Research Evaluation, 2024, **33**, rvae036
https://doi.org/10.1093/reseval/rvae036
Article

The anglicization of science in China

Kai Li [1], Xiang Zheng [2] and Chaoqun Ni [2,*,†]

[1]School of Information Sciences, University of Tennessee, 1345 Circle Park Drive, Knoxville, TN, 37996, United States
[2]Information School, University of Wisconsin-Madison, 600 N Park St, Madison, WI, 53706, United States

*Corresponding author. Email: chaoqun.ni@wisc.edu.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

The preeminence of English as the lingua franca in global science has led to English-dominant publication practices, even in non-English-speaking countries. We examine the complex dynamics of language use in scientific publications in China, a major contributor to global scientific output, and the tensions between English and the native language. By analyzing 2,209,987 multilingual publications from 183,457 projects funded by the National Natural Science Foundation of China, we reveal a strong preference for English as the publication language in China, with 66.2% of publications in English versus 33.8% in Chinese. Key projects and natural sciences and engineering projects favor English more; regional projects and social sciences projects use Chinese more. However, English has a growing prevalence over the years across all research fields, project types, and publication venues. There is a negative correlation between the shares of English usage in journals and conference proceedings. We find only a minor overlap between English and Chinese-language publications, indicating unique contributions rather than repetitive content. However, Chinese-language publications are more likely to be similar to English-language publications. For highly similar cross-language publication pairs, the Chinese version tends to be published earlier. The findings underscore the risk of underestimating China's scientific output by only counting English-language publications. We highlight the importance of creating a comprehensive multilingual database and the significant role of non-English-language research in global scientific discourse.

Keywords: multilingual publishing; scientific communication; language policy; academic multilingualism; bibliometrics; quantitative analysis.

1. Introduction

Since the latter half of the 20th century, English has become the de facto lingua franca of science, even in nations where English is not the primary language (Gordin 2015; Ramírez-Castañeda 2020). The existence of a lingua franca in sciences is critical for global scientific knowledge communication and dissemination, enabling scientists worldwide to share, access, and build their work upon a vast body of scientific literature (Baldauf 2001; Valkimadi et al. 2009; Mongeon and Paul-Hus 2016). The dominance of the English language encourages research policies that reward the use of this language (Aagaard, Bloch and Schneider 2015; Fejes and Nylander 2017; Mathies, Kivistö and Birnbaum 2020). However, scientific discoveries and advancements are not limited to English alone (González-Alcaide, Valderrama-Zurián, and Aleixandre-Benavent 2012; Amano, González-Varo and Sutherland 2016). Compared with English, native languages hold immense value in conveying and preserving indigenous knowledge and research findings and connecting research to local communities and native speakers (Kulczycki et al. 2020). Publications written in native languages are often situated in local social contexts and thus provide nuanced findings for local issues. Moreover, publishing in native languages allows researchers from peripheral language systems to contribute without language barriers, which creates possibilities for fair representation of science (Lewison 2009; Dann 2011; Faraldo-Cabana 2018). Therefore, it is vital to recognize and promote multilingual research beyond English for a more inclusive, vibrant, and efficient research environment (Uzuner 2008; Sivertsen 2018).

As one of the non-English-speaking leading producers of scientific publications (Heilbron and Gingras 2018; Tollefson 2018; Wagner, Zhang and Leydesdorff 2022; Baker 2023), China is an interesting yet underexplored actor in how different academic languages are used by researchers. Over the past decades, the world has witnessed an exponential growth of scientific publications produced by scientists in China, as documented by major English-based bibliographic databases such as the Web of Science (WoS) and Scopus (Zhou, Su and Leydesdorff 2010; Tollefson 2018). Meanwhile, a considerable amount of research in China was still published in Chinese. China established its native language publication system and databases, including Chinese Academic Journals and China National Knowledge Infrastructure (CNKI), providing native-language outlets for research dissemination. In 2020, China published ∼452,000 scientific papers in domestic Chinese journals, which mostly were in Chinese, and 553,000 papers in Science Citation Index (SCI) journals, which mostly were in English (Ministry of Science and Technology of the People's Republic of China 2022). Within this context, it becomes essential to explore how China reconciles the tension between utilizing the globally accessible English language and promoting the use of their native languages in the publication system as these countries strive to make their contributions known on a global scale.

Meanwhile, it is unclear exactly to which extent China's research has been anglicized and how Chinese researchers publish in multiple languages, mainly English and Chinese (Clarke et al. 2007). This gap in China is particularly notable in scientometrics and research evaluation, given that China has surpassed the United States as the top producer of scientific papers (Tollefson 2018). This research gap persists primarily due to the absence of data infrastructure covering publications in both Chinese and English. This study delves into the status of science anglicization in China to further understand its role in global science, based on publications from projects funded by the National Natural Science Foundation of China (NSFC). NSFC is China's primary government research funder for natural sciences and engineering (Yang 2016). These government-endorsed projects are representative of Chinese top high-quality STEM research to a large extent. Using 2,209,987 publications by projects funded by 183,457 NSFC projects between 2010 and 2015, we produced a cartographic overview, as conceptualized by Gingras (2016), of the market share of English and the native language, Chinese, used in China's scientific publications and how these patterns vary by time, field, project type, and venues, with a strong focus on understanding how the distribution of the two languages is correlated with the other factors. In addition, we investigate the overlap between local science (published in Chinese) and international science (published in English) in China using a natural language processing approach. Our approach based on the NSFC outcome database overcomes the disadvantage of the significant overrepresentation of English-language publications in common bibliographic databases.

As the first attempt of the kind, this study sheds light on the current multilingual publication practices of researchers in China, particularly in the STEM fields, and provides insights for policymakers to understand the language preferences of researchers and formulate corresponding policies. It also carries significant implications for the global scientific community. By comprehending the language diversity in scientific publishing, researchers, policymakers, and funding agencies can more effectively assess the breadth and depth of China and similar countries' scientific achievements, enabling greater collaboration, knowledge exchange, and integration of native research into global scientific endeavors.

2. Literature review

2.1 Multilingual publishing practices in non-English speaking countries

The publishing practices in multilingual contexts within non-English-speaking countries have garnered attention from bibliometric researchers. Most studies predominantly rely on bibliographic databases such as WoS (Liu 2017; Koch and Vanderstraeten 2019; Rao, Xia and Li 2020; Mironescu, Moroșanu and Bibiri 2023), with some incorporating multiple datasets, including national databases (Kulczycki et al. 2018, 2020).

Existing analysis indicates an increasing use of English in academic publishing in non-English-speaking countries. Using WoS data, Koch and Vanderstraeten (2019) demonstrated that the proportion of English-language publications in Chile has increased from 1976 to 2015 in both national and international journals. Mironescu, Moroșanu and Bibiri (2023) observed that while WoS linguistic journals in central and eastern Europe founded between 1950 and 1975 have maintained their multilingual identity, those founded between 2000 and 2010 have adopted English as the primary language of publication. Warchał and Zakrajewski (2023) utilized survey data from a case study university to reveal that a significant proportion of social sciences researchers and a considerable but smaller proportion of humanities scholars disseminate their research results in English. According to WoS data analyzed by Rao, Xia and Li (2020), publications in Chinese by Chinese scholars do not exceed one-third of those in English, and there was a general downward trend in the ratio of Chinese to English-language publications over time. Kulczycki et al. (2018, 2020) combined multiple national datasets from selected European countries, finding significant variations in the share of publications in English, with a high of 68% in Finland and a low of 17% in Poland.

Language practices also differ across disciplines. Koch and Vanderstraeten (2019) found that several Chilean social sciences and humanities journals favor Spanish for publication, targeting a continental audience with regionally relevant topics. Kulczycki et al. (2018, 2020) noted that local languages are extensively used in social sciences and humanities publications throughout Europe. Liu (2017) highlighted that in WoS, unlike in the natural and social sciences, non-English papers have consistently played a significant role in the arts and humanities since 1975. Rao, Xia and Li (2020) found that Chinese scholars publish more WoS-indexed publications in Chinese in fields such as physics, linguistics, and philosophy.

The literature discussed drivers for the increase in English-language publications in non-English-speaking countries. One of the crucial drivers is the national and institutional research evaluation practices and policies, which encourage and reward the use of the English language to gain international recognition (Heilbron and Gingras 2018). Many evaluation practices depend on cross-country bibliographic databases, which offer limited coverage of non-English languages and indigenous works (Van Leeuwen et al. 2001; Mongeon and Paul-Hus 2016; Fejes and Nylander 2017). Studies have examined the impact of performance-based national research evaluation policies and funding pressures on the increase in English publications (Ossenblok, Engels and Sivertsen 2012; Aagaard, Bloch and Schneider 2015; Korytkowski and Kulczycki 2019; Mathies, Kivistö and Birnbaum 2020). Under the policy context, researchers' considerations for language selection, such as enhancing international visibility and meeting evaluation criteria, have also been analyzed (López-Navarro et al. 2015; Stockemer and Wigginton 2019; Zheng and Guo 2019; Warchał and Zakrajewski 2023).

