NBER WORKING PAPER SERIES

WHO STANDS ON THE SHOULDERS OF CHINESE (SCIENTIFIC) GIANTS? EVIDENCE FROM CHEMISTRY

Shumin Qiu
Claudia Steinwender
Pierre Azoulay

Working Paper 30772
http://www.nber.org/papers/w30772

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022, Revised November 2024

All authors contributed equally. Address correspondence to pazoulay@mit.edu. We thank Richard Freeman, Alessandra Allocca, Pat Kline, as well as seminar audiences at USC, Yale, UC Berkeley, the CRC, Metascience 2023, ICSSI 2023, and the NBER China Working Group for insightful comments. We thank Kimberly White, Tabea Rudolf, and Lynn van den Busch for outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Shumin Qiu, Claudia Steinwender, and Pierre Azoulay. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Who Stands on the Shoulders of Chinese (Scientific) Giants? Evidence from Chemistry
Shumin Qiu, Claudia Steinwender, and Pierre Azoulay
NBER Working Paper No. 30772
December 2022, Revised November 2024
JEL No. I23, O30, O53

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.

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

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

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

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. This is not merely a choice of convenience. 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. 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. Chemistry also stands out from other disciplines in that it provides a store of fundamental scientific knowledge that is highly relevant for firms in the chemical and biopharmaceutical sectors. Finally, the focus on Chemistry allows us to relax data constraints that has hampered the study of frictions in the diffusion of scientific knowledge across borders at a very granular detail.

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.

Our analysis is related to research studying the global diffusion of scientific knowledge, which has found citations to strongly fall with distance and across country borders (Peri, 2005; Belenzon and Schankerman, 2013; Head et al., 2019; Fry and MacGarvie, 2024). Recent contributions have highlighted the increased importance of China as a producer of science (Tollefson, 2018; Xie and Freeman, 2019, 2020, 2023), as well as the growing impediments to international collaborations involving Chinese scientists (Jia et al., 2022; Aghion et al., 2023). Finally, our findings echo the work of sociologists who have previously documented of patterns of unequal citation flows between scientists located in frontier countries and those hailing from countries confined to the periphery of the global "Republic of Science" (Gomez et al., 2022).

2 Chinese Research in Chemistry

For thousands of years ancient China led the world with remarkable inventions and achievements in the chemical and metallurgical arts (Agnew, 1997). Many important empirical discoveries originated from ancient Chinese alchemy and medicinal chemistry; their translation into Western languages had a pronounced influence on modern chemical science (Leicester, 1971). For example, many historians believe that gunpowder technology, one of the most influential inventions in human history, had its origins in China (581-681 CE), and then spread to the Middle East and Europe along the Silk Road (Needham et al., 1986). Chinese pre-modern "scientists" also pioneered the manufacturing processes for salt, wine, paper, and porcelain (Li, 1948; Tsuen-Hsuin, 1985).

Although historians suggest that modern chemistry grew, at least in part, out of the work of Chinese alchemists (Leicester, 1971), chemistry as a modern science was absent in China until the 19th century, when European science was introduced through missions, trade, and wars (Li, 1948). In the late Ch'ing Dynasty (mid-to-late 19th century), during which the rulers adopted a closed-door policy with very limited communication with the outside world, Chinese chemistry (as well as other sciences) lagged far behind western countries. After wars with European countries broke open China's door, modern chemistry started to develop with the purpose of "learning from foreigners to compete with them" as China became integrated into the global "Republic of Science" (Bai, 2000). Research by Western chemists were intensively translated into Chinese and disseminated

in China: Between 1912 and 1949, 41% of chemistry articles and textbooks were translated from English, while the remainder were written by Chinese chemists.[1]

The rapid development of modern Chinese chemistry took place after the founding of the People's Republic of China, especially after the deep opening policy begun in 1978 (Bai, 2000). Between 2000 and 2017, the number of Chinese universities increased by 140% (from 753 to 1,805), and correspondingly the number of chemistry departments rose by 182% (from 243 to 686). Research faculty in Chinese universities increased by 69% during the same period, and the number of chemistry researchers tripled. Public research funding invested in chemistry also shows a 14-fold increase between 2000 and 2017, higher than the ten-fold increase observed for other fields on average.[2] Meanwhile, China has continuously expanded global collaboration and communication by funding students' graduate studies abroad, and facilitating Chinese scholars' participation in international collaboration through the funding of shorter-term stays in frontier countries. The number of state-financed students studying abroad increased five-fold, from 7,564 in 2000 to 46,347 in 2017, whereas attendance of international conferences increased almost eight-fold during the same period. Between 1978 and 2018, a total of 5.86 million students studied abroad, 82% of whom returned to China. The flow of transnational human capital, particularly the return of elite scientists, has helped create a solid foundation for Chinese scientific research. Scientists holding overseas degrees account for 37% of the total number of members of Chinese Academies of Sciences and Engineering elected between 1955 to 2009. During this period, 300 US-trained academics returned to China, a figure to be compared with 160 Soviet-trained and 80 UK-trained academics who returned during the same period (Zhao et al., 2009). In chemistry specifically, there is evidence that students receiving graduate training in the United States are among the best and brightest. Gaulé and Piacentini (2013) document that Chinese students perform about as well as the awardees of the prestigious NSF doctoral fellowship program, and far better than other foreign students.

Table A.1 demonstrates the importance of Chinese elite researchers. According to the annual *Highly Cited Researchers* (HCRs) rankings published by *Clarivate Analytics* between 2014 and 2018, chemistry ranks highest among scientific fields in terms of highly cited researchers (column 1). These 211 researchers account for 19.27% of the world's HCRs in chemistry (column 2). Table A.2 shows that Chinese chemists have become world-leading contributors compared to other countries. During 2000 to 2015, China's share of publications in chemistry was 14.96%, ranking it second only after the United States. Japan is a distant third with 7.66%, followed by Germany, India, and the United Kingdom. The ranking with respect to HCRs is similar, with the United States accounting for the largest share of the world's elite chemists (43.01%), followed by China (19.27%).[3]

[1]Source: National bibliography of the Republican period (1911-1949) produced by Beijing Library.

[2]The source for these figures, and those mentioned below is the *Compilation of University Science and Technology Statistics* produced by the Chinese Ministry of Education.

[3]China's strong position in chemistry becomes particularly striking when we compare it to China's ranking with respect to all fields. Across the sciences, the United States are clearly the most dominant nation with 8,306 HCRs

3 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.[4,5]

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), 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

amounting to almost half (46.48%) of the world's top scientists. Second by a large margin is the UK with 1,701 HCRs (9.52% of the world), and China is third with 1,104 HCRs (6.18% of the world). This makes China's HCR share in chemistry more than three times larger than the average across fields.

[4]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.

[5]The replication code and data is available on `https://github.com/ShuminQiu/ChineseShoulder`.

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.[6]

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.

4 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

[6] Our reliance on *PubMed* and *Web of Science* entails the exclusion of articles published in Chinese journals, which are not indexed by these databases (Cyranoski, 2019; Wang et al., 2021). Since home-grown Chinese journals are cited almost exclusively by Chinese researchers, the estimates of a Chinese citation discount presented below could be construed as an underestimate of citation frictions we document.

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.

4.1 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 quality 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 Qiu et al. (2024), as well as 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 B.[7] Our matching procedure splits the number of "home-debiased"

[7]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); eliminate diaspora citations from the stock of ROW citations; match on investigator-level quality covariates rather than article-level quality covariates; or do not match on citations at all.

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.[8]

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.[9] 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 D.1 in the online appendix 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.[10] The combination of matching and covariate inclusion results in fine-grained comparisons that plausibly hold quality constant across Chinese and non-Chinese publications. More details about the matching procedure are available in Appendix D.

[8] As a referee pointed out to us, the bias could run in the other direction if ROW citation counts reflect in part the size of the diaspora community residing in these "rest of the world" countries. This concern is salient given the presence of expatriate Chinese scientists throughout the world, and not simply in the US. In Table C.1 in the online Appendix, we eliminate likely diaspora citations from the ROW citation count for each cited article, with little effect on the estimates.

[9] 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.

[10] 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.

4.2 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.[11]

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 the articles in our sample.[12] 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.[13,14]

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; Duede et al., 2024). 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.

[11] This approach can be thought of as a more granular version of the "citational lensing" framework proposed by Gomez et al. (2022), which focuses on citation flows between country pairs and how they relate to the textual similarity between the articles published by authors from the same destination and origin countries.

[12] 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.

[13] Azoulay et al. (2019a, Appendix C) provide more detail regarding the use of PMRA for research purposes. To facilitate the harvesting of PubMed-related records on a large scale, we have developed an open-source software tool that queries PMRA and stores the retrieved data in a MySQL database. The software is available for download at `http://www.stellman-greene.com/FindRelated/`.

[14] Our focus on Chemistry can be justified pragmatically: Most chemistry journals are indexed by *PubMed*, making it feasible to construct citation risk sets that are plausible because they leverage literature search practices common in the chemistry domain. To our knowledge, this would not be straightforward to implement for any other discipline in which China is considered a frontier country.

4.3 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 × PhD degree year; yielding 3,847 bins); (b) investigator cumulative publications (13 bins), (c) investigator cumulative citations (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;[15] (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 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.[16] 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

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

[16] 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.

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.

5 Empirical Results

Table 2 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.

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 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.[17] 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.[18] This evidence seems hard to reconcile with explanations

[17] 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.

[18] 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.

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 2, 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.[19] 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.

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 99th 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 coun-

[19]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.

tries. 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.

6 Leveraging article citations from patents

More than a fundamental scientific discipline, the field of chemistry also forms the basis for technological advances in industry, including the biopharmaceutical sector (Adams, 1990). In addition, patenting is a common way for firms to appropriate the returns from innovation in this domain (Cohen et al., 2000), since at least the emergence of the periodic table of elements in the late 19th century (Moser, 2012).

Recently, it has become possible to track citations made by patents to the open scientific literature at scale (Marx and Fuegi, 2020; Roach and Cohen, 2013), thus providing a lens on understanding how advances in basic science percolate in industry R&D (Azoulay et al., 2019b). We leverage this novel source of data by studying the extent to which US patent inventors rely on scientific research in chemistry originating from China versus other countries.

The research design parallels the one used for the analysis of article-to-article citations. To begin, we focus on the full set of chemistry articles published between 2000 and 2018 (as in Table 1), changing the outcome variables from article citations to US patent citations with all-American inventor teams. We find that the China discount is also present in patent citations. In the raw data, Chinese articles receive 0.083 cites from US patents on average, while non-Chinese articles receive 0.255 cites. Table 3 provides the results of Poisson regressions. Column 1 shows that a Chinese article received on average 70% fewer citations by US patents compared to a article from other non-US countries. The magnitude of the estimate barely changes when we add fixed effects for the number of authors (column 2), and is roughly halved when adding publication year effects in column 3. Column 4 includes journal fixed effects in the specification, which further reduces the magnitude of the China citation discount to 25%. Focusing on within-journal variation can be thought of as a crude way to hold the quality of the underlying articles constant. Below, we go further by performing a careful matching on article quality, as well as researchers' patenting activities, based on the data set of articles written by our elite chemistry PIs.

Scientists who are active inventors themselves may have networks that reach beyond academia into industry, thereby heightening both awareness and relevance of their investigations in the eyes of R&D intensive firms (which account for the lion's share of all inventive activity, cf. Azoulay et al. 2012). Therefore, we modify the coarsened exact matching strategy used in Section 4 and enrich it with measures of PI's patenting activities, which we track by linking investigators' names to the names of inventors on USPTO patents. This mapping process is challenging, as we must guard against mistakenly assigning a patent to a scientist when the invention was actually performed by a namesake. In order to accurately attribute patents to our set of elite investigators, we conduct extensive manual checks which take into account data fields such as institutional affiliation, city, country, and research interest that overlap between the CV and the patent information. We identify 5,562 patents which list one of our 751 elite PIs as an inventor: 48.1% of Chinese investigators and 61.7% of other non-US investigators have been granted at least one patent by the USPTO.

Next, we gather a list of patent citations to our PIs' articles, which we extract from the dataset constructed by Marx and Fuegi (2020). This process results in a list of 47,831 patent-to-article citations for the 78,541 PI articles. Each article receives on average 0.609 patent citations overall, but only 0.310 patent citations from all-US inventor teams. Importantly, 87% of the articles in the sample are not cited by any US patent. As we did for the analysis of article-to-article citations, we restrict the patent-to-article citations to those from inventions with US inventors only.

A key limitation of the analysis presented below is that we do not have at our disposal a set of patents that are at risk of citing each article—the *PubMed* Related Citations Algorithm identifies topically-close articles, but there is no equivalent algorithm to identify topically-close patents: Every observation in the data corresponds to an actual citation. As a result, we aggregate the number of patent citations up to the article level and estimate Poisson models to analyze the determinants of the count of US patents for each article via Quasi-Maximum Likelihood, with robust standard errors clustered at the level of the PI country.

