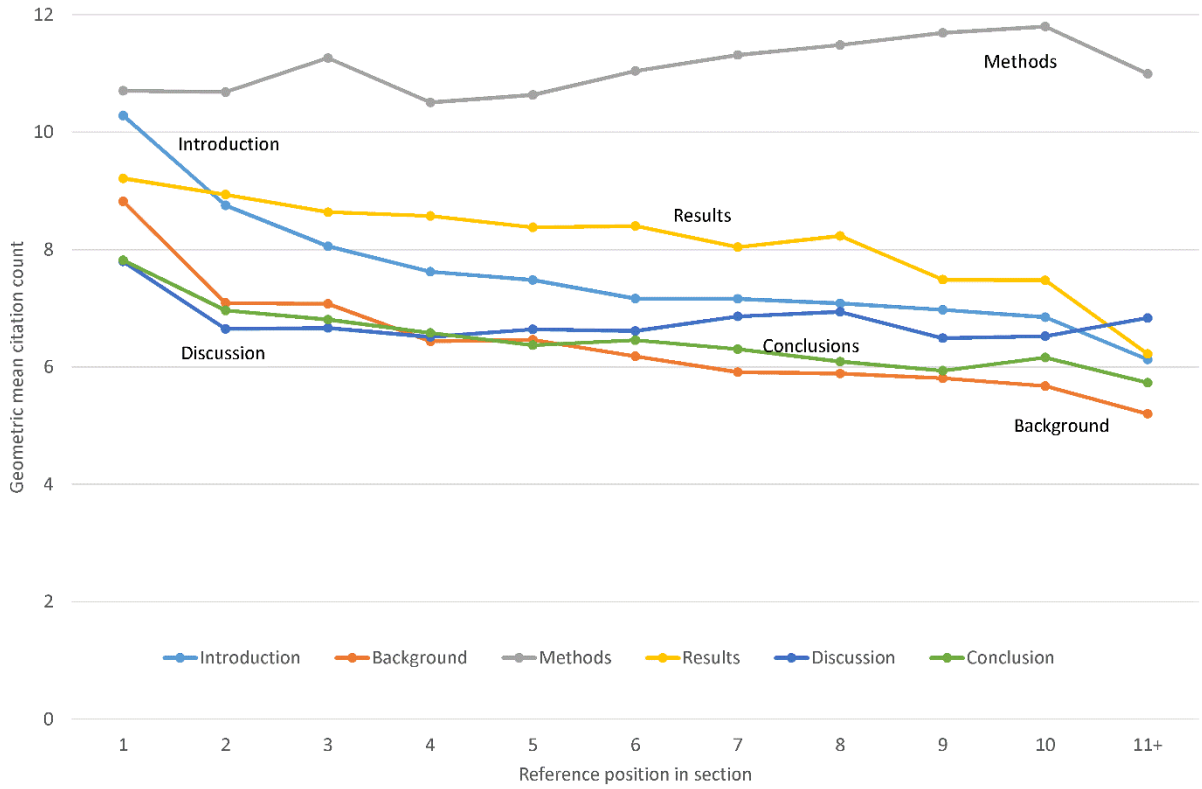


Figure 1. Median (across fields) geometric mean (within field) number of PMC Open Access citations for articles by position in the reference list within a section. Qualification: Field must have 30+ articles in each position.

Later references in the Introduction tend to be less cited in most fields (Figure 2), with the small Philosophy & Theology field being a partial exception (the second reference is slightly more cited than the first). Thus, starting the Introduction with a relatively highly cited article seems to be almost universal in academia, although the extent of this trend varies by field. A similar pattern is evident for the Background section (Figure 3).

The relative heights of the lines should not be compared between fields in Figures 2 to 7 because the average age of articles in the PubMed Open Access collection differs between fields.

