Technological Forecasting & Social Change 172 (2021) 121071

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Rethinking the disruption index as a measure of scientific and technological advances

Xuanmin Ruan [a], Dongqing Lyu [a], Kaile Gong [b], Ying Cheng [a], Jiang Li [a,*]

[a] *School of Information Management, Nanjing University, Nanjing 210023, China*
[b] *School of Journalism and Communication, Nanjing Normal University, Nanjing 210097, China*

ARTICLE INFO

Keywords:
Disruption index
Number of references
Regression analysis

ABSTRACT

Wu et al. (2019) used the disruption(*D*) index to measure scientific and technological advances in Nature. Their findings spurred extensive discussion in academia on whether we can measure the disruption (i.e., innovation or novelty) of a research paper or a patent based on the number of citations. In this paper, we calculate the *D* index of ~0.76 million publications published between 1954 and 2013 in six disciplines including both sciences and social sciences in English and Chinese. We found that the number of references has a negative effect on the *D* index of a paper with a relatively small number of references, and a positive effect on the *D* index of a paper with a large number of references. We also found that low coverage of a citation database boosts *D* values. Specifically, low coverage of non-journal literature in the Web of Science (WOS) boosted *D* values in social sciences, and the exclusion of non-Chinese language literature in the Chinese Social Sciences Citation Index (CSSCI) resulted in the inflation of *D* values in Chinese language literature. Limitations of the *D* index observed in scientific papers also exist in technological patents. This paper sheds light on the use of citation-based measurements of scientific and technological advances and highlights the limitations of this index.

1. Introduction

It is widely acknowledged that citations are one of the most crucial, simple, standard, and objective indicators for measuring scientific impact (Didegah and Thelwall, 2013; Yan et al., 2012). However, many scholars have acknowledged the abuse of citations for several reasons. The first reason is that writers' motivation for citing is sometimes negative. Second, citation counts are influenced by numerous external factors that are not related to the quality of the paper (Onodera and Yoshikane, 2015; Tahamtan et al., 2016). Another criticism is that the citation count is a one-dimensional measurement (Bu et al., 2021), whereas research should be evaluated from multiple dimensions (DORA, 2012). For example, citations can be used to measure research innovation, novelty, or disruption (Foster et al., 2015; Funk and Owen-Smith, 2017; Uzzi et al., 2013; Wang et al., 2017), which goes beyond measuring scientific impact.

Wu et al. (2019) used the disruption (*D*) index, a citation-based indicator that originated from Funk and Owen-Smith's (2017) *CD* index, to measure science and technology advances in *Nature.* They found that "large teams develop and small teams disrupt science and technology," (p. 378) which spurred extensive discussion in academia. The basic idea is that "…when the papers that cite a given article also reference a substantial proportion of that article's references, then the article can be seen as consolidating its scientific domain. When the converse is true — that is, when future citations to the article do not also acknowledge the article's own intellectual forebears — the article can be seen as disrupting its domain" (Azoulay, 2019, p. 331). Subsequently, Bornmann and Tekles (2019a, 2019b) conducted several studies on the disruption index and found that the values of *D* depend on the citation window.

In terms of the calculation of the disruption index, the number of references of the focal paper are likely related to the disruption index because, intuitively, the more references the focal paper has, the harder it is to disrupt all of the references. There could be bias in the calculation of *D* if the focal paper has only one or two references. Therefore, (Bornmann et al., 2020a,b) argued that *D* should be calculated only for papers with at least ten citations and references. To the best of our knowledge, to date, no empirical studies have validated the effect of the number of references on *D*. Therefore, it is crucial to address this issue and identify the exact relationship between the two variables.

* Corresponding author.
E-mail address: lijiang@nju.edu.cn (J. Li).

https://doi.org/10.1016/j.techfore.2021.121071
Received 16 May 2020; Received in revised form 26 June 2021; Accepted 22 July 2021
Available online 3 August 2021
0040-1625/© 2021 Elsevier Inc. All rights reserved.

2. Related works

The disruption index is a simplified form of the network-based *CD* index, which was introduced by Funk and Owen-Smith (2017). The CD index was designed to reflect the degrees of destabilization and consolidation of patents. The citation network of a focal patent includes three types of patents: the focal patent, its reference patents, and subsequent patents. The intuition behind the *CD* index is that consolidating patents should be cited together with parts of their reference patents by subsequent patents. As a result, it increases the citations of the reference patents. A patent has the highest consolidation when all of the subsequent patents cite it and parts of the reference patents simultaneously. By contrast, a destabilizing patent decreases the citations of its references. The most destabilizing patents are ones that have no co-cited subsequent patents with its predecessors. The calculation of the *CD* index is shown in Eqs. (1)–(3).

$$CD_t = \frac{1}{n_t}\sum_{i=1}^{n}\frac{-2f_{it}b_{it}+f_{it}}{w_{it}},\ w_{it}>0, \tag{1}$$

$$f_{it} = \begin{cases} 1 \ \textit{if i cites the focal patent} \\ 0 \ \textit{otherwise} \end{cases}, \tag{2}$$

and

$$b_{it} = \begin{cases} 1 \ \textit{if i cites predecessors of the focal patent} \\ 0 \ \textit{otherwise} \end{cases}, \tag{3}$$

where n_t is the number of citations in i, and w_{it} is the weight for patent i at time t.

Wu et al. (2019) simplified the *CD* index and used it to evaluate the disruptiveness of academic papers (Eq. (4)). They validated the disruption index in several ways including analyzing the disruption values of Nobel-prize-winning papers, and comparing disruption of reviews and their original articles.

$$D = \frac{n_i - n_j}{n_i + n_j + n_k}, \tag{4}$$

where n_i is the number of papers citing the focal paper only, n_j is the number of papers citing both the focal paper and parts of its references, and n_k is the number of papers citing references of the focal paper only.

Following the work of Wu et al. (2019), Bornmann and his team continued work on the disruption index. Bornmann and Tekles (2019a) calculated the disruption index of four example papers and found that the index depends on the citation window. They determined that at least a three-year citation window is needed to produce meaningful results. Bornmann and Tekles (2019b) further calculated the disruption values of papers published in *Scientometrics* and showed that the values of their disruption are concentrated around zero with the highest value $D =$ 0.13, and a paper ranks among the top 1% if the D value is higher than 0.027. They also speculated that the number of references of a paper potentially influences the D value. Later, Bornmann et al.(2020a) examined the convergent validity of the disruption index and the four variants with assessments by peers. The results showed that the newly proposed index, DI$_5$, performed better than the disruption index and other variants. In another work, they proposed a field-specific version of DI$_5$ and used both the modified DI$_5$ index and the original disruption index to evaluate the disruptiveness of papers published in *Scientometrics* (Bornmann et al., 2020b). They found that the DI$_5$ was more efficient in identifying landmark papers in the field of *Scientometrics*.

3. Data and methodology

3.1. Dataset

According to Clarivate Analytics (Martín-Martín et al., 2018), English articles dominate the Web of Science (WOS). Based on the disciplinary classification scheme introduced by Puuska et al. (2014), we randomly selected five WOS disciplines—Mathematics, Applied (Mathematics); Management; Plant Sciences; Neurosciences; and Engineering, Electrical and Electronics (Engineering)—in the disciplinary groups of natural sciences, social sciences, agriculture and forestry, medicine and health sciences, and engineering, respectively. The first dataset included 691,647 papers that were (1) indexed as article papers, (2) published in journals in the Journal Citation Report 2018 edition, (3) published from 1954 to 2011, and (4) cited at least ten times in the subsequent five years after publication. Specifically, there are 151,788 papers in Engineering, 32,206 papers in Management, 41,815 papers in Mathematics, 338,155 papers in Neurosciences, and 127,683 papers in Plant Sciences.

