# Do You Cite What You Tweet? Contextualizing the Tweet-Citation Relationship

Madelaine Hare[*], Keith MacKnight[**], Mercie Chikezie[***], Geoff Krause[****], Timothy D. Bowman[*****], Rodrigo Costas[******], Philippe Mongeon[*******]

[*]*maddie.hare@dal.ca*
https://orcid.org/0000-0002-2123-9518
Department of Information Science, Dalhousie University, Canada

[**]*kt531164@dal.ca*
https://orcid.org/0009-0001-0059-1926
Department of Information Science, Dalhousie University, Canada

****mr576353@dal.ca*
https://orcid.org/0009-0004-8412-3557
Department of Information Science, Dalhousie University, Canada

*****gkrause@dal.ca*
https://orcid.org/0000-0001-7943-5119
Department of Information Science, Dalhousie University, Canada

*****timothy.d.bowman@wayne.edu
https://orcid.org/0000-0003-0247-4771
School of Information Sciences, Wayne State University, United States

******rcostas@cwts.leidenuniv.nl
https://orcid.org/0000-0002-7465-6462
Centre for Science and Technology Studies, Leiden University, The Netherlands

*******PMongeon@dal.ca*
https://orcid.org/0000-0003-1021-059X
Department of Information Science, Dalhousie University, Canada

**Abstract**
Investigating the context in which researchers engage with social media objects facilitates a greater understanding of their research behaviour. This study shifts analytical focus from the research paper itself to the geographical, socio-topical, and individual dimensions of the Tweeter and the tweeted paper to understand if researchers cite what they tweet. Results show that Tweeters are more likely to cite papers affiliated with their same institution, papers published in journals in which they also have published, and papers in which they hold authorship. It finds that the older the academic age of a Tweeter the less likely they are to cite what they tweet, though there is a positive relationship between citations and the number of papers they have published and references they have accumulated over time. This paper sheds light on the contextual nature of the tweet-citation relationship.

## 1. Background

In the early days of altmetric research much attention was directed at the correlation between tweets (or other altmetric indicators) and citations with hopes that tweets could predict future citations and thus provide an early citation impact indicator. Over the past decades, a host of issues related to the use and interpretation of altmetrics have been the subject of scholarly attention. Haustein (2016) identifies the lack of conceptual frameworks and theoretical foundations relating to altmetrics as interfering with their interpretation and understanding. This is facilitated by altmetrics' heterogeneity (purpose, function, diversity of indicators, data source,

and selectivity that dictate their online activities and events) as another major challenge associated with their use and interpretation of their meaning (Haustein, 2016).

Haustein (2016) forwards the need for a more developed conceptualization of altmetrics, building on Lazarsfeld's (1993) argument that for any metric to be a valid indicator of a social act, the act must be conceptualized. Our study rethinks this notion and contends that to meaningfully interpret altmetrics as indicators of research behaviour, the behaviour must be contextualized. Using social media events for the purposes of scholarly metrics requires us to understand the meaning and reason for the activity, which is not easily done; as Haustein (2016) argues, the same events on the same social media platforms can occur for different reasons. For example, a researcher might tweet a paper to promote their own work, to share something relevant to their field, or for the purposes of criticism (Haustein, 2016). Creating models that differentiate between different types of social media acts is possible, as demonstrated by Mongeon (2018) in his research on analysing information shared by researchers on social media using social and topical distance. In this way, social media activity could therefore be contextualized by shifting attention from research papers or the outputs of social media (tweets) to elements relating to the users themselves.

The research of Diaz-Faes et al. (2019) demonstrates that the activity of users on Twitter in relation to scientific publications can used to group these users into different dimensions of behaviour. Classifying users in this way is another potential approach to arriving at an understanding of the context for why users are engaging with content. Diaz-Faes et al. (2019) argue for the creation of secondary social media metrics, which move away from a focus on papers and their reception on social media to focus on users, their online activity, and their interactions with social media objects.

