Full paper

STI 2022 Conference Proceedings

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

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

Proceeding Editors

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

OPEN ACCESS

Citation: Zhou, H., Sun, B., Guns, R., & Zhang, L. (2022). How do natural sciences learn from social sciences? Investigating the prevalence, origin, and location of references. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti2245). https://doi.org/10.5281/zenodo.6906971

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

How do natural sciences learn from social sciences? Investigating the prevalence, origin, and location of references[1]

Hongyu Zhou*, Beibei Sun**, Raf Guns* and Lin Zhang**

*hongyu.Zhou@uantwerpen.com; raf.Guns@uantwerpen.com
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, Antwerp, 2020 (Belgium)

** betty_sun@whu.edu.cn; zhanglin_1117@126.com
Center for Science, Technology & Education Assessment, Wuhan University, Wuhan, 430072 (China)
School of Information Management, Wuhan University, Wuhan, 430072 (China)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven, 3000 (Belgium)

Introduction

Since Comte's classification of science, the divide and interaction between natural sciences and social sciences have remained a focal question in philosophy of science, science of science, and research evaluation. The two major branches of science differ significantly in epistemology (research subject), research methods, publication patterns (language, productivity, publication channel) as well as education and profession, etc.

Despite the great cognitive and practical discrepancy, the intellectual exchange and mutual learning between the two are also intensively discussed, arguably more on social sciences benefiting from natural science. Some have claimed that social sciences, at an early age, attempted to emulate the natural science ideal (Beck, 1949), which gave rise to the cumulative development of social sciences knowledge, for example, the mathematization of economics. Such affinity became more prominent with the rise of digital scholarship, big data, and biotechnology. The recent zeitgeist in computational social sciences seems to empower many social sciences disciplines with natural-science-like research technologies to conduct quantitative studies using human behavioural big data.

On the other side of the coin, natural sciences learning from social sciences is also repetitively encouraged in philosophy, widely practiced in education, yet poorly evidenced in the actual production of science. As early as 1958, Harvard University initiated a fellowship program to provide rigorous training in social sciences to younger medical professionals and researchers (Medical and Social, 1958). As for college education, courses from social sciences can be found in the curriculum of many natural sciences disciplines with social implications, such as behavioral science and psychology in medical science, economics in engineering.

[1] This work is funded by the Flemish Government through its support to the Centre for R&D Monitoring (ECOOM), as part of the program on the study of interdisciplinarity and impact.

Besides a growing presence in education, many natural scientists also celebrated the preeminence of social sciences knowledge in the scientific research of human-influenced or human-dominated systems. Meloni (2014) argued "biology is becoming social" by illustrating how evolutionary biology, neuroscience, and molecular biology are adjusting their view take on altruism, brain, and gene, respectively, to a more social perspective. In synthetic biology, an emerging field that can contribute both rice producing beta-carotene as well as bioweapons, social scientists are invited to take a more vital role in research programs to provide discussions on ethical, legal, and social implications (ELSI), a required proponent in government-funded research from both Europe, UK and USA (Calvert & Martin, 2009). However, it is often expected for the social scientists to be 'contributors', who contribute ELSI *ex post facto*, or 'brokers' between scientists and the public, helping scientists navigate public controversy (Calvert & Martin, 2009).

Summarized from the above-discussed examples, is it possible to speculate that much natural sciences research mostly solicits assistance from social sciences after their science is made, or before as a general philosophical guide, but scantly during their research? That is to say, social sciences knowledge may be learned from, remembered, invited, but seldomly cited in the natural sciences literature. Can one trace, pinpoint and characterize the knowledge diffusion process from social sciences to natural sciences in the scientific literature? This constitutes the primary research question in this study.

We contribute a better understanding of this question by answering the following three sub-questions:

1. How many references, if any, are made from natural sciences to social sciences?
2. Which disciplines are the primary contributors?
3. What are the possible purposes or functions of such citations?