Most efforts to investigate multilingual publication practices have had to rely on limited datasets, where local publications are often not indexed. This reliance on constrained datasets and the insufficient infrastructure for gathering comprehensive publication samples have impeded a deeper understanding of the global landscape of scientific knowledge production. Moreover, as one of the top research powerhouses, China has not been sufficiently examined. One challenge is that local Chinese-language publication databases often restrict access to large data quantities (Xia, Wright and Adams 2008). Consequently, these technical obstacles make large-scale empirical studies on this subject remarkably challenging, with only a few exceptions focusing on specific research fields or universities to examine language usage patterns in China (Zhou, Su and Leydesdorff 2010; Cui and Zhang 2018; Wei and Zhang 2020). The deficiency in infrastructure constrains the capacity to perform an exhaustive assessment of the research activities undertaken by Chinese researchers and China as a nation. This is due to the challenges, and potentially the impossibility, of gathering all research publications in various languages produced by

Chinese scholars. This practical challenge serves as a major motivation for the present research.

2.2 China's research evaluation system

This research is situated within the Chinese academic system. China's modern higher education system was rebuilt in the late 1970s after the Cultural Revolution's disruptions (Horta and Shen 2020). Since then, the Chinese government has significantly invested in higher education institutions and activities, as evidenced by three national programs promoting university research: Project 211 and Project 985 were launched in the 1990s, and the Double First-Class Initiative was launched in 2015 (see Shu, Sugimoto and Larivière 2021 for definition and history review). These national programs aim to improve the research strength and international competitiveness of top Chinese Universities (Zong and Zhang 2019; Wei and Zhang 2020; Shu, Sugimoto and Larivière 2021). However, they have been criticized for adopting an elitist model that concentrates resources on a few selected universities, exacerbating inequalities and inefficiencies within the Chinese higher education system (Mohrman and Wang 2010; Ying 2011; Shu, Sugimoto and Larivière 2021).

Internationalization has become a prominent pursuit of the Chinese higher education system over the past few decades, following the 'open door' policy implemented by the Chinese government after 1978 (Huan 1986; Wei 1995). This trend is closely associated with efforts to expand scientific impact in esteemed international journals. China has increased its investment in scientific research and its internationalization progress (Marginson 2022; National Science Board, National Science Foundation 2022). As a result, SCI as well as other similar indices based on English-language publications were introduced into the Chinese higher education system and established an authoritative framework for evaluating Chinese researchers' performance (Shao and Shen 2012; Qian et al. 2020). The initial goal of introducing SCI to China was to enhance the visibility of research by Chinese researchers and to improve the quality of peer review in China (Fu, Frietsch and Tagscherer 2013; Qian et al. 2020). However, publishing in SCI journals has widely become a requirement of PhD graduation, faculty hiring, promotion, and tenure at top-tier universities in China, particularly in STEM fields (Shu et al. 2020; Zheng et al. 2022). Additionally, direct monetary rewards for publications further incentivize English-language publications in SCI journals (Shao and Shen 2012; Quan, Chen and Shu 2017). These policies foster an institutional environment that encourages the use of English (Flowerdew and Li 2009; Xu 2020; Xu, Oancea and Rose 2021).

Although not the focus of this study, a notable policy shift in the Chinese research system occurred at the beginning of the 2020s. China's top leader Xi Jinping called on Chinese scientists to 'write scientific papers on the soil of the motherland' and contribute to the national interests by localizing their research outcomes. Accordingly, the Chinese government issued policy documents discouraging the use of SCI for evaluating research performance (Ministry of Education Ministry of Science 2020), aiming to extend earlier reforms to the Chinese higher education system (Chinese Communist Party Central Committee State Council 2018). As a result, researchers, especially those funded by the government, are encouraged to publish their papers in Chinese-language journals, enabling fast research access for Chinese audiences. However, these changes will not be reflected in our data, as our dataset does not cover publications beyond 2020.

3. Data and methods

We obtained the dataset from the NSFC output portal in August 2021 (National Natural Science Foundation of China 2023). This portal is the official platform for NSFC Principal Investigators (PIs) to self-report progress and outcomes in all forms and languages, including research articles, books, reports, and patents, at the end of their projects. According to NSFC, accurately and comprehensively self-reporting research outcomes is strictly required by the NSFC. The final review and assessment of the projects are primarily based on the research outcomes reported. Every research outcome must meet three criteria: it should be authored by the PI or project participants, directly relate to the funded project, and explicitly acknowledge the financial support provided by the NSFC (National Natural Science Foundation of China 2021a). Failure to report the information on the portal may significantly affect the review results and delay the project's completion. Therefore, we believe the current dataset has reliable coverage for the years when mandatory reporting is in place. This nature of the data makes it appropriate and reliable for understanding the research landscape of NSFC funding. The platform documents critical metadata information for NSFC-funded projects and outcomes, including project title, duration, funding amount, project type, PI information, and outcome details. This study focuses on journal and conference publications as the primary outcome of research projects, which account for 90.5% of the total project outcomes as reported in the database. As the NSFC outcome portal only documents the complete information for projects funded from 2010 onwards, we focused on projects awarded by NSFC between 2010 and 2015, which means projects included in our sample ended in or before 2020 (The most prolonged duration for an NSFC project is 5 years). This study focuses on the following dimensions of the data for subsequent analyses.

Project disciplinary field: We analyzed the research outcomes of the NSFC project by the eight disciplinary departments (disciplinary fields) in NSFC at the time the data was collected, including the departments of *Mathematics and Physical Sciences (MPS)*, *Chemical Sciences (Chem)*, *Life Sciences (Life)*, *Earth Sciences (Earth)*, *Engineering & Material Science (EMS)*, *Information Sciences (Info)*, *Management Sciences (MS)*, and *Health Sciences (Health)*. It is important to note that the *Information Sciences* in NSFC is strongly situated in the knowledge domain of computer science, while *Management Sciences* houses many social science disciplines within the scope of NSFC, such as economics, public administration, sociology, and library and information science. Additionally, there are various subdepartments focusing on smaller research fields under each disciplinary department. For instance, *Mathematics* and *Physical Sciences* has 30 sub-departments, making it the largest department in terms of the number of sub-departments. In comparison, the Department of *Management Sciences* only has four sub-departments. We refer to the disciplinary department by NSFC as fields and subdepartments as subfields in this study.

Project types: The various disciplinary fields within NSFC provide funding for different project types, also referred to as project classes by NSFC. Based on a manual comparison of our data with statistics in the NSFC Annual Reports, we decided to include four major types, which consider (1) the number of projects in each category and (2) how comprehensive our dataset covers the funded projects every year being reported by NSFC. This study selected the following four types: Key Projects (重点项目; *Key*), General Projects (面上项目; *General*), Young Scientist Projects (青年科学基金项目; *Young*), and Projects for Less Developed Regions (地区科学基金项目; *Region*). These four project types account for ∼95.8% of all projects supported by NSFC in our dataset.

- *Key*: It is the most prestigious program designed for more senior researchers with previous NSFC project experience. It usually supports pivotal and impactful research with significant funding. It normally lasts for 5 years and with the most significant amount of funding.
- *General*: It is a generalized program that can be applied by any PIs. It is also the largest program in terms of the number of funded projects. The normal length of a *General* project is 4 years.
- *Young*: This program is only eligible for PIs under the biological age of 35 (male) or 40 (female) years old. A *Young* project typically lasts for 3 years.
- *Region*: This program is designed for PIs working in institutions in less-developed regions, with a strong focus to 'facilitate the construction of the regional innovation system as well as the social and economic development of the regions' (National Natural Science Foundation of China 2021b). Similar to *General*, most *Region* projects also last for 4 years.

Every project type mentioned above is supported by all the eight disciplinary departments within NSFC. It should be acknowledged that a few other important project types are not selected mostly because they are not covered by the NSFC output website and hence our downloaded dataset, such as Major Research Plan (the most prestigious project type in NSFC aiming to solve critical scientific issues, with each funded project covering multiple individual projects) and Excellent Young Scientists Fund (a more prestigious project type for young researchers than the Young Scientists Fund). However, the number of projects under these latter categories is much smaller than the selected categories, so our choice does not have a major impact on the representative of our final data sample.

Award year: We used the year when an NSFC project was first awarded as the award year. This study focused on projects awarded by NSFC between 2010 and 2015, which means projects included in our sample ended in or before 2020 (The longest duration for a project is 5 years). While the database documents the NSFC project from the late 1980s, we chose 2010 as the starting year because the NSFC outcome portal only documents the complete information for project outcomes from 2010 onwards. We thus considered 2010 as the starting year in our study to ensure consistent linkage between projects and outcomes. We chose 2015 as the end of the project year to ensure we had complete outcome records for all projects by the time we collected the data (see Supplementary Table S1 for publications excluded). This creates a sample with 185,465 projects funded by NSFC from 2010 to 2015 and 2,323,443 corresponding publications.

