https://doi.org/10.51408/issi2025_049

Are Citation Context Information Stronger Related to Peer Ratings Than Citation Counts? A Descriptive Analysis

Paul Donner[1], Stephan Stahlschmidt[2], Robin Haunschild[3], Lutz Bornmann[4]

[1]*donner@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstrasse 6a, 10117 Berlin (Germany)

[2]*stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstrasse 6a, 10117 Berlin (Germany)
University of Granada, Unit of Computational Humanities and Social Sciences (U-CHASS), EC3 Research Group,
Campus Universitario de Cartuja, 18071 Granada (Spain)

[3]*R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Information Service,
Heisenbergstrasse 1, 70569 Stuttgart (Germany)

[4]*bornmann@gv.mpg.de, L.Bornmann@fkf.mpg.de*
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)
Max Planck Institute for Solid State Research, Information Service,
Heisenbergstrasse 1, 70569 Stuttgart (Germany)

Abstract

In this study, we investigated whether citation context information is able to increase the validity of citation impact analyses to measure research quality compared to simple citation counts. We analyzed the statistical relationships of information extracted from structured citation context data in the Web of Science (Clarivate) such as the placement of citations within specific sections of an article with post-publication peer review quality ratings from Faculty Opinions (H1 connect), used as an external validity criterion for research quality. The study is based on publications in medicine and life sciences. Our findings reveal that quantitative metrics derived from citation contexts, particularly in-text citation counts, exhibit stronger correlations with expert evaluations compared to traditional citation counts. Consequently, integrating citation context data appears to improve the legitimacy and reliability of citation analyses as tools for assessing research quality.

Introduction

Implicit in conventional citation analysis, which is mostly an analysis of the times cited information from citation databases, is the assumption that all citations have equal value. A paper is cited or not – depending on its utility and merit. More detailed inspection of citations in scientific documents shows, however, that there are great differences in how literature is processed by the authors in their papers. Some papers are cited *en bloc* within a long list of other cited papers to demonstrate that there is literature available on a certain topic (mostly in the introduction of a paper) while other papers are discussed in great depth. Evidently, the former first category of cited paper has had less impact on the citing paper than the second one. Whereas the traditional citation analysis – the times cited analysis – focuses on references in the reference list of a document, "an *in-text citation* is a *mention* of a reference within the full text of a document. A reference can be mentioned one or more times in a document. Each mention is an in-text citation" (Boyack, van Eck, Colavizza, & Waltman, 2018). It is one goal of citation context analysis (CCA) to further develop traditional citation analysis and to provide more detailed insights into the use and impact of publications. Recently, Clarivate has started to systematically provide citation context information in the Web of Science (WoS) for many citing publications. This data has now attained sufficient coverage that an initial analysis has become feasible.

In this study, we explored the possibilities of using Clarivate's citation context information for more meaningful citation analyses. We investigated if CCA can improve the validity of measuring research impact (as one important dimension of research quality) by bibliometric means. The reason for a hypothesized improvement in construct validity is that CCA goes beyond reference list citation counting, quantitatively and qualitatively. The quantitative extension lies in the counting of repeated in-text citations and in the information of how many papers are cited to support a particular statement. We used the latter information to calculate a score on the level of cited papers. This score indicates the proportion of in-text citations in which the paper was cited as the single cited reference to support a statement, rather than one of several. The qualitative extension consists of taking the position of citations (e.g., in certain sections) and the text surrounding citations into account (e.g., is there a direct use of the cited paper's content, or does the cited paper serve as a background reference for a certain topic).

We hypothesized that citation context information aggregated at the level of cited publications contains additional information relevant for the assessment of publications' research quality as higher quality research is used differently in citation contexts than lower quality research. Higher quality research is expected to be utilized more often as significant citations, rather than perfunctory citations, because their influence on the citing paper's author is assumed to be greater. This could manifest in different ways, such as a higher probability to be cited in specific paper sections (Cano, 1989; Maricic, Spaventi, Pavicic, & Pifat-Mrzljak, 1998; Tang & Safer, 2008), a higher probability to be used for certain purposes (Tang & Safer, 2008), more frequent mentions in the citing paper (Zhu, Turney, Lemire, & Vellino, 2015), or more frequently being cited as the only reference in a citation context, rather than being one of a string or block of references cited together in one context (Beck, Sandbulte, Neupane, & Carroll, 2018). CCA may offer a significant improvement of the underlying basis of citation analysis by moving from a superficial reference list analysis to a more sophisticated and data-rich in-text citation analysis.

