CESifo WORKING PAPERS

10217
2023

Original Version: January 2023
This Version: December 2023

Who Stands on the Shoulders of Chinese (Scientific) Giants? Evidence from Chemistry

Shumin Qiu, Claudia Steinwender, Pierre Azoulay

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10217*

Who Stands on the Shoulders of Chinese (Scientific) Giants? Evidence from Chemistry

Abstract

China's rise in science has the potential to push forward the knowledge frontier, but mere production of knowledge does not guarantee that others are able to build on it. We ask whether chemistry research originating from China offers broad shoulders for follow-on scientists to stand on. We show that even after carefully controlling for the quality of Chinese research, Chinese scientists' articles receive on average 28% fewer citations from US researchers, relative to scientists from other countries. Only Chinese researchers with unusually deep networks in the US can overcome, at least in part, the citation discount.

JEL-Codes: I230, O300, O350.

Keywords: research and development, economics of science, innovation, international spillovers.

Shumin Qiu
East China University of Science and Technology, School of Business
China – Shanghai 200237
qiushm@ecust.edu.cn

Claudia Steinwender
Ludwig Maximilian University
Department of Economics
Ludwigstr. 28 (Vgb.)
Germany – 80539 Munich
claudia.steinwender@econ.lmu.de

*Pierre Azoulay**
MIT, Sloan School of Management
100 Main Street – E62-487
USA – Cambridge, MA 02142
pazoulay@mit.edu

*corresponding author

December 31, 2023

1 Introduction

China has overtaken the United States to become the world's largest producer of scientific publications (Tollefson, 2018; Xie and Freeman, 2019). From the standpoint of its impact on the global economy, an important question is whether, beyond its undeniable quantitative importance, Chinese research contributes to pushing the world scientific frontier outward.

Recent empirical findings lend credence to the view that the quality of Chinese research has improved in concert with the number of articles emanating from Chinese research institutions. For instance, the incidence of Chinese addresses in world-leading journals such as *Science* and *Nature* has more than doubled between 2000 and 2016 (Xie and Freeman, 2020). The average number of citations per article, and China's overall share of citations has also risen markedly (Xie and Freeman, 2019).

These stylized facts notwithstanding, the extent to which Chinese scientific knowledge offers "broad shoulders" for follow-on researchers to stand on remains an open question. The last twenty years have seen a 2.5-fold increase in the number of Chinese academic scientists (PRC National Bureau of Statistics, various years), many of them working in relatively new, less research-intensive institutions. Because of this increase in scientific labor supply, the rising impact of Chinese research could merely reflect an elevated propensity on the part of Chinese researchers to cite research "made in China."

We focus our study on the domain of academic chemistry. While China has been a rising country across a broad cross-section of scientific domains, its status as a producer of frontier scientific knowledge has stood out in a narrower set of fields, chemistry preeminent among them (see Appendix A). China's share of world publications in *Web of Science* has grown from 5.33% in 2000 to 25.94% in 2018, with chemistry, engineering and materials science being the strongest contributors to growth, as can be seen in Figure 1. In some chemistry subfields, such as organic chemistry, China even surpassed the United States in recent years to become the world's top producer of publications.

Contrasting Chinese and non-Chinese (and non-US) researchers, we study the extent to which articles of similar observable quality are differentially likely to be cited by researchers based in the US. Our preferred specifications point to a "China citation discount" equal to 28% of the baseline probability of citation. This discount is halved for Chinese researchers who received some of their scientific training in the United States. We also find evidence that this discount is not a mere reflection of clustering of Chinese researchers in particular subfields that are less likely to be cited by US scientists. Nor is it likely to reflect ethnic animus, since we do not observe a similar discount for researchers with Chinese names located outside China. In addition, a similar discount appears present in US citations to the scientific literature contained in patents.

Among the top chemistry nations (by number of publications in our sample), no other country experiences a citation discount from the US; instead, Switzerland and Germany experience small citation premia. These results are notable because our choice of setting—elite scientists, in a domain where China has a long tradition of excellence—would seem to be one without strong impediments to the diffusion of knowledge across borders.

2 Description of Data Sources

Since our goal is to investigate how research undertaken in China disseminates compared to research undertaken in other countries, we focus on the publications of the world's best researchers, understood to be those with the highest rate of publications in a defined list of chemistry journals. We focus on elite chemists from all countries, excluding the United States.[1]

We compile a list of the 31 most impactful journals in the field of chemistry (see Table B.1 in the online appendix). We consider all original research articles in these journals published between 2000 and 2018, the period of China's rise in science. We drop articles produced by teams larger than 15 coauthors. This yields a total number of 552,933 articles.

Based on the author disambiguation work of Torvik and Smalheiser (2009) and its update to the 2018 version of *PubMed*, we are able to assign each article to unique authors. We focus on last authorship position, which indicates principal investigatorship according to the publication norms in the field of chemistry. From a set of 124,966 unique last authors, we select the top 1% in terms of the number of elite journal publications, and obtain a sample of 1,250 investigators.

Researcher-level data. We focus on investigators from all countries, excluding the United States, which leaves us with 751 investigators. From CVs, we extract information about demographics (birth year, gender), PhD education (university, country, completion year),

---

[1] As explained in more detail below, we will consider the US a large frontier country whose researchers are at risk of citing articles written by Chinese and non-Chinese scientists.

post-doctoral experience (organization and time period), as well as employment spells since post-doc (organization, country, and time period). We define the “year of independence” of each researcher as the year of their first faculty employment after post-doctoral education. We use the country that is associated with their most frequent affiliation on publications after career independence to assign each scientists to a unique country—this does not necessarily correspond to the nationality of the scientist. Out of the 751 scientists, 156 (20.78%) work in Chinese institutions. They are overwhelmingly male (96%), their average doctoral degree year is 1988, and the average number of post-doctoral years is 4. 11.32% of scientists hold a PhD degree from universities located in the United States, and 49.13% of them spent their post-doc years at institutions in the United States. More details are available in Appendix B.