Ultimately, there is difficulty in correlating meaning with tweets, as the instance of tweeting is linked to various changing contexts; situating tweets within their context can aid in deciphering the likelihood of a resulting citation. Our work proposes to revisit the relationship between tweets and citations from an information behaviour perspective, shifting focus from the research paper to the Tweeters and their social media and publication behaviours as the elements of analysis. We do this by operationalizing the context of the relationship between the Tweeter and the tweeted paper by looking at geographical and socio-topical dimensions of the Tweeter and the publication. We also consider the individual characteristics of tweeting authors. Our objective is to determine if the act of a researcher tweeting an article can predict whether they will cite that same publication. An understanding of this connection can illuminate how research is disseminated through social media and engaged with by scholars; further, it could contribute to how we use altmetrics to interpret and understand the societal and research impact of scholarly work.

## 2. Research Objectives

The goal of this work is to examine the contextual characteristics associated with researchers tweeting publications and how they might predict a subsequent citation. We hypothesize that the topical proximity of the tweeted work to the Tweeter’s research will positively affect the likelihood of a future citation. We also hypothesize that geographical proximity will positively affect the likelihood of a future citation (interaction effect with topical proximity). Lastly, we hypothesize that several factors related to the tweeting authors can affect the likelihood of a future citation, such as their academic age, or their number of tweets, references, or published papers.

Specifically, our research questions are:

> **RQ1** To what extent does the act of tweeting a research publication predict the act of citing it?
> **RQ2** How are the individual characteristics of the Tweeter (academic age and total number of tweets, authored papers, distinct references) related to the likelihood of citing the tweeted paper?
> **RQ3** How is the geographical proximity (country and institutional affiliation) of the authors of the tweeted paper related to the Tweeter?
> **RQ4** What is the socio-topical relationship (journal and topic) between the Tweeter's research and the tweeted paper?

## 3. Data and Methods

We use a relational database version of the OpenAlex database (Priem et al., 2022) hosted by the Maritime Institute for Science, Technology, and Society (MISTS). In this database, OpenAlex works are assigned to a discipline from the Science Metrix classification (Archambault et al., 2011), based on the journals in which they are published. For journals in the "paper-level classification" category, we assign papers to a discipline based on the cited journals. In case of a tie, the paper is assigned to the tied disciplines. Our researcher data come from the open dataset of scholars on Twitter created by Mongeon, Bowman & Costas (2022), which includes approximately 500,000 matched OpenAlex author IDs and Twitter accounts. Finally, data on the tweeted publications and the Twitter users who tweeted them was obtained from a Crossref Event Data dump downloaded in January 2023.

By combining these datasets, we obtain tables in which each observation is a tweeted paper and includes relevant metadata about the tweet, the Twitter user and their publication record, as well as the tweeted publications and their authors**.** The final analyzed dataset totalled 7, 085, 157 tweets made between 2017 and 2019.

Table 1. Dimensions and their indicators

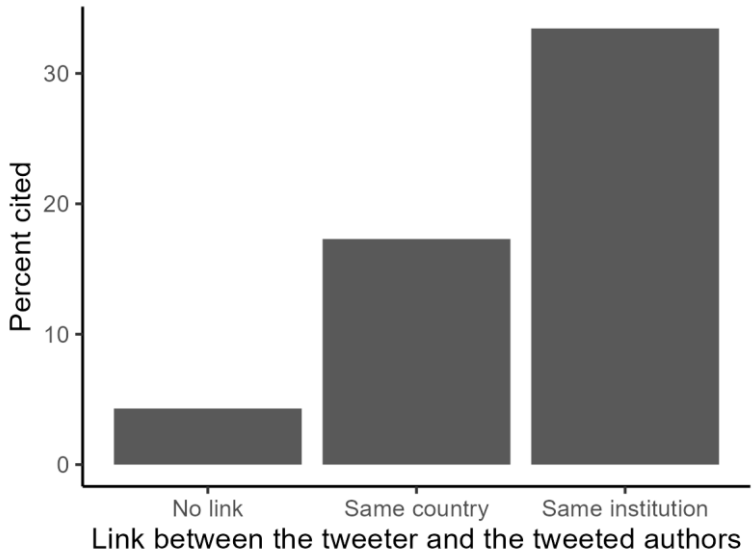
| Dimension | Indicator | Description |
|---|---|---|
| Geographical | same_county | Dichotomous variable indicating whether the tweeting author is affiliated to the same country as at least one of the authors of the tweeted publication. |
| | same_institution | Dichotomous variable indicating whether the tweeting author is affiliated to the same institution as at least one of the authors of the tweeted publication. |
| Socio-topical | same_domain | Dichotomous variable indicating whether the tweeting author has at least one publication in the same domain as the tweeted publication. |
| | same_field | Dichotomous variable indicating whether the tweeting author has at least one publication in the same field as the tweeted publication. |
| | same_subfield | Dichotomous variable indicating whether the tweeting author has at least one publication in the same subfield as the tweeted publication. |

