Journal of Comparative Economics 51 (2023) 235–258

Contents lists available at ScienceDirect

Journal of Comparative Economics

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

Melting pot or salad bowl: Cultural effects on industrial similarity during trade liberalization

Yang Jiao [a], Leilei Shen [b], Yuyun Liu [c,*]

[1] Department of Business of Public Policy, Young Harris College, 1 College Street Young Harris, GA, 30582, USA
[2] Department of Economics, Kansas State University, 327 Waters Hall, 1603 Old Claflin Place., Manhattan, KS, 66506
[3] International School of Business & Finance, Sun Yat-sen University, Zhuhai, Guangdong, China. P.R.

A R T I C L E I N F O

JEL:
F1
O1
R1
Z1

Keywords:
Cultural difference
Industrial similarity
Trade liberalization
Knowledge spillover

A B S T R A C T

This paper examines the impact of culture on the industrial similarity between two adjacent counties during the process of trade liberalization between 1998 and 2007. The findings show that trade accelerates the industrial similarities between counties. The evolution of similarity is unequal between regions and over time, as industries in neighboring counties that belong to different dialect regions become more similar than industries within the same dialect region. The rise in industrial similarity is being fueled by the emergence of high-skill-intensive businesses as a result of trade-related technology spillovers. Adjacent counties of different dialects have greater skill endowments, which facilitate information transfer.

1. Introduction

The literature on the function of formal institutions in industrial agglomeration and industrial similarity is substantial. However, little is known about the impact of informal institutions such as culture, which is defined as traditional ethnic, religious, and social beliefs, language, and values passed down from generation to generation. Using a comprehensive firm-level dataset from China that spans from 1998 to 2007, our study fills this gap by revealing the causal relationship between cultural division and industrial similarity during the process of trade liberalization. Understanding the linkage is essential to China and many African countries where trade liberalization has occurred amid cultural and linguistic diversity.

The present paper has two focuses. First, it analyzes the effects of trade shocks on industrial similarity within a region by leveraging cross-market variance in import exposure. As noted in Imbs (2004), the similarity of industries for two distinct regions measures the industrial interconnection and reliance. We also compare the similarity index with the industrial co- agglomeration index in Ellison et al. (2010), and the results show that these two indices are substantially correlated. Our findings help to understand the impact of trade on the industrial connectedness of two adjacent counties, which has received little attention in the literature. Second, this paper examines the impact of cultural division on the dynamics of industrial simi- larity during trade liberalization. We devise a novel study design to distinguish cultural impact from institutional effect. In particular, to disentangle cultural effects from other effects, such as institutional factors, we carefully select the research units – pairs of neighboring counties located in the same prefecture.

The analyses yield several classes of results. First, the empirical estimates suggest that, overall, industries in two neighboring

* Corresponding author
E-mail address: liuyuyun3@mail.sysu.edu.cn (Y. Liu).

https://doi.org/10.1016/j.jce.2022.09.002
Received 13 April 2022; Received in revised form 20 July 2022; Accepted 6 September 2022
Available online 7 October 2022

counties become more assimilated, regardless of whether they belong to the same or a different dialect division. Quantitatively, a one thousand dollar increase in regional exposure to imports per person increases the industrial similarity by approximately 9.4% for all counties in our sample. Second, cultural division plays a substantial role in shaping industrial similarity for regions that face exogenous trade shocks. Although the level of industrial similarity is higher for counties within the same dialect division at the beginning of the period, there is an additional increase in industrial similarity of 0.2 percentage points for counties with different dialects from a one thousand dollar increase in regional exposure to imports, which represents a 11.7% increase for counties with different dialects. This allows counties with different dialects to close the gap in industrial similarity compared to counties with the same dialects. Our findings are robust to a battery of sensitivity checks (see Section 4.2 for details).

Industrial location is driven by exogenous endowments such as technologies, labor, and natural resources, and the spatial pattern is formed through interindustry specialization with industries settling in locations with comparative advantages. Trade liberalization increases the scope for specialization along the lines of comparative advantage and enhances the importance of access to international markets (Melitz,2003;Melitz and Ottaviano,2008). Our results support this theory.

Since the influential paper of Coe and Helpman(1995), several studies have found that the international transmission of technology through the channel of trade has been a significant contributor to industrial agglomeration within a region. However, the transfer of technology is highly asymmetric across regions within a country, and regions whose industrial structures are more skill-intensive will facilitate such growth; therefore, industrial agglomeration occurs. We further investigate whether cultural segregation leads to differential abundance of skills across regions. The results show that two neighboring counties that belong to different dialect groups are actually equipped with a higher level of skills when compared to counties within the same dialect region. Furthermore, by reducing skills in neighboring counties with the same dialect and raising skills in neighboring counties with different dialects, trade increases the variety of skill abundance.

Trade-based information exchange has been shown to contribute to economic agglomeration on a global scale. However, there is a significant disparity in technology transfer between the regions of a nation. The amount of technology transferred when China imports goods and technology from more developed countries varies greatly depending on the abundance of local skilled labor and the skill intensity of industries, which supports the expansion of high-skilled sectors. The rise of high-skilled industries should therefore outperform the growth of low-skilled sectors. As a result, counties with more skill-intensive industrial structures would aid this growth, which leads to industrial agglomeration. Our findings imply that increased industrial similarity is mostly due to the rise of high-skill-intensive industries as a result of trade-related technology spillovers. Counties with distinct dialects have a greater skill abundance, which facilitates information spillovers and enhances high-skill industry similarities.