Publication data. We compile the full publication list for all 751 scientists in our sample between the years 2000 and 2018. To ensure that we capture only research that was influenced to a significant extent by the scientist, we restrict the publications in two ways. First, we focus on publications which list the scientist as last author. Second, we consider only articles that were published after the PI became an “independent” researcher. Overall, our sample comprises 78,541 scientific articles in chemistry. On average, each scientist published 104.58 articles as last author in the time period we consider.

Citation data. We compile a list of citations to these publications from *Web of Science*. Since we want to link citations to countries, we remove citing articles lacking country information (4.2% of citations), which results in our database comprising 2,839,144 citation records from 2000 to 2021 for the 78,541 last-authored articles. Each article in our dataset received on average 36.18 cites. To uncover the causes of differences in cross-country citation behavior, we focus on the propensity of US researchers to cite articles that originate from China versus other countries. We single out the US, because it is undoubtedly a frontier country in chemistry research that attracts collaborations and trainees from the world at large. Furthermore, its large size implies that citation linkages between the US and other countries are frequent enough to make the statistical analysis tractable. In order to ensure that we can unambiguously interpret cross-border citation linkages, we restrict the sample of citing articles to those for which all authors are affiliated with a US institution. This yields 271,194 citations records for the 78,541 focal articles, belonging to 98,915 unique citing articles from the US.

Aggregate Evidence. To provide a first descriptive look at US citations of Chinese articles, we examine the full set of articles published in the field of chemistry between 2000 and 2018 and ask whether there is any difference in the number of citations Chinese articles receive from the US compared to articles written in other, non-US countries. The Poisson regression estimates of Table 1 imply that a Chinese article receives on average 48% fewer US citations compared to an article from other non-US countries (column 1). Because citations increase over time, and the rise of Chinese research has been more recent, we control in column 2 for publication year effects, which reduces the effect to 34% in absolute value. We add journal fixed effects in column 3, which reduces the citation discount further to 24%. Of course, there remains large variation *within journals* with respect to how much follow-on research articles can inspire. It may well be the case that once we properly control for the quality of the research, the discount vanishes.

3 Empirical Strategy

To detect whether the Chinese citation discount can be explained by researchers' and articles' observable characteristics, we begin by constructing a set of control articles that are comparable in quality with each publication by elite Chinese PIs. Second, for each treated or control article, we delineate a set of articles that are *at risk of citing* the elite researchers' publications. Finally, we model the probability that each potential citation *actually cites* a Chinese-authored articles, relative to articles authored by non-Chinese PIs.

3.1 Matching Chinese with non-Chinese articles

For each article by a Chinese PI, we try to find at least one comparable article authored by a non-Chinese, non-US PI. We implement a "Coarsened Exact Matching" (CEM) procedure (Blackwell et al., 2009). The first step is to select a relatively small set of covariates on which we need to guarantee balance *ex ante*. This choice entails judgement, but is strongly guided by our desire to hold the "fertility" of cited articles approximately equal across the treatment and control groups. The second step is to create a large number of strata to cover the entire support of the joint distribution of the covariates selected in the previous step. In a third step, each observation is allocated to a unique strata, and for each observation in the treated group, control observations are selected from the same strata.

Measuring article quality. The literature traditionally uses citations to capture variation in quality. But we cannot use citations from US scientists to do so, since these correspond to our outcome variable. Further, citations often exhibit strong 'home bias': articles are disproportionately cited by scientists from the same country. Moreover, this home bias appears especially pronounced in China. The citations that China receives from itself as a share of all citations it receives from the world is 56%, the largest of all countries (see Figure C.1a in the online appendix). This may be driven by politically motivated citations, which may be especially strong in power-oriented societies such as China (Jia et al., 2019). Regardless of its underlying cause, home bias potentially makes article citation counts a less valid proxy for article quality, specifically for articles originating from China relative to articles published by researchers located in other countries.

Our preferred proxy for article quality subtracts citations originating from the US from the raw number of citations, includes citations received from foreign sources (the "rest of the world"), and applies an adjustment to the citations received from home, accounting for the extent to which each country's home share of citations is "abnormally large" (using as a benchmark the country's relative size in terms of publications). Details regarding the computation of these "debiased citations" is available in the Appendix C.[2] Our matching procedure splits the number of "home-debiased" citations into six bins: 0-25th percentile; 25-50th percentile; 50-75th percentile; 75-95th percentile; 95-99th percentile; and the top percentile.

One may worry that citations from the rest of the world (ROW) are endogenous if they are themselves affected by citations from the US. For example, if a Chinese article is cited less by US scientists, this may also lead to fewer citations by scientists in other countries. However, this spillover effect would bias our estimates *against* finding a citation discount for Chinese articles. A similar rationale would lead us to underestimate (in magnitudes) a Chinese citation discount if citations from ROW were strategic. For example, if Chinese articles are less likely to be authored by journal editors, they may receive fewer citations from ROW compared to articles from other countries, if PIs are more likely to cite editors in expectation of favorable treatment in the peer review process.

---

[2] Note that our estimation results are not very sensitive to this adjustment. In Table C.1 of the online appendix we show that our results hold when we drop home citations completely; do not discount home citations at all (i.e., match on all citations outside US); or do not match on citations at all.

Additional matching variables. In addition to the citation measure, we match on journal, publication year, number of authors (4 groups: 1-3; 4-6; 7-9; 10 or more coauthors), and year of PhD receipt. The union of all matching criteria defines a strata. Within each strata, articles are indistinguishable from the perspective of the CEM algorithm, and the matching is performed at the level of the strata.[3] This procedure yields 6,905 treated articles written by 155 Chinese PIs, and 9,287 control articles written by 402 non-Chinese PIs. On average, there are 1.34 control articles per treated article. Table 2 compares the characteristics of control and treated articles. The first four rows display the variables used for matching. By construction, the two sets of articles were published in the same year. Due to coarsened matching, the remaining variables are not identical across treatment and control group, but differences in both the mean and median are small and statistically insignificant. Treated articles received around 18 debiased non-US citations, only slightly fewer than control articles.