The Chinese Social Sciences Citation Index database (CSSCI) is an essential Chinese citation index database including 28 disciplines of social sciences and humanities. The second dataset included 72,872 Chinese papers indexed in CSSCI that were (1) indexed as article papers, (2) published in journals included in the *Annual Report for Chinese Academic Journal Impact Factors (Social Science 2020)*, (3) published from 2000 to 13, and (4) cited at least five times in the subsequent five years after publication. We lowered the standard of citations and included papers with at least five citations because the CSSCI database contains far fewer citation records than WOS.

3.2. Dependent, independent, and control variables

Since Bornmann and Tekles (2019a) argued that the "disruption index depends on length of citation window" (p. 1), we used a fixed five-year citation window to calculate D in Eq. (4) to eliminate the influence of the length of the citation window on D. We used a five-year citation window to calculate the disruption index following the work of Funk and Owen-Smith (2017).

Wu et al. (2019) demonstrated that large teams develop while small teams disrupt science and technology. Thus, we chose the number of authors as a control variable to eliminate the effect on D. We used citation counts as another control variable because Bornmann et al. (2020a) argued that D is negatively correlated with citations at a medium level. According to Funk and Owen-Smith (2017), the disruption index of patents is correlated with the publication year of patents and their citations. Therefore, the publication year of papers was considered a control variable. In addition, D is a citation-based metric that could be affected by JIF. As a result, four variables were chosen as control variables: the number of authors (*author_num*), citation counts (*citations*), *JIF*, and the publication year (*year*), as shown in Table 1.

Table 1
Summary of the dependent, independent, and control variables.

Variables	Definition
Dependent variable	
Disruption index (*D*)	The disruption index of the focal paper
Independent variable	
Reference number (*ref_num*)	The number of references of the focal paper
Control variables	
Author number (*author_num*)	The number of authors in the focal paper
JIF	The JIF of the journal in which the focal paper was published
Publication year (*year*)	The number of years between 2020 and the publication year of the focal paper
Citation counts (*citations*)	Five-year citation counts of the focal paper

Note: The JIF for the first dataset was obtained from the Journal Citation Report 2018 edition; the JIF for the second dataset was obtained from the *Annual Report for Chinese Academic Journal Impact Factors (Social Science 2020)*.

3.3. Method

We first used line charts and histograms to illustrate the distribution of D and the rough relationships between *ref_num* and D in each field. We then divided the data into two groups: (1) the low *ref_num* group which included papers with nine references at most in the five WOS fields or four references in CSSCI; and (2) the high *ref_num* group including English papers with at least ten references or Chinese papers with at least five references. The following analyses were conducted in each group, separately.

The Pearson correlation analysis was used to explore the relationship between the dependent, independent, and control variables. Three variables, including *ref_num, author_num*, and *citations*, were highly skewed and had some outliers. To solve this problem, we deemed values that were not within the range of *mean±standard deviation* as outliers and removed them from the data. We also took a logarithmic form of the three variables to transform their distribution to a normal distribution. Papers with no references were removed from the analyses of correlation and regression given that their values of D always equal one.

A multiple stepwise linear regression model was performed to further examine the effect of the number of references on the disruption index. D is a continuous variable and the distribution approximates a normal distribution with large data; thus, stepwise linear regression is considered an appropriate method. We also took a logarithmic form of *ref_num, author_num*, and *citations* as independent or control variables.

4. Results

4.1. Exploratory analysis

The change in D along with ref_num. Fig. 1 shows the three periods of the relationship between D and *ref_num*. In the first period before *ref_num* reaches ten, D decreases rapidly with the growth of *ref_num* in each field. Then, D shows a slight rise in the five fields except for CSSCI. In CSSCI, the values of D still show a decreasing trend. In the third period where the values of *ref_num* are extremely high (e.g. *ref_num*=50 in Mathematics and *ref_num*=20 in CSSCI), D fluctuates and shows no obvious pattern. This is mainly because only a few papers have an extreme number of references, and thus, the average values of D are not steady. The effect of *ref_num* on D varies in the three different periods and it is not suitable to analyze them together. As a result, we divided the paper into two groups: the high *ref_num* group which includes papers that have at least ten references in the five WOS fields or at least five references in CSSCI, and the low *ref_num* group including the remaining papers.

Distribution of D. Figs. 2 and 3 present the histograms of D in the low and high *ref_num* group, respectively. As shown in Fig. 2, the values of D roughly approximate a normal distribution. In CSSCI, the distribution of D moderately leans to the right, which means that the disruptive papers are more common in CSSCI than in the five WOS fields. In the high *ref_num* group, the values of D are more concentrated around zero and also present a roughly normal distribution which leans to the left, indicating that the number of disruptive papers is less than the developing ones in the five WOS fields. In addition, the number of developing papers is greater than that in the low *ref_num* group across fields.

Scatters of ref_num and D. We drew scatter plots of D and *ref_num* before we conducted the correlation analysis. As shown in Fig. 4a–f, in the low *ref_num* group, the values of D display a roughly symmetrical distribution hovering around the horizontal line $D = 0$. In addition, as the *ref_num* increases from zero to nine (or four in CSSCI), the scatter plots of D gradually aggregate to the line $D = 0$. The six linear fitting lines in Fig. 4a–f shows that D negatively correlates with the number of references at different degrees. The absolute value of the negative correlation coefficient is the largest in CSSCI, as shown in Fig. 4, followed by Engineering and Management. Correlations between *ref_num* and D are weak in the other three fields. By comparison, the correlations between the *ref_num* and D are much weaker in the high *ref_num* group. Fig. 5a–f shows a slightly negative relationship between *ref_num* and D in CSSCI and no obvious relations in the other five fields.

4.2. Descriptive statistics and Pearson correlations

Descriptive statistics of variables. The descriptive statistics of the dependent variable, independent variable, and control variables in the low *ref_num* group are reported in Table 2. The maximum of D is 1 in each field and the minimum values range from -0.923 to -0.431. The mean of D is higher in Management and CSSCI (0.201 and 0.545, respectively) than that in the other four fields (less than 0.100). The values of *ref_num* range from zero to nine with means that vary from 4.619 to 6.603 in the five WOS fields. By comparison, *ref_num* is much smaller in CSSCI with a range of zero to four and a mean of 1.453. The variables *author_num* and *citations* show significantly skewed distributions and large variations in the maximum values across the fields. For example, one paper can have as many as 106 authors in Engineering while the maximum value of *author_num* is only 19 in Mathematics and eight in CSSCI.