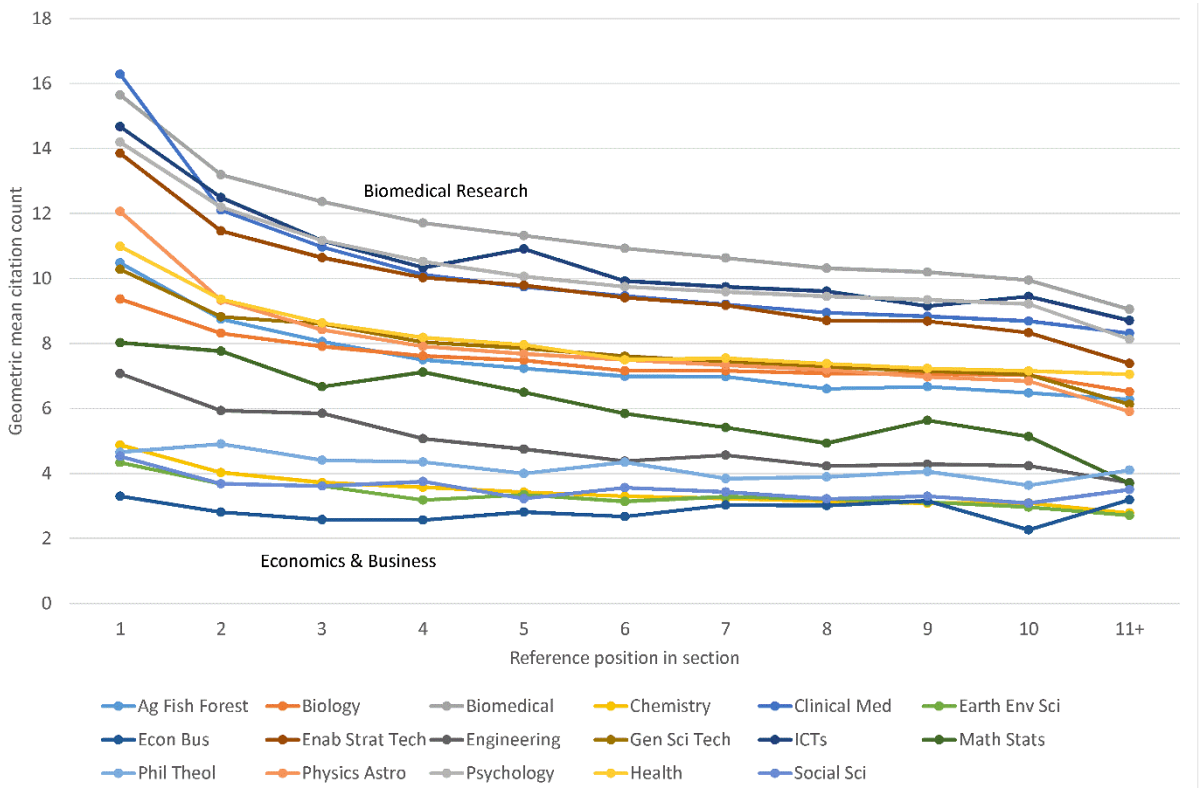


Figure 2. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Introduction**. Qualification: Field must have 30+ articles in each position.