In order to not shrink the size of the matched article sample further, we do not layer additional matching criteria; instead, our regression specifications will include relevant covariates not used for matching as controls.[4] The combination of matching and covariate inclusion results in fine-grained comparisons that plausibly hold "fertility" constant across Chinese and non-Chinese publications. More details about the matching procedure are available in Appendix D.

3.2 Definition of the risk set of citing articles

To test whether articles in the control or treatment group are cited differentially by US authors, we first need to determine which US articles are *at risk* of citing the articles in our sample. Moreover, since we would like to evaluate how social or geographic proximity shapes the propensity to cite, participation in the risk set should not mechanically reflect such factors. We deem an article eligible to be part of the citation risk set if it is *topically* related to the article in our sample.

In order to specify topical relationships between articles, we leverage the "Related Articles" function in *PubMed* to harvest journal articles that are intellectually proximate to

---

[3] As there may be different numbers of treated and control articles in different strata, CEM assigns a weight to each matched article to adjust for strata size, and we use this weight in all regression models.

[4] e.g., the cumulative number of articles published by each PI, as well as the number of citations each PI's entire corpus of work has received to date.

the articles in our sample.[5] This functionality is based on a topic-based content similarity model called *PubMed Related Citations Algorithm* or *PMRA* (Lin and Wilbur, 2007). This algorithm yields relatedness rankings and scores between any two articles based on the extent to which two articles are similar with respect to titles, abstracts, and keywords. For each article in our data, its citation risk set includes every PMRA neighbor whose authors work in US institutions and appeared after the focal article was published.

Importantly, the risk set does not include actual citations that are PMRA-unrelated. Of the 43,979 US articles actually citing the 16,192 articles in the matched sample, only 6,272 (14.27%) correspond to related records in the sense of PMRA (see Figure D.1 and the discussion in Appendix D for more details). By limiting the citation risk set to articles topically aligned with the cited articles in our sample, we exclude citations that do not reflect intellectual influence, but rather status considerations or attempts to curry favor with referees of editors (Teplitskiy et al., 2022). This makes it more likely that the effects we uncover correspond to actual "standing on shoulders" rather than mere jockeying for influence in the relevant scientific networks.

The combined risk set for the 16,192 articles in the matched sample comprises 188,753 citable/potentially citing article pairs, with each article having on average 11.7 potentially citing articles in its risk set.

3.3 Model specification

We model the probability that article i is cited by each article $j \in J_i$, the citation risk set for article i, as a function of the characteristics of article i and article pair ij, using the following linear probability model:

$$\mathbb{1}_{(j\,\text{cites}\,i)} = \beta_0 + \beta_1 China_i + \beta_2 X_{ij} + \beta_3 China_i \times X_{ij} + \varphi(i,j) + \varepsilon_{ij} \tag{1}$$

The dependent variable is an indicator variable that takes on value 1 if article j actually cites article i, and 0 otherwise. Our main regressor of interest, $China_i$, is an indicator variable for whether i's last author is located in China, whereas X is a vector of observable covariates, and $\varphi(i,j)$ corresponds to a large set of fixed effects for i and $i \times j$ characteristics. These include fixed effects for: (a) each strata emerging from the CEM procedure (the interaction of our bins for journal × publication year × number of authors × debiased non-US citations

---

[5] The articles in our sample (and most of the citations to these articles) appear in journals indexed by *PubMed*, an online resource from the National Library of Medicine that indexes more than 40,000 journals within the life sciences, including almost all the journals in which elite chemists routinely publish.

× PhD degree year; yielding 3,847 bins); (b) investigator cumulative publication bins (13 bins), (c) investigator cumulative citations bins (13 bins), (d) investigator gender, (e) topic similarity rank bins, (f) the interaction of i and j publication years (262 bins); (g) an indicator variable for the case when citing and cited articles were published in the same journal. We do not report coefficient estimates for these covariates, but they are always included. We cluster standard errors simultaneously at the level of the individual PI—to allow for arbitrary correlation of citation patterns across publications within each individual researcher—and the level of a strata—to allow for correlation of citation patterns across publications within a strata.

We include additional controls X_{ij}, and also interact these covariates with the China effect to explore whether the China citation discount is driven by particular channels. Below we describe how we constructed these covariates, with further details available in Appendix E.

Communication. A natural explanation for the China citation discount may be that there are language challenges restricting the communication between US and Chinese researchers and thus their awareness of published articles (even though we only consider articles in English journals); or that China is just far away which also reduces awareness of scientific output. We capture these potential channels by using two covariates: (a) an indicator variable for PI countries that list English among their official languages; and (b) the average geographic distance between the city of the PI and the city of the affiliations of the US citing authors.

Network. The dissemination of research may depend on PIs' access to potential citers via formal or informal networks. We include several covariates to capture the impact of this channel: (a) an indicator variable denoting whether PIs have obtained training (PhD or postdoc) in the US; (b) an indicator variable denoting whether the cited article's reprint or first author has a US affiliation; (c) an indicator variable for whether any middle author has a US affiliation;[6] (d) the log of the cumulative number of past U.S. coauthorships for each PI in the year prior the year of publication for the cited article; (e) an indicator variable that captures shared ethnicity between the PI and at least one author from article j; (e) an indicator variable for the presence of a past coauthor of the PI on article j's authorship roster; (f) an indicator variable for the presence of a common author on the authorship

---

[6] According to publication conventions in chemistry, first or reprint authors have contributed significantly to the research undertaken.

rosters of articles i and j; and (g) an indicator variable for PIs that have written editorials, which proxies for editors and other influential scientists, who may be cited more for strategic reasons.