Additionally, we perform several heterogeneous analyses. We first divide the cities into high and low trade exposure based on their position in the distribution of trade value. In cities that were more exposed to trade, industries became more assimilated. We further classify cities into high and low skills according to the educational attainment of the labor force. When a city has a higher proportion of high-skilled workers, the results show that industries become increasingly comparable and thus build more agglomerated industrial structures. Finally, we test the impact of government connection on industrial similarity by exploring the difference between capital and non-capital cities. Closer government relations signify more government control and intervention in the establishment of industry policies, which prevents capital cities from becoming increasingly similar.

As with all studies that employ a Bartik instrument or shift-share approach, it is important to verify the validity of our findings. There are two sources of violation. One arises from the shift part, or the trade shocks are not independent of local market conditions. To alleviate the concern, we conduct several sensitivity tests, including controlling for the pre-existing trends in the sectoral changes in the local labor market. We also show that pre-period changes in the outcomes of interest are uncorrelated with subsequent trade shocks predicted by the instrument. To put it in a formal test, we perform a balancing test proposed under the identification assumptions in Borusyak et al.(2018) and show that at a more aggregated industry level, exports are uncorrelated with an industry-specific weighted average of other local shocks such as changes in educational attainment, the employment share, the number of firms, regional GDP, etc.[1]

The other source is from the geographic distribution of the industrial structure. It is possible that our results are simply due to varying industrial specialization across regions. For instance, prefectures that initially specialized in industries that require large capital investment would experience a relative increase in industrial concentration over this period even without import shocks. This issue is at the heart of Goldsmith-Pinkham et al.(2020), who emphasizes how, with shift-share-style variables, identification relies on the exogeneity of the initial industry shares. We thus calculate Rotemberg weights as proposed by Goldsmith-Pinkham et al.(2020) and show that they are less concentrated in a few industries. Both validity tests suggest that our shift-share instrument yields consistent estimates.

Our paper brings together three strands of literature. It first adds to the growing literature on culture and economic outcomes, with a particular interest in the determinants of industrial agglomeration. There are two common ways in the literature to measure culture. One is based on survey questions, whose answers are aggregated at the country level to measure values and beliefs. The other approach is based on a clear division in language, religion, or ethnicity. It is increasingly recognized that culture is an important determinant of a series of economic outcomes, including international trade, distribution policy, corruption, etc. (Guiso et al.,2009; Barr and Serra, 2010;Luttmer and Singhal,2011). Our study, to the best of our knowledge, is the first to explore the impact of culture on the industrial similarity of two areas.

[1] As proved in Borusyak et al., (2018), exogenous independent shocks to many industries cause the Bartik/shift- share estimator to be consistent, even when the shares are not exogenous.

Second, this paper complements previous studies about the determinants of industrial agglom- eration in the context of China. Using the same database that we use, Lu and Tao(2009) find that industrial agglomeration in China increased steadily between 1998 and 2005, although it is still considerably lower than that of selected developed countries such as France, the United Kingdom, and the United States. Bai et al.(2004) study the impact of local protectionism, as measured by the share of state-owned enterprises (SOEs), on industrial agglomeration in China. They note that less geographic concentration is found in industries with stronger local government protection. The findings in our study lend some support to their results and provide affirmative evidence that both international trade and interconnections with the government play important roles in shaping the industrial structure of the local economy. Moreover, we show that cultural division across regions also matters.

Our paper also relates to the literature on culture and trade. One strand of the literature examines how culture affects trade. Aker et al.(2014) uses data from Niger and argues that transaction costs are higher for trade between regions with different ethnicity than for trade in homogeneous areas. On the one hand, trust is higher among people of the same ethnic group, and this reduces transaction costs. On the other hand, ethnic diversity may lead to specialization and complementarity in production that increases the motivation for trade. Another strand of the literature examines how trade affects culture. Gershman(2020) finds that the historical African slave trade contributed to the propagation of persistent witchcraft beliefs. Areas that were more exposed to the slave trade have a lower level of trust currently (Nunn and Wantchekon,2011). In areas more exposed to the African slave trade, Teso(2019) finds long-run persistence and the cultural transmission of women substituting for missing men by taking up traditionally male tasks.

This paper is structured as follows. Section 2 presents the background of cultural division and its relation to industrial location choice in China. Section3 describes the dataset, constructs the key variables, and develops the specification strategy. Section4 reports the main results, followed by an analysis that investigates and quantifies the possible channels by which culture may affect industrial concentration during the process of trade liberalization. Section6 investigates the heterogeneous effect of regions with high and low exposure to trade shocks. The last section concludes the paper.

2. Background

Chinese culture is one of the world's oldest cultures, originating thousands of years ago. The area over which this culture prevails is extremely diverse and varying, with customs and traditions that differ greatly across provinces, cities, and even towns. Language is an important aspect of culture that upholds the social norms and values within a group. Language also serves as a medium through which attitudes, ethics, and beliefs are transmitted from one generation to the next.

The Chinese language can be divided into ten different dialectal groups (Li et al.,1988). The use of dialects has long been a tradition in China (Erbaugh,1995) that has not been signifi- cantly weakened by the emergence of Putonghua as the national standardized language. The prevalence of dialects can be traced through Chinese history during which several major waves of migration occurred for geographical and political reasons as dialectal groups tended to isolate themselves both socially and geographically (Purcell,1947). These ten dialect groups' boundaries correspond to the sharpest zones of linguistic variation. The Mandarin dialect roughly occupies the northern and western regions of China. If one drew a steep diagonal line from the northeast near Beijing toward southwestern China, then the space north and west of this line represents this primarily Mandarin-speaking area. The area to the south and east of this line holds most of the main Chinese dialects.

Fig. A1 portrays a dialect atlas of China. In addition to the Mandarin dialect regions, from north to south, the main southeastern dialects are formally known as Jin (spoken in Shaanxi), Wu (spoken mostly in Jiangsu, Zhejiang, and Shanghai), Gan (spoken mostly in Jiangxi), Xiang (spoken widely in Hunan), Min and Hui (also spoken in Taiwan and the Hainan Islands), Hakka (spoken primarily in southern China and Taiwan), and Yue and Ping (Cantonese).