Related studies

Existing literature provides us with many theoretical underpinnings and methodologies to serve our research purpose. The first two sub-questions touch on the frequently discussed topic of interdisciplinary knowledge diffusion between disciplines. Such analyses are often conducted by quantifying the number/percentage of citations that certain clusters of papers make or receive from other clusters. Indicators such as the percentage of citations outside categories (Porter & Chubin, 1985) or toward categories (Broadus, 1952) have been proposed to contextually describe the dynamics of knowledge diffusion or the interdisciplinary status of a certain discipline. Zhou et al. (forthcoming) further develop these indicators by introducing a multidimensional framework in which one-dimensional citation flow is decomposed into two conceptually orthogonal dimensions, namely broadness and intensity. Nonetheless, the interdisciplinary citations from natural sciences to social sciences have not yet been analyzed through such methodology, mostly by case studies and general discussions mentioned above.

For the third sub-question, the functions of citations, especially in interdisciplinary contexts, deliver indispensable information on the motivation of knowledge solicitation endeavors. Recent advancements in comprehensive bibliographic databases and accessible full-text datasets make in-depth quantitative studies on citation function possible. For instance, citation sentences are analyzed semantically to unveil the sentiment polarity – e.g. weak, contrast, positive, neutral – by Teufel et al., (2006), function – e.g. criticizing, comparison, use, substantiating, basis, and neutral – by Jha et al., (2017), and influence – e.g. perfunctory, significant positive, and significant negative – by Hernández-Alvarez et al., (2017). Furthermore, the section structure of papers (IMRaD: Introduction, Method, Result, and

Discussion) can also be used to infer the function of citations, since references in different sections typically have a different intellectual contribution to the citing article (Bertin & Atanassova, 2014; Tahamtan & Bornmann, 2019).

This paper aims to integrate multi-source data, i.e. reference, reference section, and reference sentence, to deliver a nuanced understanding of the focal research question: how does natural sciences learn from social sciences in literature? The rest of this article is organized as follows: we first introduce the data and methodology employed and then present empirical results and discussions. The last section concludes.

Data and Methodology

To construct a dataset with full-text publications from natural sciences categorized by necessary information, we obtained 230,585 publications records and their full-text data from PubMed Central as the subject of this study. The dataset contains articles published from 2005 to 2018 in seven journals, namely, *PLoS Biology, PLoS Computational Biology, PLoS Genetics, PLoS Medicine, PLoS Neglected Tropical Diseases, PLoS One, and PLoS Pathogens*. From the retrieved full-text data (in XML format), we retrieved three types of information:

- Bibliographic information such as title, journal, publication year, and DOI;
- Bibliographic information for the references listed in the article, for instance, title, author, journal, sequence, etc..;
- Section structures (titles) and corresponding references (sequence in the reference list) mentioned within.

Since the first dataset does not provide viable classifications of the field of origin for publications and their references, a second dataset is harvested from Microsoft Academic Graph (MAG) for articles with the DOIs from the first dataset. We retrieved focal papers' field of origin, references, and the bibliographic information and field of origin of references. The employment of MAG is to serve two goals. Firstly, since PLoS One is a journal that is multidisciplinary instead of natural sciences focused, we could use the field of origin to filter out all publications that are classified into any social sciences and humanities fields, yielding 214,651 publications (93.4% of the original dataset). The second goal is to classify references to their field of origin. To achieve this, for each publication, we compare the bibliographic information, i.e. title, author name, journal, and year, for references from both data sources and find matches. In total, 10,045,358 reference pairs (95.7% of the total) are successfully recognized and classified. We adopt the second level of field classification systems in MAG which recognizes 294 sub-fields, among which 60 sub-fields are selected and grouped to represent nine social sciences disciplines following the approach from our previous study (Zhou et al., 2022).

To be able to pinpoint the location of social sciences references for all articles, one must have a unified section structure such as IMRaD (Introduction, Methodology, Results and Discussion) to make comparisons possible. We, therefore, mapped section names into six unified sections, namely, **I**ntroduction, **M**ethodology, **R**esults, **D**iscussion, **S**upplementary **I**nstruction (SI), and **O**thers. The mapping is conducted in three steps.

Firstly, for each unprocessed section name, we search for keywords listed in Table 1 and map them to according unified sections. Secondly, if one unprocessed section name is mapped with multiple unified sections, we perform the following two tasks: priorities are given to 'SI' and Results since the search keywords are more specific for them; next, we compared the current mapped section names with the last adjacent section and remove what has already occurred in

the previous section. Thirdly, the unmatched cases are handled. If it occurs at the beginning of the document, we classify it as Introduction; otherwise, it is classified as the same mapped section as its previous section.