Descriptive statistics corresponding to three alternative ways to pair Chinese with non-Chinese articles while incorporating information regarding PIs' involvement in patenting are displayed in Appendix G. Column 1a and column 1b of Table 4 display results based on a coarsened exact matching approach very similar to that used earlier in Table 2. Column 1a implies that an article written by Chinese investigators receives 30% fewer citations from US patents, relative to an article written by other non-US investigators.[20] After controlling for characteristics of articles and investigators, as shown in column 1b, Chinese publications' citation discount in US patents slightly increases to 35%.

Columns 2a and 2b repeat the analysis by adding the PI's patent inventor status to the matching variables, i.e., an indicator variable equal to one if the PI has applied for at least one patent prior

[20]Since $\exp(-0.362) - 1 = -0.30$. See the last row of Table 4 for equivalent percentage changes.

to the publication year of source article. Columns 3a and 3b report estimates based on similar patenting stock. In both cases, the China citation discount remains stable: after controlling for article and researcher characteristics, column 2b shows that an article authored by a Chinese PI receives 41% fewer citations from US patents, and in column 3b, the discount is 42%. Columns 4a and 4b further impose the coarse match on home-debiased patent-to-article citations from outside the US. Once again, Chinese publications exhibit a stable citation discount in patents compared to publications from other countries, though the effect is only statistically significant at the 10% level.

We conclude that US industrial firms (which account for the bulk of patent citations) tend to build less on scientific research originating from Chinese labs, in a fashion similar to our earlier finding that US academics (which account for the bulk of article citations) appear to discount Chinese research, relative to research originating from other countries.

7 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 quality 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.[21] 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

[21] While our evidence comes from a single scientific domain, our results are consistent with those of Fry and MacGarvie (2024), who find that Chinese COVID-19 preprints receive less attention than US preprints, holding constant a wealth of covariates predictive of quality.

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. Interestingly, Fry and MacGarvie (2024) do not find evidence that the network mechanism explains the "attention deficit" experienced by Chinese preprints. However, it seems plausible that the specific context they examine—the early phase of COVID-19 which originated in China—could make the differences in network reach among Chinese scientists less relevant.

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).

The virtues of Chemistry as a strategic research site may come at the expense of generalizability. While this discipline spans a wide range of research styles, for the most part it is organized in medium-size laboratories under the helm of a single principal investigator. It seldom features exceedingly small research teams (as in pure mathematics) or "big science" efforts for which expenditures in specialized capital are so lumpy as to fully consolidate the field into a single or a handful of large authorship teams. As such, one should refrain from applying our findings to other fields of science where team structure, the degree of intellectual clustering, or patterns of international mobility are likely to generate different citation practices. Assessing the degree to which our results extend to other settings, and the reasons they might differ, represents a fruitful area for future research.

References

Adams, James D. 1990. "Fundamental Stocks of Knowledge and Productivity Growth." *Journal of Political Economy* 98:673–702.

Aghion, Philippe, Céline Antonin, Xueping Sun, David Strömberg, Luc Paluskiewicz, Raphael Wargon, and Karolina Westin. 2023. "Does Chinese Research Hinge on US Coauthors? Evidence from the China Initiative." Working Paper, Collège de France.

Agnew, Neville. 1997. "Conservation of Ancient Sites on the Silk Road." In *Conference on the Conservation of Grotto Sites*, pp. 395–405.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin. 2019a. "Does Science Advance One Funeral at a Time?" *American Economic Review* 109:2889–2920.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven N. Sampat. 2019b. "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies* 86:117–152.

Azoulay, Pierre, Joshua S. Graff Zivin, and Bhaven N. Sampat. 2012. "The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine." In *The Rate Direction of Inventive Activity Revisited*, edited by Joshua Lerner and Scott Stern, pp. 107–155. University of Chicago Press.

Bai, Chunli. 2000. "Development and Prospect of Chinese Chemistry." *University Chemistry* 15:1–10.

Belenzon, Sharon and Mark Schankerman. 2013. "Spreading the Word: Geography, Policy, and Knowledge Spillovers." *Review of Economics and Statistics* 95:884–903.

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.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." NBER Working Paper #7552.

Cyranoski, David. 2019. "China spends millions to boost homegrown journals." *Nature* 576:346–347.

Duede, Eamon, Misha Teplitskiy, Karim Lakhani, and James Evans. 2024. "Being together in place as a catalyst for scientific advance." *Research Policy* 53:104911.

Fry, Caroline and Megan MacGarvie. 2024. "Author Country of Origin and Attention on Open Science Platforms: Evidence from COVID-19 Preprints." *Management Science* 70:5426–5444.

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

Gomez, Charles J, Andrew C Herman, and Paolo Parigi. 2022. "Leading countries in global science increasingly receive more citations than other countries doing similar research." *Nature Human Behaviour* 6:919–929.

Head, Keith, Yao Amber Li, and Asier Minondo. 2019. "Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics." *Review of Economics and Statistics* 101:713–727.

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.

Leicester, H. Marshall. 1971. *The Historical Background of Chemistry*. Courier Corporation.

Li, Chiao-Ping. 1948. *The Chemical Arts of Old China*. Journal of Chemical Education.

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.

Marx, Matt and Aaron Fuegi. 2020. "Reliance on Science: Worldwide Front-page Patent Citations to Scientific Articles." *Strategic Management Journal* 41:1572–1594.

Moser, Petra. 2012. "Innovation Without Patents: Evidence from World's Fairs." *Journal of Law and Economics* 55:43–74.

Needham, Joseph, Ping-Yü Ho, Gwei-djen Lu, and Ling Wang. 1986. *Military Technology: The Gunpowder Epic*. Cambridge University Press.

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

Peri, Giovanni. 2005. "Determinants of Knowledge Flows and Their Effect on Innovation." *The Review of Economics and Statistics* 87:308–322.

Qiu, Shumin, Claudia Steinwender, and Pierre Azoulay. 2024. "Paper Tiger? Chinese Science and Home Bias in Citations." NBER Working Paper #32468.

Roach, Michael and Wesley M. Cohen. 2013. "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research." *Management Science* 59:504–525.

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.

Tsuen-Hsuin, Tsien. 1985. "Chemistry and Chemical Technology: Paper and Printing." In *Science and Civilization in China*, edited by Joseph Needham, volume 5.1, pp. 132–184. Cambridge University Press, England.

Wang, Jing, Willem Halffman, and Hub Zwart. 2021. "The Chinese Scientific Publication System: Specific Features,Specific Challenges." *Learned Publishing* 34:79–289.

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.

Zhao, Deguo, Yanhou Cai, Lingfeng Wang, and Yongjun Feng. 2009. "Survey Report of Academicians of Chinese Academy of Sciences and Chinese Academy of Sciences."

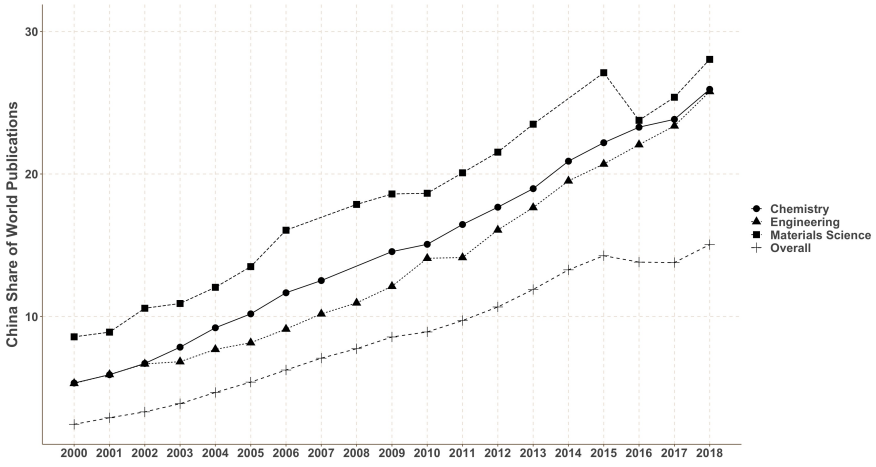


Figure 1: China's share of world publications, 2000-2018

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.

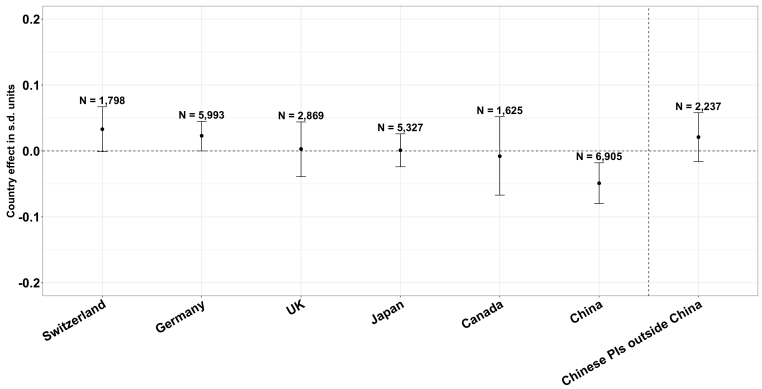


Figure 2: Heterogeneous effects of Chinese PIs on US citations, by source country

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 2. 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: 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, cited PI cumulative publications (13 bins), cumulative citations (13 bins) that each PI's entire corpus of work has received to date, 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.

Table 3: Effect of Chinese investigatorship on the number of US Patent citations [Poisson Model]

	(1)	(2)	(3)	(4)
Chinese investigator	-1.124*	-1.171**	-0.506**	-0.284**
	(0.103)	(0.105)	(0.078)	(0.056)
No. of Articles	658,621	658,621	658,621	651,872
Number of Author Fixed Effects		yes	yes	yes
Publication Year Fixed Effects			yes	yes
Journal Fixed Effects				yes
Pseudo R^2	0.018	0.020	0.166	0.255
% increase	-68%	-70%	-40%	-25%

Note: The dependent variable is the number of US patent citations, i.e., patent citations from all-US inventor teams. 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. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 4: Effect of Chinese investigatorship on the number of US Patent citations [Poisson Model]

Matching on:	article-to-article citations (baseline)		baseline + patent inventor status		baseline + patent stock categories		baseline + patent inventor status + patent-to-article citations	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Chinese Investigator	-0.362*	-0.430*	-0.367*	-0.533*	-0.369†	-0.552*	-0.482*	-0.425†
	(0.159)	(0.197)	(0.182)	(0.236)	(0.197)	(0.254)	(0.216)	(0.224)
Investigator from English-speaking Country		0.030		-0.048		-0.150		0.031
		(0.225)		(0.270)		(0.284)		(0.286)
Investigator with US Training		-0.190		-0.156		-0.196		0.139
		(0.172)		(0.197)		(0.206)		(0.196)
US Cited Author(s)		-0.318		-0.637*		-0.908**		-1.043**
		(0.265)		(0.323)		(0.300)		(0.303)
Subfield Home Research Intensity		-0.090		0.867		0.646		0.553
		(0.809)		(0.915)		(0.969)		(0.895)
Subfield Foreign Research Intensity		-0.791		-1.049		-0.492		0.675
		(0.673)		(1.049)		(1.046)		(0.848)
Subfield USA Research Intensity		1.814		2.370		1.377		-0.936
		(1.322)		(2.003)		(2.138)		(1.981)
Importance of Subfield for Investigator		-0.418†		-0.363		-0.332		-0.093
		(0.239)		(0.229)		(0.227)		(0.211)
Importance of Investigator for the Subfield		-2.249†		-1.718		-1.387		-0.254
		(1.346)		(1.401)		(1.423)		(1.316)
Ellison/Glaeser Index of Scholarly Focus		-4.774		-5.811		-4.080		-1.503
		(3.364)		(4.408)		(4.397)		(5.002)
Retraction-heavy Subfield		0.034		-0.264		-0.432		-1.585**
		(0.404)		(0.479)		(0.584)		(0.484)
Investigator Publication Stock (log)		0.127		0.023		-0.075		-0.552*
		(0.155)		(0.184)		(0.208)		(0.237)
Investigator Citation Stock (log)		-0.046		0.001		0.031		0.310*
		(0.108)		(0.123)		(0.140)		(0.143)
Pseudo R^2	0.178	0.188	0.207	0.217	0.203	0.213	0.189	0.201
Cited Articles	16,192	16,192	16,089	16,089	13,448	13,448	14,915	14,915
Investigators	557	557	545	545	527	527	542	542
% increase	-30%	-35%	-31%	-41%	-31%	-42%	-38%	-35%

Note: The dependent variable is the cumulative number of US patent citations received by a published article, i.e., patent citations from all-US inventor teams. Table D.1, Table G.1, Table G.2 and Table G.3 respectively provide descriptive statistics for control and treated articles that form the estimation samples used in Columns 1a and 1b, columns 2a and 2b, columns 3a and 3b, and columns 4a and 4b. All specifications include fixed effects for the cited article's publication year, journal, number of authors, as well as indicator variables for the investigator's highest degree year and investigator gender (coefficients not reported). Coefficients derive from a Poisson specification estimated via quasi–maximum likelihood. Robust standard errors in parentheses are clustered at the level of the investigator. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

ONLINE APPENDIX FOR:

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

Appendix A: Chinese Chemistry Research in Context

Table A.1: China's highly cited researchers (HCRs) and China's publications across fields

Field	No. of China's HCRs in field	China's share of world HCRs in field (%)	China's share in world publications in field (%)
Chemistry	211	19.27	14.96
Materials Science	205	26.12	19.99
Engineering	164	18.94	14.78
Computer Science	59	9.83	13.13
Physics	57	7.33	14.78
Mathematics	49	10.06	15.93
Geosciences	45	5.77	10.97
Molecular Biology/Biochemistry/Genetics	29	1.36	10.00
Plant/Animal Science	20	2.03	9.79
Agricultural Sciences	18	2.62	9.34
Pharmacology/Toxicology	10	1.44	10.42
Environment/Ecology	9	1.18	10.34
Neuroscience/Behavior	9	1.09	4.88
Microbiology	5	0.88	6.65
Immunology	5	0.83	6.27
Clinical Medicine	2	0.10	5.33
Psychiatry/Psychology	1	0.16	2.24

Notes: (1) Highly Cited Researchers (HCRs) are selected based on their production of multiple highly cited articles that rank in the top 1% by citations in a field and year (in the Web of Science database); (2) we count the number of HCRs of each country without dropping duplicates (i.e., the same person on the HCR list of different years is counted repeatedly) and for researchers who are affiliated with more than one institution, we defined their affiliation (country) based on their primary institution in the year when the HCR report was issued; (3) the share is computed based on English-language research articles published in each field between 2000 and 2015. Articles are attributed to countries on the basis of the share of institutional addresses located in the country, relative to the total number of institutional addresses.