In this study, we posed the research question whether CCA improves the measurement of research impact as one aspect of research quality compared to the usual citation count analysis. To answer this question, we compared peer ratings of focal papers on the platform Faculty Opinions (FO, provided by H1, https://connect.h1.co), formerly F1000, with information derived from the citation contexts of focal papers in citing papers indexed in the WoS. Since peer ratings may be the best way of assessing the quality of focal papers (Bornmann, 2011), the correlation of the ratings with simple citation counts on the one hand and outcomes of the CCA on the other hand may reveal possible improvements by the consideration of citation context information in enriched citation analyses compared to simple citation counting.

Datasets and methods

Faculty Opinions dataset

FO is a medicine and life sciences post-publication appraisal and recommendation service. FO expert members ('peers') rate papers on a 3-level ordinal scale ('good', 'very good', 'excellent') to express their perceived quality level of a paper. Note that neither low-quality nor ordinary quality publications are rated as such. Peers must regard contributions as good or better to recommend them for consideration in the FO database. They do so publicly under their own name within the FO subscription service and usually provide a concise explanation of the importance of rated publications. Given its unique nature as a large-scale dataset on concise peer reviews, FO data has been applied extensively in bibliometric and altmetric research and we refer to Williams (2017) for an in-depth description of the platform and resulting data (this description is still current although the operator changed).

H1 provided us with a dataset of 246,245 peer ratings of scientific publications from their service for this study, current as of November 2023. We excluded 282 records: FO members can provide a dissent rating which are exceedingly rare. These express disagreement with an existing recommendation but did not fit into the three-level quality scale and were therefore excluded. For each publication year from 2001 on, there are more than 3000 annual FO recommendations. The peak publication year was 2012 with over 16,000 recommendations. Papers received on average 1.2 recommendations and 16% of papers received more than one recommendation. The most common rating score was 'good', with 51%, while 39% of ratings were judged 'very good', and 10% were rated 'excellent'.

Using the official publication date of the rated paper and the date of its recommendation, we computed how long it typically took for a recommendation to be made. The average passed time is 222 days, with a standard deviation of 765 days. However, about 14% of the recommendations were posted before the recorded official publication date. Although the typically short time interval lets us assume that FO ratings are unlikely to be affected by citation count information searched by FO members, it is possible that citing authors were partially informed and influenced by FO ratings.

Web of Science citation context data

We use an April 2024 snapshot of WoS that includes the SCIE, SSCI, AHCI, CPCI-S, and CPCI-SSH and which is licensed through, and made available by, the German Kompetenznetzwerk Bibliometrie (Schmidt et al., 2024). Citation context data is available in the WoS since 2021 on a large scale under the feature name of Enriched

Cited References. This includes currently a numeric value between 0.0 and 1.0 for the relative position of the reference in the text of a paper, the original and a standardized section title, as well as the inferred reference function. Contrary to the section classification building upon the well-studied introduction, methods, results, and discussion (IMRaD) structure (Sollaci & Pereira, 2004), the citation functions constitute a classification developed by Clarivate with five classes: ‘background’, ‘basis’, ‘differ’, ‘support’, and ‘discuss’.

Matching and resulting analytical dataset

We constructed an analytical dataset by matching WoS data with citation context information to FO data. As we wanted to study the associations of citation context variables with quality assessments, our study is necessarily limited to those publications for which any citation context data is available. This study therefore does not include any uncited document records. It also does not include citation information from citing publications without citation context data. We first restricted the dataset to publications of the years 2020, 2021, and 2022 as these currently have the best relative coverage of citation context data. The used citation context data were from citing publications of any publication years. We also limited the data to papers with the document type ‘article’, since papers with different document types can be cited differently (Lundberg, 2007). For this restricted WoS dataset, we continued with the matching to the FO data.

WoS and FO records were matched primarily by the DOI. For FO records without DOIs, matching was done by exact match on journal title, volume, issue, and first page. We also wanted to include additional papers that have not been recommended by FO members but have been published in the same journals as the recommended papers. For identifying the papers without FO rating, we selected all unrated WoS records of document type ‘article’ published between 2020 and 2022 in journals which ever had published a rated paper in the entire FO dataset. For the purposes of our study, the publications without any FO rating but published in these journals were assigned the rating level ‘unrated’. Table 1 summarizes the numbers of records in the different datasets.