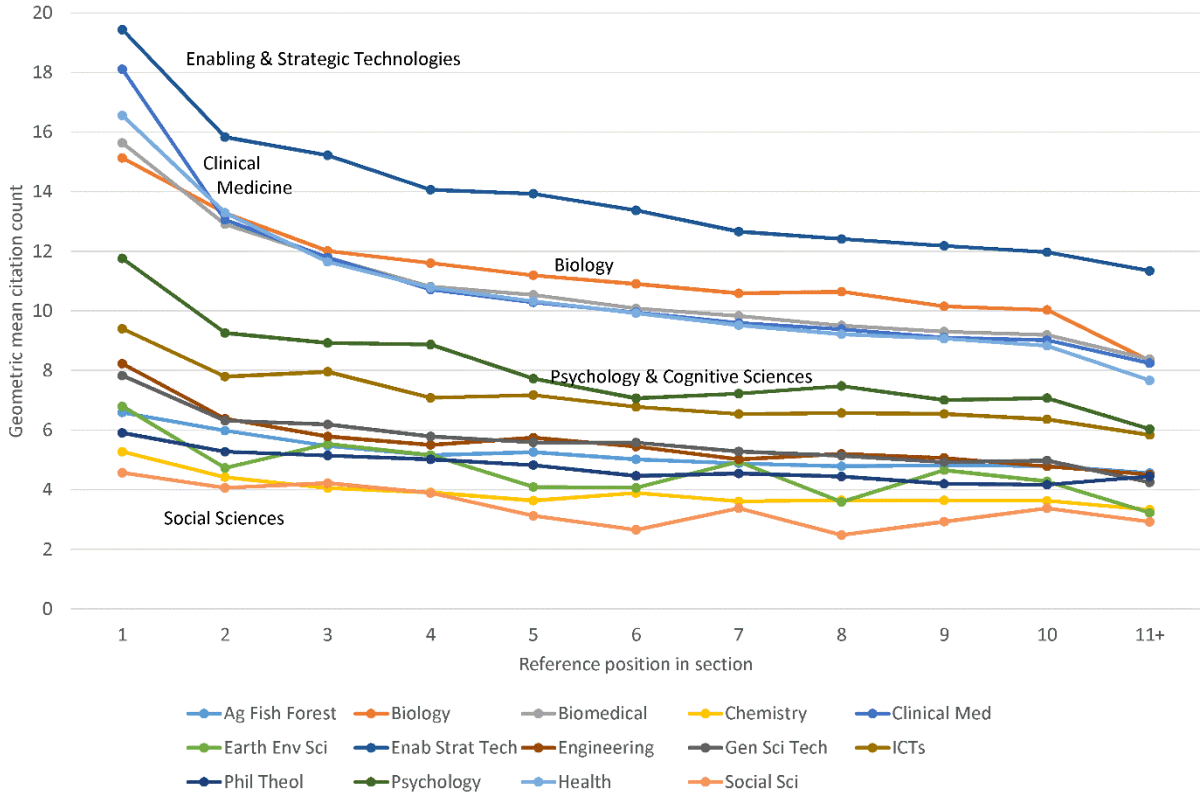


Figure 3. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Background**. Qualification: Field must have 30+ articles in each position.

Whilst some fields have a tendency for later Methods citations to be more cited, for most fields, position has little effect on average citation rates (Figure 4). Later citations in Biology, for example, might be for widely used standard algorithms or tools, such as those in DNA sequencing.