Table A.2: China's research in chemistry compared to other countries/regions

Country/Region	Chemistry			All fields		
	Share of Articles	No. of HCRs	Share of HCRs	Share of Articles	No. of HCRs	Share of HCRs
United States	19.31	471	43.01	25.17	8,306	46.48
China	14.96	211	19.27	9.20	1,104	6.18
Japan	7.66	30	2.74	6.32	420	2.35
Germany	5.55	77	7.03	5.06	1,023	5.73
India	4.68	4	0.37	2.95	27	0.15
United Kingdom	4.20	31	2.83	5.70	1,701	9.52
France	3.94	27	2.47	3.57	474	2.65
South Korea	3.28	33	3.01	2.83	163	0.91
Italy	2.96	3	0.27	3.32	274	1.53
Spain	2.95	26	2.37	2.55	309	1.73
Canada	2.47	20	1.83	3.34	451	2.52
Taiwan	1.49	2	0.18	1.74	86	0.48
Australia	1.48	22	1.83	2.44	581	3.25
Iran	1.37	1	0.09	1.09	46	0.26
Switzerland	1.15	33	3.01	1.08	420	2.35
Netherlands	1.06	7	0.64	1.75	465	2.60
Sweden	0.97	1	0.09	1.23	163	0.91
Belgium	0.74	3	0.27	0.88	197	1.10
Czech	0.67	7	0.64	0.49	26	0.15
Israel	0.59	10	0.91	0.80	58	0.33
Singapore	0.56	19	1.74	0.53	164	0.92
Denmark	0.53	7	0.64	0.68	163	0.91
Hong Kong	0.43	12	1.10	0.57	144	0.81
South Africa	0.29	3	0.27	0.45	32	0.18
Ireland	0.26	5	0.46	0.33	79	0.44
Saudi Arabia	0.25	32	2.92	0.23	272	1.52
Rest of World	16.22	0	0.00	15.71	722	4.04

Notes: (1) We list 26 countries and a residual "rest of the world" category that jointly include all HCRs in chemistry, ranked by the share of chemistry articles they produce; (2) the share of chemistry articles is computed based on English-language original research articles in chemistry during the period 2000-2015. Articles are attributed to countries on the basis of the share of institutional addresses located in the country, relative to the total number of institutional addresses; (3) the figures correspond to *Highly Cited Researchers* reports during the 2014-2018 period, and excludes the social sciences and business categories when tallying the number of HCRs in a country; (4) we count the number of HCRs of each country without dropping duplicates (i.e., the same person on the HCR list of different years is counted repeatedly); for researchers who are affiliated to more than one institution, we defined their affiliation (country) based on their primary institution in the year when the *Highly Cited Researchers* report was issued.

Appendix B: Details on Dataset Construction

Elite chemistry journals. We select 31 journals to create a list of "elite" chemistry publication outlets. The starting point for the list is the 2020 Edition of Clarivate's *Journal Citation Reports*, but the final list displayed in Appendix Table B.1 also reflects a number of *ad hoc* adjustments that were made after consultations with a number of academic chemists.

Scientists and their publications. To generate a list of elite scientists, we tabulate the total number of publications in the 31 chemistry journals mentioned above, between the years 2000 and 2018, counting only "original articles" (i.e., excluding reviews, editorials, etc.) where each scientist appears in last-authorship position, which identifies them as a PI according to the norms prevailing in Chemistry. The aggregation at the individual level is performed leveraging the *author-ity* disambiguation project in PubMed. We then retain the top percentile of the author-level publication count, and eliminate from this list authors affiliated with US institutions.

In order to obtain individual characteristics for them, we contacted each scientist through e-mail requests. In this process, we received 254 responses (32.77% response rate). For those who do not respond, we collected information through their laboratory or institution faculty page, their *Who's Who* profile, and Google searches. This process yielded sufficient information for an additional 497 researchers. 24 investigators were dropped because we couldn't obtain sufficient data for them from any of the sources. Our final sample thus consists of 751 investigators.

We restrict our sample of articles to last-authored articles published after 2000 (because there were very few Chinese scientists actively publishing in international journals before that time) and prior to 2018 (so that every article in the sample will have at least three years to be cited, as our citation data ends in 2021). Although we select scientists on the basis of their publication count in the set of 31 "elite" Chemistry journals, our sample includes all articles published by these scientists, including articles appearing in multi-discipline journals such as *Science*, *Nature*, or the *Proceedings of the National Academy of Sciences*.[i]

Assigning individual scientists to countries. In our sample, a scientist is French not because she was born in France, or educated in France, but because she is affiliated with a French institution. In practice, we use the country that is associated with the most frequent affiliation on publications after career independence to assign each scientists to a unique country. Scientists may move to different countries over the course of their career, and we worry about the selectivity bias that may arise from assigning each publication to the country in which it was produced, since opportunities to move could plausibly be related to productivity. In practice, the country assignment procedure does not matter very much: For 95% of publications, the country of the first faculty employment is the same as the country of the affiliation at the time of publication, and our results are robust to this alternative definition. We also examined whether our results are robust to assigning scientists to the country in which their first non post-doctoral position is located, once again with little change to the results.

Recall that only the cited article's last author is constrained to be a non-US researcher. First- and middle-authors in articles published by non-US PIs can be affiliated with US research institutions, and it is plausible that such coauthorships elevate the propensity of citation by other US-based researchers. In order to assign countries to specific authors, we need to link the address lines listed on articles to authors; this is possible for 80% of the articles in our database. For these articles we find that in most cases (95%), the first authors are linked to the first address record. Therefore, for the remaining 20% of articles for which we cannot accurately assign countries to specific authors, we use the first address line to define the first author's country.

[i] Specifically, 71.11% of articles are published in one of the 31 elite Chemistry journals listed in Appendix Table B.1, whereas 28.89% are published in other journals.

Table B.1: Elite journal list

Journal Name	2020 JIF	No. of Cites from 2020, Received by Articles 2011-2020	JIF without Self Cites	Total Articles 2018-2019
Nature Materials	43.841	112,429	43.542	194
Nature Nanotechnology	39.213	75,845	38.84	143
Nature Chemistry	24.427	41,139	24.235	150
Journal of the American Chemical Society	15.419	609,264	14.394	2,442
Angewandte Chemie-International Edition	15.336	410,750	13.98	3,599
Nature Chemical Biology	15.04	27,429	14.81	171
ACS Central Science	14.553	11,097	14.38	197
Nano Letters	11.189	177,909	10.776	1,148
Chemical Science	9.825	65,945	9.359	1,344
Chemistry of Materials	9.811	124,153	9.311	1,004
Journal of Medicinal Chemistry	7.446	85,946	6.629	889
Analytical Chemistry	6.986	156,738	6.087	1,992
Chemical Communications	6.222	213,293	5.879	3,020
Organic Letters	6.005	107,262	5.203	1,843
Macromolecules	5.985	111,371	5.158	1,055
Chemistry-A European Journal	5.236	110,572	4.867	2,075
Inorganic Chemistry	5.165	103,059	4.62	1,854
Journal of Biological Chemistry	5.157	397,474	4.969	1,295
Dalton Transactions	4.39	80,783	4.004	1,698
Journal of Organic Chemistry	4.354	101,397	3.94	1,501
Langmuir	3.882	125,608	3.523	1,608
Organometallics	3.876	36,815	3.509	494
Journal of Chemical Physics	3.488	231,490	2.739	2,008
Journal of Biochemistry	3.387	9,099	3.288	106
Analytical Biochemistry	3.365	42,956	3.268	303
Journal of Physical Chemistry B	2.991	111,523	2.68	1,130
Bioorganic & Medicinal Chemistry Letters	2.823	42,591	2.702	638
Journal of Physical Chemistry A	2.781	63,916	2.469	1,060
Tetrahedron	2.457	45,774	2.356	452
Tetrahedron Letters	2.415	60,350	2.277	771
Chemical Physics Letters	2.328	50,930	2.127	993

Note: Journals are ranked by Journal Impact Factor according to the 2020 edition of Clarivate's *Journal Citation Report*.

Appendix C: Home-debiased Citations

Measuring and adjusting econometric estimates for variation in the *quality* of scientific articles is a perennial challenge in the economics of science literature. And yet, this study must tackle this challenge head on since it is necessary to maintain the assumption that articles authored by Chinese and non-Chinese PIs offer the same potential for follow-on journals once we have created a matched dataset of articles at risk of being cited. To increase the plausibility of this assumption, our approach is to exploit the propensity of these articles to be cited outside the United Sates (recall that citations from the US will be our outcome variable, and as a result these American citations should play absolutely no role in the matching procedure).

A natural starting point might be to consider that articles published in the same journal and year, and who receive the same amount of citations from non-US sources are of equivalent quality from the standpoint of US-citers. As noted in Section 3.1 of the manuscript, citation home-bias threatens the validity of this approach. To fix ideas, consider two articles published in the journal *Nature* in the same year. Paper A has a Chinese PI, while Paper B has a Spanish PI, and both articles received 100 citations from non-US sources. Because China has a much larger pool of potential citers than Spain, and because Chinese researchers and Spanish researchers might differ in the extent to which they tend to cite locally, we might suspect that these two articles do not offer the same broad shoulders to follow-on US researchers in spite of the apparent parity in citations received from non-US sources.[ii]

One approach would be to exclude citations from home when making these comparisons, and in fact we follow that strategy in a robustness check (see Table C.1). This strikes us as too blunt of an approach in general, as citations received from home surely convey some information regarding the fertility potential of any scientific article. The challenge is to find a procedure that appropriately "handicaps" the non-US citation count by adjusting it to reflect the extent to which in-country citations are abnormally large or "home biased"—a term we borrow from the international trade literature.

To create a reasonable benchmark, we use a dartboard approach and consider a hypothetical situation in which the total citations in the world (where i and j index countries),

$$world_cites \coloneqq \sum_i \sum_j citations_{ij},$$

is distributed across countries according to the size of all citing and cited countries. Then, a benchmark for the citations of country i from country j can be determined by the expected citations of country i's articles by country j's articles based on the share of citations country i receives from the world, $\sum_j citations_{ij} \,/\, world_cites$, multiplied by the size of the potential citing country based on its publication share, $pubs_j \,/\, \sum_k pubs_k$:

$$citations_{ij}^{benchmark} \coloneqq world_cites \times \frac{\sum_j citations_{ij}}{world_cites} \times \frac{pubs_j}{\sum_k pubs_k} \tag{C.1}$$

Figure C.1a plots the benchmark share of home citations according to this expression, i.e., $citations_{ii}^{benchmark} \,/\, \sum_j citations_{ij}$, as a second, light blue bar. China's benchmark share is also large, 18.5%, due to China's size in terms of publications, but what is even more striking is the gap between the benchmark share and the home citation share of total citations: its level is 37.7%, by far the largest of all countries, even though all of them exhibit home bias to some extent.