Publication year: Research outcomes usually appear after a project is awarded. This study uses the year a publication was published as the publication year. A small proportion (108,234, 4.6%) of publications in our sample lack the information or have an incorrect publication year (e.g. a paper was published 10 years before its supporting project was funded). To address these issues, we limited the publication sample to those published between 2010 and 2020. It is noted that 41.2% of publications excluded are in Chinese, which is relatively close to the ratio among publications during the period.

Language detection: Language detection is a crucial step for this project. We used the Python package *lingua* (version 1.1.3) to detect the language of publications based on publication titles in our dataset (Stahl 2023). This package uses *n*-grams of sizes one to five to calculate the Bayesian probability of a text string belonging to a language. We used *lingua* to estimate the language of publications based on publication title information provided by the NSFC outcome portal. We applied this package to our corpus using its default setting to identify all languages used by researchers in China. Initially, it identified 53 languages used in publications by authors in China, with English and Chinese being the dominant ones (99.8%, see Supplementary Table S2). After manually reviewing a random selection of 200 publications for each identified language, we discovered that all titles categorized under languages other than Chinese were in fact, in English. A deeper probe showed that the incorrect identification of publication languages resulted from special non-English words appearing in the titles. For instance, the title 'Three homoclinic solutions for second-order p-Laplacian differential system' was mistakenly labeled as Latin. Consequently, we manually re-categorized all publications from the other eight languages as English. Moreover, we conducted a manual review of a random subset of 400 publications that the *lingua* package identified as English or Chinese. We detected no errors in these samples. Given that Chinese and English overwhelmingly dominate our dataset's publication languages, this research primarily concentrated on these two, representing more than 99.9% of all publications in the dataset. After restricting languages, our final sample consists of 2,209,987 publications under 183,457 projects that supported these publications (see Figure 1 and Supplementary Table S3 for granular breakdown).

Similar publication detection across languages: To understand if there are highly similar publications in the two languages, we compared the pairwise similarity between all English- and Chinese-language publications within one project based on their titles. This analysis is specifically concerning the similarity of topics and content, without considering the overlap of full texts. As such, instances of (self-)plagiarism within the main body of the text were not included in our scope. We used titles because a significant fraction of publications do not have abstracts available in the dataset. However, we also conducted a supplementary analysis by combining and comparing titles and abstracts on a subset of our dataset where intelligible abstracts were available, leading to similar results (see Section 4).

First, for every title, we used sentence-BERT (SBERT) to generate a sentence vector embedding in semantic meaning (Reimers and Gurevych 2020). SBERT uses BERT to produce contextualized word embeddings for all input sentence tokens and then combines the embeddings of individual tokens in the sentence through a pooling operation to produce a single vector representation for the entire sentence. Given that our texts contain Chinese and English, we used the pre-trained model, *paraphrase-multilingual-MiniLM-L12-v2*, to generate embeddings in consistent spaces across multiple languages (Reimers and Gurevych 2020). This

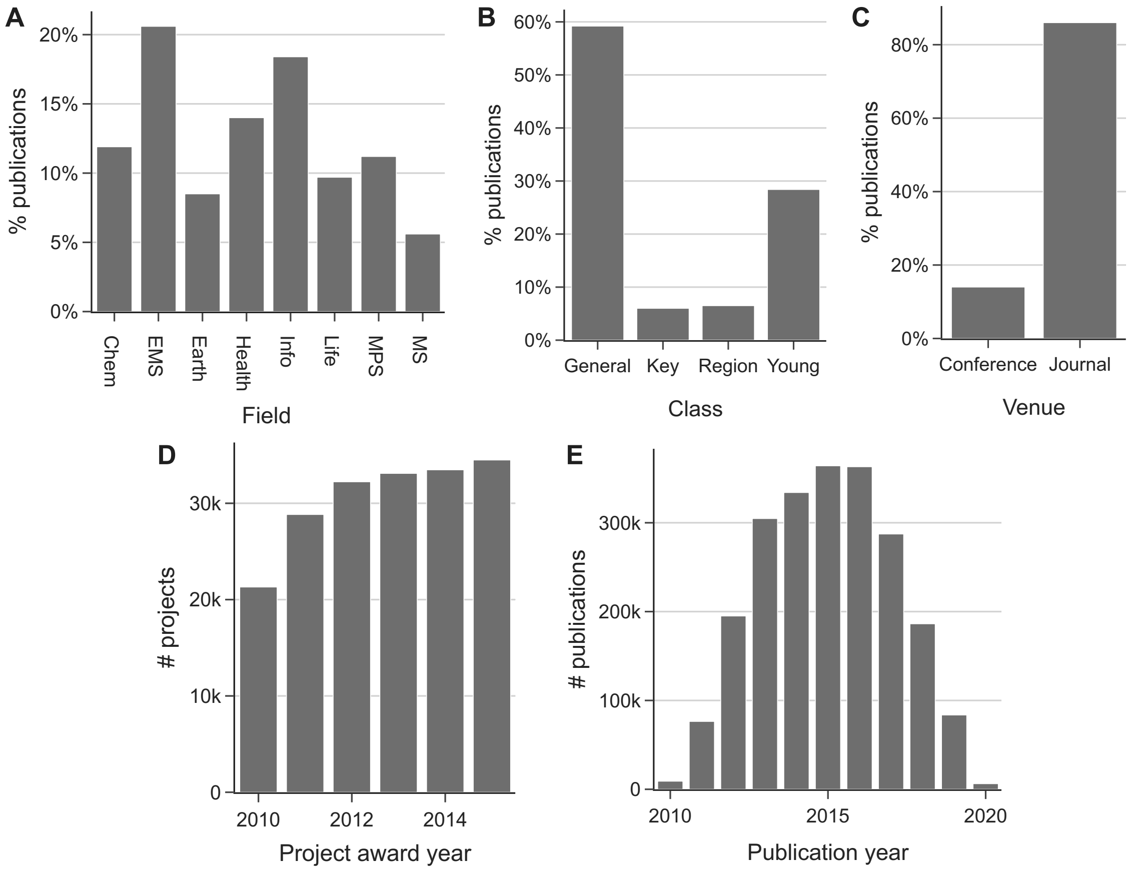

Figure 1. Distribution of data. (A) Publication percentage by field. (B) Publication percentage by project type. (C) Publication percentage by publication venue. (D) Temporal distributions of projects. (E) Temporal distributions of publications.

model was trained on a vast corpus of multilingual paraphrase data, supporting more than 50 languages. During training, sentences with similar meanings, even from different languages, will have closely aligned vector representations within the embedding space. This ensures that their semantical similarity can be numerically computed even if two texts are in different languages.

Before training, we cleaned the publication titles by removing hypertext markup language (HTML) tags, HTML entities, and other noises. We also removed all ending punctuation because most titles typically end with no punctuation. We avoided further preprocessing since BERT-based models were originally trained without such modifications. Changes like altering letter cases or removing stop words might misrepresent the essence of a title, especially given their typically short lengths. We used the Python library Sentence Transformers (v2.2.2), to transform each title into a vector embedding. These vectors, each composed of 384 numerical values, capture the semantics of the titles. To quantify the similarity, we computed a cosine similarity score between each title pair embedding vectors. A similarity score closer to 1 indicates a high similarity, while a similarity score closer to -1 indicates dissimilarity. The computation was performed using the resources of the Center for High Throughput Computing (2006).

Moreover, as a validity check of the similarity score, we did a proportional sampling of 5000 cross-language title pairs from all pairs while keeping the original distribution of fields and project award years. We binned the similarity scores into 20 equal-length intervals and applied variable sampling weights to each bin to ensure that every level of similarity is representative in the 5000 pairs (see Figure 2A). We then used the large language model Generative Pre-trained Transformer 4 (GPT-4) model to annotate a five-point similarity scale from 1 (very different) to 5 (very similar) by designed prompts and compared the results (see Figure 2B and Supplementary notes). The annotated results are highly correlated with the similarity scores (Spearman's rank correlation = 0.814, $P < 0.001$). To determine the threshold of high similarity, we used the Gini impurity, a common measure for finding the best split in decision tree algorithms (see Supplementary notes). Given a similarity threshold, Gini impurity calculates the likelihood that a pair labeled as 'very similar' by GPT-4 actually has a similarity lower than that threshold. A low Gini impurity means that the threshold correctly categorizes most GPT-4-labeled 'very similar' pairs into the highly similar group above the threshold. As a result, the threshold of 0.8 was selected as it yielded the lowest Gini impurity (0.15) among all threshold candidates (see Figure 2C).