Table 3 reports the descriptive statistics of variables in the high *ref_num* group. The ranges of values of D are more concentrated around zero compared with those in the low *ref_num* group. The mean of D in

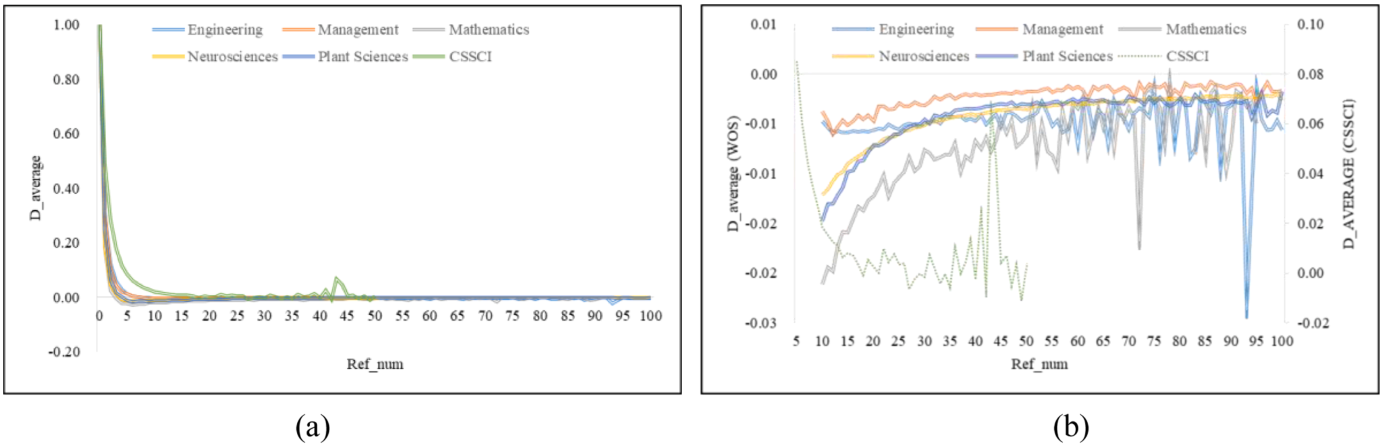


Fig. 1. The change of *D* with *ref_num*. (**a**) The change of D along with *ref_num* among papers with 100 references at most. The average values of D rapidly decrease when *ref_num* is small and then remain steady across fields. (**b**) the change of D along with *ref_num* among papers with at least ten references in the five WOS fields or five references at most in CSSCI. There is a slight rise and then a distinct fluctuation of D along with the increase of *ref_num* in the five fields except for CSSCI. The average value of D decreases dramatically in CSSCI and fluctuates when *ref_num* is greater than 25.

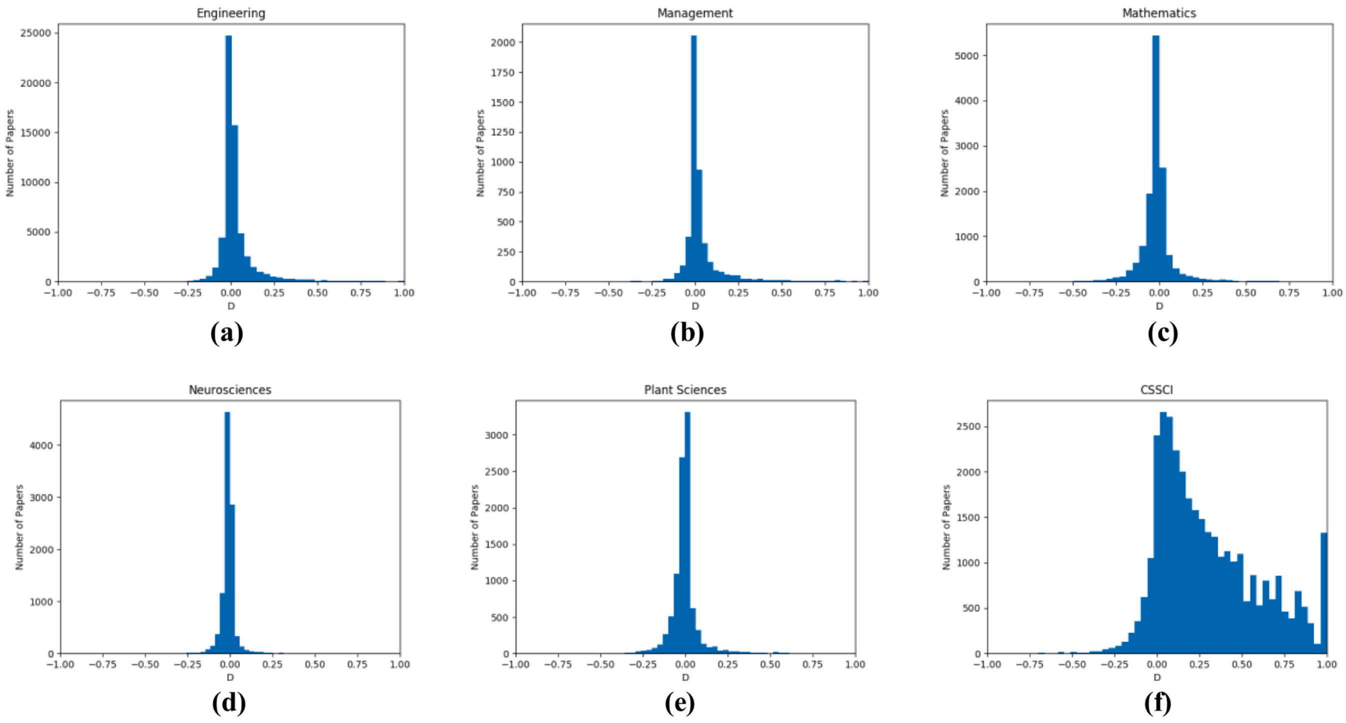


Fig. 2. The distribution of *D* in the low *ref_num* group. (**a–e**) The values of *D* in the five WOS fields roughly approximate a normal distribution, among which the majority of papers are located in the range of −0.10 to 0.10. (**f**) The values of *D* in CSSCI show a skewed distribution, which indicates that the imbalance between the proportion of disruptive papers ($D>0$) and developing papers ($D<0$) is greater than the five fields of WOS.