Geographic topical clustering. To the extent that researchers in certain countries concentrate in different subfields, it is important for the analysis to control for these country-level specialization patterns.[7] We define the subfield of each source article as the set of its PMRA-neighbors, counting only the neighbors whose similarity score is above 0.5 and which appeared before the source article. Using these PMRA-derived subfields, we construct three subfield-level covariates: (i) the subfield's *home-research intensity* corresponds to the sum of the PMRA-relatedness scores for the articles in the subfield whose researchers are from the PI's country; (ii) the subfield's *foreign-research intensity* corresponds to the sum of the PMRA-relatedness scores for the articles in the subfield whose researchers are not from the PI's country and not from the US; and (iii) the subfield's *US-research intensity* corresponds to the sum of the PMRA-relatedness scores for the articles in the subfield whose researchers are from the US.

Investigator's intellectual focus. It is possible that investigators who concentrate their research in specific subfields receive higher recognition and thus more citations from the US, and researchers may be differently focused across countries. For this purpose we use the subfield definition based on PMRA and specify the three following measures: (i) *the subfield's importance for the investigator*; i.e., the number of articles for a given PI that belong to the subfield of the focal article divided by the total number of articles authored by the PI; (ii) *the investigator's importance for the subfield*, i.e., the number of articles of a given PI that belong to the focal article's subfield divided by the total number of articles in the subfield; (iii) *the investigator's research portfolio focus*, computed as an index to measure a PI's topical concentration across articles.

Reputation. US researchers may be hesitant to cite articles if they appear in subfields with a reputation for questionable ethical standards. We construct an indicator variable which denotes whether a subfield is "retraction heavy," i.e., whether there exists, among a cited article's PMRA neighbors, at least one article that has been retracted.

---

[7] The existence of such patterns is not mere speculation on our part. For instance, Borjas and Doran (2015) document the persistence of Russian influence in certain mathematical subfields even after the dissolution of the Soviet Union.

4 Empirical Results

Table 3 reports the estimation results. The analysis progresses from estimating the main effect of Chinese location, to attempts of making the citation discount disappear by including additional covariates, to exploring potential channels through which Chinese research may come to experience this discount.

Main effect of Chinese location. The specification in column 1 only includes the Chinese investigator indicator variable in addition to baseline controls. Consistent with the correlations in the aggregate data mentioned earlier, Chinese articles face a lower probability of being cited in US research. The discount is large in magnitude—equal to one fourth of the mean citation rate—as well as precisely estimated.[8]

Additional controls. The comparison between columns 2 and 1 reveals that while some of the control variables affect the likelihood of US citations, the magnitude of the Chinese PI effect appears impervious to their inclusion in the specification. For example, the communication controls do not affect the China discount, and also do not explain citation patterns: Articles from English-speaking countries do not receive more citations from the US; and PIs from cities farther away from the citing author's US city do not appear to receive fewer US citations.

Investigators with US training are cited more by US authors, so a US education probably increases the reach of PIs' US network. Past as well as current coauthors are also more likely to cite, most likely because they are more aware of the focal article (as we already control for topical relatedness when we construct the risk set). In contrast, the presence of US coauthors in the cited paper, the number of past US coauthorships for the focal investigator, a common ethnicity by cited and citing authors, or being an editorial author does not in general increase the propensity of being cited by US researchers. Overall, differential network reach does not appear to confound the citation discount experienced by Chinese PIs.

The spatial clustering of research fields has significant effects on citations from the US, but is not correlated with the China effect. The intellectual focus of the PI matters: articles that are written in subfields that are closely related to the other publications of the PI are associated with an increased rate of US citations. The same is true for articles in subfields

---

[8] Tables G.1 and G.5 in the online appendix reports similar results pertaining to US citations contained in patents rather than scientific articles.

for which the PI is an important contributor globally. Articles belonging to retraction-heavy subfields are cited relatively less, but this effect is statistically indistinguishable from zero.

Across all specifications, we observe a statistically significant and negative “China effect”: articles written by Chinese PIs receive significantly fewer citations from US scientists than articles written by non-Chinese PIs. The magnitude of the effect is empirically meaningful: Since the baseline probability of being cited by a US article is low in our sample (3.2%), the probability of a Chinese-authored article being cited is 28.1% lower than the baseline probability (based on the estimates from column 2, our baseline).

One may wonder whether the Chinese citation discount exists because the emergence of Chinese science is quite recent. In this case, one may expect the China discount to become smaller over time. Figure F.1 in the online appendix presents estimates the Chinese discount separately by cited article publication year. There is no discernible pattern in the discount over time, but it is negative in almost all years, and statistically significant for many of the years. Overall, it does not seem that we can expect the Chinese citation discount to be a transitory phenomenon.

Another question is whether China’s experience is unique, or whether other countries suffer from the same bias. In fact, the choice of China to define the treated group of articles is arbitrary. Would we find similar evidence of a discount if we chose to make researchers from other countries with a storied legacy in chemistry research pivotal? In Figure 2 we replicate our analysis by making the articles from PIs located in other top chemistry countries the treated group. Among the six countries that have at least 1,500 articles in the matched sample, no other country experiences a significant citation discount, and the magnitude of the discount is also largest for China. Switzerland and Germany, two countries which are renowned for their important chemical industries, experience citation premia of 74% and 59%, respectively.

Examining ethnic animus as a channel. So far, we have identified a stable China discount across a variety of specifications that is unique to China. Is this a reflection of animus towards Chinese researchers, rather than reduced awareness or a reduced integration into the US research community?

We complement Figure 2 by treating 40 investigators with Chinese names working outside of China as if they belonged to an independent country, and ask whether the research originating from this aggregate is discounted by US researchers.[9] We find that this is not the case, and in fact these investigators receive, if anything, a small (and imprecisely estimated) premium relative to investigators located in other countries.[10] This evidence seems hard to reconcile with explanations stressing the role of animus, which should affect ethnically Chinese investigators regardless of their location.