Table 1. Overview of datasets.

dataset	*records*
(1) WoS items ever cited with any citation context information	31,219,721 items
(2) WoS articles from 2020 to 2022 with citation context information	4,570,945 articles
(3) items with FO recommendations (publications with the same DOIs in WoS were discarded)	192,328 items, 246,245 recommendations
(4) matched data of (2) and (3)	13,617 articles, 15,771 recommendations
(5) analytical dataset: (4) extended with unrated publications	1,531,556 articles, 15,771 recommendations

Variables and statistics used

We processed the citation context data and calculated variables on the level of cited items as follows:

- Ordinary citation counts and number of in-text citations: For example, an item cited by three papers, which is referenced in these papers 2, 1, and 5 times, has a citation count of 3, but 8 in-text citations.
- Relative shares of citation contexts of normalized sections: We calculated for each cited item the relative shares of citation contexts of normalized sections, as defined by Clarivate ('introduction', 'methods', 'results', and 'discussion'). We additionally defined the section as 'missing' when no section information was available. For instance, a cited item with 5 in-text citations, of which 4 are in the introduction and 1 in the discussion section, would have variable values of 0.8 for share of introduction section, 0.2 for share of discussion section and 0.0 for the shares of the other categories.
- Relative shares of citation functions: In the same manner, the relative shares of citation functions, as defined by Clarivate ('discuss', 'background', 'basis', 'support', and 'differ') were calculated.
- Relative share of an item being cited as a single reference: A new variable was created for the relative share of an item being cited as a single reference: The share was calculated from the relative position data. References cited closely together within a citing paper were identified as those whose positions were within 1% of a paper's page of each other. This normalization for paper length in pages is necessary: A difference of, say, 0.05 on the 0.0 to 1.0 scale of a relative position is a very small distance for two references in a two-page paper but a large distance in a 40-page paper. The 1% of a page parameter value was found experimentally to provide satisfactory results by testing different parameter values on how well they identify multi-citation clusters in sample articles. The single reference share expresses what proportion of an item's citation contexts is not in such multi-reference citation contexts, usually a string of multiple references to support a single claim or statement. It quantifies the share of citation contexts in which an item is the only reference cited to support a statement.

For the descriptive analyses of associations between the citation context variables, the polyserial correlation coefficient was used, which is designed for the quantification of associations between ordered categorical and numeric variables.

Results

Table 2 shows the polyserial correlations between the citation context variables and FO ratings, in two variants. First, four ordinal levels ('unrated', 'good', 'very good', and 'excellent') were used. Second, we restricted the analyses to FO rated items (i.e., 'good', 'very good', and 'excellent'). By using the restriction to that subset, we can show more clearly which citation context variables could potentially differentiate quality at the high end. Multiple ratings for one item were not aggregated but treated as independent observations, so this view on the dataset has more observations than publication records. The results in Table 2 reveal that the number of citations and in-text citations are moderately associated with better ratings taking into account cited publications without FO rating as a supplementary fourth quality level. When excluding unrated items, the number of in-text citations also exhibits slightly higher agreements with the FO ratings than the number of citations. The coefficients for the citation section, citation function, and share as single reference variables in the table are much smaller than those for the number of (in-text) citations. Size and direction of these coefficients are inconsistent across the two calculation variants, with the exception of the 'results' section and 'differ' function shares.

Table 2. Polyserial correlation coefficients between FO ratings and citation context variables.

		including unrated papers (n=1,533,710)	*excluding unrated papers (n=15,745)*
number of citations		0.44	0.07
number of in-text citations		0.47	0.08
share as single reference		0.00	0.01
citation section	introduction	−0.06	0.03
	results	0.06	0.05
	methods	−0.03	−0.01
	discussion	0.03	−0.07
	missing	0.02	0.01
citation function	discuss	0.05	−0.03
	background	−0.04	0.03
	basis	−0.02	0.01
	support	0.00	−0.02
	differ	−0.02	−0.04

In order to have a more detailed insight into the relationship of citation (context) variables and experts' ratings, average values of the citation context variables for the four quality rating levels are presented in Table 3. The average values show that only the relationships between rating categories and citations and in-text citations are monotonically increasing. Averages for section and function shares and shares as single references only differentiate in some cases when comparing unrated to rated levels, e.g., for the 'results' section or the 'discuss' function.