The impact of language on economic activities has been extensively studied in the literature, including economic growth (Alesina et al.,2003), public goods provision (Desmet et al.,2020), and international trade (Melitz and Toubal,2014). Using Chinese data, Liu et al.(2020) found an inverted U-shaped relation between dialect distance and internal migration. Despite the importance of language in shaping an individual's social and economic outcomes, questions re- garding how the division of different linguistic zones affects industrial concentration, especially within a county, remain under-researched.

Some anecdotal evidence suggests a strong relationship between cultural (dialect) division and the location choices of industry (Liu et al.,2021). One of the examples listed here is the geographic division of the fireworks industry. The fireworks industry in China originated in Liuyang City, Hunan Province. In Liuyang, there are still traces of historical gunpowder research and development at the foot of a mountain on the banks of the Liuyang River that are preserved for future generations to revere. Li Tian, a craftsman who lived in the Tang Dynasty in southern Liuyang, was famous for his invention of firecrackers. Because of Tian's invention, Liuyang became synonymous with fireworks and is commonly known as the hometown of firecrackers and fireworks. Since then, "Liuyang Fireworks" branded products are widely recognized in China. The city's hundreds of fireworks companies produce more than 50 percent of China's fireworks, according to Liuyang government data. Other cities, such as Liling and Pingxiang, with close geographic proximity to Liuyang, developed a similar industrial structure with a substantial share of fireworks companies. However, Liuyang's neighboring city, Changsha, does not have a large fireworks industry. A closer examination of these cities reveals that the primarily used dialect in Liuyang, Liling, and Pingxiang is Gan, while the dialect used in Changsha is Xiang. This phenomenon is not unique in China, and other examples include the cotton, chemical fibers, textiles, and coloring industries in Jiangsu Province and the coal and steel industries in Liaoning Province.

3. Data, Defining County Pairs, Key Variables, and the Specification

3.1. Data

Several data sources are used in our analysis. The data used to create each county's industrial structure comes from the National Bureau of Statistics of China's Annual Survey of Industrial Firms (ASIF), which was conducted from 1998 to 2007. The ASIF is China's most comprehensive and representative firm-level dataset, with surveyed businesses accounting for the majority of the country's industrial value-added. The dataset is used to compute matrices in the national income account (for example, GDP) and the significant statistics reported in China Statistical Yearbooks. The dataset includes extensive business information such as industry affiliation, location, and all operational and performance items from the accounting statements that we are interested in, such as output, intermediate materials, book values, and employment.

The number of firms increases from 165,118 in 1998 to 336,768 in 2007, mainly because the manufacturing industry in China grew rapidly over the sample period. These businesses are spread across China's 31 provinces (including four municipalities), 344 cities, and 2,829 counties, in 464 four-digit manufacturing industries. This set of data includes all SOEs. Non-SOEs are only questioned if they have yearly revenues of 5 million RMB (equal to 771,010 US dollars in 2020) or higher. However, because these large enterprises account for roughly 80% of revenues, or a majority of the industrial structure, there is minimal selection bias in our analysis.

The World Trade Flows dataset, created by Feenstra et al.(2005), contains the trade statistics. From 1998-2007, this dataset contains import data by nation and the 4-digit SITC (revision 2) industry code. The data include imports and exports by industry at the country level, which is our dependent variable. The values are reported in nominal US dollars and translated to real US dollars by using the Bureau of Labor Statistics' consumer price index.

3.2. Defining County Pairs

Notably, the administrative division and dialect divisions in China do not always overlap.[2] Our research approach takes advantage of this distinction by comparing the locations within the same administrative unit that belong to the same or different dialect divisions. The county, followed by the city and the province, is China's most disaggregated administrative entity. We concentrate on the industrial structures of two neighboring counties that are administered and managed by the same metropolis. The control group consists of counties that speak the same dialect and are located in the same dialect region, while the treatment group consists of counties that speak different dialects and are located in various dialect regions. Another reason that we choose a pair of neighboring counties that are part of the same city is to keep the protectionism effect separate from local government, as most policies are enacted at the city level.

Defining neighboring counties is essential, but the task is quite challenging because over the past thirty years, Chinese administrative boundaries and codes of counties, cities, or even provinces have experienced significant changes. New counties are established, while existing counties may have been combined into larger ones or even merged into cities due to the urbanization process. As noted in Lu and Tao(2009), between 1998 and 2005, the number of counties in China increased by 366 (from 2,496 to 2,862), and the number of changes in county codes was 648.[3][3]

After adjusting the administrative information for all counties, we should be able to identify the dialect division of each county and compare it with its neighboring county via the following three steps. In the first step, each county is matched with a unique dialect according to Xu (1999). As mentioned above, this book is based on a survey of Chinese regional dialects in 1986. Next, we restrict matches to counties that are governed and managed by the same city and that share the same administrative border, namely, neighboring counties in the same city. In our sample, 2,460 pairs of neighboring counties belong to the same city. In the next step, we separate neighboring counties according to whether they share the same dialect. Approximately 80 percent of the county pairs belong to the control group and share the same dialect.

[2] As noted in Fig. A1, the dialect division line sometimes separates one province into two or more dialects, or two or more provinces belong to the same dialect division.

[3] To mitigate the major changes in the administrative division, we exclude counties that experienced major changes in area expansion or boundaries and borderline changes. There are 2,244 of the 2,835 counties that did not have major changes in their administrative division between 1986 and 2007. We focus on this period for two main reasons. First, our dialect measurement is based on a survey from 1986. Second, the measurement of our outcome variable – the industrial structure – is based on a firm-level dataset for the period between 1998 and 2007. Focusing on this period helps to accurately match all counties with their unique dialect division and substantially alleviates bias due to the measurement errors that result from administrative changes.