Table 1. Keywords and mapped unifying section

Mapped section	Keywords
INTRODUCTION	introduction, background, motivation
METHODOLOGY	method(s), material(s), methodology, dataset, database
RESULTS	result(s)
DISCUSSION	discussion(s), conclusion(s), limitation(s), summary
SI (Supplementary Instruction)	supporting, supplementary, appendix, competing interests, data availability, corresponding author
OTHERS	correction(s)

To address the three sub-questions raised in the previous section, three aspects of interdisciplinary citation patterns from social sciences to natural sciences are analyzed respectively, namely, prevalence, origin, and location.

The *prevalence* of knowledge flow describes the influence of social sciences knowledge on natural science, which is quantified by three indicators in this study. Firstly, the reference percentage is defined as the percentage of social sciences references in all the references cited by our natural sciences publication set. It answers the question "how much weight does social sciences account for in the entire knowledge base of natural science". In addition, the broadness quantifies the percentage of publications in our defined publication set that cite at least one reference from the social sciences, which simplifies the breadth of influence of knowledge. The third indicator, reference intensity, captures the percentage of social sciences references relative to all references made from a subset of the publications that cite social sciences at least once. It signifies the influence intensity of knowledge. We show in our previous studies that the latter two variables capture two distinct aspects of the first variable; detailed mathematical definitions can be found in Zhou et al. (forthcoming).

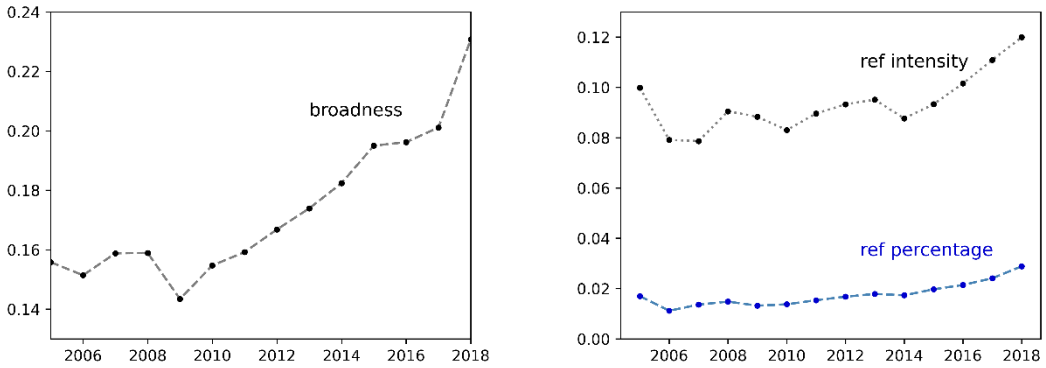
We then move on to trace the concrete disciplinary *origin* of knowledge borrowed from social sciences by looking into the number of references to nine social sciences disciplines[2] over time. The analysis concludes by presenting the sectional distribution (*location*) of references to nine social sciences disciplines in our publication set. References made in the Methodology section are discussed specifically and compared with the overall distribution of all references to infer disparities among the functional role of social sciences disciplines.

[2] See Zhou et al. (2022) for a detailed categorization on nine social science fields and subfields included.

Results and discussion

The prevalence of knowledge from social sciences in natural science

Figure 2. The prevalence of social sciences knowledge in natural science.