Heterogeneous effects. So far we have established that Chinese elite chemists experience a citation discount from the US on average. We now ask whether some Chinese PIs can overcome the discount, whether the discount is less severe in some subfields, or whether some US researchers are less biased against Chinese research. We check this by allowing the China effect to vary across a number of characteristics.

For example, in column 3 of Table 3, we test whether strong networks of Chinese PIs help Chinese articles overcome the citation discount. Recent research by Xie and Freeman (2023) suggests that US-trained Chinese researchers obtain more citations per paper and have a higher proportion of their publications in high-impact journals than other Chinese PIs. Gaulé and Piacentini (2013) document that Chinese students receiving graduate training in the United States are among the best and brightest, performing about as well as the awardees of the prestigious NSF doctoral fellowship program, and far better than other foreign students. In contrast, we find that US education is not enough for Chinese researchers to overcome the US citation discount, although it reduces its magnitude by more than half. We do not find significant effects from having US coauthors, neither for Chinese articles nor for those from other countries. We also examine whether US-based researchers who have ethnic roots in China help diffuse Chinese research to the US, as suggested by recent research (Xie and Freeman, 2020). We do find a positive interaction effect with citer ethnicity, so that US researchers with a Chinese name do not cite Chinese articles less than articles from other countries.[11] This is consistent with the view that Chinese researchers have access to US-based ethnic-Chinese researchers in their network, but not to other US researchers.

---

[9] In addition to 155 PIs from mainland China in our estimation sample, 40 investigators have Chinese names but work outside of mainland China, including 13 in Taiwan, 8 in Hong Kong, 8 in Singapore, 4 in Japan, 3 in Canada, 2 in Switzerland, and one in both the UK and Sweden. To fix ideas about the relative size of this aggregate, Canada contributes 40 investigators to the estimation sample, while Switzerland contributes 38.

[10] Note that this is not driven by a positive selection of Chinese PIs into Europe or Japan, as this result holds if we just consider Chinese PIs based in Hong Kong, Taiwan, or Singapore.

[11] The China effect for PIs who are cited by US authors with Chinese ethnicity is the sum of the main estimate for China combined with the corresponding interaction effect. The magnitude of the combined effect in an imprecisely estimated -0.0046.

Column 4 checks whether the specialization of China in certain subfields has implications on being cited in US research. We do not find effects that are significantly different from zero. In column 5 we study whether Chinese PIs who are especially focused in their research overcome the China bias. We find that this is the case for Chinese PIs for whom the focal article's subfield is important, i.e., Chinese PIs that are publishing in their area of expertise. However, the China effect only disappears for the most focused PIs, i.e., those in the 99^{th} percentile of the subfield importance distribution. It could be the case that US researchers are more aware of focused Chinese PIs, or that the research by these specialized Chinese PIs is taken more seriously.

Because of the relatively high frequency of retraction scandals that have afflicted Chinese scientific teams (Liao et al., 2018; Huang, 2017), we speculate that non-Chinese scientists could deem knowledge and ideas that originate in China to be less reliable than those originating in other countries. In column 6, we test this conjecture by interacting the Chinese PI indicator with a dummy that indicates the existence of retractions in the focal article's subfield. We do find evidence of an additional citation discount imposed on Chinese articles belonging to "retraction-heavy" subfields, but the corresponding estimate is not statistically significant, and the magnitude of the Chinese PI effect barely changes when controlling for "perceived" quality in this way.

Column 7 allows for all interaction effects to enter the specification simultaneously, with similar results. Overall, these specifications point towards an obdurate citation discount experienced by articles published by elite Chinese chemists, one that can only be overcome in a handful of contingencies.

5 Conclusion

The inclusion of Chinese scientists in the global "Republic of Science" has gathered pace over the past two decades. An increasing body of evidence points to a gradual bridging of the gap that long existed between the impact of Chinese published scientific output and that of frontier countries (Xie and Freeman, 2019). Observers note—with a mix of awe and trepidation—that Chinese scientists are about to overtake US scientists in at least one domain: Artificial Intelligence (O'Meara, 2019).

Our study sheds light on the propensity to cite research emanating from Chinese laboratories by pairing Chinese and non-Chinese articles well matched on attributes that plausibly capture the scientific “fertility” of each publication. Focusing on elite researchers in a single domain—chemistry—we uncover the existence of a sizable citation discount for Chinese articles, relative to non-Chinese articles. What explains the relative underciting of Chinese science by US scientists?

One possibility is that in spite of our best efforts, systematic differences in citation potential subsist between treated and control articles, even after carefully matching on journal and citations received from non-US sources. Another possibility is that the discount reflects animus directed at Chinese scientists, but this hypothesis does not sit well with the evidence that it is not apparent for researchers with Chinese names doing research outside China. Yet another possibility is that the discount reflects hostility or skepticism directed at Chinese institutions rather than Chinese scientists. Skepticism aimed at research originating from Chinese institutions could reflect perceptions of lower reliability for Chinese-produced knowledge due to the number of well-publicized cases of scientific misconduct in China (Huang, 2017; Liao et al., 2018). However, this conjecture does not receive support in our analysis, since the discount for publishing in retraction-heavy subfields is small on average, and scarcely larger for Chinese scientists. Perhaps foreign scientists harbor resentments against Chinese institutions because they are hostile to China’s political institutions in general. We cannot rule this explanation out as there are no foreign researchers in Chinese institutions within our sample.

One final possibility is that US scientists are simply less aware of Chinese research, perhaps because Chinese scientists, even if they belong to the elite, have less access to the networks that provide broad exposure to research findings. This explanation is most consistent with the evidence we present, since the discount is (i) partly overcome by returnees who completed their scientific training in the US; (ii) absent for US citing authors with a Chinese name; and (iii) reduced for Chinese PIs that are very specialized in their subfield.