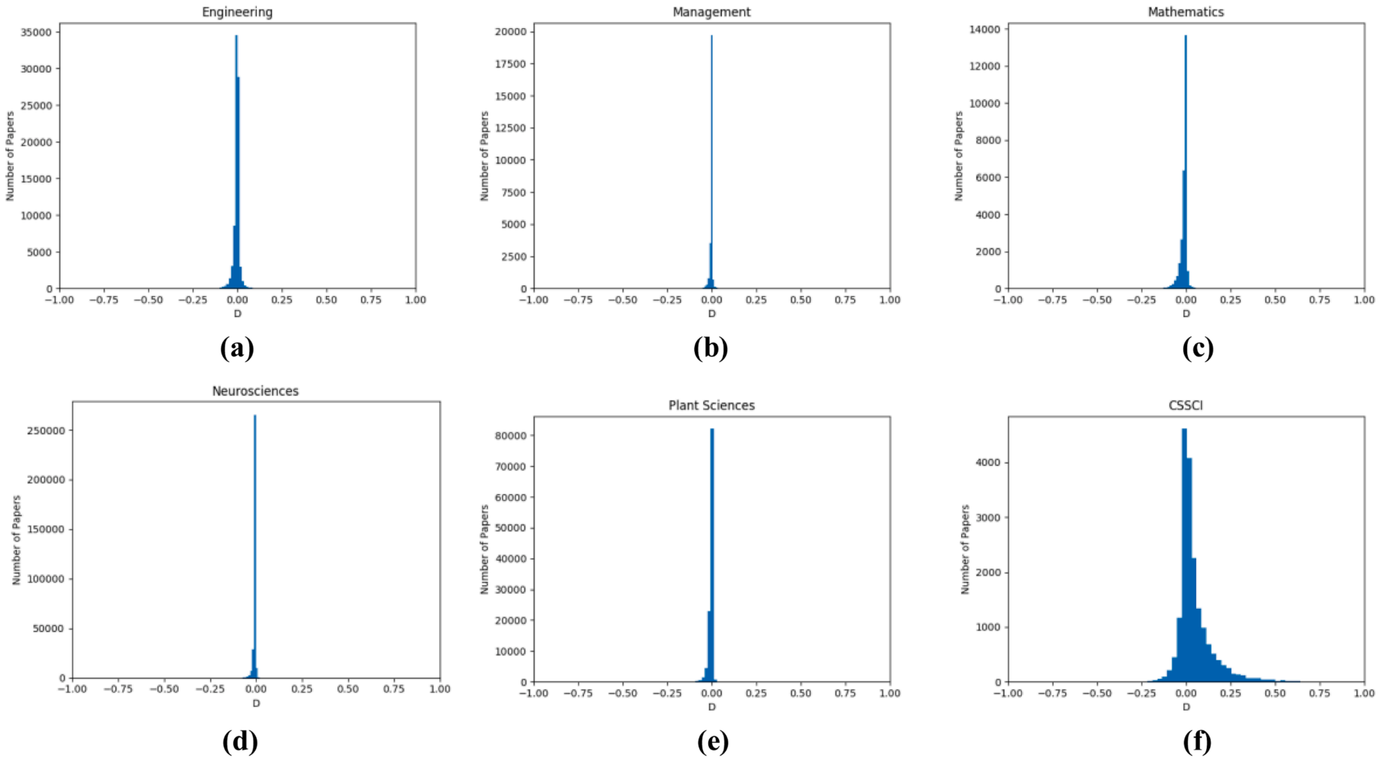
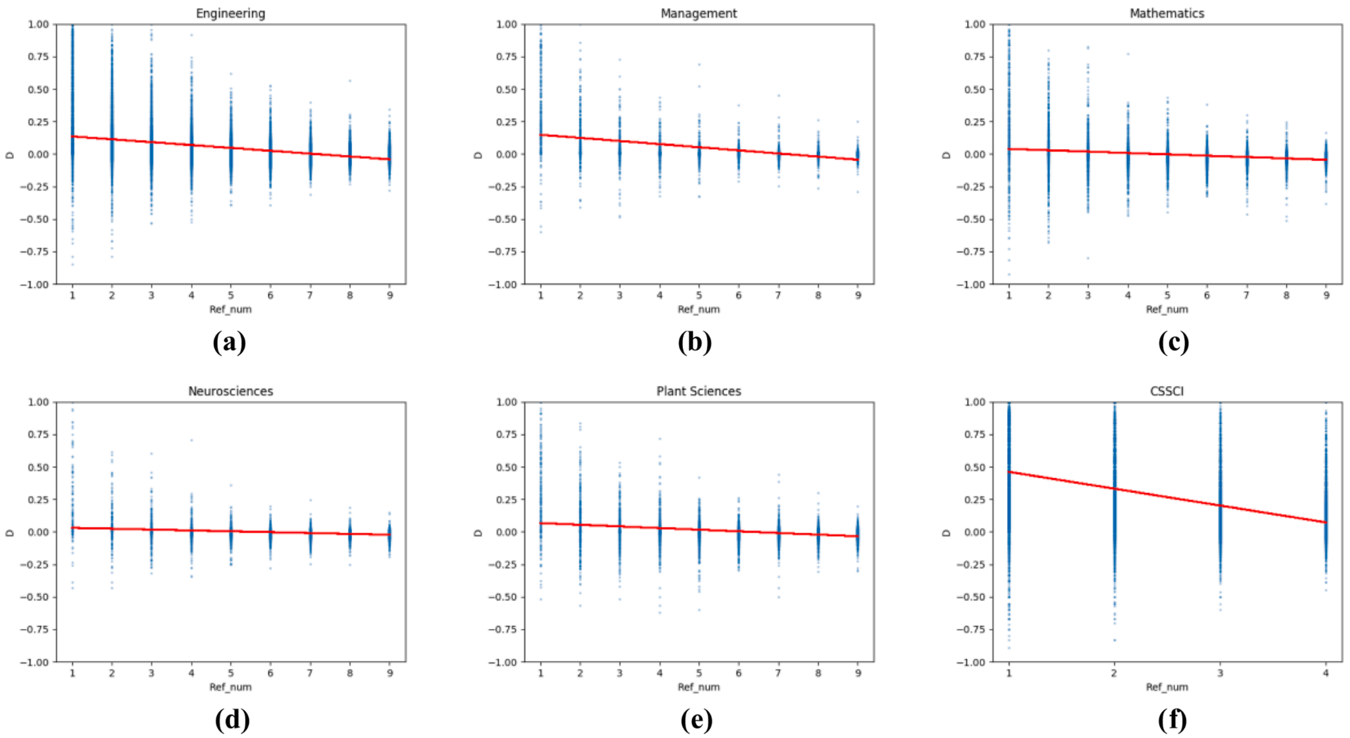


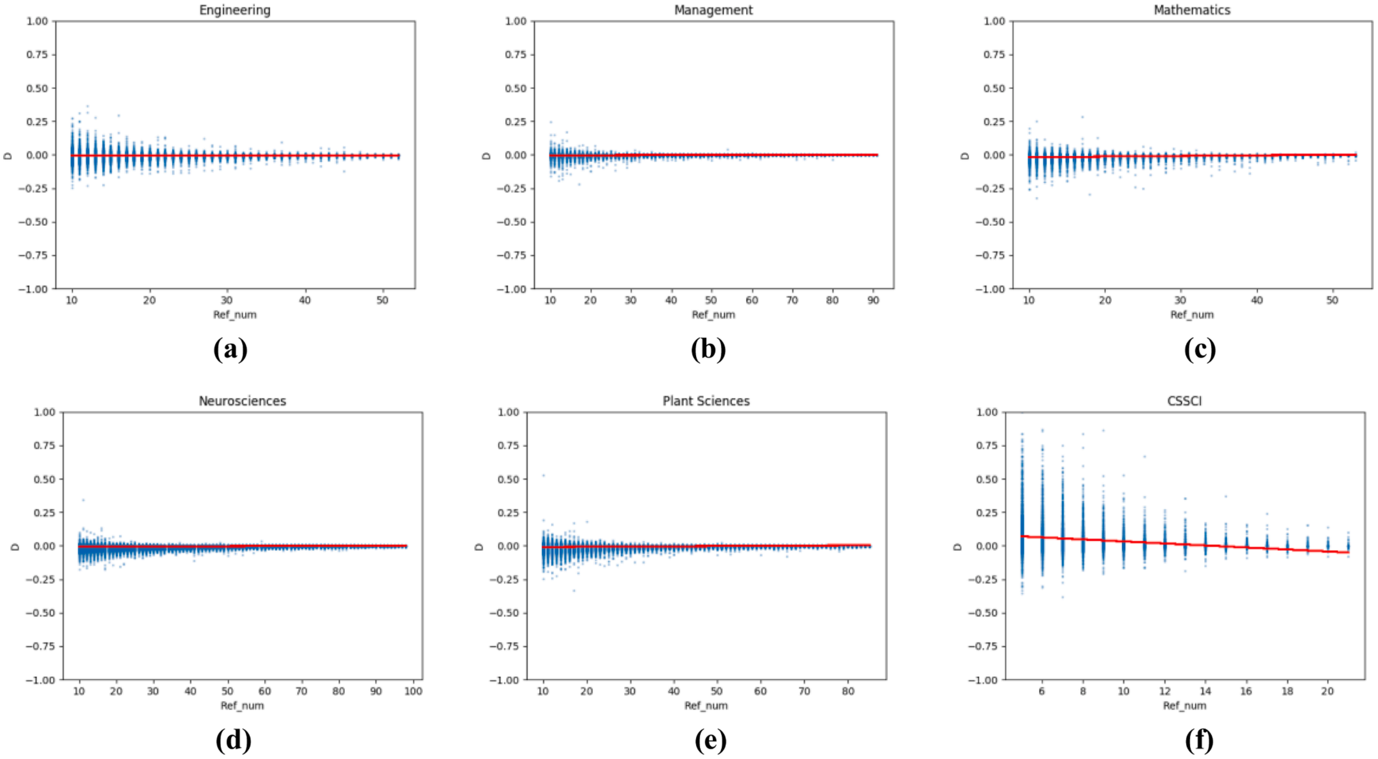
Fig. 3. The distribution of *D* in the high *ref_num* group. (**a–e**) The values of *D* in the five WOS fields roughly approximate a normal distribution, among which the majority of papers are located in the range of −0.05 to 0.00. (**f**) The values of *D* in CSSCI show a roughly normal distribution, and the majority of papers are located in the range of −0.2 to 0.2.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Scatters of *ref_num* and *D* in the low *ref_num* group. **(a–f)** Scatters of *ref_num* and *D* with their linear fitting curves in the low *ref_num* group the six fields, respectively. The linear fitting curves show a significantly negative relationship between *ref_num* and *D* in CSSCI, Engineering, and Management and a weak negative correlation in the other three fields.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Scatters of *ref_num* and *D* in the high *ref_num* group. **(a–f)** Scatters of *ref_num* and *D* with their linear fitting curves in the high *ref_num* group in the six fields, respectively. The linear fitting curves show a slightly negative relationship between *ref_num* and *D* in CSSCI, and no obvious relations in the other five fields.