4. Results

4.1 English dominates the scientific publishing language landscape in China

Our results show an overwhelming dominance of English in the scientific publishing enterprise in China. Among all 2,209,987 NSFC project-associated publications in our analytical sample, 1,462,536 (66.2%) are in English, while 747,451 (33.8%) are in Chinese. Regarding individual NSFC projects, 54,924 projects (29.6%) exclusively published

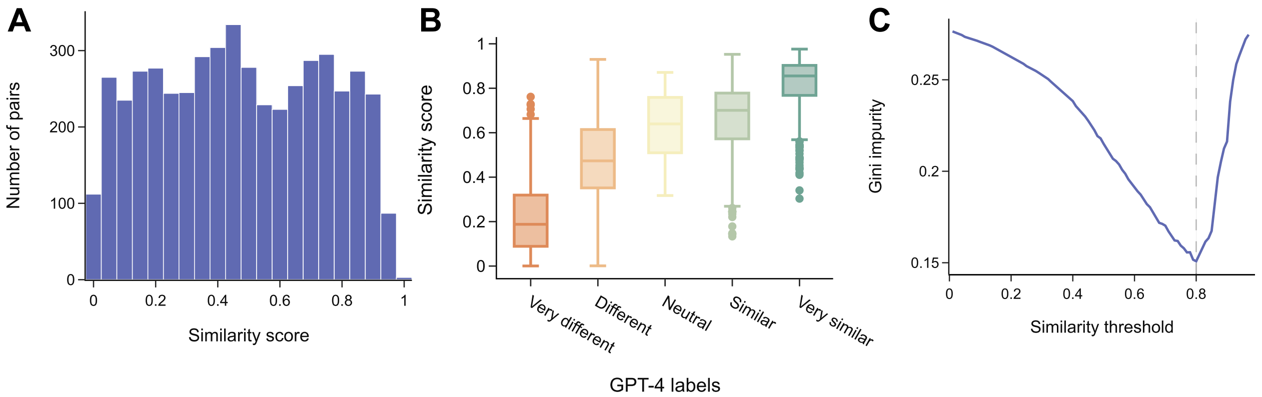

Figure 2. Robustness check of cosine similarity score and high similarity threshold. (A) Distribution of title pair numbers across binned similarity scores. By using the weighted sampling strategy, the distribution of similarity scores in the sample tends toward a uniform distribution rather than a normal distribution. (B) Box plots for similarity score distributions across five-point similarity scales labeled by GPT-4. (C) Gini impurity of 'very similar' pairs by GPT-4 by different thresholds of similarity. The dashed vertical line denotes the lowest Gini impurity (0.15) at the similarity threshold of 0.8, suggesting that this threshold reaches the best agreement between the GPT-4 'very similar' label and high similarity scores above the threshold.

papers in English, compared with 17,331 projects (9.3%) in Chinese. Hence, most projects have publications covering both languages. On average, each NSFC project publishes ~7.97 journal publications in English. These findings add to the evidence that English is the lingua franca in scientific research in many geographic regions and countries (Valkimadi et al. 2009; Mongeon and Paul-Hus 2016).

The extent of English's predominant role in scientific publishing varies depending on the academic field (see Figure 3A). Our results show that *Chemical Sciences* (83.8%) has the highest share of English-language publications among the eight fields, followed by *Mathematics and Physical Sciences* (79.9%). *Management Sciences* has the lowest share of English-language publications (36.3%), echoing existing evidence that social sciences are more localized than other fields (Stockemer and Wigginton 2019). Nonetheless, there tend to be vast differences in the use of languages on the level of subfields (NSFC classification, see Figure 3B). For example, *Traditional Chinese Medicine*, a subfield in *Health Science*, has more than 80% of publications published in Chinese, despite the much lower share of Chinese-language publications in Health Sciences (40%).

The share of English-language publications by NSFC projects also varies by the project types, or *project class* as named by NSFC (see Figure 3A). About 79.3% of publications produced by *Key* projects are in English, while 41.4% of publications produced by *Region* projects are in English. The *General* and *Young* projects have similar shares of English publications, which are lower than *Key*. Furthermore, across the eight disciplinary fields, *Region* projects have the lowest share of English-language publications among all four project types, while *Key* projects have the highest. Given the intrinsic 'laddered' hierarchical structure of project types, our results suggest a possible positive correlation between the tendency to publish in English and the level or prestige of project types.

4.2 The growing prevalence of English in China's scientific publishing

We observed the increasing dominance of English in the publications produced by China in the past decade (see Figure 4). Figure 4A and B highlight an increase in the proportion of English-language publications across all fields and project types, respectively. Overall, from 2010 to 2020, the share of English-language publications increased by 26.8 percentage points (pp), from 53.4% to 80.3%. *Earth Sciences* experienced the highest increase rate (41.0 pp) among the eight fields, while *Mathematics and Physical Sciences* (13.1 pp) were the lowest. *General* has the highest increase rate among the four project types (29.4 pp), while *Key* (15.8 pp) is the lowest. This observation suggests a potential shift in the significance of the two academic languages used in the Chinese scientific ecosystem in the past decade, and the increasing prominence of English as the primary publishing language among Chinese researchers is a universal phenomenon regardless of field or project type. The speed of change in a category is generally negatively correlated with the share of publications in English-language, which may indicate a saturation point to be reached. We also examined trends of English-language publications based on project progress, i.e. the year of English-language publications relative to the year the project started. Our results show a notable trend toward an increase in the proportion of English-language publications in the later years of the project, which is observed across projects in all fields and project types (see Figure 4C and D). Additionally, we observed that newer projects tend to have higher ratios of English-language publications throughout the project progress, in addition to this parallel trend.

4.3 Language disparities between journal and conference publications

A divergence exists in language preferences between journal and conference publications in China. English-language publications hold a more substantial share among conference proceedings, accounting for 72.7%, as opposed to 65.1% among journal articles. It is noteworthy, however, that this discrepancy should not be taken at face value. Within the dataset, the field of *Information Science* played a pivotal role, contributing 63.2% of the overall conference proceedings. Impressively, the overwhelming majority, at 92.7%, of these publications were in English (see Figure 5A). In contrast, English conference proceedings in fields outside *Information Science* constituted ~60% of the total. Moreover, there is a negative correlation (Pearson $r = -0.233$, $P < 0.01$) between the proportions of English-language publications in journals and conference proceedings. This correlation persists irrespective of the presence of the field of *Information Science*.

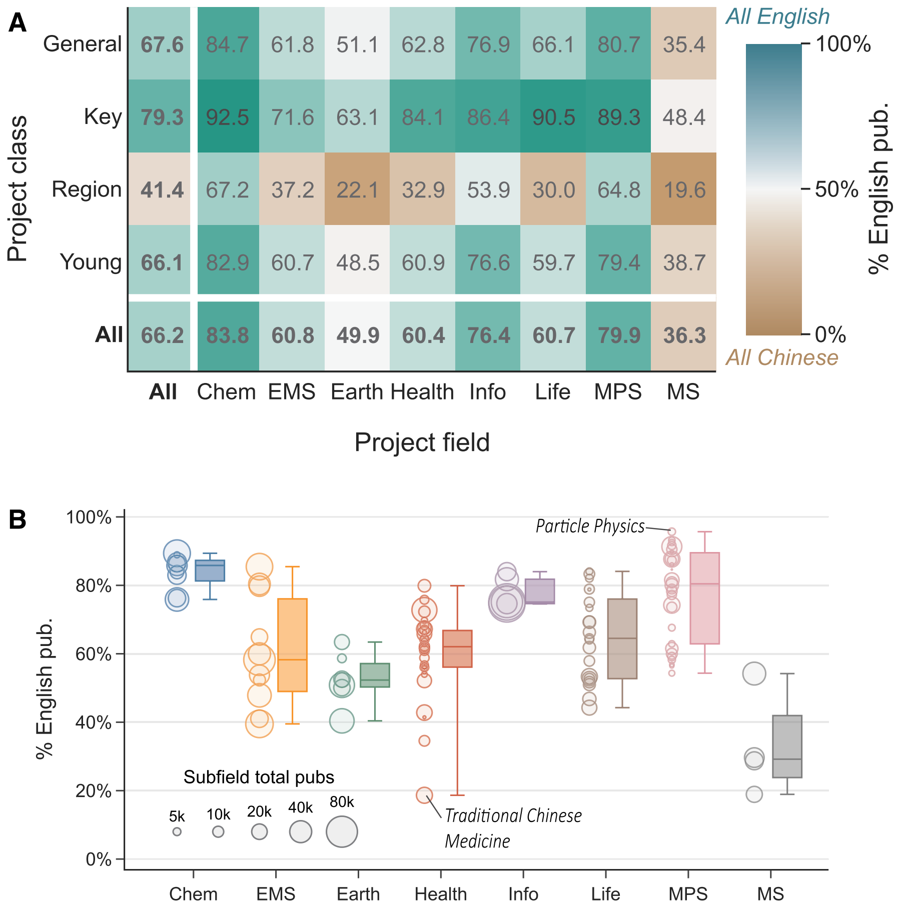

Figure 3. Varying share of English-language publications among NSFC project publications. (A) By project field and type. The number in each cell denotes the share of English-language publications in each combination of project type and field. The numbers in bold represent the cumulative percentages of English-language publications for each project field/type. (B) By subfields. Each bubble corresponds to one of the 123 subfields and is sized proportional to the total number of publications within that subfield. Box plots show the distribution of the share of publications in English across subfields, categorized by their broader fields.