In practice, we adjust the abnormally large number of home citations by using a country-specific debiasing factor δ_h which we define as the number of benchmark home citations given by equation (C.1) divided by the actual home citations:

$$\delta_h \quad \coloneqq \quad \frac{citations_{hh}^{benchmark}}{citations_{hh}^{actual}}$$

To get to our final quality measure, we therefore change the raw citation measure of quality given by an article's total citations, $citations_p = citations_{p,h} + citations_{p,US} + citations_{p,ROW}$, in two ways. First, we take out citations

[ii] Qiu et al. (2024) document that China exhibits the largest citation home bias across all major countries and in nearly all scientific fields studied.

from the US, and second, we scale home citations by the country-specific debiasing factor δ_h:[iii]

$$quality_p \coloneqq \delta_h citations_{p,h} + citations_{p,ROW}$$

Figure C.2 plots the histogram of home-debiased, non-US citations for the matched sample of articles whose propensity to be cited in the US we will examine later.

A description of the matching procedure which takes the count of home-debiased citations as an input follows in the next section. However, our estimation results are not very sensitive to the particular manner in which we match (or do not match) on cited-article citations. Table C.1 replicates the benchmark specification reported in the second column of Table 2. Although we only report the coefficient estimate and standard errors for the main China effect, all other covariates are in fact included across all columns displayed. The first column reproduces the results of Table 2 exactly, matching cited articles on ROW plus home citations debiased as explained above. The second column generates a new matched sample excluding all citations from home when matching. The third column uses a sample that results from a match on all non-US citations (i.e., home citations are not discounted in this analysis). The fourth column uses a sample that results from a matching procedure that does not take article-level citations into account at all.

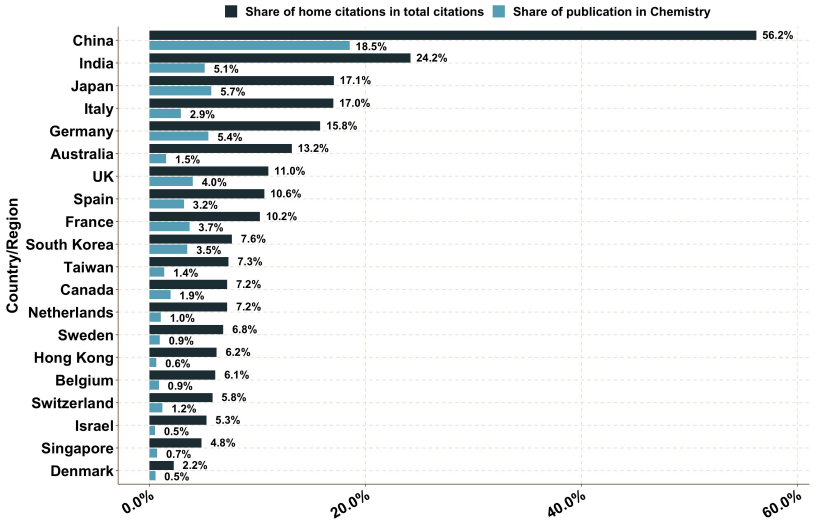
The fifth column adds two additional covariate to the benchmark list of matching covariates, but leaves out citations: (i) the presence of funding acknowledgement (regardless of funder) in the cited article (as a binary indicator variable), and (ii) the cumulative number of "top publications" for the PI, defined as those appearing in Nature, Science, and PNAS, which are the most prestigious general-interest journals.[iv] We coarsen the count into 6 distinct bins: zero top publication, exactly one top publication, exactly two top publications, three or four top publications, five or six top publications, or 7 or more top publications.

The sixth column investigates whether matching on ROW citations artificially inflates the measured quality of articles from some countries, owing to the size of their scientific diaspora. This concern is especially salient for China in light of the large number of expatriate Chinese scientists in the English-speaking world and beyond (although this could be true for other countries as well). We ascertain an author's ethnicity using Torvik and Agarwal's (2016) *Ethnea* algorithm and dataset, which infers a name's ethnicity from the frequency of affiliation locations for that name in *PubMed* using a multinomial logistic model. For each ROW citing paper, we look at the authors' names and we check whether one or more authors share the same ethnicity as the PI of the cited paper. Whenever that is the case, we exclude the citing article from the stock of ROW citations. We then build our matched sample of article anew, leaving everything else the same. Using this modified ROW citation count, we can examine the magnitude of the China effect in the benchmark specification after purging the ROW matching citations from citing authors belonging to PI's home country diaspora.

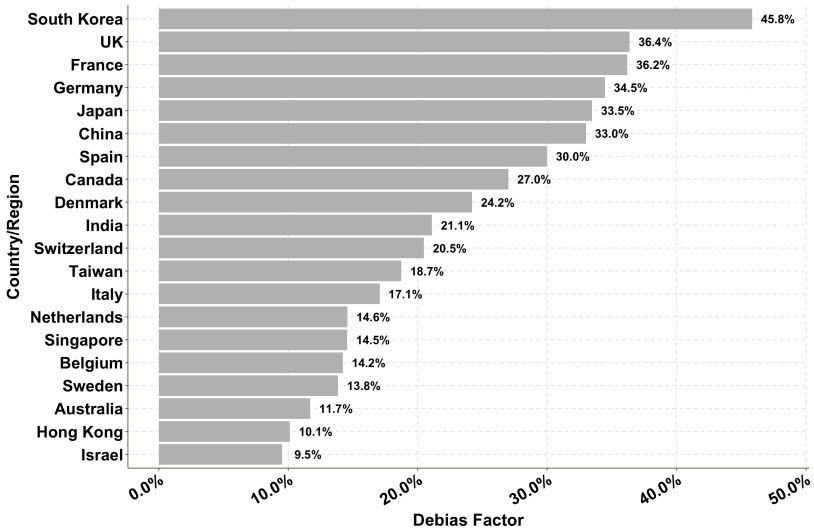
As can be readily observed, the coefficients and standard errors are remarkably similar across these six specifications, highlighting the robustness of the core finding.

[iii] The distribution of δ_h across countries is given in Figure C.1b.

[iv] In these three outlets, chemists do not only compete with other chemists for journal space; they also compete with physicists, biologists, etc. Our belief is that the cumulative count of past publications in these three journals is a crude but viable way to distinguish "superstar PIs" from mere "star PIs" (recall that every PI in the sample belongs to the academic elite in the sense that are in the top percentile of the distribution according to the overall number of publications appearing in about 30 top chemistry journals between 2000 and 2018).



(a) Actual home citation share versus benchmark share



(b) Debias factors by country/region

Figure C.1: Home bias in citations

Note: These figures include the top 20 countries according to citations; in the analysis all countries are included.

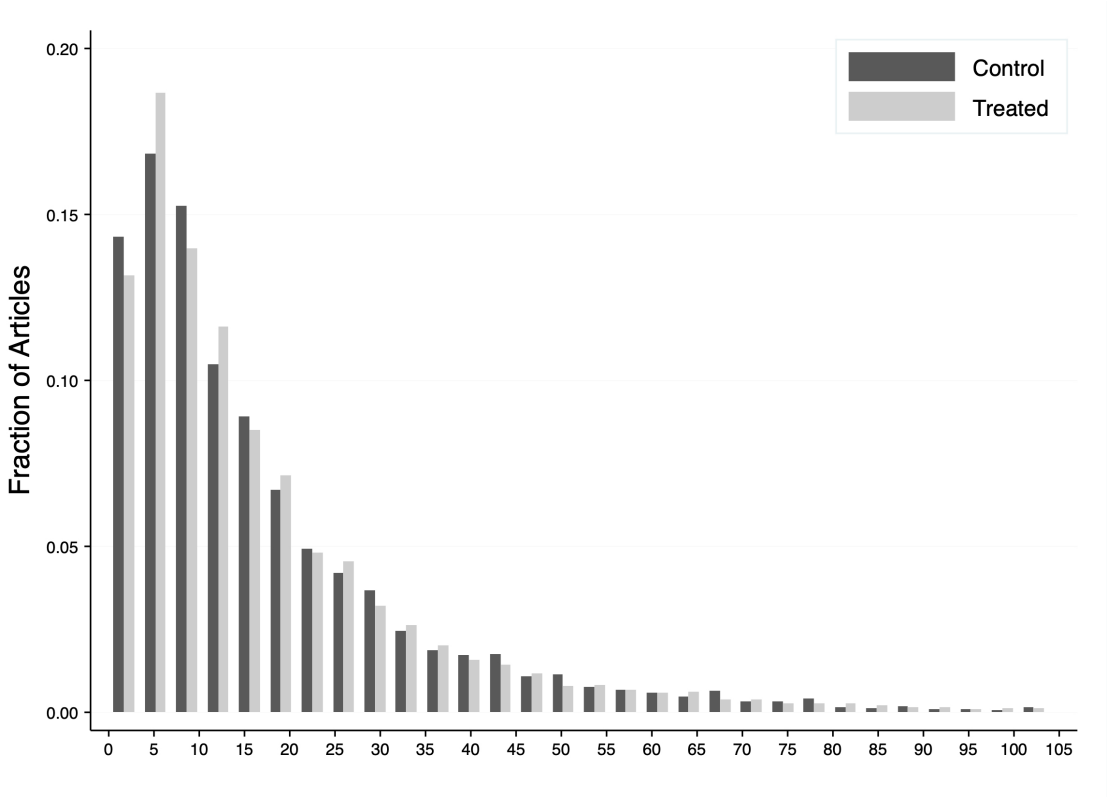


Figure C.2: Distribution of non-US citations for control and treated articles

Note: The histogram for the number of home-debiased citations excluding US citations is displayed above. The histogram excludes publications with 104 or more citations (approximately 1% of the sample).

Table C.1: Matching based on different quality measures

	Match on Debiased Citations Outside US	Match on Citations Excluding Home	Match on All Citations Outside US	Do not Match on Citations	Match on Funded Publication + PI's Top Publications	Match on Home-debiased Citations+ ROW non-US citations - ROW diaspora citations
Chinese Investigator	-0.009**	-0.006*	-0.012**	-0.009**	-0.012**	-0.009**
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Mean of Dependent Variable	0.032	0.031	0.035	0.035	0.034	0.031
s.d. of Dependent Variable	0.177	0.173	0.184	0.184	0.180	0.175
China effect in s.d. units	-0.049	-0.034	-0.065	-0.049	-0.066	-0.052
Adjusted R^2	0.084	0.079	0.087	0.069	0.072	0.081
No. of Investigators	557	560	553	567	521	559
No. of Cited Articles	16,192	15,502	16,103	27,065	13,963	16,180
No. of Citing Articles	71,409	70,819	71,903	100,819	69,120	71,450
No. of Citing/Cited Article Pairs	188,753	180,947	192,754	343,864	181,507	188,293

Note: The dependent variable is an indicator variable that equals 1 if the related article cites the source article, and 0 otherwise. Fixed effects included as in the baseline, i.e., column 2 of Table 2. Standard errors in parentheses are two-way clustered at the investigator and strata level. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Appendix D: Matched sample of cited articles and citation risk set

Creating a matched sample of articles at risk of being cited. We create the matched sample using a Coarse Exact Matching procedure (CEM). The procedure is coarse because we do not attempt to precisely match on matching covariate values; rather, we divide the support of the joint distribution of the covariates into a finite number of strata, and we match a treated observation if and only if a control observation can be recruited from this strata. An important advantage of CEM is that the analyst can guarantee the degree of covariate balance *ex ante*, but this comes at a cost: the more fine-grained the partition of the support for the joint distribution (i.e., the higher the number of strata), the larger the number of unmatched treated observations.

To illustrate, an important matching covariate in our design is the number of home-debiased citations (excluding US citations). For the coarsened matching we divide this measure into six bins: $0-25^{th}$ percentile; $25-50^{th}$ percentile; $50-75^{th}$ percentile; $75-95^{th}$ percentile; $95-99^{th}$ percentile; and the top percentile.[v]

Table D.1 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.

Citation risk set. To facilitate the harvesting of citation risk sets in *PubMed* on a large scale, we have developed an open-source software tool that queries *PubMed* and PMRA and stores the retrieved data in a MySQL database. The software is available for download at `http://www.stellman-greene.com/FindRelated/`. Prior research leveraging the intellectual linkages between articles generated by PMRA include Azoulay et al. (2015, 2019), and Myers (2020). Appendix C in Azoulay et al. (2019) describes the algorithm in detail. Importantly, the PMRA algorithm does *not* take citations into account—it only takes as input titles words, abstract words, and keywords.

For each source publication, the citation risk set consists of topically-related articles published after the source. On average, only 10% of these articles actually cite the source. The proportion of actual citation linkages that correspond to intellectual linkages in the sense of PMRA does not vary much across country of origin, as indicated by Figure D.1. Moreover, the citation risk set excludes any actual citing article that is not topically related to the source in the sense of PMRA. Table D.2 shows that leaving these unrelated citations in the risk set does not alter our substantive conclusions.