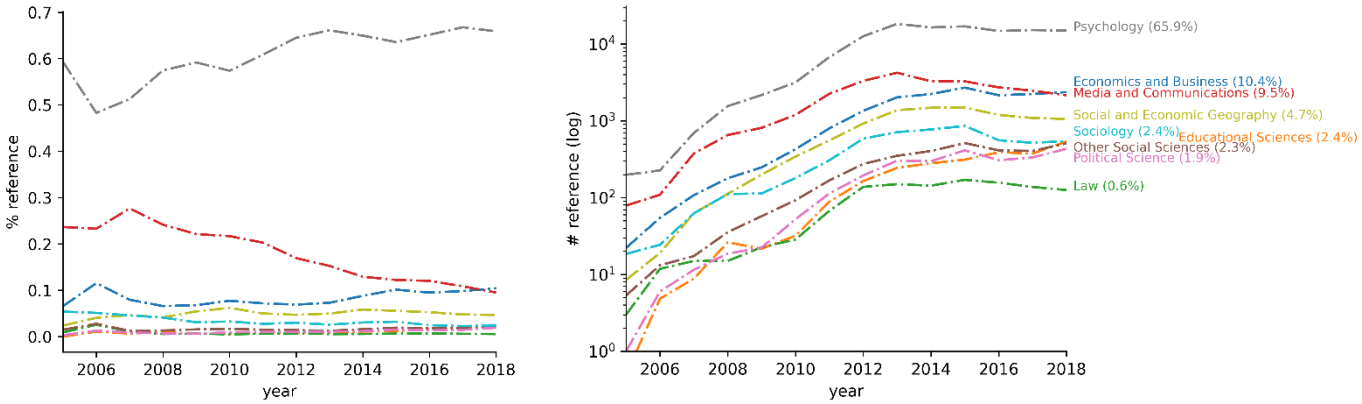
(a) **(b)**

Figure 2 (a) shows the broadness of influence of social sciences knowledge on natural sciences and its temporal evolution. An increasing percentage of publications from natural sciences are employing social sciences knowledge to support their research purpose, rising from 15.6% in 2005 to 23.1% in 2018; that is to say, almost one-fourth of the natural sciences publications in our dataset utilized knowledge from social sciences in 2018. Such a widening interest is also accompanied by more intense usage and a greater representation of the knowledge base by social sciences. In 2018, social sciences accounted for 2.9% of all references' fields of origin, increasing from 1.7% in 2005. For those who actually cite social sciences, the reference intensity goes from 9.9% to 11.9% during 13 years. Therefore, the rising prevalence of knowledge from social sciences in natural sciences presents not only relatively broader interest but also more intense citation habits.

The disciplinary origin

Where does the observed rising prevalence come from, i.e. which social sciences disciplines play a more important role in the knowledge base of natural science? The second analysis answers these questions by looking into the distributions of references from nine social sciences disciplines and their evolution, as shown in Figure 3. Psychology contributes the largest amount of references throughout the entire studied period, with a growing proportion among all social sciences references from 48.3% in 2006 to 65.9% in 2018. Economics & Business and Media & Communications constitute the second source of knowledge, occupying another 20% of references together in 2018. However, the two disciplines exhibit different evolution patterns, as Economics & Business is increasing in percentages yet Media & Communications decreased from 23.7% in 2005 to 9.5% in 2018. Social & Economic Geography contributes 4.7% in 2018. The rest of the disciplines make less than 3% of references each.

Figure 3. Disciplinary origin of social sciences knowledge in natural science.



(a) **(b)**

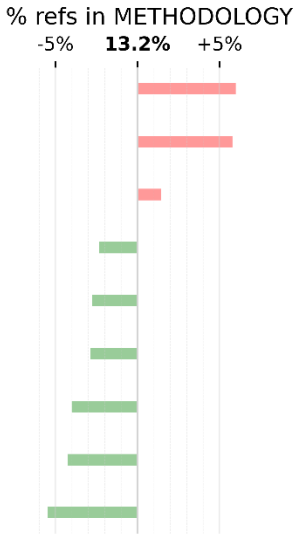
The section location

The section location of references is indicative of the purpose of knowledge borrowing in the design of research by authors. For all the references we recognized in our dataset, most reside in the Introduction and Discussion sections, with 37.9% and 36.7%, respectively. 13.2% of references can be found in the Methodology section, whereas the Results section holds the least (12.2%). Here we delve into the distribution of references in sections to different social sciences disciplines to uncover if they deliver different functions in their relationships with natural science. The left panel of Table 2 presents the section location of all social sciences references. In the right panel, we specifically analysed the social sciences references in the Methodology section, compared with all references regardless of disciplines as the baseline. Disciplines with a higher proportion of references in the Methodology section compared to the baseline are denoted in red color and the height of bars represents the difference from the baseline (13.16%).