This observation stands in contrast to the conventional expectation that languages should exhibit a consistent pattern of usage in journal and conference publications within the same research domain. This contrast is particularly pronounced in *Chemistry*, where only 40% of conference publications are in English, whereas over 80% of journal publications use English. On the temporal dimension, there is a generally consistent upward trend for both types of venues regarding the adoption of English in the past decade (see Figure 5B). In most instances, the shifts in language usage between journal and conference publications follow a parallel trajectory, regardless of the field differences. This suggests that Chinese researchers increasingly depend on English throughout the course of their projects. Furthermore, this intricate interplay between field-specific influences and language choices underscores the nuanced nature of language preferences in academic publications in China.

4.4 Content discrepancy between English and Chinese-language publications

Recognizing the importance of multilingual scientific publishing in addressing language barriers for researchers, a critique within China's scientific evaluation system focuses on cross-language duplicate publishing, which publishes the same content in multiple languages to boost productivity (Tucker et al. 2011; Qi et al. 2013; Chen et al. 2018). However, there is a lack of robust, systematic evidence supporting these criticisms (Teixeira da Silva 2020). This section examines the content similarity between English and Chinese-language publications within individual projects.

Our results show that highly similar publications are relatively scarce. Out of the 8,085,117 pairs of publication titles created by 1,061,192 English-language publications and 706,361 Chinese-language publications from 117,097 NSFC projects, ∼0.4% of the pairs are highly similar ones (similarity score > 0.8, see Section 3). These highly similar pairs account for 2.0% of English and 3.0% of Chinese-language publications in the data. These discoveries indicate that researchers predominantly produce unique content in each of the two languages, implying less need for concern regarding cross-language duplicate publishing. Moreover, this also highlights the need to acknowledge the essential nature of a multilingual publishing system to proficiently convey scientific information to a wide range of audiences.

We further examined similar publication pairs by analyzing the peak similarity score for each publication (see Figure 6A). A peak similarity score is the highest similarity score among a publication's all possible cross-language pairs, representing the most similar content in the other language for that publication. Because the pair matching was done

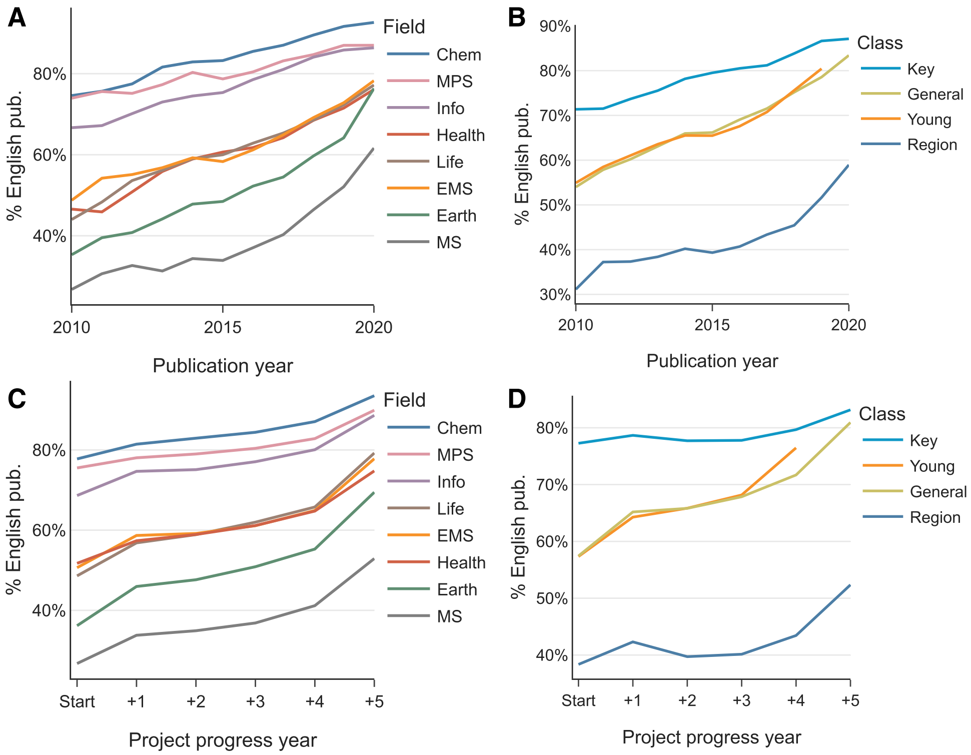

Figure 4. Temporal trend of the share of English-language publications. (A) Over the publication years by field. (B) Over the publication years by project type. (C) Over the progress of projects by field. (D) Over the progress of projects by project type. 'Start' is the beginning year of an NSFC project. Note that *Key* projects usually span 5 years, *General* and *Region* projects for 4 years, and *Young* projects for 3 years.

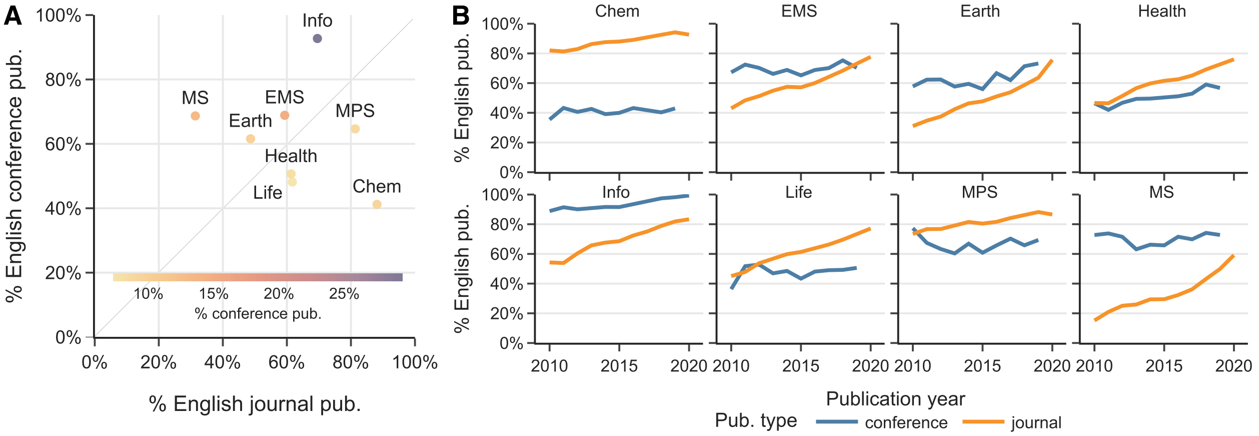

Figure 5. English usage by publication types (journal vs conference proceeding). (A) Share of English-language publications among journal/conference publications by field. A field's marker is colored in proportion to the total percentage of conference publications among all publications in that field. (B) Share of English-language publications among journal/conference publications over publication years by field.

with replacement, the similarity score distributions for English and Chinese-language publications can be asymmetric. The average peak similarity score for publications in Chinese stands at 0.55, compared to 0.52 for those in English (t-test $P < 0.001$). This suggests that even though English and Chinese-language publications are generally distinct, Chinese-language publications tend to be more similar to their English counterparts than vice versa. When broken down by field, most fields show a higher percentage of Chinese-language publications with at least one highly similar match (larger than 0.8) compared to English-language publications (see Figure 6B). The exceptions are *Earth Sciences* and *Management Science*. This data suggests that more Chinese-language publications are likely being adapted to English to expand their international readership.

We acknowledge the potential discrepancy between paper titles and their actual content, which could impact the accuracy of our initial analysis based solely on titles. To verify the integrity of our findings, we conducted a supplementary analysis on a select subset of our dataset. This subset included 36,643 English and 20,726 Chinese papers from 15,355 projects, each with an intelligible abstract available. Using the combined texts of both titles and abstracts in our similarity analysis, we found that the outcomes aligned closely with our primary results, reinforcing the validity of our conclusions (see Supplementary Figure S1).