Table 2
Pearson correlation analysis and descriptive statistics of the low *ref_num* group.

Variables	Correlation coefficient						Descriptive statistics			
	D	Ref_num	Author_num	JIF	Citations	Year	Minimum	Maximum	Mean	Standard deviation
Engineering										
D	1	−0.504***	0.020***	0.045***	0.058***	0.010*	−0.846	1.000	0.077	0.234
Ref_num		1	0.023***	0.009*	0.054***	−0.113***	0	9	5.388	2.575
Author_num			1	−0.173***	0.020***	−0.185***	1	106	3.737	2.613
JIF				1	0.176***	−0.075***	0.171	17.730	3.739	2.655
Citations					1	−0.113***	10	1085	22.474	26.196
Year						1	1960	2011	/	/
Management										
D	1	−0.520***	−0.032*	−0.016	0.040**	−0.003	−0.600	1.000	0.201	0.383
Ref_num		1	0.070***	0.060***	0.032*	−0.101***	0	9	4.619	3.104
Author_num			1	−0.054***	−0.002	−0.173***	1	50	2.097	1.288
JIF				1	0.107***	0.301***	0.214	10.632	3.935	1.874
Citations					1	−0.123***	10	743	19.197	17.931
Year						1	1954	2011	/	/
Mathematics										
D	1	−0.280***	−0.013	0.048***	0.007	0.113***	−0.923	1.000	0.011	0.190
Ref_num		1	0.088***	0.021*	0.054***	−0.199***	0	9	5.699	2.450
Author_num			1	0.032***	0.052***	−0.214***	1	19	2.092	0.986
JIF				1	0.092***	0.056***	0.149	7.224	1.848	0.945
Citations					1	−0.080***	10	659	17.701	14.719
Year						1	1954	2011	/	/
Neurosciences										
D	1	−0.303***	−0.021*	0.002	−0.031**	−0.067***	−0.431	1.000	0.022	0.180
Ref_num		1	0.102***	−0.005	0.050***	−0.043***	0	9	6.603	2.297
Author_num			1	0.002	0.062***	−0.372***	1	98	3.447	2.606
JIF				1	0.069***	0.076***	0.269	21.126	4.072	3.050
Citations					1	−0.004	10	588	21.193	21.193
Year						1	1954	2011	/	/
Plant Sciences										
D	1	−0.359***	0.021*	−0.062***	0.011	−0.041***	−0.619	1.000	0.067	0.271
Ref_num		1	0.073***	0.113***	0.060***	−0.061***	0	9	6.006	2.671
Author_num			1	−0.084***	0.035***	−0.468***	1	25	3.125	1.922
JIF				1	0.142***	0.331***	0.214	14.006	3.268	1.581
Citations					1	0.016	10	161	15.835	8.878
Year						1	1954	2011	/	/
CSSCI										
D	1	−0.488***	−0.099***	−0.036***	0.019***	0.088***	−0.889	1.000	0.545	0.411
Ref_num		1	0.093***	0.038***	0.049***	−0.162***	0	4	1.453	1.380
Author_num			1	0.044***	0.047***	−0.133***	1	8	1.761	0.947
JIF				1	0.290***	0.102***	0.034	10.787	2.722	2.141
Citations					1	0.002	5	382	9.175	8.472
Year						1	2000	2013	/	/

Note. The Pearson correlation analysis is based on the logarithmic form of *ref_num, author_num,* and *citations*. The descriptive statistics are based on the original values of variables.
*$p < .05$; **$p < .01$; ***$p < .001$.

CSSCI (0.047) is slightly higher than those in the other five fields (approximates zero). The values of *ref_num* show a significant variation across the six fields. On average, papers include more than 30 references in Management, Neurosciences, and Plant Sciences while those in Engineering and Mathematics have less than 20 references. Compared with the five WOS fields, papers in CSSCI only have eight references, on average. The variables *author_num* and *citations* have skewed distributions with a mean range from two to five and from 21 to 30 in the five WOS fields, respectively. In addition, papers are cited 11.933 times, on average, in CSSCI, which is much smaller than citations in the five WOS fields.

Pearson correlations of the dependent variable and the independent variable. Before we conducted the regression analysis, we examined the correlations between *D* and *ref_num*. As shown in Tables 2 and 3, *ref_num* is statistically significantly correlated with *D*. However, the signs are reversed for the two different groups. The statistical significance of coefficients is expected because of the large sample size. Thus, we interpret the correlation coefficients according to the recommendations by Cohen (1988), i.e., a small effect ($r = 0.1$), medium effect ($r = 0.3$), large effect ($r = 0.5$), and very large effect ($r = 0.7$). In the low *ref_num* group, *D* is negatively correlated with *ref_num* at a large effect level in Engineering (r=−0.504), Management (r=−0.520), a medium effect in Neurosciences (r=−0.303), Plant Sciences (r=−0.359), and CSSCI (r=−0.488), and a small effect level in Mathematics (r=−0.280). In contrast, the correlations between *D* and *ref_num* in the high *ref_num* group are positive at a small effect level except for Engineering ($r = 0.016$) and CSSCI (r=−0.256). A small and negative correlation is found between *D* and *ref_num* in CSSCI.