3.3. Key Variables

3.3.1. Industrial Similarity

In contrast to the main measures for industrial agglomeration used in the literature, we employ different approaches to analyze the closeness of the industrial structures of two nearby counties.[4] Our approach borrows a concept proposed by Imbs(2004), which states that the comovements of business cycles across regions are driven by the similarity in the industrial structure. If firms and industries that produce products with higher proximity are more likely to interact with one another in various ways, then by including reliance on similar inputs (such as raw materials, labor, or machinery) and dependence on technologies, their industrial interconnectedness may be gauged by the closeness of the industrial structures of the regions. Therefore, the value of the measurements not only reflects firms' interconnectedness and their increased economies of scale but also manifests the dynamic process of industrial movement in two distinct regions.

To cross-validate our similarity index, we construct four different measurements and use them to describe the industrial similarity or industrial proximity (SC_{ij}) between counties i and j. [5] The first and main measure follows the concept of Imbs(2004) and Baxter and Kouparitsas (2005), which assesses the summation of the correlation of sectoral shares:

$$SC1_{ij} = \frac{\sum_{n=1}^{N} s_{in} \times s_{jn}}{\sqrt{\sum_{n=1}^{N} s_{in}^2} \times \sqrt{\sum_{n=1}^{N} s_{jn}^2}} \tag{1}$$

where N is defined as one of the 425 four-digit manufacturing industries, and s_{in} and s_{jn} are the GDP shares of industry n relative to the total manufacturing GDP of county i and county j, respectively. The variable $SC1_{ij}$ takes values in the interval [0,1]. A greater similarity in sectoral structure leads to larger values of $SC1_{ij}$. If s_{in}=s_{jn} for all industries, which means that the shares of each industry are the same across the two counties, then $SC1_{ij}$ is equal to 1.

Our industrial similarity index for two adjacent regions is similar but somewhat different from the coagglomeration index in Ellison et al.(2010). First, they both assess the intensity of indus- trial concentration within a region. Second, industrial similarity is measured at the county-pair and represents the comovement of the industrial structure of two areas. To compare these two measurements, following the approach in EGK(2010)'s paper, we construct a coagglomeration index that is used to measure the extent of industrial clustering within a region.[6] As illustrated in Fig. A2 in the Appendix, these two indices are highly correlated and prove that our similarity index can be considered as an alternative measure for coagglomeration. Furthermore, the similarity index combines the information between inputs and outputs linkage when measuring the comovement of industrial structures for two adjacent regions. Therefore, it is a preferable measure in our context.

Our second measure of industrial similarity is similar to that in Finger and Kreinin(1979):

$$SC2_{ij} = \sum_{n=1}^{N} \min\left(s_{in}, s_{jn}\right) \tag{2}$$

$SC2_{ij}$ also takes values in the interval [0,1]. When the two counties share no industry, it equals 0; if the shares of each industry are the same across the two counties, then it is 1. This similarity measure is the most restrictive of those considered.

The third and fourth measurements are two dummy variables. If the main industry of one county, the industry with the largest GDP share, is also developed in the other county, then SC3ij equals 1 and is 0 otherwise. The variable SC3ij reflects the similarity between the two counties in terms of the main industry. The last measurement, SC4ij, is as follows. If sin=sjn for at least one industry, then SC4ij equals 1 and is 0 otherwise. SC4ij takes the value of 1 as long as the two counties share any industry and takes the value of 0 if the industries developed in county i are different from those in county j. Thus, SC4ij is the strictest industrial difference and the loosest indicator of industrial similarity.

The summary of the statistics of the four industrial similarity measurements between 1998 and 2007 are reported in Table 1. The industrial structures for two adjacent counties are more assimilated in a geographic region according to all measures over the sample period. The industrial proximity for neighboring counties in the same dialect region is higher than that for neighboring counties located in different dialect regions, but the growth rate for counties in the same dialect region is lower. Note that the sample size is substantially larger for counties in the same dialect region than for those in different dialect regions. From a rough comparison between two years, we find an increasing (decreasing) trend for the same (different) group between 1998 and 2007, mainly because of the administrative restructuring of counties in China.

[4] Industrial agglomeration and regional specialization, which are occasionally used interchangeably in the liter- ature (Bai et al.,2004), can be used to measure the geographic concentration of industries. These two conceptions, according to Long and Zhang(2012), are substantially different from one another. Regional specialization analyzes the share of total production concentrated within a small number of industries, whereas in-dustrial agglomeration assesses the interconnection of industries within an area. Despite their popularity, neither metric captures the relationships between enterprises and sectors in different locations.

[5] Note that industrial similarity and industrial proximity are used interchangeably throughout the paper.

[6] The equation and approach are provided in the technical notes of Appendix B.

Table 1
Summary Statistics of Industrial Similarity

	1998				2007			
	Same		Different		Same		Different	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
SC1	0.22	0.25	0.17	0.26	0.23	0.25	0.23	0.26
SC2	0.17	0.17	0.14	0.20	0.18	0.17	0.18	0.19
SC3	0.61	0.49	0.47	0.50	0.64	0.48	0.64	0.48
SC4	0.91	0.28	0.70	0.46	0.91	0.28	0.85	0.35
Observations		1757		414		1876		393

Notes: This table lists the summary statistics of four measurements of industrial similarity used in this paper. It also separates neighboring counties that belong to the same or different dialect regions. SC1 follows the methodology used in Imbs (2004) and Baxter and Kouparitsas (2005), while SC2 follows Finger and Kreinin (1979). SC3 and SC4 are two dummies that indicate whether two counties have same leading industry or at least one industry in common, respectively.