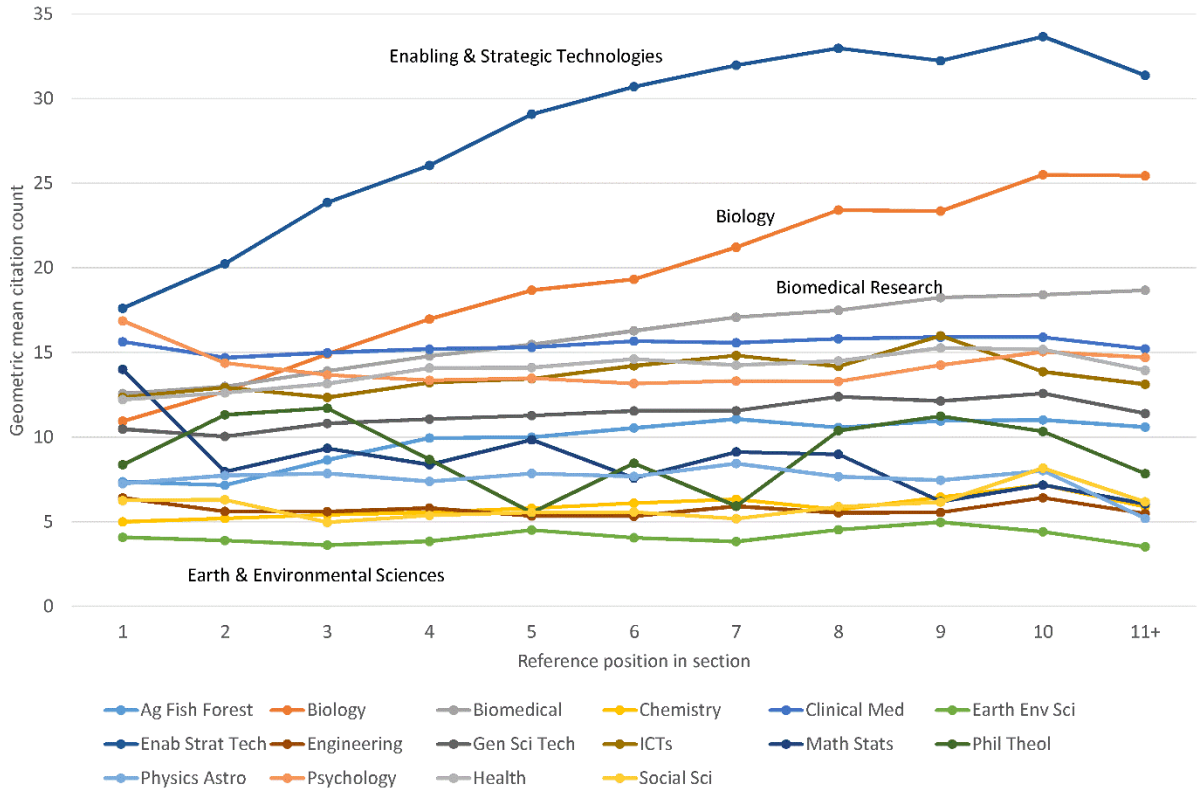


Figure 4. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Methods**. Qualification: Field must have 30+ articles in each position.

Average citation counts in the Results section are little affected by their order (Figure 5), except for Psychology & Cognitive Sciences, ICTs and Public Health, which have clear decreasing trends.

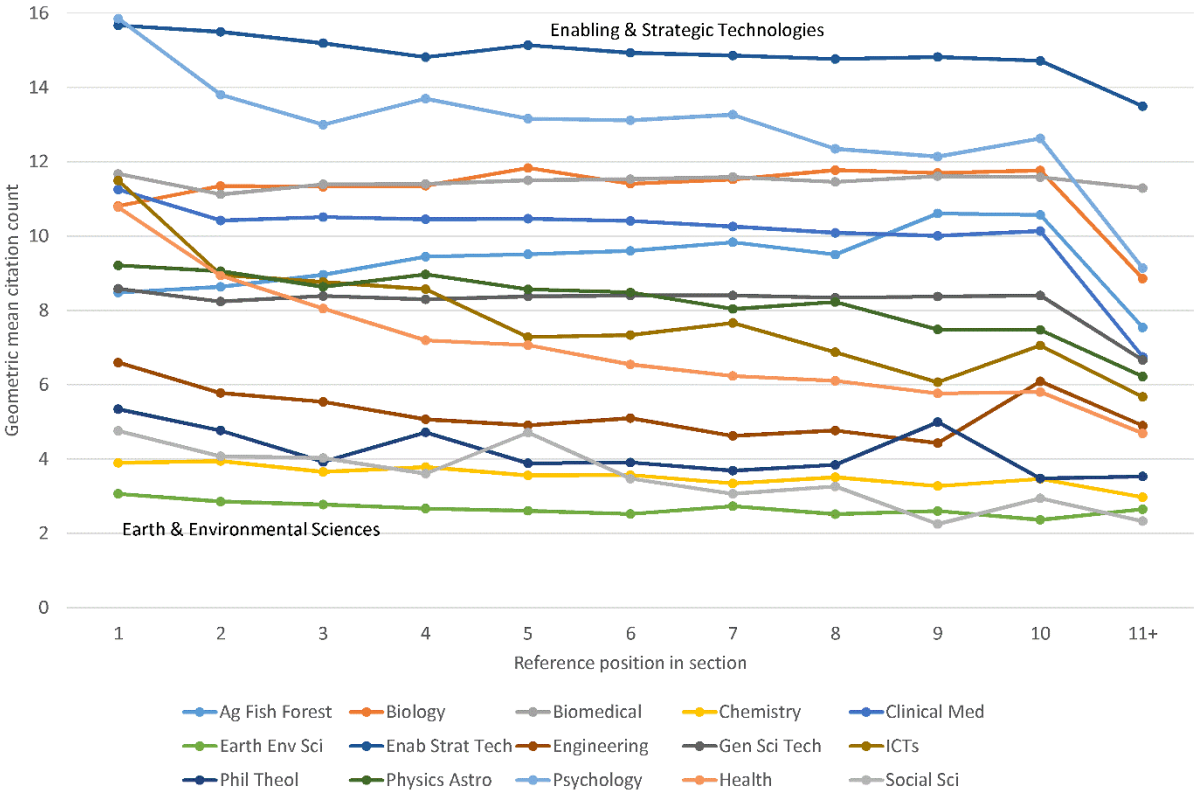


Figure 5. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Results**. Qualification: Field must have 30+ articles in each position.