| | | |
|---|---|---|
| | same_journal | Dichotomous variable indicating whether the tweeting author has at least one publication in the same journal as the tweeted publication. |
| | co-authorship | Dichotomous variable indicating whether the tweet was a co-author on a tweeted paper. |
| | self-tweet | Dichotomous variable indicating whether a Tweeter tweeted their own work. |
| Individual | academic_age | Variable calculated by subtracting the first year of publication from the year of the tweet. |
| | n_tweeted_papers | The total number of tweeted papers by a Tweeter. |
| | n_papers | The total number of publications of a Tweeter. |
| | n_distinct_references | The total number of distinct references a Tweeter cited cumulatively |

## 4. Results

*Geographical dimensions*

Figure 1. Citation rates by geographic affiliation




The results of our analysis show the relationship between citation rates and country and institution. Figure 1 shows that 17.3% of tweeted papers within our dataset which were created in the same country as the Tweeter were later cited by that Tweeter, while 33.5% of papers from the same institution as the Tweeter were also later cited. 4.3% of the tweeted papers with no identified geographical tie between the Tweeter and the tweeted paper are cited by the Tweeter. As indicated in Figure 1, authors are more likely to cite papers they tweet if the paper was affiliated with the same institution of the Tweeter. This likelihood is halved if the paper is affiliated with the same country of the Tweeter, though this affiliation does positively influence their likelihood to cite the tweeted publication. Papers from the same institution as the Tweeter are also somewhat likely to have a degree of topical proximity to the work of the Tweeter, which may affect their likelihood of being cited.

*Socio-topical dimensions*

Figure 2. Citation rates by socio-topical characteristics

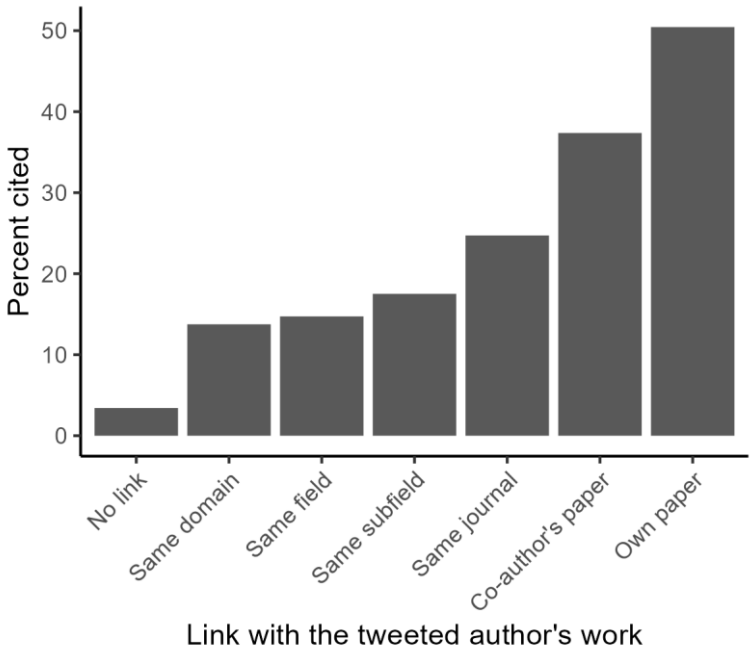



Figure 2 shows the relationship between various socio-topical dimensions and cited papers. Our results indicate that a publication is more likely to be cited if it was written by the Tweeter. Similarly, a tweet is more likely to result in a citation if the Tweeter was a co-author of a paper. If a tweeting author has at least one publication in the same journal as the tweeted paper, this also impacts the likelihood of a citation. This may be related to the factor of topical proximity, as academic journals cited papers are published in are likely to contain papers with similar topics as the Tweeter. Further, sub-fields have significant bearing on whether a work will be cited, whereas same field and domain possess relatively equal, but lesser influence. Our results, therefore, indicate that if the topic or discipline of the paper is the same as that of the Tweeter it is more likely to be cited, instantiated by the greatest socio-topical influence on whether a tweet is likely to be cited being whether that is authored or co-authored by a Tweeter.