We also report the summary statistics of all other control variables in Table A1 in the Appendix. In this table, we separate them by county *i* and county *j* and report them individually. Notably, counties in the same dialect region differ significantly from counties in other dialect regions. In the same dialect county pairs, the number of employed workers and the number of businesses are higher. They also have a higher degree of light intensity, which is used as a substitute for GDP measurement. In the same dialect county pairs, the average level of education is likewise slightly higher. In general, counties in the same dialect region are wealthier and have larger skilled labor forces than counties in other dialect regions.

3.3.2. Trade Exposure at the County-pair Level

The measure of China's local market exposure to trade is one of the important explanatory factors in our empirical studies. We focus on the exposure to import shocks between 1998 and 2007, which measures the extent to which an area is initially specialized in industries that later see a large increase in the value of imports, to capture the impact of trade liberalization. Our trade exposure metric is computed for each pair of adjacent counties. We focus on import exposure throughout the paper since our results are only significant with import shocks. This suggests that import competition is the major conduit for altering the industrial structure.

Following Autor et al.(2013) (ADH (2013) hereafter), we measure the e ffects of international trade on the local market of a pair of counties *i* and *j* on a per-worker basis while weighting the importance of sectorial trade in a particular local market by using the share of national sectoral employment for a particular labor market. Our basic measure of China's local market trade exposure can be described as follows:

$$\Delta T_{ijt} = \sum_{s} \frac{L_{ijst}}{L_{st}} \frac{\Delta T_{st}}{L_{ijt}} \tag{3}$$

where ΔT_{st} stands for the change in trade between China and the world in industry (sector) *s* between 1998 and 2007. For this reason, we refer to 1998 as the base year. The variable L_{ijt} represents employment in local market *ij* for the base year, while L_{st} represents China's employment in industry *s* in the base year. L_{ijst}/L_{st} represents the share of county pair *ij*'s employment in industry *s* in the base year. Both L_{ijst} and L_{st} are calculated based on ASIF data.

We then construct the instruments for changes in trade exposures as follows. First, we create China's predicted imports from each country by following the gravity model strategy used in Feenstra et al.(2019). We obtain trade flows at the 6-digit HS classification from UN Comtrade, MFN and preferential tariffs from the World Integrated Trade system (WITS) TRAINS Database. Next, we aggregate China's predicted trade flows with the world at the four-digit 2002 China Industry Classification (CIC). Finally, we construct the predicted change in trade exposure per worker as the instrument for the change in trade exposure.[7]

3.4. Identification

3.4.1. Empirical Strategy

To investigate the effects of dialect on industrial similarity during trade liberalization in China, our estimation strategy should be able to identify the effects of import expansion induced by the changes in access to foreign markets on industrial similarity. The identification is based on the assumption that changes in imports are unrelated to prefecture-specific production shocks; thus, the changes are not dialect-specific. In particular, we evaluate the cultural effect on industrial proximity by estimating the following equation:

$$\Delta SC_{cij} = \beta_1 \Delta \text{ Import }_c \times \text{ DDialect }_{ij} + \beta_2 \Delta \text{ Import }_c + \beta_3 X_{ij} + \varepsilon_{cij} \tag{4}$$

where $\Delta SC_{cij} = \Delta SC_{cij,2007} - \Delta SC_{cij,1998}$ measures the change in industrial similarity between two neighboring counties (*i* and *j*) located in city (*c*) between 1998 and 2007. Δ*Importc* measures the city's (c's) exposure to foreign imports as constructed above. $DDialect_{ij}$ is a

[7] The details of instrument construction are mentioned in the Technical Note of Appendix B.1.

dummy variable that indicates whether counties i and j belong to different dialect regions.

The first differencing helps eliminate the difference in the initial industrial structures across county pairs; that is, the analysis used in our study controls for all time-invariant differences across county pairs. Moreover, the identification in Eq.(4) comes from the cross-city variation in the same sample period, which helps control for the time effects that are common to all prefectures, such as eased access to foreign markets and improvement in the regulation of the financial market. The remaining estimation biases of $\Delta Importc$ could be caused by the endogenous changes in imports from other countries between 1998 and 2007 (such as the nonrandom distributions of the input price and technological shocks) that generate time-varying effects on ΔSC_{cij}. To account for the market size and its impact on industrial similarity, our main regression is weighted by the joint geographic area of counties i and j. To accommodate potential heteroskedasticity and serial correlation, we cluster the standard errors at the city level following the suggestion by Bertrand et al.(2004).

In addition to the baseline estimation presented above, in the robustness checks, we add a set of controls X_c for the geographic endowment, economic determinants and demographic composition of the labor market that might independently affect county pairs' industrial integration in the initial period. In particular, our geographic control contains land elevation and soil suitability. The economic determinants include total employment, the number of firms, and light intensity as a proxy for the GDP values of paired counties. The labor market demographics incorporate the share of the population that is illiterate, share of people aged 15 years and older, and average educational attainment.

3.4.2. Validity of Bartik Instruments

Our basic specification in Eq.(4) relates industrial similarity to the import shocks. The trade shock is based on the Bartik (or shift-share) approach that does not explore import expansion at the local level but rather uses a weighted average of national import expansion. This approach relies on the assumption that the other time-varying, region-specific determinants of the out- come variable are uncorrelated with a prefecture's initial industry composition. As indicated in other papers (such as Imbert et al.,2018;Goldsmith-Pinkham et al.,2020;Bombardini and Li, 2020), this is the key threat to identification, and we address this issue in many ways.

The first approach is to control for pre-existing trends in sectoral changes at the county pair level to account for the possibility that a pair of counties with more trade-exposed industries may be on a different trajectory in terms of overall industrial structure outcomes. The second approach is to check whether our results are primarily driven by regions that are incapable of ex ante setting their investment behavior to address foreseen trade shocks. Specifically, we examine whether current industrial changes reflect the anticipated import shocks (see section 4.2.4 for details).