Average citation counts in the Discussion section show a slight decline in average citation counts overall in most fields, as citation position increases (Figure 6). Two fields display erratic shapes, probably due to low numbers in both cases.

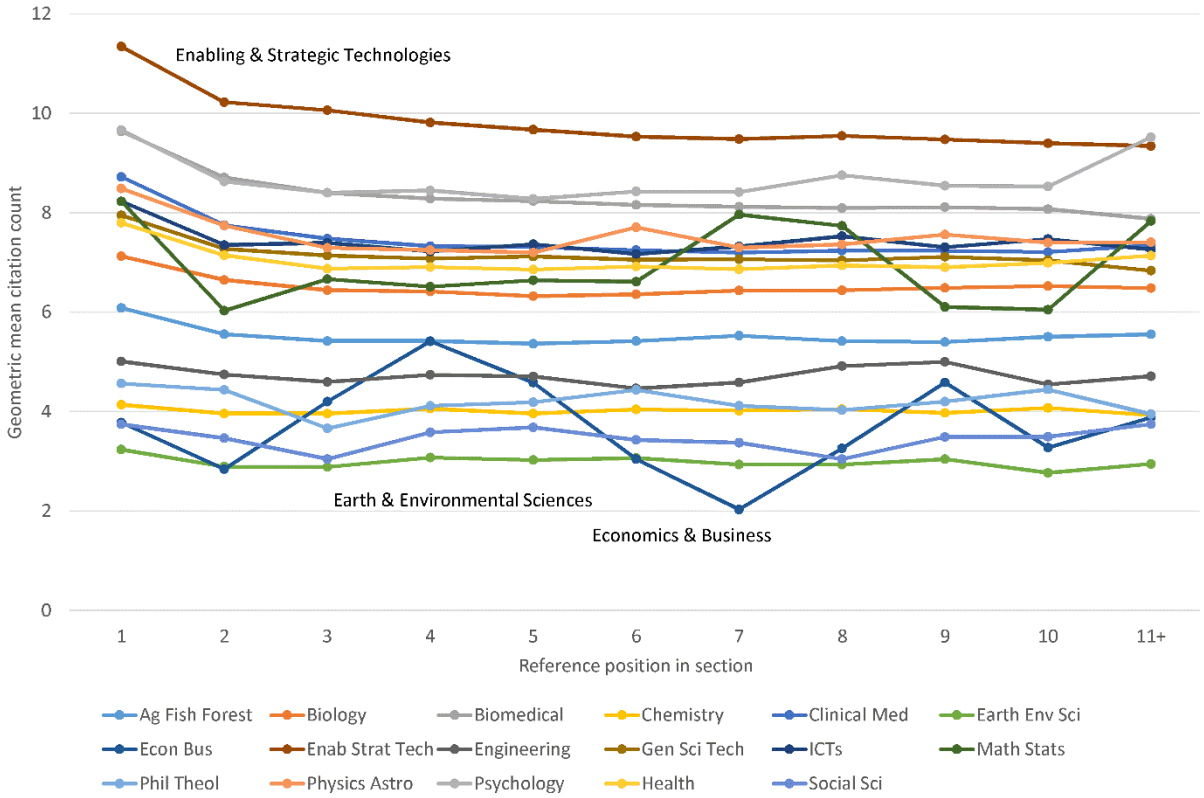


Figure 6. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Discussion**. Qualification: Field must have 30+ articles in each position.

In the Conclusions (Figure 7), three fields have a decreasing trend in average citation counts as citation position increases. The first article is the highest cited in most fields.