*Individual dimensions*

Figure 3. Citation rates by academic age

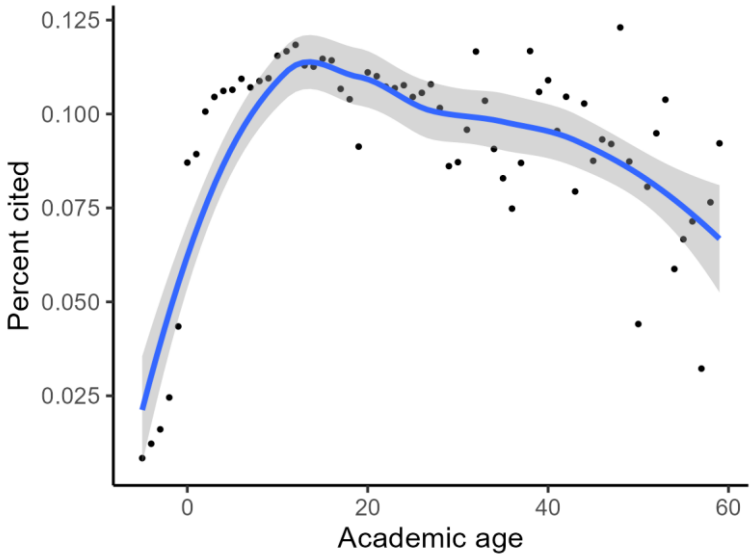



Figure 3 shows that the academic age of the Tweeter has a negative correlation with whether they will cite what they tweet. Our results indicate that early career researchers have a higher likelihood of later citing the papers they share on Twitter, this likelihood peaks around the tenth year, and as academic age increases, researchers are less likely to cite publications they tweet.

Figure 4. Citation rates by total number of tweeted papers

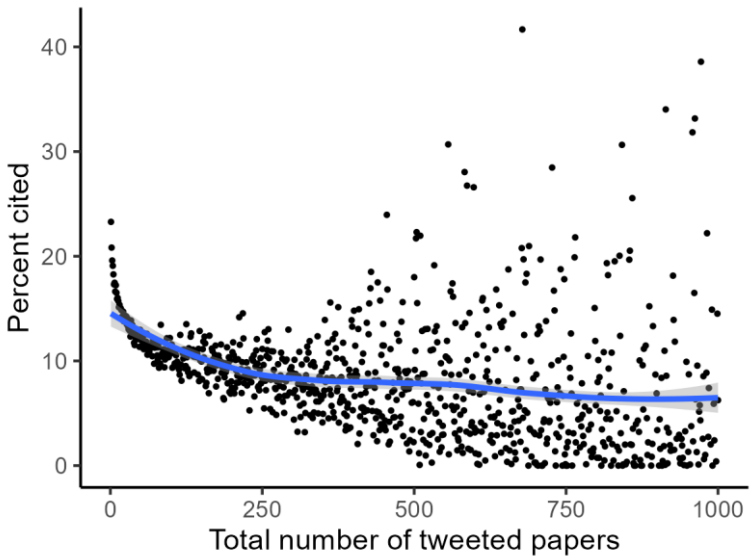
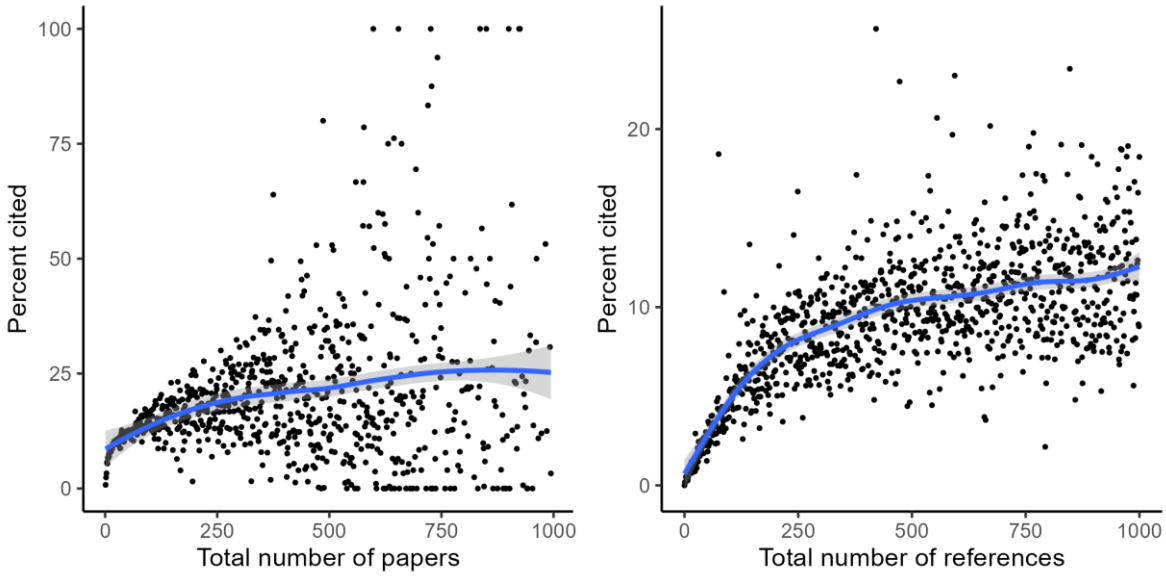