The third approach is to take the complementary view of the identification requirements proposed by Borusyak et al.(2018). According to them, the Bartik-style instrument can be generally formulated as $\sum_k s_{ikt} g_{ikt}$, where g_{ikt} denotes the shock experienced by industry k in period t, and s_{ikt} measures the exposure of city i to the shock. The exposure to each shock is captured by the initial employment share of the industry in local employment, i.e., $L_{ik,t-1}/L_{i,t-1}$. This identification presumes that an industry-level shock is uncorrelated with a weighted average of the unobservable local unobserved shocks, with the weights reflecting the importance of the industry in the local economy. Following this logic, in Table A3 of the Appendix, we perform the balancing test suggested by Borusyak et al.(2018) and show that a weighted average of observable local shocks (e.g., the changes in workforce educational attainment, composition of the working age population, employment share of the local market, and number of firms) is uncorrelated with import values at the industry level.

The last approach is to calculate the "Rotemberg weights" associated with each industry as suggested by Goldsmith-Pinkham et al.(2020). The industry-specific Rotemberg weights measure the importance of each industry in determining the coefficient of interest and capture the degree of sensitivity to misspecification when the exogeneity assumption for the initial industry composition fails. Table A4 of the Appendix shows the highest Rotemberg weights for five industries. These industries are the electronics and electronic equipment manufacturing, vegetable oil processing, coking, chemical fiber manufacturing, and steel-making industries. These top five industries account for 56% of the positive weights in the Bartik estimator, which means that the weights are less concentrated in a few industries. To further assess the sensitivity of our baseline results to any specific industries, we reconstruct the trade shocks and corresponding instruments but exclude one 2-digit CSIC sector at a time. [8]

4. Main Results

4.1. Baseline Results

To address the potential endogeneity issue, we follow Feenstra et al.(2019) in using an instrumental variable (IV) estimation strategy. The results are reported in Table 2, with trade exposure measured by the import value.[9] The validity of our IV estimations

[8] We individually exclude 2-digit CSIC sectors instead of 4-digit sectors because an industry is more accurately measured in an aggregate classification; therefore, this will reduce the measurement error. Moreover, using an aggregate classification alleviates endogeneity and the computational burden.

[9] We also evaluate the model with the export value in Table A2 in the Appendix, but the results are not statistically significant; therefore, we believe that import competition is the main force that shapes the industrial patterns in China. For this reason, throughout the paper, our analysis is primarily based on imports.

Table 2
Baseline Results

	(1)	(2)	(3)	(4)	(5)
Diff Dialect × Δ Imports	0.200** (0.098)	0.150* (0.091)	0.179** (0.079)	0.186** (0.079)	0.234** (0.093)
Δ Imports	0.210*** (0.030)	0.205*** (0.029)	0.186*** (0.028)	0.205*** (0.030)	0.166*** (0.035)
Identification Test of Instruments:					
Kleibergen-Paap rk LM statistic	13.115	12.936	14.036	14.116	15.310
P-value	0.0000	0.0003	0.0002	0.0002	0.0001
Kleibergen-Paap rk Wald F statistic	60.857	46.305	47.789	47.016	42.026
10% maximal IV size	7.030	7.030	7.030	7.030	7.030
First Stage: Diff Dialect × Δ $\widehat{\text{Imports}}$	1.407*** (0.307)	1.109*** (0.015)	1.109*** (0.010)	1.113*** (0.150)	1.102*** (0.109)
Δ $\widehat{\text{Imports}}$	0.311*** (0.108)	0.011 (0.011)	0.014 (0.023)	0.013 (0.011)	0.008 (0.013)
Observations	2,052	2,036	2,036	2,036	2,035
R-squared	0.009	0.021	0.031	0.034	0.007
Econ Controls	NO	YES	YES	YES	YES
Demo Controls	NO	NO	YES	YES	YES
Geo Controls	NO	NO	NO	YES	YES
City-Time Trend	NO	NO	NO	NO	YES

Notes: The results reported are weighted by the total geographic areas of paired counties. Econ Control includes total employment, number of firms, and light intensity as a proxy for GDP for both paired counties. Demo Control includes the share of the population that is illiterate, the share of people aged 15 and older, and average educational attainment. Geo Control contains land elevation and soil suitability. Standard errors are clustered at the prefecture city-level. *** p<0.01, ** p<0.05, * p<0.1.

hinges on the satisfaction of the relevance condition and the exclusion restriction. The relevance condition is confirmed by the highly significant correlation in the first-stage regression reported in the middle panel of Table 2. Moreover, any concern of a weak instrument is ruled out by the result of the Kleibergen-Paap rk Wald F statistic. The first column does not control any covariate and shows that between 1998 and 2007, trade liberalization increased industrial similarity for two neighboring counties. However, when compared to counties from different dialect regions, counties from the same dialect area have a lower rate of industrial proximity expansion.

Even though the first difference estimation eliminates the initial, time-invariant difference in the industrial structure at the county-pair level, it is still important to take into account other confounding factors that might change. As a result, we conduct several regressions by using four different sets of confounding variables, and the results are shown in the remaining columns of Table 2.

Labor market pooling, as mentioned in Frenken et al.(2007), contributes to the process of industrial similarity. We add these variables, and it has a small impact on our baseline results; the estimates in column (2) decline somewhat when including the factors of the total employment, number of firms, and light intensity of the paired counties. Another determinant is labor force specialization in the local economy (Rosenthal and Strange,2003). We calculate the share of the population that is illiterate, the share of people aged 15 years and older, and the average educational attainment to represent labor productivity at the county level based on the annual yearbook of the Chinese National Bureau of Statistics. The results in column (3) increase after adjusting for the labor market's demographic composition.