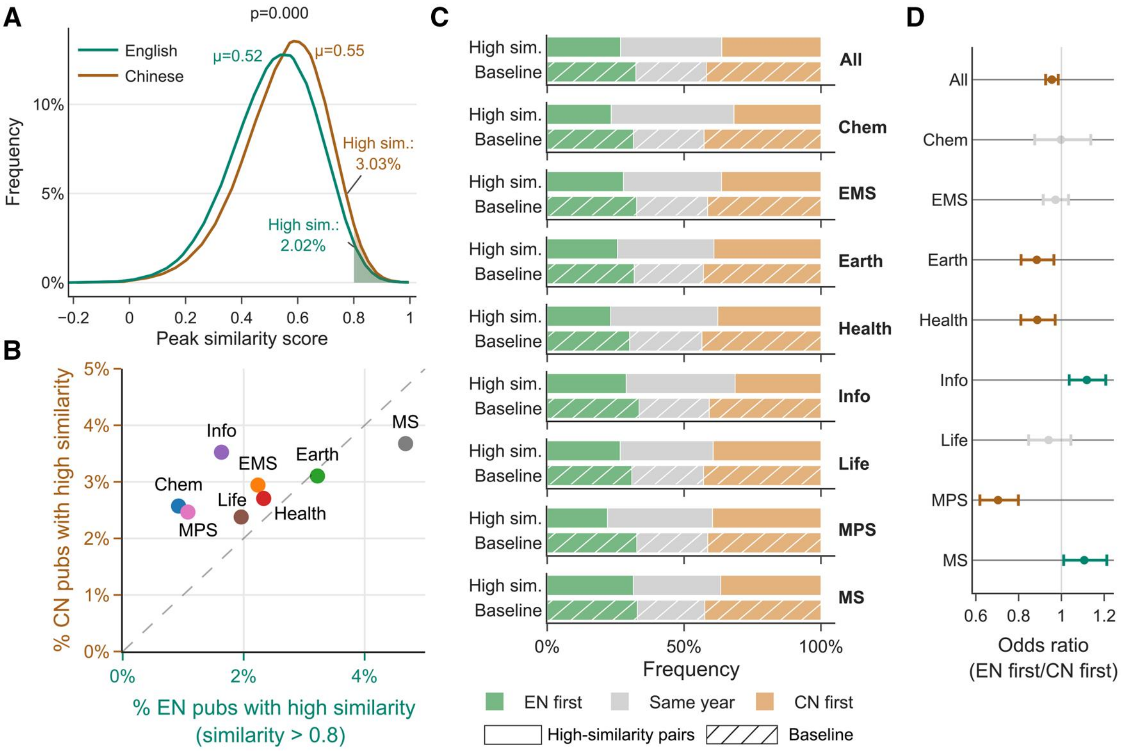

Figure 6. Distribution and publication year difference of similar papers in different languages. (A) Distributions of peak similarity scores for English and Chinese-language publications. The peak similarity score is the highest similarity score for each publication among all its cross-language pairs. It represents the most similar cross-language publication match. *p* value is calculated by t test. (B) Share of English (X axis) and Chinese (Y axis) publications with highly similar matches (any title similarity > 0.8) by field. (C) Comparison of publication years in highly similar and all cross-language pairs (baseline, marked with diagonal lines) by field. (D) Odds ratios for the number of English-first pairs over Chinese-first pairs between the high-similarity publication subset and the baseline. An odds ratio above 1 suggests a publication is more likely to be published in English first among its bilingual similar versions. A confidence interval including 1 denotes no significant difference ($P \geq 0.05$). *P* values and confidence intervals are calculated using Chi-Square tests.

Moreover, to analyze which language tended to be published first while controlling the topics, we compared the timing of English and Chinese-language publications for each pairwise year sequence of highly similar publication pairs (see Figure 6C and D). Compared with the baseline consisting of all publication pairs, the highly similar pairs have more publications published in the same year across fields (37.0% vs 25.7% of total pairs). Among the highly similar pairs that were published in different years, more Chinese-language publications were published before the similar English ones overall, which is also a pattern that holds for the fields of *Earth Sciences*, *Health Science*, and *Mathematics and Physical Sciences*, while the opposite trend appears in *Information Science* and *Management Sciences* (see Figure 6D). This evidence suggests that English is less likely to become the earlier publication language when publishing similar contents in bilingual versions, which corresponds to the finding in Figure 4 that there is often an increased proportion of publications in English as projects progress.

5. Discussion

Leveraging the distinctive NSFC outcome database, we analyzed China's scientific publication language spectrum. Our results show the absolute dominance of English for scientific publishing by scientists in China, but any assessment of the scientific productivity of China as a nation should also not overlook publications in Chinese. We found that China published more publications in English (66.2%) than in Chinese (33.8%) during the period of investigation. This pattern holds for most fields, except for *Management Sciences*, where English-language publications account for 36.3% of the total. The prevalence of English as the publication language experienced a steady increase during the period of investigation, with slightly different rates across the eight fields.

In addition to the overall landscape, a more detailed breakdown reveals that the hard natural sciences and engineering fields are more inclined toward English-language publication compared to the social sciences fields. Moreover, English has played a more significant role in our publication sample in high-level projects. For example, higher-level projects, such as the *Key* and *General* projects, are more likely to be published in English. On the other hand, the *Region* project, where PIs are in less developed regions, published more papers in Chinese. This indicates that projects of a higher level may prioritize publishing in English to reach a broader audience and establish a wider influence in their respective fields. In addition, researchers have moved to publish more English publications over the years and over the progress of their projects, which shows the increasing English usage in China's research.

The findings above confirm that science in China has been increasingly involved in an English-dominant mode of scientific knowledge production and dissemination. Particularly on the country level, our findings supplement existing evidence primarily situated in social sciences and humanities fields (Kulczycki et al. 2018, 2020; Mironescu, Moroșanu and Bibiri 2023). This language transition reflects the Chinese scientific community's increasing adoption of English-language publication norms since the late 1980s (Flowerdew and Li 2009). Later on, academic institutions and funding bodies in China perceived SCI as a pivotal tool for evaluating both research outputs and individual researchers (Qian et al. 2020). This prevailing culture and policy stance is instrumental in amplifying the favor for English among researchers (Ossenblok, Engels and Sivertsen 2012). Nonetheless, as mentioned above, in 2020, the Chinese government started advocating a reduced dependency on English-language citation indices, urging more publications written in the native language. Future research can monitor the language usage trends and examine the effects of this policy shift on the broader Chinese research ecosystem.

Additionally, our findings highlight a marked disparity in language preferences between journal publications and conference proceedings. In general, our results reveal a greater prevalence of English-language publications in conference papers than journal articles in some fields, except for *Mathematics and Physical Sciences*, *Chemistry*, *Life Science*, and *Health Sciences*. This discrepancy can be attributed to the tendency in these fields to present summaries and abstracts at conferences, while full-length papers are typically found in journals. Under these circumstances, Chinese scientists are more inclined to prioritize publishing journal articles in English, despite the challenges associated with publishing in a non-native language (Flowerdew 1999; Flowerdew and Li 2009; Bortolus 2012). This trend is particularly evident in the remarkably high percentage of English-language conference papers in the *Information Sciences* field (92.7%), where conferences are typically the primary venues for full-length papers and are highly regarded in China's academic environment. The higher rate of English-language publications among journal articles than conference proceedings further underscores the potential influence of China's publishing incentives, aligning with previous findings (Franzoni, Scellato and Stephan 2011).

Our findings reveal a minimal similarity between English and Chinese-language publications produced by NSFC projects. Only 2.0% of English and 3.0% of Chinese-language publications have a cross-language match of substantial similarity under the same project. It suggests that concerns regarding the duplicate publishing issue, which assumes that publications in both languages convey similar scientific content, may be less serious (Teixeira da Silva 2020). However, since we did not examine the full text, we do not exclude the possibility that publications with different topics have duplicate texts. On the contrary, the limited similarity between English and Chinese-language publications indicates that the current understanding of China's research landscape, primarily drawn from English-language publications, fails to paint a complete picture. Chinese-language research remains overlooked, and consumers of English-language publications miss out on locally communicated scientific insights within the Chinese community. This implies that the current assessment of China as a major powerhouse of scientific research may still underestimate the full extent of its endeavors.

The data further indicate that, across most fields, Chinese-language publications often precede those in English within individual projects even if they have highly similar content. A majority of individual projects spanning varied fields mirror this tendency, with *Information Sciences* as an exception. This field is prominently characterized by English-language conference papers, potentially reflecting distinct publication practices within this domain. Several potential factors could underlie this observed trend. One is the publication delay, defined as the period between the submission of an article and its eventual publication (Luwel and Moed 2006). English-based journals may exhibit prolonged publication delays compared to their Chinese counterparts, leading to earlier publication timelines for manuscripts submitted to Chinese journals. However, this hypothesis needs to be tested due to the absence of robust evidence of publication delays between English and Chinese-based journals. Furthermore, Chinese scientists may confront barriers to publishing in English-language venues. The time and effort needed to navigate language-related challenges might decelerate publishing in English (Flowerdew and Li 2009; Ramírez-Castañeda 2020). The resultant delay may, to some extent, account for the staggered temporal distribution in the appearance of Chinese and English-language publications. Future research is needed to unravel the factors behind the observed dynamics. Investigating the role of publication delays, language barriers, and other possible variables will render a more nuanced understanding of China's research landscape.

6. Conclusion

This study sheds light on research evaluation in several ways. By analyzing the language usage in publications from China, we have confirmed that the anglicization trend observed in other countries also occurred in China in the 2010s, and it is more pronounced in high-level projects and natural sciences and engineering fields. Although Chinese is still mainstream in social science, English's predominance is also rising. This indicates that the policy tools and research evaluation metrics used by the Chinese government and institutions, which heavily rely on global English-dominant bibliographic databases at least until 2022, have significantly altered researchers' publishing practices to publish in international, English-language journals, potentially at the expense of local language dissemination and diverse scholarly communication. It may contribute to shaping a research culture that values more English-language publications than local language publications, as implied by the popularity of English in high-level projects. Our findings advocate reviewing current reward systems to foster a more balanced approach that values both global engagement and local relevance in scholarly work.