Inclusion of control covariates in the econometric specifications. All the coefficients reported in this article reflect estimates net of the inclusion of a large number of baseline covariates. These include fixed effects for: (a) each strata defined in the coarsened exact matching algorithm (the interaction of our bins for journal × publication year × number of authors × debiased non-US citations × PhD degree year; yielding 3,847 bins); (b) investigator cumulative publication (13 bins), (c) investigator cumulative citations (13 bins), (d) investigator gender, (e) topic similarity rank bins corresponding to the topical similarity between each source article and each potentially citing article in its risk set,[vi] (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.

[v] When creating these bins, we compute a separate empirical distribution of citations for each publication year.

[vi] We group similarity ranks between articles into 26 bins, with finer bins for rankings below 150 (10 ranks per bin), and gradually create larger bins for higher ranks (151-170, 171-190, 191-210, 211-230, 231-260, 261-300, 301-350, 351-450, 451-650, 651-1000, above 1000).

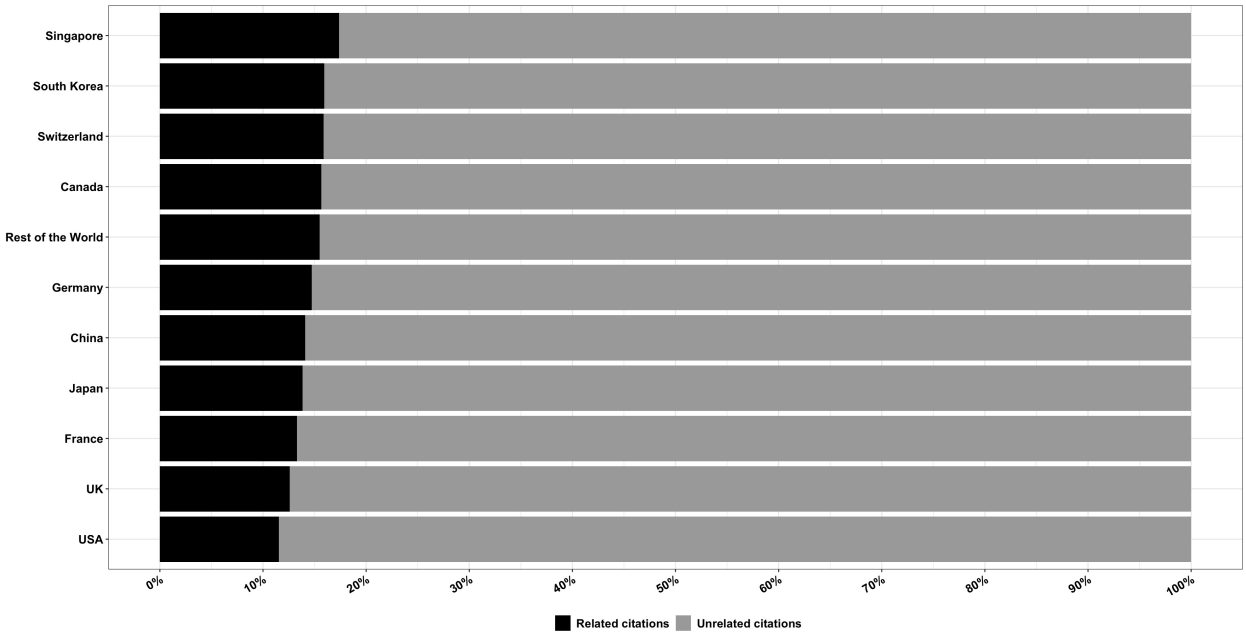


Figure D.1: Ratio of related citations to overall citations, by country

Table D.1: 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 D.2: Include unrelated actual citations in risk set

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Chinese Investigator	-0.006**	-0.006**	-0.009**	-0.006*	-0.007**	-0.006**	-0.009*
	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.004)
Communication							
Investigator from English-speaking Country		-0.001	-0.000	-0.001	-0.001	-0.001	-0.000
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Log (Avg. Distance)		0.002	0.001	0.002	0.001	0.001	0.001
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Network							
US First/Reprinted Cited Author		0.009	0.011	0.009	0.009	0.009	0.011
		(0.008)	(0.010)	(0.008)	(0.008)	(0.008)	(0.009)
US Cited Author in Other Positions		0.003	0.001	0.003	0.003	0.003	0.002
		(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)
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		0.000	-0.003+	0.000	0.000	0.000	-0.003+
		(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Citing Author is Investigator's Past Collaborator		0.039**	0.039**	0.039**	0.039**	0.039**	0.039**
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Common Coauthor		0.095**	0.095**	0.095**	0.095**	0.095**	0.095**
		(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Investigator is an Editorial Author		-0.001	0.002	-0.001	-0.002	-0.001	0.002
		(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)
Geographic Topical Clustering							
Subfield Home Research Intensity		0.008	0.010	0.033	0.009	0.009	0.037
		(0.007)	(0.007)	(0.037)	(0.007)	(0.007)	(0.037)
Subfield Foreign Research Intensity		-0.021**	-0.022**	-0.026**	-0.022**	-0.022**	-0.027**
		(0.006)	(0.006)	(0.008)	(0.006)	(0.006)	(0.008)
Subfield USA Research Intensity		0.035*	0.036*	0.045+	0.036*	0.035*	0.049+
		(0.017)	(0.017)	(0.027)	(0.017)	(0.017)	(0.026)
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		0.009	0.009	0.009	0.009	0.009	0.010
		(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.009)
Ellison/Glaeser Index of Scholarly Focus		0.060	0.057	0.060	0.054	0.060	0.054
		(0.043)	(0.043)	(0.042)	(0.052)	(0.043)	(0.052)
Reputation							
Retraction-heavy Subfield		-0.003	-0.003	-0.003	-0.003	-0.000	-0.000
		(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Interactions with Network							
Chinese Investigator × Investigator with US Training			0.004				0.004
			(0.003)				(0.003)
Chinese Investigator × US First/Reprinted Cited Author			-0.015				-0.014
			(0.016)				(0.016)
Chinese Investigator × US Cited Author in Other Positions			0.005				0.005
			(0.007)				(0.007)
Chinese Investigator × Log (Cumulative No. of US Coauthorships)			-0.001				-0.001
			(0.001)				(0.001)
Chinese Investigator * Citation from Same Ethnicity			0.005**				0.006**
			(0.002)				(0.002)
Chinese Investigator × Investigator is an Editorial Author			-0.015				-0.015
			(0.015)				(0.015)
Interactions with Geographic Topical Clustering							
Chinese Investigator × Subfield Home Research Intensity				-0.028			-0.030
				(0.040)			(0.039)
Chinese Investigator × Subfield Foreign Research Intensity				0.012			0.012
				(0.013)			(0.013)
Chinese Investigator × Subfield USA Research Intensity				-0.026			-0.031
				(0.030)			(0.030)
Interactions with Investigator's Intellectual Focus							
Chinese Investigator * Importance of Subfield for Investigator					0.004		0.004
					(0.003)		(0.003)
Chinese Investigator × Importance of Investigator for the Subfield					0.000		-0.002
					(0.019)		(0.019)
Chinese Investigator × Ellison/Glaeser Index of of Scholarly Focus					0.017		0.007
					(0.084)		(0.086)
Interactions with Reputation							
Chinese Investigator × Subfield Retraction Intensity						-0.005	-0.006
						(0.005)	(0.005)
Mean of Dependent Variable	0.185	0.185	0.185	0.185	0.185	0.185	0.185
s.d. of Dependent Variable	0.388	0.388	0.388	0.388	0.388	0.388	0.388
China effect in s.d. units	-0.014	-0.016	-0.023	-0.015	-0.019	-0.016	-0.024
Adjusted R^2	0.837	0.838	0.838	0.838	0.838	0.838	0.838
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	80,088	80,088	80,088	80,088	80,088	80,088	80,088
No. of Citing/Cited Pairs	226,456	226,456	226,456	226,456	226,456	226,456	226,456

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. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Appendix E: Measurement of Covariates

This appendix provides details on the construction of some important covariates used as controls or as interaction effects with the effect of Chinese PI-ship.

Geographic distance. To capture citation frictions induced by geographic separation, we compute the log average distance between the PI city of affiliation and all cities listed in address fields for the citing article.

Citing authors' ethnic origins. To identify the ethnic origin of US-based scholars, we map a scholar's last name to its ethnic origins based on the algorithm pioneered by Nguyen (2019). Her novel algorithm computes the probability that a last name corresponds to a particular ethnic origin based on the de-anonymized full population samples of US Censii between 1910 and 1940. A last name can be assigned to multiple origin countries with different probabilities, which are based on the relative frequencies of surname-origin pairs appearing in the population samples of US Censii. We use the country with the maximum probability to identify the ethnic origin of each last name. More details regarding the name-ethnicity mapping can be consulted in Nguyen (2019).

Investigator focus. Inspiring ourselves from Ellison and Glaeser (1997), we propose an index which 'normalizes' the PI's topical distribution of articles by the topical distribution of articles in the underlying population. For each PI and year, we consider the set of last-authored articles published by the PI in the previous five years, and assign each article to several research topics m based on the MeSH (Medical Subject Headings) keywords. Define N_{it} as the number of last-authored articles that PI i published in the five years before t; $Share_{imt}$ as the share of the MeSH terms of the N_{it} articles that are assigned to MeSH research topic m, and x_{mt} as the share of MeSH terms m in publications that appeared in year t. We define the Ellison-Glaser index as:

$$eg_index_{it} \coloneqq \frac{N_{it}}{N_{it}-1} \times \frac{\sum_m (Share_{imt} - x_{mt})^2}{1 - \sum_m x_{mt}^2} - \frac{1}{N_{it}-1} \tag{E.1}$$

Retraction-heavy subfields. There is widespread skepticism regarding the quality of Chinese academic research (e.g., Xie and Freeman 2019). Research undertaken by Chinese scientists is sometimes alleged to consist of imitations or replications rather than original research, and other times assumed to be less well written (National Science Board, 2020). However, the systematic evidence is considerably more ambiguous than anecdotes would imply. Packalen (2019) uses Natural Language Processing techniques to analyze the novelty of research across nations, and concludes that Chinese research has in large part caught up to the world scientific frontier. Conversely, Bornmann et al. (2018) show that Chinese research is less relevant for elite publications, as measured by their prevalence (relative to China's publication share) in the references of highly-cited articles.

In addition, we attempt to measure the wedge between *perceived* quality and *actual* quality with subfield-related information, namely, the extent to which an article's subfield is "retraction-heavy." Recall that for each cited article (control or treated), we have at our disposal a set of intellectual neighbors in the sense of PMRA. Confining ourselves to the subset of these neighbors published up to the year of publication for each cited article,[vii] we count the number of instances where one of these neighbor articles is retracted, which receives a specific flag in the *PubMed* metadata. These events are rare (the average is 0.039, with a minimum of 0 and a maximum of 5 in our sample), so in practice we dichotomize the covariate, with each cited article in a subfield containing at least one retracted neighbor (3.67% of the articles) or zero (96.33% of the articles).

[vii] Although we only count retractions of articles that were published before the source, the retraction event can occur either before, in the same year, or after the publication of the source.

Appendix F: Additional results and robustness checks

Heterogeneity of China citation discount over time. 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 presents estimates of the China 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.

One might expect the China citation discount to have grown in magnitude since 2018, when the U.S. Department of Justice started the China Initiative which resulted in investigations of several hundred researchers that were collaborating with Chinese scientists (Jia et al., 2022). In fact, the point estimate in 2018 in Figure F.1 is negative, but we would need to wait several more years in order to study this effect systematically.

Heterogeneity by cited article quality. We ask whether the magnitude of the China citation discount varies with the underlying quality of the article. Are articles who appear in the most prestigious journals—or receive the most citations from non-US sources—impervious to the frictions we have documented in the sample on average?

We measure quality in two alternative ways in Figure F.2. In Figure F.2a we estimate heterogeneous effects by journal quality, as measured by its Journal Impact Factor (JIF). Notice that since our PIs are star scientists, the majority of the articles we consider are published in the top decile. Nonetheless, the discount is driven by journals in the top 5%.[viii] In Figure F.2b we assess quality by the debiased citations it received from non-US sources. We interact the Chinese PI indicator with indicator variables for the position of the focal article in the citation distribution of all other articles with the same publication year. We create 11 percentile bins, allowing for more heterogeneity at the top of the distribution.[ix] To allow for comparability of the estimates across percentile bins—which have a different baseline probability of being cited by US authors—we plot the coefficients measured in units of standard deviations. We find that the bias is especially pronounced in the middle of the distribution, i.e., articles that lie between the 50th and 70th percentile of the impact distribution, but the effect is negative throughout.

Heterogeneity by cited PI seniority. One may also expect that the China citation discount would be dampened in the case of more senior PIs, as awareness of a PI's research would increase the longer they had been active in the research community. In Table F.1 we proxy for seniority in different ways. Columns 1 and 2 proxy for seniority using the age of the PI. One may expect that the bias falls as PIs get older and better known; and this seems to be true, as the bias is not significantly different from zero for PIs who completed their PhD before 2014. However, columns 3-6 show a different pattern when we measure seniority based on the stock of publications or total citations; in this case, the bias is stronger for the more accomplished PIs.