The regional variations in resource endowments have a significant impact on the placement choice of businesses (Ellison and Glaeser,1999). We employ two proxies for the intensity of resource allocation, land elevation, and soil suitability to adjust for the effects of resource endowments. Column (4) shows the estimates after accounting for resource endowments. The magnitude of the variable of interest – the interaction between counties with the same dialect and imports – is still positive but larger than the baseline, which indicates that geographic variation in resource endowment continues to play an important role in determining China's current industrial structure.

Using the estimates from column (4) as an example, we can compute the economic scale of the impact as follows: for counties with different dialects, a $1,000 increase in county pair exposure to imports results in a 0.02 percentage point increase in industrial similarity or a 9.5 percent increase from the mean. This bridges the gap between counties with distinct dialects and counties with the same dialects in terms of industrial similarity. Moreover, there is an additional increase in industrial similarity of 0.02 percentage points for counties with different dialects from a one thousand dollar increase in regional exposure to imports, which represents a 11.7% increase for counties with different dialects.

4.2. *Sensitivity Analysis*

In this subsection, we conduct a battery of sensitivity checks to address various estimation concerns.

4.2.1. *Controlling for Other Industrial Reforms*

In Appendix Table A6, we include two control variables, namely, the share of SOEs in all domestic enterprises and the number of foreign-invested firms, to account for two ongoing policy reforms in the early 2000s, namely, SOE reform and the loosening of foreign direct investment laws. These additional controls have no effect on our primary findings.

SOE reform remains a major aspect of China's economic reform effort in the 1990s because of their significant contribution to the economy, and its implementation and scope are linked to the structure of local industry. However, following a large loss and commercial failure in the late 1990s, the second stage of SOE reform was marked by ownership transformation (*gaizhi*) and an emphasis on the privatization of SOEs (Song,2015). Since then, the share of SOEs in the country's gross industrial output, for example, fell from one-half in 1998 to one-quarter in 2011 (Gang and Hope,2013).

We use changes in the output share of SOEs between 1998 and 2007 to reflect the severity of SOE reform across different regions. In the original sample period, a substantial shift in the value indicates not only more intense reform but also a greater contribution of SOEs to the local economy. After correcting for changes in the percentage of SOEs in all domestic businesses, we continue to find statistically significant estimates with similar magnitudes, as indicated in column (2). In addition, SOE reform had a positive impact on industrial agglomeration, in part because SOE reform significantly improved resource allocation efficiency (Lu and Tao,2009).

4.2.2. Alternative Measurements of Industrial Similarity and Trade Exposure

Our primary measure of industrial resemblance is based on the idea of regional industry comovement, which may underestimate or overstate the reality of industry coagglomeration. We explore many alternative measures of industrial proximity, as stated above, to see if the conclusions are robust. We calculate the total of two counties' minimum industrial share values (SC2), (ii) determine whether two counties share the greatest industry (SC3), and (iii) ascertain whether two counties share any industry (SC4) (SC4). The third measure is the strictest and takes the smallest number, as shown in Table A7, while the last is the most lenient. The regression findings with these alternate measures are presented in this table. The results are consistent with the baseline findings regardless of how we calculate industrial similarity.

As a condition for joining the WTO, China implemented a large and widespread tariff reduction between 1992 and 1997, but there have been few changes since then (Lu and Yu,2015). By exploiting regional variation in the industrial structure, we can construct another measurement of exposure to substitute for tariff reduction. In this measurement, counties that host industries that experience more tariff reductions should be more exposed to international trade. Conse- quently, we use another measure of trade exposure that maps the extent of tariff reductions across industries with the industrial composition of the county. Table A8 displays the results. The results indicate that the growth rate of industrial similarity for counties with different dialects is faster when they experience a tariff reduction upon WTO accession. In particular, a 10% decrease in tariffs reduces the similarity gap between counties from the same and different dialect groups by 17.5% (see the last column of Table A8).

4.2.3. Sensitivity Test with Alternative Trade Periods

Following ADH (2013), we undertake two placebo tests to examine any pretreatment tendencies as a supplementary robustness check. To begin, we regress the change in industrial similarity between 1998 and 2007 on the change in trade exposure between 1990 and 1997. Table A9 shows the estimation findings in the first three columns. We are unable to obtain a statistically and economically viable approximation. This finding supports our hypothesis that the pretreatment trade liberalization patterns do not predict the future industrial coagglomeration trends, which implies that our hypothesis is correct. Second, we regress changes in the sample period's industrial similarity on changes in import exposure from 2008 to 2017. The results are presented in the last three columns of Table A9, and as expected, the negligible estimates show minimal indication of reverse causation and little trace of an anticipated response in the industrial structure to future trade shocks, which confirms our identification.

4.2.4. Impact of Original Industrial Intensity

Some may argue that whether the counties will become more similar or not depends on the pre-shock industry specializations. Following the approach in Goldsmith-Pinkham et al.(2020), we calculate the top five concentrated industries based on the highest Rotemberg weights. The top five industries with the highest concentrations include the Electronics and Electronic Equipment, Vegetable Oil Processing, Coking, Cotton and Chemical Fiber, and Steel-making Industries. After gradually excluding these five businesses from the calculation, we recalculate the industry similarity index, which is shown in the table below. For instance, SC_{R1} removes the Electronics and Electronic Equipment Industry from the construction of the similarity index, and SC_{R2} removes both the Vegetable Oil Processing and Electronics and Electronic Equipment Industries.

The point estimates for both the trade term and the interaction term decrease but remain statistically significant, as seen in Table A9. Additionally, the estimate of trade is quite stable, which indicates that trade has a significant influence on determining the degree of industrial similarity between two adjacent counties.