Figure 4 show a negative correlation between the total number of a researcher's tweeted papers and the rate of citation. Authors are less likely to cite what they tweet if they are highly active Tweeters.

Figure 5. Citation rates by total number of papers (left) and distinct references (right)




Our analysis finds that the individual characteristics the Tweeter have an impact on whether a tweeted a paper will later be cited. Figure 5 shows that the total number of papers a Tweeter has published in their academic career has a weak but positive correlation with their likelihood of citing tweeted papers. A stronger correlation is evident in the first 100 papers, indicating that researchers who are more prolific are more likely to cite what they tweet, but this tapers off around 250 publications. Further, the cumulative number of distinct references a Tweeter has made also has a positive correlation with their likelihood to cite what they tweet. As a Tweeter's career progresses, however, this positive trend levels off, again indicating that early career researchers are more likely to cite publications they have tweeted than those with lengthier careers.

## 5. Discussion

The results of our study indicate that various geographic, socio-topical, and individual dimensions relating to an author and a tweeted publication influence the likelihood of a tweeted publication being cited. Tweeters are more likely to cite works that are affiliated with their same institution, possibly indicating greater topical similarity or relevance, or possible intellectual involvement with colleagues within their own institution. This finding aligns with our socio-topical results, which show that Tweeters are more likely to cite work they author or co-author; and these collaborations are more likely to occur with colleagues within their same institution or country. In this way, cited tweets are influenced by geographic proximity and privy, perhaps, to the institutional dynamics of scholarship. This also sheds light on the heterogenic uses of social media; scholars citing their own work may indicate how Twitter can be used as a platform for increasing the visibility of one's own scholarship.

Moreover, the topical similarity of a tweeted paper to one's own work and field of study is influential on the relationship between the tweet and its eventual citation. That subfield has greater influence than fields or domains show that Tweeters are citing works specifically relevant to their work, and less if they only relate in a more general sense to their disciplinary area. Tweeters are also more likely to cite papers published in journals in which they too have published, demonstrating the disciplinary circles that influence how scholars interact with work.

Finally, individual dimensions depicted in our results illuminate how academic age and the characteristics of a Tweeter's scholarly career (total number of tweeted papers, published papers, and distinct references) influence their citation activity. The plateau of citation rates depicted in the total number of published papers and distinct references aligns with the negative trend shown in the Figure 3 depicting academic age. As careers progress and researchers publish more, they are less likely to cite what they tweet. This may indicate that researchers may be more active on Twitter at the start of their careers and aim to make their work and scholarly presence more visible to their peers, and later in their career they are less likely to engage with work on social media, correlating to a drop in their likelihood to cite tweeted papers. Interestingly, the more papers a researcher tweets, the less likely they are to cite them, potentially indicating that less frequent Tweeters engage more deeply with the papers they choose to disseminate on social media, or those that tweet a great deal may engage with work on Twitter for a diverse range of reasons not always in relation to their own work, substantiating Bowman et al.'s (2015) contention that Tweeters tweet for both professional and personal purposes.