Pearson correlations of the dependent variable and control variables. In the high *ref_num* group, the *citations* of a paper are negatively associated with its disruption index at a small effect level in Mathematics (r=−0.154) and Neurosciences (r=−0.246), and a weak effect level in the other three WOS fields, which implies that less-cited papers tend to disrupt science more, while more-cited papers develop science more. In contrast, the correlation between *D* and *citations* is positive at a weak effect level in CSSCI ($r = 0.024$). Estimates in Table 3 indicate that there is a small and positive association between *author_num* and *D* in Neurosciences ($r = 0.121$) and Plant Sciences ($r = 0.148$), a weak and positive correlation in Engineering ($r = 0.029$), Management ($r = 0.024$) and Mathematics ($r = 0.081$), and a weak and negative correlation in CSSCI (r=−0.056). The small and negative relationship between *D* and *year* in the WOS fields indicates that papers published recently are generally more disruptive than those published earlier. The reverse is the case in CSSCI where *year* correlates positively with *D* at a weak level.

Table 3
Pearson correlation analysis and descriptive statistics of the high *ref_num* group.

Variables	Correlation analysis						Descriptive statistics			
	D	Ref_num	Author_num	JIF	Citations	Year	Minimum	Maximum	Mean	Standard deviation
Engineering										
D	1	0.016***	0.029***	0.056***	−0.064***	−0.150***	−0.639	0.365	−0.005	0.018
Ref_num		1	−0.009**	0.054***	0.105***	−0.073***	10	455	19.130	10.816
Author_num			1	−0.156***	0.000	−0.161***	1	73	3.606	2.306
JIF				1	0.210***	−0.070***	0.171	17.730	4.249	2.808
Citations					1	−0.087***	10	2075	27.557	36.912
Year						1	1960	2011	/	/
Management										
D	1	0.128***	0.024***	−0.029***	−0.072***	−0.257***	−0.218	0.246	−0.003	0.010
Ref_num		1	0.113***	0.125***	0.224***	−0.285***	10	321	32.792	19.245
Author_num			1	−0.046***	0.058***	−0.173***	1	143	2.394	1.495
JIF				1	0.220***	0.281***	0.214	12.289	4.403	2.167
Citations					1	−0.127***	10	616	26.528	24.361
Year						1	1954	2011	/	/
Mathematics										
D	1	0.198***	0.081***	0.125***	−0.154***	−0.066***	−0.324	0.281	−0.014	0.021
Ref_num		1	0.082***	0.099***	0.140***	−0.154***	10	523	19.953	10.849
Author_num			1	0.034***	0.058***	−0.172***	1	16	2.359	1.084
JIF				1	0.149***	0.047***	0.149	7.224	2.027	0.978
Citations					1	−0.067***	10	1254	21.499	23.383
Year						1	1954	2011	/	/
Neurosciences										
D	1	0.297***	0.121***	−0.026***	−0.246***	−0.286***	−0.178	0.343	−0.005	0.008
Ref_num		1	0.123***	0.093***	0.185***	−0.395***	10	639	38.546	19.872
Author_num			1	0.107***	0.098***	−0.396***	1	168	4.487	2.852
JIF				1	0.267***	−0.036***	0.113	33.162	4.556	3.211
Citations					1	−0.057***	10	2094	29.028	28.316
Year						1	1954	2011	/	/
Plant Sciences										
D	1	0.254***	0.148***	0.023***	−0.072***	−0.242***	−0.333	0.526	−0.006	0.012
Ref_num		1	0.230***	0.229***	0.309***	−0.451***	10	444	32.402	17.544
Author_num			1	0.063***	0.151***	−0.469***	1	140	4.220	2.628
JIF				1	0.348***	0.093***	0.200	14.006	4.114	1.863
Citations					1	−0.122***	10	1344	23.546	19.628
Year						1	1954	2011	/	/
CSSCI										
D	1	−0.256***	−0.056***	−0.034***	0.024**	0.097***	−0.385	1.000	0.047	0.101
Ref_num		1	0.039***	0.106***	0.115***	−0.184***	5	83	8.274	4.248
Author_num			1	0.061***	0.069***	−0.169***	1	6	2.057	0.998
JIF				1	0.378***	0.044***	0.044	10.787	3.309	2.362
Citations					1	−0.019**	5	325	11.933	13.169
Year						1	2000	2013	/	/

Note. The Pearson correlation analysis is based on the logarithmic form of *ref_num, author_num,* and *citations*. The descriptive statistics are based on the original values of variables.
*p < .05; **p < .01; ***p < .001.

Table 2 also demonstrates that *D* correlates weakly with almost all of the control variables (r<0.100) with different signs in the low *ref_num* group.

Pearson correlations among the independent and control variables. Tables 2 and 3 report the Pearson correlation coefficients of each pair of independent and control variables in the low and high *ref_num* group, respectively. Most pairs of variables present a correlation at a weak or small effect level, and only a few variable pairs demonstrate medium-level correlations. For example, the correlation between *year* and *author_num* (r=−0.468) as well as *JIF (r* = 0.331) in Plant Sciences is at a medium level, as shown in Table 2.

4.3. Linear regression analysis

Table 4 reports the results of the stepwise regression models based on the low *ref_num* group (model 1) and the high *ref_num* group (model 2) in each field. Overall, the F test of each model shows that there is a statistically significant linear relationship between independent/control variables and the dependent variable (p<.001). The values of variance inflation factor (VIF) of the independent and control variables are far smaller than ten, indicating that there are no multicollinearity issues among the variables (Menard 2002). The values of Durbin-Watson are approximately 2 in all models, which indicates that there is no autocorrelation detected in the models.

Models of the low ref_num group. In the low *ref_num* group, the adjusted R^2 of the model is relatively high in Engineering (R^2=0.265), Management (R^2=0.277) and CSSCI (R^2=0.244), and it is smaller in Mathematics (R^2=0.085), Neurosciences (R^2=0.098) and Plant Sciences (R^2=0.134). *Ref_num* has a statistically significant and negative effect on *D* in all six fields. According to the absolute values of standardized coefficients, *ref_num* is the most critical factor among all of the selected variables. In addition, *ref_num* accounts for 7.8%−27.1% variation in *D* across fields while the changes of R^2 of the control variables are less than 7%, which provide additional evidence that the influence of *ref_num* on *D* is non-negligible and other variables contribute less to the model.

Models of the high ref_num group. Table 4 illustrates that the adjusted R^2 of model 2 is much less than that of model 1 in the five fields except for Neurosciences and Mathematics. In the five WOS fields, *ref_num* has a statistically significant and positive effect on *D*. Adding *ref_num* to the model increases less than 1% R^2 in Engineering and Management, which means that *ref_num* plays a less important role in the two fields. By comparison, the effect of *ref_num* is greater in the other four fields, and R^2 increases by at least 3.9% by adding it to the model. In addition, according to the change in R^2, variables like *year* in Engineering and

Table 4
Linear regression analysis in each field.