Similar to the section structure of all references, all social sciences disciplines are cited most heavily in the Introduction section, however with a significantly higher percentage. Discussion, with a similar functional role as the Introduction, varies in proportions for different social sciences disciplines. The most prominent discrepancy can be observed in the Results and Methodology section. . The presence of social sciences disciplines in references from the Results section is significantly lower than average, except for law almost reaching the average level. On the other hand, a diverging pattern is found for reference patterns in Methodology, having three disciplines with greater presence and the rest tan average. Economics & Business contributes the highest percentage(21.5%) of knowledge to methodology in natural sciences publications followed by Other Social Science fot 19.00% and Sociology for 14.65%. The other six disciplines all show les presence in Methodology this is most notably the case for Educational Sciences (7.73%).

Table 2. Sectional location (%) of social sciences knowledge in natural science.

	INTR	METH	RESU	DISC
Economics and Business	42.51	21.50	6.95	29.01
Other Social Science	45.41	19.00	5.76	29.84
Sociology	41.64	14.65	4.92	38.79
Law	45.16	10.85	10.30	33.69
Political Science	46.86	10.42	5.72	36.99
Psychology	43.12	10.34	4.86	41.67
Social and Economic Geography	56.02	9.19	3.04	31.74
Media and Communications	41.78	8.95	8.09	41.17
Educational Sciences	47.31	7.73	4.36	40.60



Conclusion

This paper examines how social sciences knowledge affects natural sciences by empirically investigating the prevalence, origin, and location of academic references to social sciences from a set of natural sciences publications. Results show that social sciences knowledge accounts for 2.9% of the knowledge base for natural sciences and influences 23.1% of publications in 2018. Psychology, Economics & Business, and Media & Communications contribute the majority of knowledge. Most social sciences disciplines are primarily cited in the Introduction and Discussion section, except for Economics & Business, Other social sciences, and Sociology exhibiting a higher presence in Methodology. The discovered growing presence and various functions of social sciences knowledge in natural sciences deliver several political implications for education, research evaluation, and science policy. For instance, college educators should provide students from natural sciences with more necessary trainings in relevant social sciences disciplines to help them navigate the cognitive and practical discrepancies between the two, such as epistemology and methodology. Publishers from natural sciences should be prepared to expand the reviewer pool with social scientists when necessary to avoid defective quality control on the utilization of social sciences knowledge in submissions. Funding agencies need to better understand the role and significance of social sciences knowledge in the practice of natural sciences research and provide ample support to encourage its dissemination and application.

The study has several limitations. The dataset we employ mostly covers publications from life sciences, omitting some other natural sciences disciplines such as physics. Additionally, the context of learned social sciences knowledge is not thoroughly investigated. In the future study, we will improve this study by expanding datasets and conducting topic modelling or keyword extractions to deliver nuanced and contextual analyses.

References

Beck, L. W. (1949). The “Natural Science Ideal” in the Social Sciences. *The Scientific Monthly*, *68*(6), 386–394.

Bertin, M., & Atanassova, I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. *BIR@ECIR*.

Broadus, R. N. (1952). An Analysis of Literature Cited in the American Sociological Review. *American Sociological Review*, *17*(3), 355–357. https://doi.org/10.2307/2088083

Calvert, J., & Martin, P. (2009). The role of social scientists in synthetic biology. *EMBO Reports*, *10*(3), 201–204.

Hernández-Alvarez, M., Soriano, J. M. G., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, *23*(4), 561–588. https://doi.org/10.1017/S1351324916000346

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

Medical and Social. (1958). Medicine and Social Science. Editorial. *New England Journal of Medicine*, *258*(8), 400–401. https://doi.org/10.1056/NEJM195802202580813

Meloni, M. (2014). How Biology Became Social, and What it Means for Social Theory. *The Sociological Review*, *62*(3), 593–614. https://doi.org/10.1111/1467-954X.12151

Porter, A. L., & Chubin, D. E. (1985). An indicator of cross-disciplinary research. *Scientometrics*, *8*(3), 161–176. https://doi.org/10.1007/BF02016934

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103–110.

Zhou, H., Guns, R., & Engels, T. C. E. (2022). Are social sciences becoming more interdisciplinary? Evidence from publications 1960–2014. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24627

Zhou, H., Guns, R., & Engels, T. C. E. (forthcoming). Indicating interdisciplinarity: A multidimensional framework to characterize interdisciplinary knowledge flow.