Table 3. Average values of (in-text) citations, citation context variables, and share as single reference across rating categories (n=1,533,710).

		FO rating level			
citation (context) variable		*unrated*	*good*	*very good*	*exceptional*
citations		6.2	26.2	29.4	76.8
number of in-text citations		10.0	43.2	50.0	125.2
share as single reference		0.46	0.46	0.47	0.46
share of citation section	introduction	0.43	0.37	0.37	0.39
	results	0.11	0.13	0.15	0.15
	methods	0.10	0.08	0.09	0.08
	discussion	0.34	0.39	0.37	0.35
	missing	0.02	0.03	0.03	0.03
share of citation function	discuss	0.38	0.43	0.42	0.41
	background	0.46	0.43	0.42	0.45
	basis	0.12	0.11	0.11	0.11
	support	0.04	0.04	0.04	0.04
	differ	0.00	0.00	0.00	0.00

Discussion

Using citation context data that is available in the WoS since 2021 on a large scale, we investigated in this study whether CCA enhances the validity of the measurement of research impact, a critical aspect of research quality, compared to traditional citation count analysis. We conducted a quantitative analysis comparing peer ratings of papers from the FO platform, with citation context information from the papers' citations indexed in the WoS. Given that peer ratings may be a superior measure of the papers' quality, examining the correlation between these ratings, and both simple citation counts and CCA outcomes, could highlight potential enhancements from integrating citation context information into citation analyses versus relying solely on citation counts.

Our investigations of the association of research quality, in terms of FO ratings, and variables derived from citation context information, have brought to light intriguing findings. In general, our results show that the number of in-text citations associate more strongly with FO ratings than regular citation counts. The number of in-text citations thus exhibit higher construct validity as a proxy variable for research quality than citation counts. On the other hand, the correlational analysis has not shown any clear associations of the other investigated citation context variables with FO ratings.

This study is subject to some limitations. It is limited in scope to medicine and life sciences as covered by FO. The generalizability of our findings is difficult to assess due to well-known differences of citation practices across fields of science: "there are large field-level differences that are reflected in position within the text, citation interval (or reference age), and citation counts of references" (Boyack et al., 2018).

A technical limitation of our study is given by the limited availability of in-text citations. As in-text citations are much more frequent for recent citing years (at the time of this study), we focused on recent literature. It is not guaranteed that our results are transferable to older citing years.

Acknowledgments

Access to WoS bibliometric data has been supported via the German Kompetenznetzwerk Bibliometrie, funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A). Stephan Stahlschmidt's contribution has been funded by the German Federal Ministry of Education and Research (grant number: 01PH20006B). We would like to thank H1 staff for providing data access and Clarivate for detailed descriptions of the citation context information.

Refereces

Beck, J., Sandbulte, J., Neupane, B., & Carroll, J. M. (2018). A study of citation motivations in HCI research. Retrieved December, 2, 2024, from https://osf.io/preprints/socarxiv/me8zd

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*, 199-245. doi: 10.1002/aris.2011.1440450112.

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. doi: 10.1016/j.joi.2017.11.005.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284-290. doi: 10.1002/(SICI)1097-4571(198907)40:4%3C284::AID-ASI10%3E3.0.CO;2-Z.

Lundberg, J. (2007). Lifting the crown: Citation *z*-score. *Journal of Informetrics, 1*(2), 145-154. doi: 10.1016/j.joi.2006.09.007.

Maricic, S., Spaventi, J., Pavicic, 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. doi: 10.1002/(SICI)1097-4571(19980501)49:6%3C530::AID-ASI5%3E3.0.CO;2-8.

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., . . . Stahlschmidt, S. (2024). The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis. Retrieved from https://doi.org/10.5281/zenodo.13935407 doi:10.5281/zenodo.13935407

Sollaci, L., & Pereira, M. (2004). The Introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey. *Journal of the Medical Library Association, 92*, 364-367.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation, 64*(2), 246-272. doi: 10.1108/00220410810858047.

Williams, A. E. (2017). F1000: An overview and evaluation. *Information and Learning Science, 118*(7/8), 364-371. doi: 10.1108/ILS-06-2017-0065.

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. doi: 10.1002/asi.23179.