Heterogeneity by Chinese institution prestige. Since our finding seems to point to a discount strongly tied to location in China, a natural follow-up is to ask whether it applies to Chinese investigators regardless of institution. Out of the 155 Chinese investigators in the estimation sample, 49 are employed by one of China's eight elite universities,[x] 53 by the Chinese Academy of Sciences (a country-wide network of intramural research institutions similar in spirit to the Max Planck Institutes in Germany), and 53 in other institutions. Table F.2 breaks the overall China effect across these three types of institutions. The evidence points to the discount applying across the board, although the discount for CAS researchers is smaller and more imprecisely estimated.

Heterogeneity by *citing* article prestige. One can gain some insight into the relevance of the China citation discount for the global research frontier by examining whether it disproportionately manifests itself in influential scientific journals. Figure F.4 parses out the China effect by honing in on citations from US articles appearing in journals belonging to different quality tiers, as measured by their Journal Impact Factor (JIF). We find the effect to

[viii] The top 5% of Chemistry journals according to JIF are *Journal of the American Chemical Society*, *Angewandte Chemie International Edition*, *Nature Chemistry*, *Analytical Chemistry*, *Accounts of Chemical Research*, *Chemistry*, *Chemical Science*, *Chemical Society Reviews*, *The Journal of Physical Chemistry Letters*, *Nano Letters*, *Nanoscale*, *Nature Protocols*, *Small*, *Nucleic Acids Research* and *ChemSusChem*.

[ix] More specifically, we create percentile bins at the following cutoffs: 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, and 95th percentile.

[x] Tsinghua University, Peking University, Fudan University, Zhejiang University, Nanjing University, Shanghai Jiao Tong University, University of Science and Technology of China, and Wuhan University.

be uniformly negative, with very similar magnitudes across the board. While these missed intellectual connections have the potential to slow down scientific progress, more research is needed to quantify these effects.

Citation dynamics. In Figure F.3 we restrict (potential) citing articles to be published within 3 or 5 years of the focal article. One may expect the citation discount to become stronger, the larger the time window we consider for citations to take place, if initial underciting amplifies over time. However, the pattern looks fairly stable as we increase the time window from 3 years to 5 years to all years in our data, but standard errors become smaller as we accrue observations in the estimation sample.

Another look at elite PIs with Chinese names outside China. In Table F.3 we test whether elite PIs that have Chinese names but are working outside of China (but not in the US) experience the same bias. Interestingly, the bias is specific to PIs working in China, and not present for PIs with Chinese names that undertake their research in other countries. The difference between the coefficients is statistically significant across all models of this table (in column (2) for example, the p-value of the corresponding t-test is 0.012).[xi]

[xi] PIs with Chinese names that are working in non-Chinese research institutions may not be suitable members of the control group. In Table F.4, we check whether our results change when dropping these PIs from the estimation sample. Reassuringly, we find that our results are unaffected.

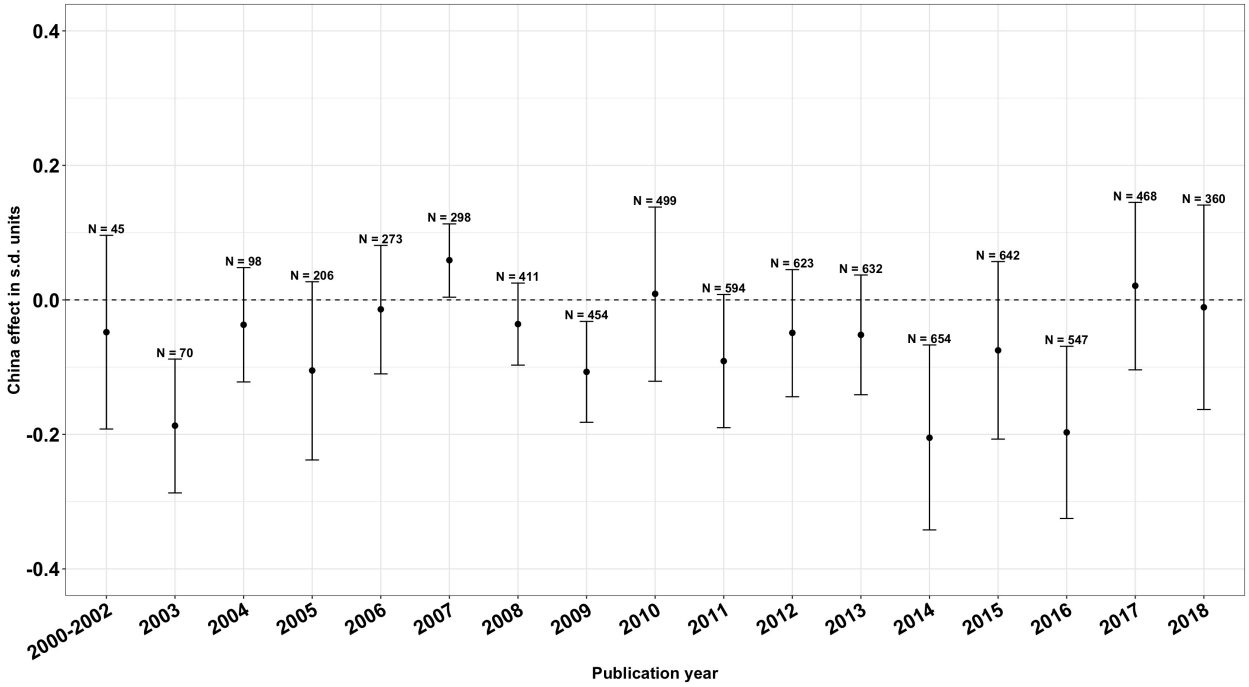
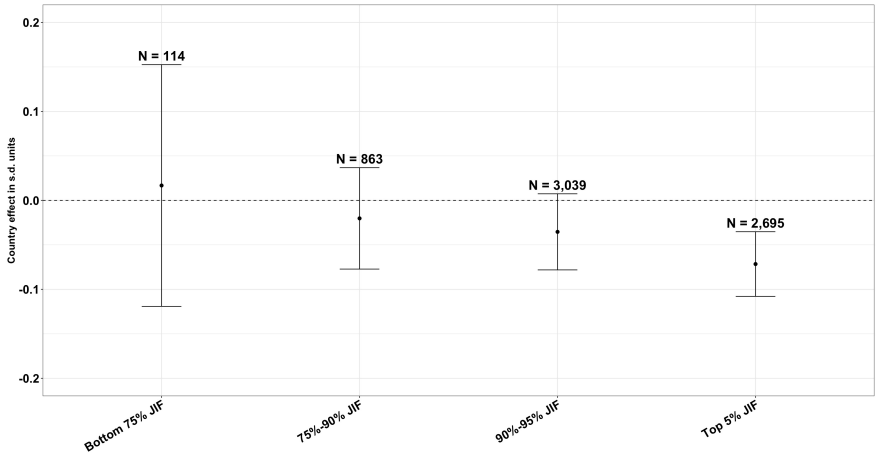
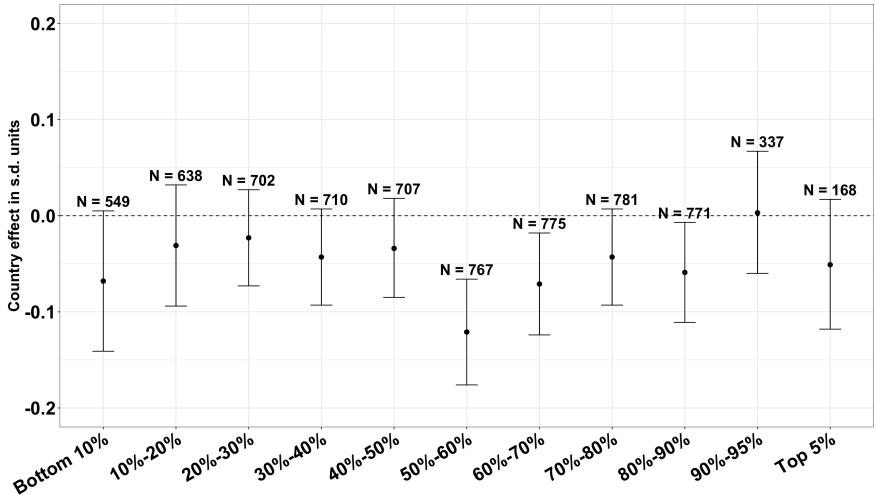


Figure F.1: The China effect over time



(a) by Journal Impact Factor (JIF)



(b) by home debiased citations outside US

Figure F.2: Heterogeneous effect of Chinese PIs on US citations, by cited article quality

Note: The dark dots in the above plots correspond to coefficient estimates stemming from a linear probability model in which the dependent variable is an indicator variable that equals 1 if the related article cited the source article, and 0 otherwise. In subfigure (a), the covariates of interest are 4 interaction terms between the China indicator variable and indicator variables for five quantiles of the distribution of source articles' Journal Impact Factor (JIF). The figure only shows source articles in chemistry journals, though an interaction between the China effect and an indicator variable for non-chemistry journals is included in the corresponding specification. In subfigure (b), the covariates of interest are twelve interaction terms between the China indicator variable and indicator variables for various quantiles of the distribution of home-debiased non-US citations received. In both figures, the corresponding specification also includes all the covariates included in column 2 of Table 2. 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 subgroup is indicated above the corresponding coefficient estimate.

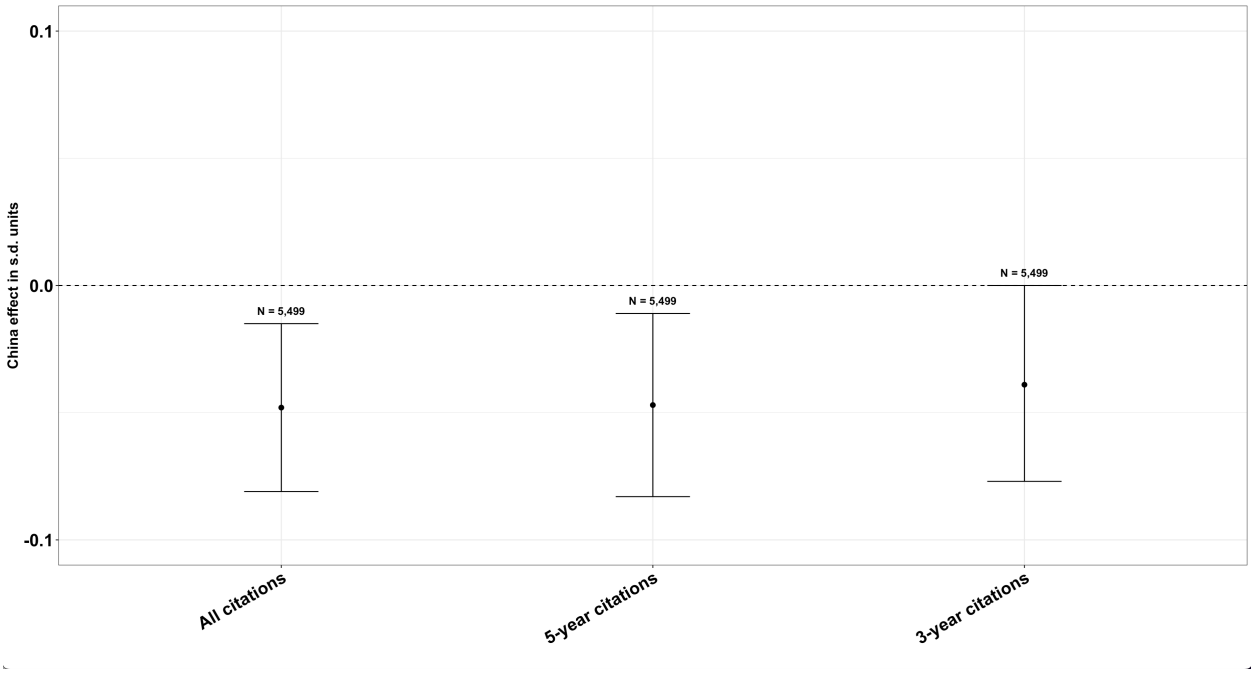


Figure F.3: Restricting citations to different time windows after publication

Note: We estimate the same specification as in column 2 of Table 2. We restrict the sample to cited articles that were published between 2000 and 2016, and consider either citations from all publication after the publication of the focal article, or from publications that occur within 5 or 3 years after the publication of the focal article, respectively.