## 6. Conclusion

As the use of altmetrics develops and informs a deeper understanding of a new type of research impact, the contextualization of the relationship between altmetric activity and Twitter users is necessary to their meaningful interpretation. This study's analysis of over 7 million unique tweets reveals the geographic, socio-topical, and individual characteristics that influence the likelihood of researcher's citing what they tweet. It has shown that the papers of cited tweets are most often affiliated with the same institution as the Tweeter, that Tweeters more often cite papers they author or co-author, and that as their career advances they are less likely to cite their tweets, though their citation rates increase with their total number of papers published and distinct references incurred. Finally, if they are prolific Tweeters, they are also less likely to cite what they tweet. Our findings have implications extending beyond Tweeter behaviour; they elicit more consideration of the true meaning of altmetric activity, shifting attention from papers and the tweets as units of analysis to the researchers engaging with work in both social and scholarly realms. The influence of Twitter on research engagement can be understood by contextualizing Tweeters and their connection to individual works to add multiple dimensions to their altmetric activity.

### *Limitations*

This study possesses several limitations. First, this study only considers Twitter counts and does not analyze other forms of altmetrics. The Open Dataset of Scholars on Twitter used to match Twitter users with researchers is a limited dataset of authors with at least one publication. Our dataset created with CrossRef Event data only considers papers with DOIs. Additionally, errors with OpenAlex disambiguation may incorrectly attribute authors to publications. Further, this study does not consider the influence of time on the causation or correlation of tweets and subsequent citations. Finally, by gathering OpenAlex data from a May 2022 data dump, citations accumulated past that period are excluded from our data.

### *Further research*

Further research that aims to contextualize relationships between altmetric events and citations may wish to broaden the scope of an altmetrics analysis by bringing in other forms of altmetric data; discussions that aim to compare different social media metrics could use a similar approach which considers geographic, socio-topical, and individual dimensions of altmetric activities (Haustein et al., 2014). Emerging altmetric data sources such as Mastodon could

provide insights on the migration of researchers to new venues for the purposes of disseminating knowledge. Additionally, other characteristics not included in the individual dimensions analyzed in this paper could be considered, such as retweets. Further studies might choose to consider journal impact factor or highly cited papers to analyze socio-topical dimensions from an impact perspective. Content-level analysis of tweets could also be performed to better understand the causal aspects of a Tweeter's decision to engage with a work, shedding light on whether a work was tweeted for purposes of promotion, sharing, criticism, or other reasons. Disciplinary characteristics could also be investigated in more detail to determine if certain disciplines have higher or lower rates of citations. Further, authorship order could be an enlightening aspect of future analyses, illuminating whether Tweeters are more likely to cite works in which they are first author, and how academic age may intersect with these elements.

**Open science practices**

The dataset analyzed in this paper uses the Open Dataset of Scholars on Twitter created by Philippe Mongeon, Timothy Bowman, and Rodrigo Costas. This is a dataset of paired OpenAlex author_ids (https://docs.openalex.org/about-the-data/author) and tweeter_id. The dataset includes 492,124 unique author_ids and 423,920 unique tweeter_ids forming 498,672 unique author-tweeter pairs. It is available here: https://zenodo.org/record/7013518#.ZDlmpHZKi5c and the following preprint provides details about the matching process and links to R scripts: https://doi.org/10.48550/arXiv.2208.11065
The dataset and R scripts produced for this analysis will be made available on Zenodo. The authors also intend to publish the final article in a fully open access publication so that it may reach as wide of an audience as possible.

**Acknowledgments**

The authors wish to thank the Social Sciences and Humanities Research Council for the funding to conduct this study.

**Author contributions**

Madelaine Hare: Writing – original draft, Writing – review and editing, Visualization, Formal analysis.
Keith MacKnight: Writing – original draft, Writing – review and editing, Visualization.
Mercy Chikezie: Writing – original draft, Writing – review and editing, Visualization.
Geoff Krause: Formal analysis, Software, Data Curation, Visualization.
Timothy Bowman: Conceptualization, Writing – original draft, Data curation.
Rodrigo Costas: Conceptualization, Writing – original draft, Data curation.
Philippe Mongeon: Conceptualization, Software, Supervision, Data Curation, Resources.