Variables	Model1 Coefficients_un	Coefficients_s	VIF	Change of R^2	Model2 Coefficients_un	Coefficients_s	VIF	Change of R^2
Engineering								
Ref_num	−0.115	−0.513***	1.015	0.254	0.001	0.012**	1.016	0.000
Author_num	0.007	0.031***	1.075	0.001	0.000	0.015***	1.058	0.000
JIF	0.002	0.040***	1.075	0.001	0.000	0.066***	1.081	0.004
Citations	0.019	0.074***	1.046	0.007	−0.003	−0.092***	1.061	0.006
Year	0.000	−0.031***	1.069	0.002	0.000	−0.150***	1.046	0.023
Constant	0.157	***	/	/	0.004	***	/	/
Adjusted R^2	0.265				0.033			
F test	4437.732***				562.862***			
Durbin-Watson	2.006				2.000			
Management								
Ref_num	−0.119	−0.530***	1.020	0.271	0.001	0.074***	1.175	0.006
Author_num	/	/	/	/	0.000	−0.021**	1.036	0.000
JIF	0.002	0.029*	1.136	0.001	0.000	0.069***	1.209	0.004
Citations	0.016	0.047***	1.039	0.003	−0.002	−0.138***	1.121	0.011
Year	−0.001	−0.059***	1.147	0.002	0.000	−0.276***	1.271	0.066
Constant	0.187	***	/	/	0.004	***	/	/
Adjusted R^2	0.277				0.088			
F test	454.699***				494.389***			
Durbin-Watson	2.005				2.005			
Mathematics								
Ref_num	−0.060	−0.271***	1.046	0.078	0.011	0.202***	1.053	0.039
Author_num	0.006	0.022***	1.053	0.000	0.003	0.065***	1.037	0.005
JIF	0.006	0.047***	1.015	0.003	0.003	0.136***	1.035	0.018
Citations	0.007	0.021*	1.018	0.000	−0.009	−0.210***	1.044	0.034
Year	0.001	0.062***	1.095	0.003	0.000	−0.044***	1.059	0.002
Constant	0.041	***	/	/	−0.025	***	/	/
Adjusted R^2	0.085				0.097			
F test	249.757***				584.599***			
Durbin-Watson	2.047				2.006			
Neurosciences								
Ref_num	−0.046	−0.304***	1.010	0.092	0.004	0.279***	1.230	0.088
Author_num	−0.002	−0.023*	1.170	0.000	0.001	0.046***	1.207	0.002
JIF	/	/	/	/	0.000	0.022***	1.087	0.000
Citations	/	/	/	/	−0.004	−0.318***	1.113	0.094
Year	0.000	−0.089***	1.160	0.007	0.000	−0.175***	1.391	0.032
Constant	0.093	***	/	/	−0.005	***	/	/
Adjusted R^2	0.098				0.216			
F test	367.793***				17,426.041***			
Durbin-Watson	1.961				2.004			
Plant Sciences								
Ref_num	−0.077	−0.366***	1.010	0.129	0.006	0.223***	1.443	0.064
Author_num	0.004	0.022*	1.286	0.000	0.001	0.056***	1.308	0.003
JIF	/	/	/	/	0.000	0.045***	1.235	0.002
Citations	0.010	0.033***	1.006	0.001	−0.004	−0.182***	1.224	0.025
Year	0.000	−0.054***	1.283	0.004	0.000	−0.141***	1.631	0.019
Constant	0.117	***	/	/	−0.011	***	/	/
Adjusted R^2	0.134				0.113			
F test	381.390***				2878.103***			
Durbin-Watson	1.968				2.002			
CSSCI								
Ref_num	−0.274	−0.484***	1.011	0.238	−0.072	−0.251***	1.055	0.066
Author_num	−0.035	−0.055***	1.011	0.003	−0.009	−0.041***	1.036	0.002
JIF	−0.004	−0.031***	1.094	0.001	−0.001	−0.033***	1.181	0.001
Citations	0.037	0.055***	1.095	0.002	0.013	0.069***	1.177	0.003
Year	/	/	/	/	0.002	0.047***	1.070	0.003
Constant	0.439	***	/	/	0.158	***	/	/
Adjusted R^2	0.244				0.074			
F test	2764.880***				291.485***			
Durbin-Watson	1.936				1.924			

Note. For the five fields in WOS, model 1 is based on papers that have less than 10 references, and model 2 is based on papers whose *ref_num* is greater than 9. For CSSCI, model 1 is for papers with less than 5 references and model 2 fits the data that have at least 5 references. *Coefficients_un* refers to unstandardized coefficients while *Coefficients_s* refer to standardized coefficients.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Management, *citations* in Mathematics, Neurosciences, and Plant Sciences also play a crucial role in the model.

5. Discussion and conclusions

5.1. The number of references of a paper has a non-negligible effect on its disruption

The disruption index, which was originally proposed by Funk and Owen-Smith (2017) to capture the disruptiveness of new inventions on technology streams, was used by Wu et al. (2019) to measure the disruptiveness of scientific publications, patents, and software products. In this study, we calculated the D of ~0.76 million publications published between 1954 and 2013 in six disciplines, covering both sciences and social sciences, as well as both English and Chinese publications. We found that the disruption index of a paper is significantly influenced by the number of references, especially among papers with a relatively small number of references. The number of references negatively affects its disruptiveness and accounts for 7.8%−27.1% variation of D in the low *ref_num* group. This finding is consistent with Funk and Owen-Smith (2017) indicating that patents are more consolidated when they have more references. If a focal paper cites more references, it may be easier for future work to cite the references and thus decrease the disruptiveness (Funk and Owen-Smith, 2017). By contrast, in the high *ref_num* group, *ref_num* has a slightly positive effect on D in the five WOS fields, but the effect is smaller than that in the low *ref_num* group. The effect of *ref_num* is even negligible in Engineering and Management. This finding supports the arguments of Bormann (2020a, 2020b) that the disruption index should be used only for papers with sufficient references.

5.2. Disruption values are higher in social sciences

We find that the two social sciences fields compared to sciences in our dataset are characterized differently based on their D values. As shown in Table 2, the average D values in Management and CSSCI are significantly higher in the low *ref_num* group, i.e., 0.201 in the former and 0.545 in the latter. By contrast, the values are less than 0.100 in the other four fields. In addition, the percentages of the papers with $D = 1$ are higher in the social sciences, i.e., 3.1% and 27.5% in Management and CSSCI, respectively, whereas the percentages are less than 2.0% in the other four fields.

The main reason for this difference is attributed to the lack of citations in non-journal publications in social sciences in WOS and CSSCI. Due to the nature of social sciences and sciences (including engineering), they are significantly different in teams of ontology epistemology and methodology. Xie (2015) noted that “research of sciences aims to discover the truth in ‘the world of being’ while the target of social sciences is to understand ‘the world of becoming’”. It is also widely acknowledged that non-journal literature plays a more critical role in scientific communication in social sciences than in sciences. For example, Lariviere et al. (2006) quantified the share of citations in both journal and non-journal literature by using citation data from the Science Citation Index (SCI), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index databases (AHCI) from 1981 to 2000. They found that less than 50% of the citations referred to journal articles in most social sciences and humanities fields, while the percentage was higher than 70% in most science and engineering fields.

Remarkably, the mean of D values in Management is roughly the same as that in the other four fields in the high *ref_num* group, which means that the lack of non-journal publications has little effect on papers with a large number of references. This finding seems reasonable because if a focal paper only cites a small number of references, the lack of non-journal publications might dramatically reduce the number of references and therefore influence its disruptiveness. We speculate that the lack of non-journal citations in WOS and CSSCI boosts D. In an extreme case where the focal paper (being cited at least once, i.e., $i{>}0$) exclusively cited books, it has $j = k = 0$. Hence, it has $D = 1$. Therefore, we should use D to measure publications in social sciences with caution, especially papers that do not cite a sufficient number of references, given that citations from non-journal literature are not indexed in citation databases.