Moreover, through the analysis of similarity and publication time, we argue that relying solely on global bibliographic databases results in the loss of numerous publications written in Chinese, which can make distinct contributions from English-language publications and be published earlier. This phenomenon has also been highlighted for other major languages such as Spanish, and ignoring knowledge delivered in non-English-language publications can cause biases in our understanding of the current progress of science (Amano, González-Varo and Sutherland 2016). It highlights the need to create an all-encompassing national scientific database, including publications in various languages, to comprehensively understand a country's research endeavors and its role in the global science landscape.

This research provides insightful findings but presents several limitations. First, the scope of our dataset is limited both in academic fields and temporal range, concentrating solely on research funded by the NSFC within the 2010s, excluding arts and humanities. A direction for future scrutiny is the trajectory of Chinese research across all academic fields following the significant shift in national research policies from 2020. Second, we overlooked research supported by agencies other than NSFC. Third, exploring academic language use patterns across countries is a vital future exploration to deepen our comprehension of the subject matter. Fourth, full records of abstracts and main texts of the publications can extend our analysis but are difficult to collect comprehensively. Lastly, individual researcher-level attributes have not been integrated into our analysis due to the characteristics of our dataset. Merging author-specific data from external sources into our dataset can provide a more detailed understanding of how individual characteristics may affect language choice in publishing.

Acknowledgements

This research benefited from the service provided by the Center for High Throughput Computing at University of Wisconsin-Madison, and the Publons data provided by Clarivate Analytics.

Author contributions

K.L., X.Z., and C.N. designed research, collected and processed data, performed analysis, and wrote the paper.

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Funding

Funding support for this article was provided by the Vilas Life Cycle Professorships by the University of Wisconsin-Madison, The Vice Chancellor for Research and Graduate Education (VCRGE) by the University of Wisconsin-Madison.

Conflict of interest statement. None declared.

References

Aagaard, K., Bloch, C., and Schneider, J. W. (2015) 'Impacts of Performance-Based Research Funding Systems: The Case of the Norwegian Publication Indicator', *Research Evaluation*, 24: 106–17.

Amano, T., González-Varo, J. P., and Sutherland, W. J. (2016) 'Languages Are Still a Major Barrier to Global Science', *PLoS Biology*, 14: e2000933.

Baker, S. (2023) 'China Overtakes United States on Contribution to Research in Nature Index', *Nature*, <https://www.nature.com/articles/d41586-023-01705-7> accessed 15 July 2023. DOI: 10.1038/d41586-023-01705-7.

Baldauf, R. B. (2001) 'Speaking of Science: The Use by Australian University Science Staff of Language Skills', in U. Ammon (ed.) *The Dominance of English as a Language of Science, Contributions to the Sociology of Language [CSL]*, Vol. 84, pp. 139–66. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110869484.139.

Bortolus, A. (2012) 'Running Like Alice and Losing Good Ideas: On the Quasi-Compulsive Use of English by Non-Native English Speaking Scientists', *AMBIO*, 41: 769–72.

Center for High Throughput Computing (2006) 'Center for High Throughput Computing'. <https://chtc.cs.wisc.edu/> accessed 22 April 2024. DOI: 10.21231/GNT1-HW21.

Chen, W., Xing, Q.-R., Wang, H., and Wang, T. (2018) 'Retracted Publications in the Biomedical Literature with Authors from Mainland China', *Scientometrics*, 114: 217–27.

Chinese Communist Party Central Committee State Council (2018) 'Opinions on the deepening of reform on project review, talent evaluation and institution assessmentitle'. <http://www.xinhuanet.com/politics/2018-07/03/c_1123074267.htm> accessed 23 April 2024.

Clarke, A., Gatineau, M., Grimaud, O., Royer-Devaux, S., Wyn-Roberts, N., Le Bis, I., and Lewison, G. (2007) 'A Bibliometric Overview of Public Health Research in Europe', *European Journal of Public Health*, 17 Suppl 1: 43–9.

Cui, T., and Zhang, J. (2018) 'Bibliometric and Review of the Research on Circular Economy through the Evolution of Chinese Public Policy', *Scientometrics*, 116: 1013–37.

Dann, G. M. S. (2011) 'Anglophone Hegemony in Tourism Studies Today', *Enlightening Tourism: a Pathmaking Journal*, 1: 1–30.

Faraldo-Cabana, P. (2018) 'Research Excellence and Anglophone Dominance: The Case of Law, Criminology and Social Science', in K. Carrington, R. Hogg, J. Scott, and M. Sozzo (eds) *The Palgrave Handbook of Criminology and the Global South*, pp. 163–81. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-65021-0_9

Fejes, A., and Nylander, E. (2017) 'The Economy of Publications and Citations in Educational Research: What about the "Anglophone Bias"?', *Research in Education*, 99: 19–30.

Flowerdew, J. (1999) 'Problems in Writing for Scholarly Publication in English: The Case of Hong Kong', *Journal of Second Language Writing*, 8: 243–64.

Flowerdew, J., and Li, Y. (2009) 'English or Chinese? The Trade-off between Local and International Publication among Chinese Academics in the Humanities and Social Sciences', *Journal of Second Language Writing*, 18: 1–16.

Franzoni, C., Scellato, G., and Stephan, P. (2011) 'Changing Incentives to Publish', *Science*, 333: 702–3.

Fu, J., Frietsch, R., and Tagscherer, U. (2013) 'Publication activity in the Science Citation Index Expanded (SCIE) database in the context of chinese science and technology policy from 1977 to 2012', Discussion Papers 'Innovation Systems and Policy Analysis'. Karlsruhe: Fraunhofer Institute for Systems and Innovation Research (ISI).

Gingras, Y. (2016) *Bibliometrics and Research Evaluation: Uses and Abuses*. Cambridge: MIT Press.

González-Alcaide, G., Valderrama-Zurián, J. C., and Aleixandre-Benavent, R. (2012) 'The Impact Factor in Non-English-Speaking Countries', *Scientometrics*, 92: 297–311.

Gordin, M. D. (2015) *Scientific Babel: How Science Was Done Before and After Global English*, Illustrated edn. Chicago; London: University of Chicago Press.

Heilbron, J., and Gingras, Y. (2018) 'The Globalization of European Research in the Social Sciences and Humanities (1980–2014): A Bibliometric Study', in J. Heilbron, G. Sorá, and T. Boncourt (eds) *The Social and Human Sciences in Global Power Relations*, pp. 29–58. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-73299-2_2.

Horta, H., and Shen, W. (2020) 'Current and Future Challenges of the Chinese Research System', *Journal of Higher Education Policy and Management*, 42: 157–77.

Huan, G. (1986) 'China's Open Door Policy, 1978-1984', *Journal of International Affairs*, 39: 1–18.

Koch, T., and Vanderstraeten, R. (2019) 'Internationalizing a National Scientific Community? Changes in Publication and Citation Practices in Chile, 1976–2015', *Current Sociology*, 67: 723–41.

Korytkowski, P., and Kulczycki, E. (2019) 'Examining How Country-Level Science Policy Shapes Publication Patterns: The Case of Poland', *Scientometrics*, 119: 1519–43.

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A., and Zuccala, A. (2018) 'Publication Patterns in the Social Sciences and Humanities: Evidence from Eight European Countries', *Scientometrics*, 116: 463–86.

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., and Sivertsen, G. (2020) 'Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study', *Journal of the Association for Information Science and Technology*, 71: 1371–85.

Lewison, G. (2009) 'The Percentage of Reviews in Research Output: A Simple Measure of Research Esteem', *Research Evaluation*, 18: 25–37.

Liu, W. (2017) 'The Changing Role of non-English Papers in Scholarly Communication: Evidence from Web of Science's Three Journal Citation Indexes', *Learned Publishing*, 30: 115–23.

López-Navarro, I., Moreno, A. I., Quintanilla, M. Á., and Rey-Rocha, J. (2015) 'Why Do I Publish Research Articles in English instead of my Own Language? Differences in Spanish Researchers' Motivations across Scientific Domains', *Scientometrics*, 103: 939–76.

Luwel, M., and Moed, H. (2006) 'Publication Delays in the Science Field and Their Relationship to the Ageing of Scientific Literature', *Scientometrics*, 41: 29–40.

Marginson, S. (2022) '"All Things Are in Flux": China in Global Science', *Higher Education*, 83: 881–910.

Mathies, C., Kivistö, J., and Birnbaum, M. (2020) 'Following the Money? Performance-Based Funding and the Changing Publication Patterns of Finnish Academics', *Higher Education*, 79: 21–37.