**Competing interests**

The authors declare that they have no competing financial or non-financial interests.

**Funding information**

This research was conducted using funds from the Social Sciences and Humanities Research Council.

**References**

Archambault, E., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, 66–77.

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics, 103*(3), 1123–1144. https://doi.org/10.1007/s11192-015-1565-y

Bowman. (2015). Differences in personal and professional tweets of scholars. *Aslib Journal of Information Management, 67*(3), 356–371. https://doi.org/10.1108/AJIM-12-2014-0180

Costas, R., Mongeon, P., Ferreira, M. R., van Honk, J., & Franssen, T. (2020). Large-scale identification and characterization of scholars on Twitter. *Quantitative Science Studies, 1*(2), 771–791. https://doi.org/10.1162/qss_a_00047

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do Altmetrics" Correlate With Citations? *Journal of the Association for Information Science and Technology, 66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

De Winter, J. C. F. (2015). The relationship between tweets, citations, and article views for PLOS ONE articles. *Scientometrics, 102*(2), 1773–1779. https://doi.org/10.1007/s11192-014-1445-x

Diaz-Faes, A. A., Bowman, T. D., & Costas, R. (2019). Towards a second generation of 'social media metrics': Characterizing Twitter communities of attention around science. *PLoS ONE, 14*(5). https://doi.org/10.1371/journal.pone.0216408

Didegah, F., Bowman, T. D., & 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*, *69*(6), 832–843. https://doi.org/10.1002/asi.23934

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research, 13*(4), e123. https://doi.org/10.2196/jmir.2012

Haustein. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics, 108*(1), 413–423. https://doi.org/10.1007/s11192-016-1910-9

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., and Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *Information Technology*, vol. 56, no. 5, 2014, pp. 207-215. https://doi.org/10.1515/itit-2014-1048

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics, 101*(2), 1027–1042. https://doi.org/10.1007/s11192-014-1229-3

Ke, Q., Ahn, Y.-Y., & Sugimoto, C. R. (2017). A systematic identification and analysis of scientists on Twitter. *PLoS One, 12*(4), e0175368. https://doi.org/10.1371/journal.pone.0175368

Mongeon, P. (2018). Using social and topical distance to analyze information sharing on social media. *81st Annual Meeting of the Association for Information Science & Technology*.

Association for Information Science and Technology. https://doi.org/10.1002/pra2.2018.14505501043

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (arXiv:2205.01833). arXiv. https://doi.org/10.48550/arXiv.2205.01833

Robinson-Garcia, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy, 45*(6), 815–826. https://doi.org/10.1093/scipol/scy024

Shu, F., Lou, W., & Haustein, S. (2018). Can Twitter increase the visibility of Chinese publications? *Scientometrics, 116*(1), 505–519. https://doi.org/10.1007/s11192-018-2732-8

Singh, L. (2020). A Systematic Review of Higher Education Academics' Use of Microblogging for Professional Development: Case of Twitter. *Open Education Studies, 2*(1), 66–81. https://doi.org/10.1515/edu-2020-0102

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology, 68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics, 115*(3), 1231–1240. https://doi.org/10.1007/s11192-018-2715-9

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE, 8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics, 12*(1), 237–248. https://doi.org/10.1093/femsle/fny049

Wooldridge, J., & King, M. B. (2019). Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology, 70*(3), 271–282. https://doi.org/10.1002/asi.24122

Wuchty, S., Jones, B. B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Zhang, L., & Wang, J. (2018). Why highly cited articles are not highly tweeted? A biology case. *Scientometrics, 117*(1), 495–509. https://doi.org/10.1007/s11192-018-2876-6

Zuccala, A. A., Verleysen, F. T., Cornacchia, R., & Engels, T. C. E. (2015). Altmetrics for the humanities Comparing Goodreads reader ratings with citations to history books. *Journal of Information Management, 67*(3), 320–336. https://doi.org/10.1108/AJIM-11-2014-0152