Is the China citation discount likely to be a transitory phenomenon? If awareness and networking are its root causes, current US-China tensions, as well as the disruption of scientific travel induced by the COVID-19 pandemic, may further solidify the lower awareness of foreign citers vis-à-vis research produced in China (Jia et al., 2022).

From a global welfare perspective, our results naturally lead to the question of which discoveries (if any) are missing, or slowed down, by virtue of the citation frictions we uncover. While we were not able shed light on it in the present study, this remains a fascinating topic for further research.

References

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "`cem`: Coarsened Exact Matching in Stata." *Stata Journal* 9:524–546.

Borjas, George J. and Kirk B. Doran. 2015. "Cognitive Mobility: Labor Market Responses to Supply Shocks in the Space of Ideas." *Journal of Labor Economics* 33:S109–S145.

Gaulé, Patrick and Mario Piacentini. 2013. "Chinese Graduate Students and US Scientific Productivity." *Review of Economics and Statistics* 95:698–701.

Huang, Echo. 2017. "China Publishes More Science Research with Fabricated Peer Review Than Everyone Else Put Together." *Quartz*. Available at `https://qz.com/978037/china-publishes-more-science-research-with-fabricated-peer-review-than-everyone-else-put-together` [Published: May 8, 2017; Accessed: June 6, 2023].

Jia, Ruixue, Huihua Nie, and Wei Xiao. 2019. "Power and Publications in Chinese Academia." *Journal of Comparative Economics* 47:792–805.

Jia, Ruixue, Margaret E. Roberts, Ye Wang, and Eddie Yang. 2022. "The Impact of U.S.-China Tensions on U.S. Science." Working Paper, UCSD.

Liao, Qing-Jiao, Yuan-Yuan Zhang, Yu-Chen Fan, Ming-Hua Zheng, Yu Bai, Guy D. Eslick, Xing-Xiang He, Shi-Bing Zhang, Harry Hua-Xiang Xia, and Hua He. 2018. "Perceptions of Chinese Biomedical Researchers Towards Academic Misconduct: A Comparison Between 2015 and 2010." *Science and Engineering Ethics* 24:629–645.

Lin, Jimmy and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* 8:1–14.

O'Meara, Sarah. 2019. "Will China Lead the World in AI by 2030." *Nature* 572:427–428.

Teplitskiy, Misha, Eamon Duede, Michael Menietti, and Karim R. Lakhani. 2022. "How Status of Research Papers Affects the Way They Are Read and Cited." *Research Policy* 51:104484.

Tollefson, Jeff. 2018. "China Declared World's Largest Producer of Scientific Articles." *Nature* 553:390.

Torvik, Vetle I. and Neil R. Smalheiser. 2009. "Author Name Disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3:1–29.

Xie, Qingnan and Richard B. Freeman. 2019. "Bigger Than You Thought: China's Contribution to Scientific Publications and Its Impact on the Global Economy." *China & World Economy* 27:1–27.

Xie, Qingnan and Richard B. Freeman. 2020. "The Contribution of Chinese Diaspora Researchers to Scientific Publications and China's 'Great Leap Forward' in Global Science." NBER Working Paper #27169.

Xie, Qingnan and Richard B. Freeman. 2023. "Creating and Connecting US and China Science: Chinese Diaspora and Returnee Researchers." NBER Working Paper #31306.

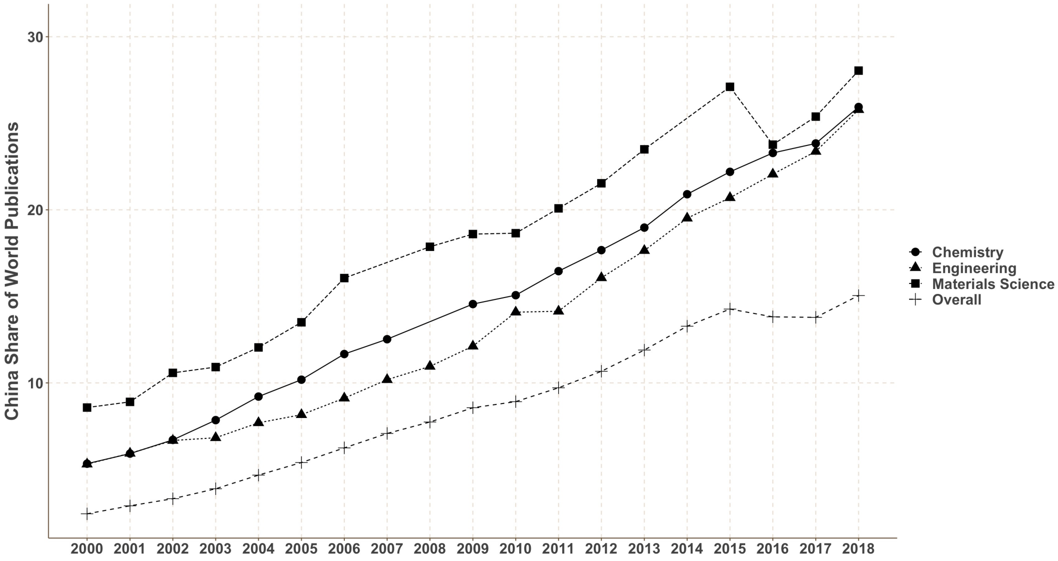

Figure 1: China's share of world publications, 2000-2018

Note: The share of publication is computed based on the share of Chinese addresses in English-language research articles in *Web of Science*, 2000-2018.