Ministry of Education Ministry of Science (2020) 'Some opinions on the appropriate use of SCI-related indicators and re-orientation of research assessment'. <http://www.moe.gov.cn/srcsite/A16/moe_784/202002/t20200223_423334.html> accessed 22 April 2024.

Ministry of Science and Technology of the People's Republic of China (2022) 'Statistical Analysis of Chinese Scientific and Technological Papers in 2020'. <https://most.gov.cn/xxgk/xinxifenlei/fdzdgknr/kjtjbg/kjtj2022/202209/P020220920391756277580.pdf> accessed 22 April 2024.

Mironescu, A., Moroșanu, A., and Bibiri, A.-D. (2023) 'The Regional Dynamics of Multilingual Publishing in Web of Science: A Statistical Analysis of Central and Eastern European Journals and Researchers in Linguistics', *Scientometrics*, 128: 1133–62.

Mohrman, K., and Wang, Y. (2010) 'China's Drive for World-Class Universities', in L. M. Portnoi, V. D. Rust, and S. S. Bagley (eds) *Higher Education, Policy, and the Global Competition Phenomenon*, pp. 161–76. New York: Palgrave Macmillan US. DOI: 10.1057/9780230106130_12.

Mongeon, P., and Paul-Hus, A. (2016) 'The Journal Coverage of Web of Science and Scopus: A Comparative Analysis', *Scientometrics*, 106: 213–28.

National Natural Science Foundation of China (2021a) 'Filling instructions'. <https://www.nsfc.gov.cn/Portals/0/fj/czsm.docx> accessed 20 July 2024.

National Natural Science Foundation of China (2021b) 'National Natural Science Fund Guide to Programs 2021'. <https://www.nsfc.gov.cn/english/site_1///pdf/NationalNaturalScienceFundGuidetoPrograms2021.pdf> accessed 4 Dec. 2022.

National Natural Science Foundation of China (2023) 'Big Data Knowledge Management Service Portal'. <https://kd.nsfc.gov.cn/> accessed 6 Nov. 2023.

National Science Board, National Science Foundation (2022) *Science and Engineering Indicators 2022: The State of U.S. Science and Engineering (No. NSB-2022-1)*. Alexandria, VA: National Science Foundation. <https://ncses.nsf.gov/pubs/nsb20221> accessed 30 Oct. 2023.

Ossenblok, T. L. B., Engels, T. C. E., and Sivertsen, G. (2012) 'The Representation of the Social Sciences and Humanities in the Web of Science—A Comparison of Publication Patterns and Incentive Structures in Flanders and Norway (2005–9)', *Research Evaluation*, 21: 280–90.

Qi, X., Ren, W., Liu, L., Yang, Z., Yang, M., Fan, D., and Han, G. (2013) 'Prevalence of Covert Duplicate Publications in Budd-Chiari Syndrome Articles in China: A Systematic Analysis', *The American Journal of Medicine*, 126: 633–9.e2.

Qian, J., Yuan, Z., Li, J., and Zhu, H. (2020) 'Science Citation Index (SCI) and Scientific Evaluation System in China', *Humanities and Social Sciences Communications*, 7: 108.

Quan, W., Chen, B., and Shu, F. (2017) 'Publish or Impoverish: An Investigation of the Monetary Reward System of Science in China (1999-2016)', *Aslib Journal of Information Management*, 69: 486–502.

Ramírez-Castañeda, V. (2020) 'Disadvantages in Preparing and Publishing Scientific Papers Caused by the Dominance of the English Language in Science: The Case of Colombian Researchers in Biological Sciences', *PLoS One*, 15: e0238372.

Rao, G., Xia, E., and Li, Q. (2020) 'Investigation of Language Choice among International Academic Articles and the Use of Chinese in the Recent Decade', *Applied Linguistics*, 2: 37–51.

Reimers, N., and Gurevych, I. (2020) 'Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation', in B. Webber, T. Cohn, Y. He, and Y. Liu (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–25. Presented at the EMNLP 2020, November, Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.365.

Shao, J., and Shen, H. (2012) 'Research Assessment and Monetary Rewards: The Overemphasized Impact Factor in China', *Research Evaluation*, 21: 199–203.

Shu, F., Quan, W., Chen, B., Qiu, J., Sugimoto, C. R., and Larivière, V. (2020) 'The Role of Web of Science Publications in China's Tenure System', *Scientometrics*, 122: 1683–95.

Shu, F., Sugimoto, C. R., and Larivière, V. (2021) 'The Institutionalized Stratification of the Chinese Higher Education System', *Quantitative Science Studies*, 2: 327–34.

Sivertsen, G. (2018) 'Balanced multilingualism in science', BiD: textos universitaris de biblioteconomia i documentació, 40. DOI: 10.1344/BiD2018.40.25

Stahl, P. M. (2023) 'pemistahl/lingua-py'. <https://github.com/pemistahl/lingua-py> accessed 6 Nov. 2023.

Stockemer, D., and Wigginton, M. J. (2019) 'Publishing in English or Another Language: An Inclusive Study of Scholar's Language Publication Preferences in the Natural, Social and Interdisciplinary Sciences', *Scientometrics*, 118: 645–52.

Teixeira da Silva, J. A. (2020) 'The Ethics of Publishing in Two Languages', *Scientometrics*, 123: 535–41.

Tollefson, J. (2018) 'China Declared World's Largest Producer of Scientific Articles', *Nature*, 553: 390.

Tucker, J. D., Chang, H., Brandt, A., Gao, X., Lin, M., Luo, J., Song, P., Sun, K., and Zhang, X. (2011) 'An Empirical Analysis of Overlap Publication in Chinese Language and English Research Manuscripts', *PLoS One*, 6: e22149.

Uzuner, S. (2008) 'Multilingual Scholars' Participation in Core/Global Academic Communities: A Literature Review', *Journal of English for Academic Purposes*, 7: 250–63.

Valkimadi, P. E., Karageorgopoulos, D. E., Vliagoftis, H., and Falagas, M. E. (2009) 'Increasing Dominance of English in Publications Archived by PubMed', *Scientometrics*, 81: 219–23.

Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., and Van Raan, A. F. J. (2001) 'Language Biases in the Coverage of the Science Citation Index and Its Consequencesfor International Comparisons of National Research Performance', *Scientometrics*, 51: 335–46.

Wagner, C. S., Zhang, L., and Leydesdorff, L. (2022) 'A Discussion of Measuring the Top-1% Most-Highly Cited Publications: quality and Impact of Chinese Papers', *Scientometrics*, 127: 1825–39.

Warchał, K., and Zakrajewski, P. (2023) 'Multilingual Publication Practices in the Social Sciences and Humanities at a Polish

University: choices and Pressures', *International Journal of Multilingualism*, 20: 801–24.

Wei, S.-J. (1995) 'The Open Door Policy and China's Rapid Growth: Evidence from City-Level Data', in Ito T. and Krueger A. (eds) *Growth Theories in Light of the East Asian Experience*, pp. 73–104. Chicago: University of Chicago Press.

Wei, F., and Zhang, G. (2020) 'Measuring the Scientific Publications of Double First-Class Universities from Mainland China', *Learned Publishing*, 33: 230–44.

Xia, J., Wright, J., and Adams, C. E. (2008) 'Five Large Chinese Biomedical Bibliographic Databases: accessibility and Coverage', *Health Information & Libraries Journal*, 25: 55–61.

Xu, X. (2020) 'Performing under "the Baton of Administrative Power"? Chinese Academics' Responses to Incentives for International Publications', *Research Evaluation*, 29: 87–99.

Xu, X., Oancea, A., and Rose, H. (2021) 'The Impacts of Incentives for International Publications on Research Cultures in Chinese Humanities and Social Sciences', *Minerva*, 59: 469–92.

Yang, W. (2016) 'National Natural Science Foundation of China: Funding Excellent Basic Research for 30 Years', *Nature*, 537: 1–5.

Ying, C. (2011) 'A Reflection on the Effects of the 985 Project', *Chinese Education & Society*, 44: 19–30. DOI: 10.2753/CED1061-1932440502.

Zheng, Y., and Guo, X. (2019) 'Publishing in and about English: Challenges and Opportunities of Chinese Multilingual Scholars' Language Practices in Academic Publishing', *Language Policy*, 18: 107–30.

Zheng, X., Zhou, W., Ni, C., and Wang, C. (2022) 'The Influencing Mechanism of Research Training on Chinese STEM Ph.D. students' Career Interests', *Asia Pacific Education Review*, DOI: 10.1007/s12564-022-09775-4.

Zhou, P., Su, X., and Leydesdorff, L. (2010) 'A Comparative Study on Communication Structures of Chinese Journals in the Social Sciences', *Journal of the American Society for Information Science and Technology*, 61: 1360–76.

Zong, X., and Zhang, W. (2019) 'Establishing World-Class Universities in China: Deploying a Quasi-Experimental Design to Evaluate the Net Effects of Project 985', *Studies in Higher Education*, 44: 417–31.

Research Evaluation, 2024, 33, 1–13
https://doi.org/10.1093/reseval/rvae036
Article