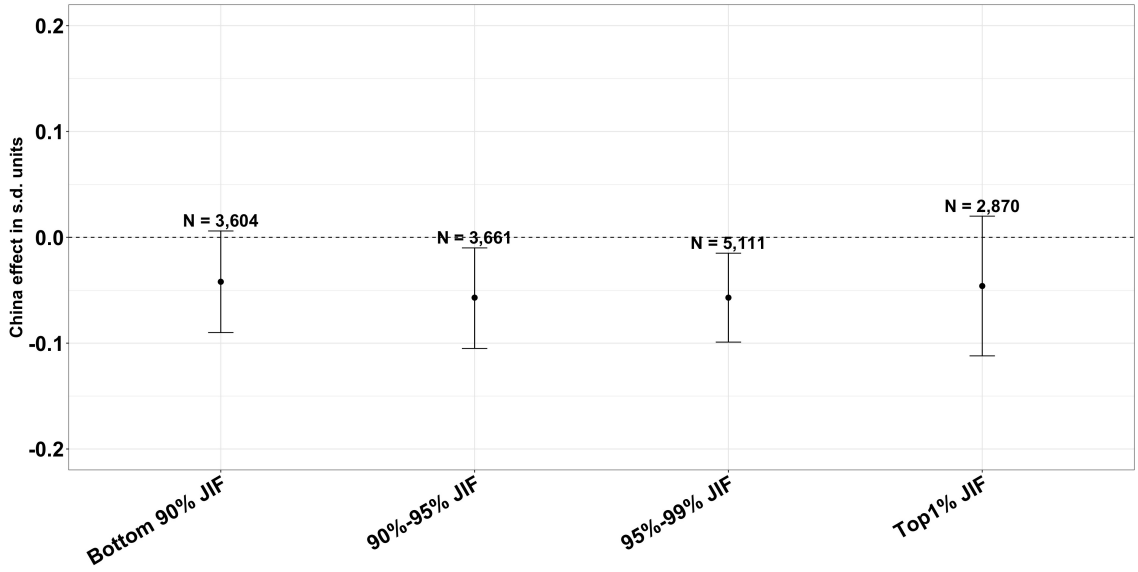


Figure F.4: Heterogeneous effects of Chinese PIs on US citations, splitting citing journals by Journal Impact Factor

Note: The dark dots in the above plots correspond to coefficient estimates stemming from a linear probability model in which the dependent variable is an indicator variable that equals 1 if the related article cited the source article, and 0 otherwise. The regressions are respectively based on four subsamples that splits citing journals into four quantiles of the distribution of citing articles' Journal Impact Factor (JIF). The corresponding specification also includes all the covariates included in column 2 of Table 5. 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 subgroup is indicated above the corresponding coefficient.

Table F.1: Heterogeneity by investigator seniority

	PhD Year		Publication Stock		Citation Stock	
	after 2014	before 2014	below Median	above Median	below Median	above Median
Chinese Investigator	-0.012**	-0.005	-0.005†	-0.015**	-0.003	-0.016**
	(0.005)	(0.003)	(0.003)	(0.006)	(0.003)	(0.006)
Mean of Dependent Variable	0.035	0.030	0.027	0.042	0.026	0.042
s.d. of Dependent Variable	0.183	0.171	0.163	0.201	0.159	0.201
China Effect in s.d. Units	-0.067	-0.029	-0.032	-0.074	-0.022	-0.078
Adjusted R^2	0.087	0.081	0.075	0.116	0.066	0.111
No. of Investigators	238	319	527	488	541	498
No. of Cited Articles	7,843	8,349	8,281	7,683	7,757	8,259
No. of Citing Articles	42,023	51,612	54,428	35,301	53,811	35,259
No. of Citing/Cited Article Pairs	82,629	106,121	127,994	60,521	119,719	68,851

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. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table F.2: Heterogeneity by investigator affiliation

	(1)	(2)	(3)	(4)	(5)
All Chinese PIs	-0.009**				
	(0.003)				
Chinese PIs, Elite 8 Only		-0.012**			-0.013**
		(0.004)			(0.003)
Chinese PIs, CAS Only			-0.002		-0.005†
			(0.004)		(0.003)
Chinese PIs, Non-elite				-0.013**	-0.009*
				(0.005)	(0.004)
Mean of Dependent Variable	0.032	0.036	0.032	0.031	0.032
s.d. of Dependent Variable	0.177	0.185	0.176	0.174	0.177
Adjusted R^2	0.084	0.104	0.085	0.088	0.084
No. of Investigators	557	426	407	400	557
No. of Cited Articles	16,192	6,287	7,082	6,451	16,192
No. of Citing Articles	71,409	41,631	44,611	36,865	71,409
No. of Citing/Cited Article Pairs	188,753	74,733	82,180	68,025	188,753

Note: The dependent variable is an indicator variable that equals 1 if the related article cites the source article, and 0 otherwise. All specifications as in the baseline, i.e., column 2 of Table 2. Standard errors in parentheses are two-way clustered at the investigator and strata level. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table F.3: China discount for researchers based in China vs. researchers with a Chinese name

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Chinese Investigator	-0.008**	-0.008**	-0.011**	-0.008*	-0.010**	-0.008**	-0.012*
	(0.002)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.005)
Chinese PIs outside China	0.000	0.004	0.004	0.004	0.004	0.004	0.004
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Communication							
Investigator from English-speaking Country		-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log(Avg. Distance)		0.003	0.002	0.003	0.003	0.003	0.002
		(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	0.005
		(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**	0.167**	0.166**	0.166**	0.166**	0.167**
		(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Investigator is an Editorial Author		-0.002	-0.001	-0.003	-0.003	-0.002	-0.001
		(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)
Geographic Topical Clustering							
Subfield Home Research Intensity		0.010	0.011	0.039	0.010	0.010	0.045
		(0.009)	(0.009)	(0.048)	(0.009)	(0.009)	(0.048)
Subfield Foreign Research Intensity		-0.022**	-0.023**	-0.029**	-0.022**	-0.022**	-0.030**
		(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		0.065	0.055	0.065	0.045	0.065	0.043
		(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				0.003
			(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				-0.036			-0.038
				(0.050)			(0.050)
Chinese Investigator × Subfield Foreign Research Intensity				0.019			0.018
				(0.016)			(0.016)
Chinese Investigator × Subfield USA Research Intensity				-0.036			-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					-0.013		-0.015
					(0.033)		(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						-0.005	-0.006
						(0.007)	(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	-0.044	-0.047	-0.071	-0.047	-0.056	-0.046	-0.075
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 and a investigator gender indicator variable (coefficients not reported). Standard errors in parentheses are two-way clustered at the investigator and strata level. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table F.4: Removing investigators with Chinese names that work outside China

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Chinese Investigator	-0.008**	-0.007*	-0.010*	-0.007*	-0.010*	-0.006*	-0.013*
	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)
Communication							
Investigator from English-speaking Country		-0.001	-0.001	-0.001	-0.001	-0.001	-0.000
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log(Avg. Distance)		0.000	-0.000	0.000	0.000	0.000	-0.000
		(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
Network							
Investigator with US Training		0.005*	0.002	0.005*	0.005*	0.005*	0.001
		(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
US First/Reprinted Cited Author		0.012	0.016	0.012	0.012	0.012	0.017
		(0.011)	(0.013)	(0.010)	(0.010)	(0.011)	(0.012)
US Cited Author in Other Positions		0.004	0.003	0.004	0.004	0.004	0.003
		(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)
Log(Cumulative No. of US Coauthorships)		0.001	0.002	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		0.051**	0.051**	0.051**	0.051**	0.051**	0.051**
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Common Coauthor		0.169**	0.169**	0.168**	0.169**	0.169**	0.169**
		(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Investigator is an Editorial Author		-0.002	-0.002	-0.002	-0.003	-0.002	-0.002
		(0.011)	(0.013)	(0.011)	(0.011)	(0.011)	(0.013)
Geographic Topical Clustering							
Subfield Home Research Intensity		0.005	0.007	0.064	0.007	0.007	0.071
		(0.009)	(0.009)	(0.049)	(0.009)	(0.009)	(0.049)
Subfield Foreign Research Intensity		-0.022**	-0.023**	-0.034**	-0.023**	-0.023**	-0.036**
		(0.008)	(0.008)	(0.011)	(0.008)	(0.008)	(0.012)
Subfield USA Research Intensity		0.047*	0.049*	0.060†	0.048*	0.046*	0.064†
		(0.022)	(0.022)	(0.033)	(0.022)	(0.022)	(0.033)
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*	0.034†	0.033*	0.032
		(0.016)	(0.016)	(0.016)	(0.019)	(0.016)	(0.019)
Ellison/Glaeser Index of Scholarly Focus		0.036	0.026	0.036	0.000	0.036	0.001
		(0.055)	(0.056)	(0.055)	(0.066)	(0.055)	(0.065)
Reputation							
Retraction-heavy Subfield		-0.002	-0.002	-0.002	-0.002	0.005	0.005
		(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)
Interactions with Network							
Chinese Investigator × Investigator with US Training			0.009*				0.009*
			(0.004)				(0.004)
Chinese Investigator × US First/Reprinted Cited Author			-0.028				-0.027
			(0.019)				(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			0.007†				0.008†
			(0.004)				(0.004)
Chinese Investigator × Investigator is an Editorial Author			-0.002				-0.002
			(0.020)				(0.020)
Interactions with Geographic Topical Clustering							
Chinese Investigator × Subfield Home Research Intensity				-0.070			-0.072
				(0.050)			(0.050)
Chinese Investigator × Subfield Foreign Research Intensity				0.029†			0.029†
				(0.017)			(0.017)
Chinese Investigator × Subfield USA Research Intensity				-0.039			-0.042
				(0.037)			(0.037)
Interactions with Investigator's Intellectual Focus							
Chinese Investigator × Importance of Subfield for Investigator					0.006†		0.005†
					(0.003)		(0.003)
Chinese Investigator × Importance of Investigator for the Subfield					-0.003		-0.001
					(0.034)		(0.034)
Chinese Investigator × Ellison/Glaeser Index of of Scholarly Focus					0.094		0.067
					(0.107)		(0.109)
Interactions with Reputation							
Chinese Investigator × Retraction-heavy Subfield						-0.011†	-0.013†
						(0.007)	(0.007)
Mean of Dependent Variable	0.033	0.033	0.033	0.033	0.033	0.033	0.033
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	-0.038	-0.065	-0.040	-0.052	-0.035	-0.077
Adjusted R^2	0.078	0.083	0.083	0.083	0.083	0.083	0.083
No. of Investigators	531	531	531	531	531	531	531
No. of Cited Articles	15,122	15,122	15,122	15,122	15,122	15,122	15,122
No. of Citing Articles	68,963	68,963	68,963	68,963	68,963	68,963	68,963
No. of Citing/Cited Article Pairs	175,776	175,776	175,776	175,776	175,776	175,776	175,776

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 and a investigator gender indicator variable (coefficients not reported). Standard errors in parentheses are two-way clustered at the investigator and strata level. $^{\dagger}p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Appendix G: Leveraging article citations from patents

Tables G.1, G.2, and G.3 provide descriptive statistics corresponding to three alternative ways to pair Chinese with non-Chinese articles while incorporating information regarding PIs' involvement in patenting. In Table G.1, we require that the PIs have the same patent inventor status (an indicator variable) in addition to a match on the baseline set of covariates as in Table D.1. Table G.2 imposes a more stringent requirement by breaking down the patenting stock up to the focal article's year of publication into four groups: no patent application, exactly one patent application, exactly two patent applications, and three or more patent applications. Finally, Table G.3 reflects the most demanding set of requirements: not only must the articles share the same baseline covariates and the same patent inventor status, they must also match based on a coarse measure of home-debiased patent-to-article citations from outside the United States.[xii] Specifically, we match on articles' home-debiased patent citations from outside the US, split into four bins: 0, 1, 2, or 3+ citations.

[xii] This follows a method similar to that used to adjust for home bias in article citations. To generate the patent-to-article citation benchmark, we follow equation (1) but replace the article-to-article citations $citations_{ij}$ with patent-to-article citations from country i's articles to country j's patents, and the publication share $pubs_j \,/\, \sum_k pubs_k$ with the country j's granted patent share in the USPTO data.