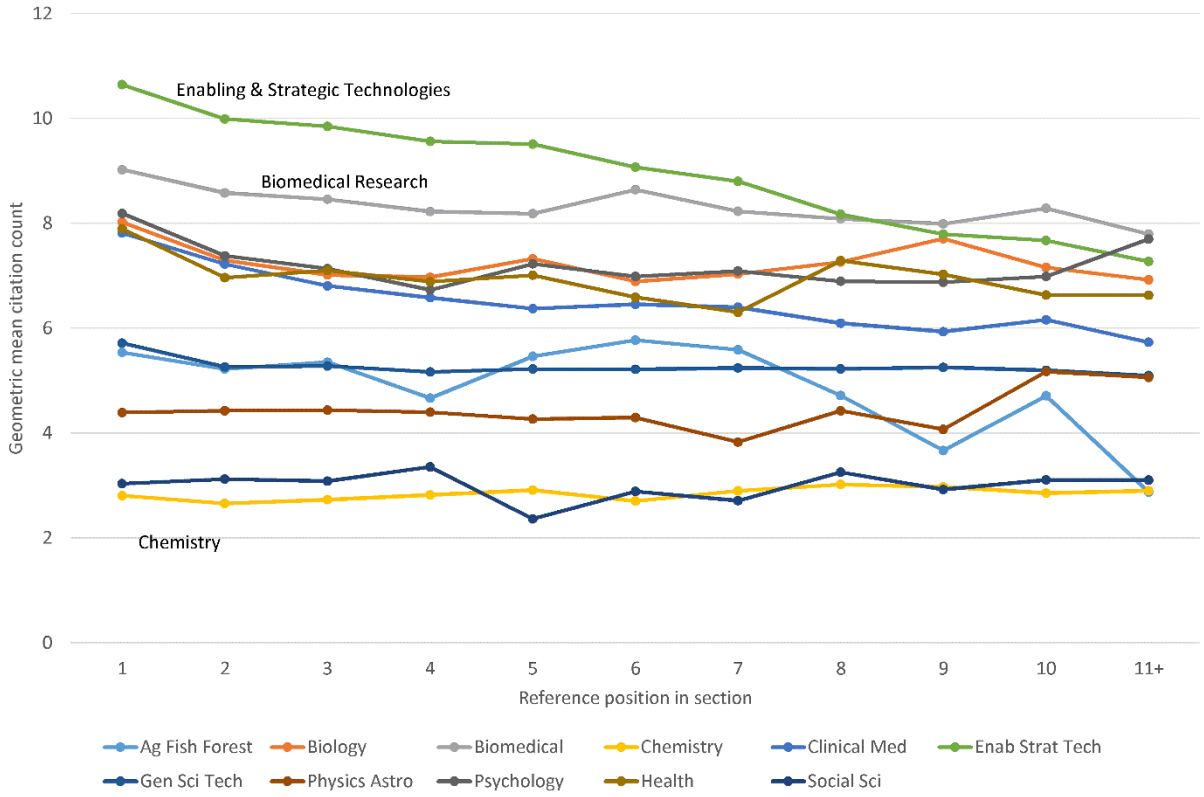


Figure 7. Geometric mean number of PMC Open Access citations for articles by position in the reference list within the **Conclusions**. Qualification: Field must have 30+ articles in each position.

## Discussion

The results are limited by using citation counts from within the collection, which lowers the numbers and has a greater impact on articles in fields that are not well represented in the PMC Open Access collection. The lower numbers will tend to reduce the differences between average citation counts in different reference positions. The results are also restricted by the source, the PMC Open Access collection, which is biased towards biomedical research and restricted to open access documents. The focus on journal articles may also affect fields differently, for example if some tend to use books rather than journal articles for general citations. Some article types do not use section headings, such as letters, and these are important in the natural sciences. Another limitation is that some authors may discuss prior research in chronological order even if this does not follow a trend from general to specific. This would tend to make earlier papers in the Introduction and Background sections more cited (because they are older) without them necessarily playing a different role. The results therefore do not prove the early cited papers make a lesser contribution to the cited paper, although they are consistent with this hypothesis.

The combining of articles from different years into a single analysis will also tend to weaken the differences between fields because more recent papers may cite articles that are currently uncited but that may later become highly cited. An exploration of average (median or geometric mean) age for cited documents for the same collection of articles did not find a strong pattern for papers cited earlier in sections to be older, however.