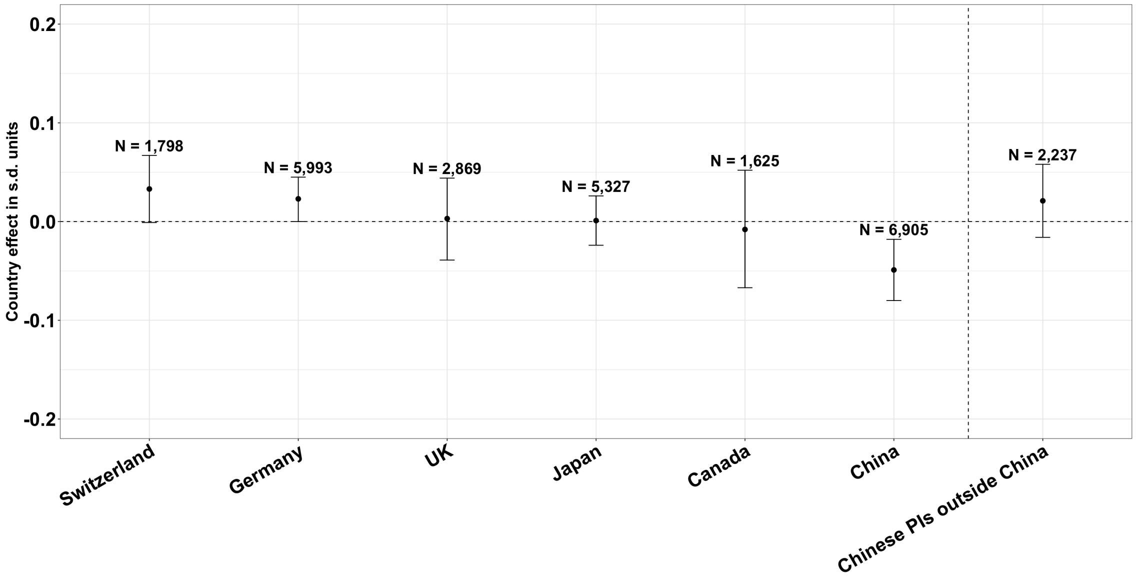

Figure 2: Heterogeneous effects of Chinese PIs on US citations, by source country

Note: We replace China with Switzerland, Germany, Canada, UK, Japan and Canada, respectively, to generate new treated and control groups, and estimate the country effect for each treated country with the same specification used in column 2 in Table 3. The dark dots in the above plots correspond to country effects measured in standard deviation units for each treated country. The 95 percent confidence interval (the corresponding standard errors are two-way clustered at the investigator and matching strata levels) around these estimates is plotted with vertical lines. The number of treated articles for each country is indicated above the corresponding coefficient estimate.

Table 1: Effect of Chinese investigatorship on the number of US citations

	(1)	(2)	(3)
Chinese investigator	-0.662	-0.416	-0.269
	(0.010)	(0.097)	(0.052)
No. of Articles	658,621	658,621	658,471
Pub. Year Fixed Effects		yes	yes
Journal Fixed Effects			yes
Pseudo R^2	0.024	0.089	0.261
% increase	-48%	-34%	-24%

Note: The dependent variable is the number of US citations, i.e., citations from articles with only US-based authors. All regressions include fixed effects for the number of authors. The sample includes all articles in the field of chemistry between 2000 and 2018, provided their authorship team hails from a single country (articles with geographically-mixed authorship teams are excluded). Coefficients derive from a Poisson specification estimated via quasi–maximum likelihood. Robust standard errors in parentheses are clustered by the country of the cited article.

Table 2: Summary statistics of control and treated articles

	Control articles (N= 9,287)					Treated articles (N= 6,905)				
	Mean	Median	Std. Dev.	Min	Max	Mean	Median	Std. Dev.	Min	Max
Characteristics used for matching:										
Source Article Publication Year	2011.879	2012	3.872	2000	2018	2011.879	2012	3.872	2000	2018
Source Article Number of Authors	5.034	5	1.975	1	15	5.174	5	1.926	1	15
Investigator PhD Degree Year	1993.447	1994	6.277	1958	2008	1993.660	1994	6.212	1952	2007
Number of Debiased Citations Outside US	18.568	11.603	24.103	0	625	17.926	11.645	21.653	0	598
Characteristics not used for matching:										
No. of Total Citations	24.177	16	29.489	1	726	33.191	21	40.570	0	1,125
No. of Citations Outside US	20.661	13	25.585	0	638	30.187	19	37.014	0	1,017
US First/Reprinted Cited Author	0.018	0	0.133	0	1	0.006	0	0.077	0	1
US Cited Author in Other Positions	0.037	0	0.189	0	1	0.016	0	0.127	0	1
Cumulative No. of US Coauthorships	8.462	4	14.787	0	130	5.610	3	7.213	0	62
Retraction-heavy Subfield	0.028	0	0.164	0	1	0.043	0	0.204	0	1
Subfield Foreign Research Intensity	17.542	14.104	16.456	0	170.860	14.427	11.631	13.167	0	190.219
Subfield Home Research Intensity	1.639	0.757	2.452	0	38.926	11.278	7.993	12.402	0	158.569
Subfield USA Research Intensity	4.402	3.053	4.927	0	64.233	4.440	3.036	5.186	0	79.131
Importance of Investigator for the Subfield	0.061	0.038	0.076	0	1	0.050	0.029	0.062	0	0.542
Importance of Subfield for Investigator	0.124	0.052	0.329	0	10	0.131	0.042	0.359	0	8.500
Ellison/Glaeser Index of Scholarly Focus	0.017	0.014	0.015	-0.026	0.270	0.015	0.013	0.012	-0.020	0.142
Female Investigator	0.038	0	0.191	0	1	0.046	0	0.209	0	1
Investigator with US Training	0.611	1	0.488	0	1	0.476	0	0.499	0	1
Investigator is Editorial Author on Elite Journals	0.032	0	0.176	0	1	0.010	0	0.102	0	1
Investigator from English-speaking Country	0.342	0	0.474	0	1	0	0	0	0	0
Investigator of Chinese Names Working outside China	0.088	0	0.284	0	1	0	0	0	0	0
Investigator Publication Stock	82.698	70	62.321	0	738	91.051	73	72.270	0	496
Investigator Citation Stock	1,305.696	729	1,638.410	0	13,082	1,163.350	643	1,435.532	0	11,243

Note: The stock of publications and citations for each investigator are assessed at the end of the year prior to the year of publication for each source article.