Table G.1: Summary statistics for control and treated articles matched on baseline covariates & a patent inventor status indicator

	Control articles (N= 8,810)					Treated articles (N= 7,279)					Difference	
	Mean	Median	Std. Dev.	Min	Max	Mean	Median	Std. Dev.	Min	Max	Mean	Median
Characteristics used for matching:												
Source Article Publication Year	2012.571	2013	3.768	2000	2018	2012.571	2013	3.768	2000	2018	0	0
Source Article No. of Authors	5.082	5	1.923	1	15	5.230	5	1.861	1	15	-0.148	0
Investigator PhD Degree Year	1993.737	1994	5.712	1958	2008	1993.940	1994	5.628	1952	2007	-0.203	0
No. of Debiased Article-to-Article Citations Outside US	15.701	10	18.723	0	328.856	15.251	9.867	17.345	0	345.264	0.450	0.133
Investigator is an Inventor	0.307	0	0.461	0	1	0.307	0	0.461	0	1	0	0
Characteristics not used for matching:												
No. of Total Article-to-Article Citations	20.554	14	23.133	0	387	28.400	18	32.565	0	580	-7.846**	-4*
No. of Article-to-Article Citations Outside US	17.711	11.750	20.058	0	333.300	25.969	16.607	29.793	0	521.303	-8.258**	-4.857*
No. of Debiased Patent-to-Article Citations Outside US	0.091	0	0.644	0	27.310	0.067	0	0.515	0	25	0.024*	0
US First/Reprinted Cited Author	0.019	0	0.135	0	1	0.006	0	0.076	0	1	0.013**	0
US Cited Author in Other Positions	0.035	0	0.183	0	1	0.013	0	0.115	0	1	0.022**	0
Retraction-heavy Subfield	0.027	0	0.163	0	1	0.040	0	0.196	0	1	-0.013**	0
Subfield Foreign Research Intensity	15.940	12.961	14.777	0	170.860	13.150	10.808	11.946	0	148.523	2.790**	2.153*
Subfield Home Research Intensity	1.507	0.710	2.288	0	38.926	10.752	7.704	11.512	0	158.569	-9.245**	-6.994*
Subfield USA Research Intensity	3.739	2.511	4.422	0	64.233	3.825	2.514	4.731	0	76.577	-0.086	-0.003
Importance of Investigator for the Subfield	0.063	0.038	0.078	0	1	0.053	0.031	0.069	0	1	0.010**	0.007*
Importance of Subfield for Investigator	0.111	0.045	0.285	0	10	0.113	0.036	0.328	0	7.333	-0.002	0.009*
Ellison/Glaeser Index of Scholarly Focus	0.016	0.014	0.014	-0.026	0.270	0.014	0.013	0.011	-0.020	0.142	0.002**	0.001*
Female Investigator	0.042	0	0.200	0	1	0.047	0	0.212	0	1	-0.005	0
Investigator with US Training	0.580	1	0.494	0	1	0.462	0	0.499	0	1	0.118**	1
Investigator from English-speaking Country	0.329	0	0.470	0	1	0	0	0	0	0	0.329**	0
Investigator is Editorial Author	0.036	0	0.187	0	1	0.010	0	0.101	0	1	0.026**	0
Investigator Publication Stock	86.758	74	63.579	0	1,068	99.953	82	75.845	0	482	-13.195**	-8*
Investigator Article-to-Article Citation Stock	1,676.827	1033	1,970.622	0	25,629	1,550.776	1004	1,700.989	0	12,811	126.051**	29*
Investigator Patent Stock	2.666	0	8.442	0	144	1.954	0	8.107	0	102	0.712**	0
Investigator Patent-to-Patent Citation Stock	15.531	0	78.958	0	1,581	4.481	0	17.065	0	185	11.050**	0

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. The baseline set of matching covariates includes journal, publication year, number of authors (four bins: 1-3, 4-6, 7-9, 10 or more authors), home-debiased article-to-article citations outside US (six bins: 0-25th percentile; 25-50th percentile; $50-75^{th}$ percentile; 75-95th percentile; 95-99th percentile; and the top percentile), and the investigator PhD degree year (in three-year bins). In addition, the two groups of articles are matched on patent inventor status, i.e., an indicator variable equal to one if the PI has applied for at least one patent prior to the publication year of source article.

Table G.2: Summary statistics for control and treated articles matched on baseline covariates & investigator patent stock

	Control articles (N= 7,211)					Treated articles (N= 6,237)					Difference	
	Mean	Median	Std. Dev.	Min	Max	Mean	Median	Std. Dev.	Min	Max	Mean	Median
Characteristics used for matching:												
Source Article Publication Year	2,012.391	2013	3.802	2000	2018	2,012.391	2013	3.802	2000	2018	0	0
Source Article No. of Authors	5.009	5	1.844	1	15	5.158	5	1.781	1	15	-0.149	0
Investigator PhD Degree Year	1,993.796	1994	5.641	1958	2008	1,993.988	1994	5.571	1952	2007	-0.192	0
No. of Debiased Article-to-Article Citations Outside US	15.478	9.930	18.641	0	328.856	15.010	9.709	17.248	0	345.264	0.468	0.221
Investigator Patent Stock	1.738	0	7.441	0	144	1.675	0	7.891	0	90	0.063	0
Characteristics not used for matching:												
No. of Total Article-to-Article Citations	20.288	13	23.026	0	387	27.901	18	32.345	0	580	-7.613**	-5*
No. of Article-to-Article Citations Outside US	17.467	11.333	19.968	0	333.300	25.501	16.222	29.538	0	521.303	-8.034**	-4.889*
No. of Debiased Patent-to-Article Citations Outside US	0.089	0	0.658	0	27.310	0.065	0	0.500	0	25	0.024*	0
US First/Reprinted Cited Author	0.019	0	0.136	0	1	0.005	0	0.073	0	1	0.014**	0
US Cited Author in Other Positions	0.035	0	0.184	0	1	0.012	0	0.110	0	1	0.023**	0
Retraction-heavy Subfield	0.028	0	0.164	0	1	0.039	0	0.193	0	1	-0.011**	0
Subfield Foreign Research Intensity	15.944	12.897	14.782	0	170.860	13.183	10.778	12.109	0	148.523	2.761**	2.119*
Subfield Home Research Intensity	1.513	0.705	2.297	0	33.455	10.640	7.501	11.414	0	158.569	-9.127**	-6.796*
Subfield USA Research Intensity	3.771	2.511	4.449	0	64.233	3.805	2.503	4.735	0	76.577	-0.034	0.008
Importance of Investigator for the Subfield	0.063	0.039	0.078	0	1	0.055	0.032	0.071	0	1	0.008**	0.007*
Importance of Subfield for Investigator	0.119	0.049	0.304	0	10	0.121	0.038	0.351	0	7.333	-0.002	0.011*
Ellison/Glaeser Index of Scholarly Focus	0.017	0.014	0.015	-0.026	0.270	0.014	0.013	0.011	-0.020	0.142	0.003**	0.001*
Female Investigator	0.040	0	0.197	0	1	0.038	0	0.192	0	1	0.002	0
Investigator with US Training	0.581	1	0.493	0	1	0.443	0	0.497	0	1	0.138**	1
Investigator from English-speaking Country	0.325	0	0.469	0	1	0	0	0	0	0	0.325**	0
Investigator is Editorial Author	0.036	0	0.186	0	1	0.009	0	0.094	0	1	0.027**	0
Investigator is an Inventor	0.192	0	0.394	0	1	0.192	0	0.394	0	1	0	0
Investigator Publication Stock	83.350	71	62.728	0	1,068	97.746	80	75.292	0	482	-14.396**	-9*
Investigator Article-to-Article Citation Stock	1,581.116	925	1,966.909	0	25,629	1,469.358	928	1,636.790	0	12,811	111.758**	-3
Investigator Patent-to-patent Citation Stock	10.573	0	69.895	0	1,581	3.775	0	15.864	0	185	6.798**	0

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. The baseline set of matching covariates includes journal, publication year, number of authors (four bins: 1-3, 4-6, 7-9, 10 or more authors), home-debiased article-to-article citations outside US (six bins: 0-25th percentile; 25-50th percentile; 50-75th percentile; 75-95th percentile; 95-99th percentile; and the top percentile), and the investigator PhD degree year (in three-year bins). In addition, the two groups of articles are matched on the indicator variable for the investigator's patent stock up to the year of publication of the source article, divided into four coarse bins: zero patents, exactly one patent, exactly two patents, and three or more patents.

Table G.3: Summary statistics for control and treated articles matched on baseline covariates, patent inventor status, and non-US patent-to-article citations

	Control articles (N= 8,072)					Treated articles (N= 6,843)					Difference	
	Mean	Median	Std. Dev.	Min	Max	Mean	Median	Std. Dev.	Min	Max	Mean	Median
Characteristics used for matching:												
Source Article Publication Year	2,012.793	2013	3.677	2000	2018	2,012.793	2013	3.677	2000	2018	0	0
Source Article No. of Authors	5.093	5	1.925	1	15	5.246	5	1.857	1	15	-0.153	0
Investigator PhD Degree Year	1,993.807	1994	5.616	1958	2008	1,994.011	1994	5.518	1952	2007	-0.204	0
No. of Debiased Article-to-Article Citations Outside US	14.702	9.621	17.229	0	328.856	14.251	9.422	15.397	0	223.677	0.451	0.199
No. of Debiased Patent-to-Article Citations Outside US	0.004	0	0.086	0	4.473	0.003	0	0.082	0	4	0.001	0
Characteristics not used for matching:												
No. of Total Article-to-Article Citations	19.269	13	21.265	0	387	26.681	17	29.589	0	404	-7.412**	-4*
No. of Article-to-Article Citations Outside US	16.664	11	18.505	0	333.300	24.478	16	27.231	0	392.157	-7.814**	-5*
US First/Reprinted Cited Author	0.019	0	0.136	0	1	0.006	0	0.075	0	1	0.013**	0
US Cited Author in Other Positions	0.035	0	0.183	0	1	0.013	0	0.111	0	1	0.022**	0
Retraction-heavy Subfield	0.028	0	0.164	0	1	0.039	0	0.194	0	1	-0.011**	0
Subfield Foreign Research Intensity	15.825	12.890	14.684	0	170.860	12.991	10.673	11.851	0	148.523	2.834**	2.217*
Subfield Home Research Intensity	1.498	0.710	2.267	0	33.455	10.789	7.716	11.568	0	158.569	-9.291**	-7.006*
Subfield USA Research Intensity	3.640	2.455	4.275	0	64.233	3.735	2.450	4.664	0	76.577	-0.095	0.005
Importance of Investigator for the Subfield	0.063	0.039	0.078	0	1	0.054	0.032	0.070	0	1	0.009**	0.007*
Importance of Subfield for Investigator	0.109	0.045	0.284	0	10	0.110	0.034	0.319	0	7.333	-0.001	0.011*
Ellison/Glaeser Index of Scholarly Focus	0.016	0.014	0.014	-0.026	0.270	0.014	0.013	0.011	-0.020	0.142	0.002**	0.001*
Female Investigator	0.042	0	0.201	0	1	0.047	0	0.211	0	1	-0.005	0
Investigator with US Training	0.579	1	0.494	0	1	0.463	0	0.499	0	1	0.116**	1
Investigator from English-speaking Country	0.329	0	0.470	0	1	0	0	0	0	0	0.329**	0
Investigator is Editorial Author	0.039	0	0.193	0	1	0.011	0	0.102	0	1	0.028**	0
Investigator Publication Stock	88.644	76	64.117	0	1,068	102.533	84	76.418	0	482	-13.889**	-8*
Investigator Article-to-Article Citation Stock	1,729.294	1080	1,997.227	0	25,629	1,602.790	1,067	1,722.249	0	12,811	126.504**	13*
Investigator Patent Stock	2.599	0	8.179	0	144	1.879	0	7.724	0	102	0.720**	0
Investigator Patent-to-Patent Citation Stock	14.798	0	76.915	0	1,581	4.481	0	17.241	0	185	10.317**	0

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. The baseline set of matching covariates includes journal, publication yearnumber of authors (four bins: 1-3, 4-6, 7-9, 10 or more authors), home-debiased article-to-article citations outside US (six bins: 0-25th percentile; 25-50th percentile; 50-75th percentile; 75-95th percentile; 95-99th percentile; and the top percentile), and the investigator's PhD degree year (in three-year bins). In addition, the two groups of articles are matched on patent inventor status (as in Table G.1), as well as home-debiased patent-to-article citations outside the US, divided into four coarse bins: zero citation, 1 citation, 2 citations, and 3 or more citations.

References

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin. 2019. "Does Science Advance One Funeral at a Time?" *American Economic Review* 109:2889–2920.

Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics* 97:1118–1136.

Bornmann, Lutz, Caroline Wagner, and Loet Leydesdorff. 2018. "The Geography of References in Elite Articles: Which Countries Contribute to the Archives of Knowledge?" *PLoS one* 13:e0194805.

Ellison, Glenn and Edward L. Glaeser. 1997. "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy* 105:889–927.

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.

Myers, Kyle. 2020. "The Elasticity of Science." *American Economic Journal: Applied Economics* 12:103–34.

National Science Board. 2020. "Research and Development: U.S. Trends and International Comparisons."

Nguyen, Kieu-Trang. 2019. "Trust and Innovation within the Firm: Evidence from Matched CEO-Firm Data." Working Paper, Northwestern University.

Packalen, Mikko. 2019. "Edge Factors: Scientific Frontier Positions of Nations." *Scientometrics* 118:787–808.

Qiu, Shumin, Claudia Steinwender, and Pierre Azoulay. 2024. "Paper Tiger? Chinese Science and Home Bias in Citations." NBER Working Paper #32468.

Torvik, Vetle I. and Sneha Agarwal. 2016. *Ethnea—An Instance-based Ethnicity Classifier Based on Geo-coded Author Names in a Large-scale Bibliographic Database*. International Symposium on Science of Science, Library of Congress, March 22-23, Washington, DC.

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.