Highly cited references are most likely to be found in the Methods section. This could be because the need for strong supporting references is greatest when justifying or

explaining methods, leading to conservatism in the selection of Methods references. It could also be due to standard methods becoming accepted, with the inventing paper being cited for them. There may also be more standard works for methods, such as review articles, that attempt to guide methods choices because researchers may be less expert in methods than in the topic of their research.

The results are consistent with prior research suggesting that perfunctory citations are most likely to be found in the Introduction (Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, 1998; Tang & Safer, 2008), if it is accepted that such perfunctory citations are likely to be highly cited. The results also confirm previous findings of disciplinary differences in the relationship between average citation counts and position within the citing paper (Boyack, van Eck, Colavizza, & Waltman, 2018).

## Conclusions

There is a tendency for the first paper in the Introduction, Background, Discussion and Conclusion to be relatively highly cited in most fields, although the effect of order is not huge. Thus, whilst it seems to be common to start these sections with a classic citation – and particularly for the Introduction and Background - this practice is not universal and the current paper has not directly tested whether the first papers tend to be general. Nevertheless, the findings add weight to previous research suggesting that early citations in these sections can be perfunctory.

Based on the above argument, it seems that highly cited papers may not make direct contributions to scholarship in a way that is proportional to their citation counts. This adds to the strength of evidence that the citation counts of highly cited papers should be viewed with suspicion. Thus, techniques that do not greatly weight highly cited articles, such as percentile (Bornmann, Leydesdorff, & Mutz, 2013) and log-based (Thelwall, 2017) citation indicators may be preferable to total or average citation counts. Alternatively, highly cited papers may tend to perform a different role, such as agenda setting, by opening avenues for research. For example, an article about the international spread of malaria is frequently cited at the start of biomedical papers about Malaria. Subject experts would need to determine if this paper is agenda setting by encouraging new Malaria research or whether it is a convenient context-setting citation that has had little influence on future research. Thus, there are three possible conclusions from the current paper.

- If highly cited papers tend to be initial general references that add little to the citing paper, research evaluators should either (a) avoid citation impact formulae that overvalue highly cited papers by treating their citations as equal to the citations of less cited papers and use instead log based or percentile indicators, or (b) attempt to detect general low value citations, perhaps through their order or position in the citing article text, and award them a lower citation score. Option (b) would be difficult to apply in practice. It would require a formula to estimate the reduced score to be given to articles based on their citation position and may require expert judgement to decide whether individual articles that are frequently cited early tend to play a lesser role for the citing paper.
- If highly cited papers near the start of a section tend to play a different, but valuable, role in science, such as agenda setting, then their citation counts could either be accepted at face value or recorded separately as evidence of a different type of impact.

# References


Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (pp. 66-77). South Africa: Durban.

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. Journal of Informetrics, 7(1), 158-165.

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.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. Journal of the American Society for Information Science, 51(7), 635-645.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

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.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. Journal of Informetrics, 7(4), 887-896.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? Journal of the Association for Information Science and Technology, 68(4), 984-998.

Klitzing, N., Hoekstra, R., & Strijbos, J. W. (2018). Literature practices: processes leading up to a citation. Journal of Documentation, 75(1), 62-77.

Lin, C. S. (2018). An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. Scientometrics, 116(2), 797-813.

Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. Journal of the American Society for Information Science, 49(6), 530-540.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. Science, 159(3810), 56-63.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social studies of science, 5(1), 86-92.

NCBI (2015). Journal Archiving and Interchange Tag Library NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015). https://jats.nlm.nih.gov/archiving/tag-library/1.1/index.html

Shadish, W. R., Tolliver, D., Gray, M., & Sen Gupta, S. K. (1995). Author judgements about works they cite: three studies from psychology journals. Social studies of Science, 25(3), 477-498.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. Journal of Documentation, 64(2), 246-272.

Thelwall, M. & Fairclough, R. (2015). Geometric journal impact factors correcting for individual highly cited articles. Journal of Informetrics, 9(2),263–272.

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. Journal of Informetrics, 11(1), 128–151.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, did we op. cit. your idem? Journal of Academic Librarianship, 1(6), 19-21.