Table 3: Estimating the China location discount (or premium) on the rate of US citations [Linear Probability Model]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Chinese Investigator	-0.008	-0.009	-0.011	-0.009	-0.010	-0.008	-0.012
	(0.002)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.005)
Communication							
Investigator from English-speaking Country		-0.001	-0.001	-0.000	-0.001	-0.001	-0.000
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log(Avg. Distance)		0.003	0.003	0.003	0.003	0.003	0.003
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Network							
Investigator with US Training		0.005	0.002	0.005	0.005	0.005	0.002
		(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
US First/Reprinted Cited Author		0.011	0.015	0.011	0.011	0.011	0.015
		(0.011)	(0.013)	(0.010)	(0.010)	(0.011)	(0.012)
US Cited Author in Other Positions		0.005	0.004	0.005	0.006	0.005	0.004
		(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)
Log(Cumulative No. of US Coauthorships)		0.001	0.001	0.001	0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Citation from Same Ethnicity		0.003	-0.001	0.003	0.003	0.003	-0.001
		(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
Citing Coauthor is Investigator's Past Collaborator		0.050	0.049	0.050	0.050	0.050	0.049
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Common Coauthor		0.166	0.166	0.166	0.166	0.166	0.166
		(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Investigator is an Editorial Author		-0.003	-0.001	-0.003	-0.003	-0.003	-0.001
		(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)
Geographic Topical Clustering							
Subfield Home Research Intensity		0.009	0.010	0.031	0.010	0.010	0.037
		(0.009)	(0.009)	(0.048)	(0.009)	(0.009)	(0.048)
Subfield Foreign Research Intensity		-0.022	-0.023	-0.028	-0.022	-0.022	-0.029
		(0.007)	(0.007)	(0.010)	(0.007)	(0.007)	(0.010)
Subfield USA Research Intensity		0.043	0.045	0.057	0.044	0.043	0.061
		(0.020)	(0.020)	(0.030)	(0.020)	(0.020)	(0.030)
Investigator's Intellectual Focus							
Importance of Subfield for Investigator		0.004	0.004	0.004	0.001	0.004	0.001
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Importance of Investigator for the Subfield		0.033	0.031	0.033	0.038	0.033	0.037
		(0.016)	(0.016)	(0.016)	(0.019)	(0.016)	(0.019)
Ellison/Glaeser Index of Scholarly Focus		0.063	0.054	0.063	0.044	0.063	0.041
		(0.053)	(0.053)	(0.053)	(0.064)	(0.053)	(0.063)
Reputation							
Retraction-heavy Subfield		-0.003	-0.003	-0.003	-0.003	-0.000	0.000
		(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Interactions with Network							
Chinese Investigator × Investigator with US Training			0.007				0.007
			(0.004)				(0.004)
Chinese Investigator × US First/Reprinted Cited Author			-0.026				-0.024
			(0.020)				(0.019)
Chinese Investigator × US Cited Author in Other Positions			0.004				0.004
			(0.010)				(0.010)
Chinese Investigator × Log(Cumulative No. of US Coauthorships)			-0.002				-0.001
			(0.002)				(0.002)
Chinese Investigator × Citation from Same Ethnicity			0.008				0.008
			(0.004)				(0.004)
Chinese Investigator × Investigator is an Editorial Author			-0.005				-0.004
			(0.020)				(0.020)
Interactions with Geographic Topical Clustering							
Chinese Investigator × Subfield Home Research Intensity				-0.029			-0.031
				(0.050)			(0.050)
Chinese Investigator × Subfield Foreign Research Intensity				0.018			0.017
				(0.016)			(0.016)
Chinese Investigator × Subfield USA Research Intensity				-0.035			-0.040
				(0.036)			(0.035)
Interactions with Investigator's Intellectual Focus							
Chinese Investigator × Importance of Subfield for Investigator					0.006		0.006
					(0.003)		(0.003)
Chinese Investigator × Importance of Investigator for the Subfield					-0.014		-0.015
					(0.034)		(0.034)
Chinese Investigator × Ellison/Glaeser Index of of Scholarly Focus					0.054		0.035
					(0.104)		(0.105)
Interactions with Reputation							
Chinese Investigator ×Retraction-heavy Subfield						-0.004	-0.006
						(0.006)	(0.007)
Mean of Dependent Variable	0.032	0.032	0.032	0.032	0.032	0.032	0.032
s.d. of Dependent Variable	0.177	0.177	0.177	0.177	0.177	0.177	0.177
China effect in s.d. units	-0.043	-0.049	-0.072	-0.049	-0.057	-0.048	-0.077
Adjusted R^2	0.079	0.084	0.084	0.084	0.084	0.084	0.084
No. of Investigators	557	557	557	557	557	557	557
No. of Cited Articles	16,192	16,192	16,192	16,192	16,192	16,192	16,192
No. of Citing Articles	71,409	71,409	71,409	71,409	71,409	71,409	71,409
No. of Citing/Cited Article Pairs	188,753	188,753	188,753	188,753	188,753	188,753	188,753

Note: The dependent variable is an indicator variable that equals 1 if the related article cites the PI's article, and 0 otherwise. All regressions include fixed effects for rank bins of each citing article j with respect to its topic similarity to article i; fixed effects for the interaction of citing and cited article publication year; fixed effects for each CEM strata; fixed effects for the investigator's highest degree year, a investigator gender indicator variable, and an indicator if citing and cited articles are published in the same journal (coefficients not reported). Standard errors in parentheses are two-way clustered at the investigator and strata level.