5.3. Disruption values are higher in the Chinese language citation database

We found that the coverage of the citation database significantly affects the values of D, i.e., the low coverage of a citation database boosted D. English articles dominate WOS based on Clarivate Analytics (Martín-Martín et al., 2018). The distribution of D values of Chinese papers (i.e., CSSCI papers) is different from that of the five English-dominated fields in WOS. For example, the proportion of papers with $D = 1$ is significantly larger in Chinese literature, i.e., 27.5%, whereas the number is less than 4% in the other five English-dominated fields. Hence, we speculate that D is also higher in the publications in the Chinese language citation database. For another extreme example, a Chinese paper (being cited at least once, i.e., $i{>}0$) has $D = 1$ if it exclusively cited English Language literature, because CSSCI does not indicate who cited these English language references. Thus, $n_j{=}n_k{=}0$. Note that this limitation of the D index we observed in scientific papers also exists in technological patents.

5.4. Rethinking the disruption index

The disruption (D) of science and technology advances is ideally measured by a citation index that covers all scientific and technological literature based on the design of the disruption index. However, even the world’s largest academic search engine, Google Scholar, which contains roughly 389 million documents including articles, citations, and patents as of January 2018 (Gusenbauer, 2019), covers only a fraction of the literature. Therefore, the disruption index is unavoidably biased by using a citation database such as WOS or CSSCI. Furthermore, this study provides evidence that the low coverage of the citation database results in a boost in D, i.e., the low coverage of non-journal literature of WOS resulted in a boost in D values in social sciences, and the exclusion of non-Chinese language literature in CSSCI resulted in a boost in D values in Chinese language papers.

As a citation-based indicator, the D index has limitations that are similar to other citation-based indicators. For example, it depends on the length of the citation window (Bornmann and Tekles, 2019a), and it is not applicable to cross-disciplinary or cross-language comparisons. In addition, it could be influenced by dozens of publication-related factors (Xie et al., 2019; Tahamtan et al., 2016). For example, a paper’s D value is influenced by the number of references, as shown in this study.

There are at least two limitations in our study. One is that the Chinese papers we collected are limited to a social science. Therefore, whether the D index is applicable to sciences in Chinese language publications or to other non-English language publications requires further investigation. The D index in English publications is also limited to the five sampled disciplines. Future studies should investigate a broader range of disciplines and languages to derive more detailed insights in the relationship between *ref_num* and D, as well as the effect of the coverage of database on D.

CRediT authorship contribution statement

Xuanmin Ruan: Formal analysis, Data curation, Writing – original draft, Methodology. **Dongqing Lyu:** Investigation, Data curation, Writing – review & editing. **Kaile Gong:** Investigation, Data curation, Resources. **Ying Cheng:** Conceptualization, Writing – review & editing. **Jiang Li:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Acknowledgments

We acknowledge the use of the Web of Science data by Clarivate Analytics provided by the Indiana University Network Science Institute and the Cyberinfrastructure for Network Science Center at Indiana University. We also acknowledge the National Natural Science Foundation of China Grants 71874077 and 71673242 for financial support.

We thank Dr. Dongbo Shi and anonymous reviewers for constructive comments.

References

Azoulay, P., 2019. Small-team science is beautiful. Nature 566 (7744), 330–332. https://doi.org/10.1038/d41586-019-00350-3.

Bornmann, L., & Tekles, A. (2019a). Disruption index depends on length of citation window. Profesional De La Informacion, 28(2). doi: UNSP e28020710.3145/epi.2019.mar.07.

Bornmann, L., Tekles, A., 2019b. Disruptive papers published in scientometrics. Scientometrics 120 (1), 331–336. https://doi.org/10.1007/s11192-019-03113-z.

Bornmann, L., Devarakonda, S., Tekles, A., Chacko, G., 2020a. Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. Quant. Sci. Stud. 1 (3), 1242–1259.

Bornmann, L., Devarakonda, S., Tekles, A., Chacko, G, 2020b. Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). Scientometrics 123 (2), 1149–1155. https://doi.org/10.1007/s11192-020-03406-8.

Bu, Y., Waltman, L., Huang, Y., 2021. A multidimensional framework for characterizing the citation impact of scientific publications. Quant. Sci. Stud. 2 (1), 155–183. https://doi.org/10.1162/qss_a_00109.

Cohen, J., 1988. Statistical Power Analysis For the Behavioral Sciences, 2nd ed. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, USA.

Didegah, F., Thelwall, M., 2013. Which factors help authors produce the highest impact research? Collaboration, journal and document properties. J. Inform. 7 (4), 861–873. https://doi.org/10.1016/j.joi.2013.08.006.

DORA (2012). San Francisco Declaration on Research Assessment. https://sfdora.org/read/.

Foster, J.G., Rzhetsky, A., Evans, J.A., 2015. Tradition and innovation in scientists' research strategies. Am. Sociol. Rev. 80 (5), 875–908.

Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. Manag. Sci. 63 (3), 791–817. https://doi.org/10.1287/mnsc.2015.2366.

Gusenbauer, M., 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 118 (1), 177–214.

Lariviere, V., Archambault, E., Gingras, Y., Vignola-Gagne, E., 2006. The place of serials in referencing practices: comparing natural sciences and engineering with social sciences and humanities. J. Am. Soc. Inf. Sci. Technol. 57 (8), 997–1004. https://doi.org/10.1002/asi.20349.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., López-Cózar, E.D., 2018. Google scholar, web of science, and Scopus: a systematic comparison of citations in 252 subject categories. J. Inform. 12 (4), 1160–1177.

Menard, S.W., 2002. Applied Logistic Regression Analysis, 2nd ed. Sage, Thousand Oaks.

Onodera, N., Yoshikane, F., 2015. Factors affecting citation rates of research articles. J. Assoc. Inf. Sci. Technol. 66 (4), 739–764. https://doi.org/10.1002/asi.23209.

Puuska, H.M., Muhonen, R., Leino, Y., 2014. International and domestic co-publishing and their citation impact in different disciplines. Scientometrics 98 (2), 823–839. https://doi.org/10.1007/s11192-013-1181-7.

Tahamtan, I., Afshar, A.S., Ahamdzadeh, K., 2016. Factors affecting number of citations: a comprehensive review of the literature. Scientometrics 107 (3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2.

Uzzi, B., Mukherjee, S., Stringer, M., et al., 2013. Atypical combinations and scientific impact. Science 342 (6157), 468–472.

Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Res. Policy 46 (8), 1416–1436.

Wu, L.F., Wang, D.S., Evans, J.A., 2019. Large teams develop and small teams disrupt science and technology. Nature 566 (7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9.

Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To Better Stand On the Shoulder of Giants. Paper presented at the ACM.

Xie, J., Gong, K.L., Li, J., Ke, Q., Kang, H.C., Cheng, Y., 2019. A probe into 66 factors which are possibly associated with the number of citations an article received. Scientometrics 119 (3), 1429–1454. https://doi.org/10.1007/s11192-019-03094-z.

Xie Y. (2015, March 6). The relationship between social science and natural science. The Chinese Version of Scientific American. https://huanqiukexue.com/plus/view.php?aid=25178.

Xuanmin Ruan is a graduate student at the School of Information Management, Nanjing University. Her research interest is scientometrics.

Dongqing Lyu is a doctoral student at the School of Information Management, Nanjing University. Her research interest is scientometrics.

Kaile Gong is an assistant professor at the School of Journalism and Communication, Nanjing Normal University. His research interest is informetrics.

Ying Cheng is currently a professor at the School of Information Management, Nanjing University, Nanjing, China. His research interests include information retrieval and information behavior.

Jiang Li is currently a professor at the School of Information Management, Nanjing University, Nanjing, China. His research interests cover Scientometrics and science of science. He is in the editorial boards of both *Journal of Informetrics* and *Scientometrics*.