4.2.5. Intermediate Input Import Shock for Upstream Industries

Our initial measurement of exposure to trade shocks partly reflects the local industrial structure; therefore, we also create another trade shock that measures import competition in the interme- diate input sectors and evaluate such intermediate input shocks from the upstream industries of a local economy. To compute the import trade shocks, we first examine the input-output table and identify the upstream industries for all intermediate inputs. The weights for the upstream industries are then determined based on the percentage of industrial value added by the intermediate inputs. The employment percentage used for upstream industries is then calculated by multiplying these weights by the employment share for each industry. These intermediate inputs, which account for employment sharing, replace those in Eq.3.

The results with intermediate input import shocks are presented in Table A10. The results appear to confirm our findings that intermediate input shocks increase industrial similarity and that the growth of similarity is more noticeable for counties in different dialectal regions.

4.2.6. Excluding Newly Founded Firms

In the last sensitivity test, we are concerned about the choice of corporate location, which is not random because enterprises may prefer to locate their headquarters or manufacturing facilities in a specific dialect region. The endogenous site selection could result in bias, which is especially important for new businesses formed after trade liberalization. New firms may choose a location with lower manufacturing costs, a convenient supply chain network, or an industrial cluster. If counties with different dialects develop a more similar industrial structure, then this may impact a firm's choice of location. As a result, our estimates are convoluted. To alleviate the concern, we exclude firms founded after 2001 and re-construct the industrial similarity index. Excluding these firms hardly alters our baseline results in columns (1) and (2), as seen in the Table A11.

5. Mechanism

5.1. Trade and Industrial Coagglomeration

The studies on industrial distribution that have already been conducted mainly concentrate on Europe and North America, with an emphasis on the effects of trade liberalization on geo- graphical patterns. Neoclassical trade models, new trade models, and new economic geography models serve as the foundation for the theoretical studies of industrial location and spatial concentration.

According to the theory of neoclassical trade models, industrial location is driven by exogenous endowments such as technologies, labor, and natural resources, and the spatial pattern is formed through interindustry specialization with industries that settle in locations with comparative advantages (Kim,1995). New trade models argue that internal scale economies provide regions with incentives to specialize even without comparative advantages and make firms concentrate their production in a few locations (Krugman, 1980). In new economic geography models, geographic concentration is driven by the interaction of transportation costs and internal scale economies. Demand linkages represent incentives for producers to locate close to buyers, whereas cost linkages generate incentives for consumers to locate close to suppliers (Krugman, 1991). High trade costs prevent tendencies toward concentration, while medium trade costs allow forward and backward linkages to cause industrial agglomeration (Krugman and Venables, 1995).

Globalization and trade are important catalysts for change in developing nations, and they provide financial resources, technologies, management skills, and markets that are required to convert and rebuild old industrial systems inherited from the emerging economy (Lu and Yu,2015;Brandt et al.,2017;Baccini et al.,2019). Trade liberalization increases the scope for specialization along the lines of comparative advantage and enhances the importance of access to international markets (Melitz,2003;Melitz and Ottaviano, 2008). Our findings provide evidence that reduced trade costs during the process of trade liberalization enhance industrial agglomeration within a region and therefore lead to a greater extent of industrial similarity for two regions that are located close to one another.

5.2. Skill Abundance and Technology Spillovers

As noted in Michaels et al.(2019), human interaction plays an important role in the agglomer- ation of economic activities through knowledge spillovers. Indeed, some knowledge is easily shared through communication among people of the same dialects, but such dissemination does not form a higher level of skills. In contrast, some more complex skills and knowledge can be spread among people with different dialects despite the communicative barriers, such as coding. Accordingly, we would expect the abundance of skills to vary across counties located in the same/different dialect regions, and consequently, skills would organically develop as a result of cultural differences.

To determine if cultural differences cause skills to vary across regions, we develop a number of task metrics that indicate skill intensity. We first manually match the ASIF professions to specific task values. As a result, we can generate a task indicator for each occupation. Following Peri and Sparber(2009), we create five work skill measures that utilize the O*NET Work Activities and Work Context Importance scales: cognition, perception, manual skill, strength, and body. Afterwards, the occupation skill indicator is weighted by the percentage of workers in the same occupation in a given county.[10]

The five skill metrics in 1998 and 2007 are shown in Panel B of Table 3. A greater score suggests that a county uses this task input more frequently. As anticipated, in both years, the skills are significantly higher in neighboring counties from different dialect areas than in their same- dialect counterparts. However, between 1998 and 2007, counties with distinct dialects improve their cognitive and perceptual skills, while counties with the same dialect enhance their manual, physical, and body-related skills. The findings in Panel A of Table 3 are used to investigate the impact of cultural division and trade on the abundance of skills. These results demonstrate that trade enhances the diversity of skill abundance by lowering skills in neighboring counties with the same dialect and increasing skills in neighboring counties with different dialects.

Numerous research conducted since the seminal publication of Coe and Helpman(1995) has discovered that trade-based international technology transfer has significantly aided in the industrial agglomeration of a region. However, technology transfer is highly unequal between countries or between areas within a country. When a developing country, such as China, imports goods and technology from more advanced countries, the amount of technology transferred varies substantially depending on the skill intensity of the

[10] The US Occupational Information Network (O*NET), which provides a database of occupation-specific descriptors for 974 jobs, provides statistics on the task composition and skill requirements of occupations. This dataset assigns numerical values to 52 unique employee abilities (referred to as "tasks" or "skills" in this dataset) that are required by each occupation.

Table 3
Impact of Dialect Division and Trade on Skill Abundance

	Cognition (1)	Perception (2)	Manual (3)	Strength (4)	Body (5)
Panel A:					
Diff Dialect × Δ Imports	0.195*** (0.069)	0.208*** (0.077)	-0.176*** (0.024)	-0.167*** (0.033)	-0.168*** (0.033)
Δ Imports	(0.302)	(0.221)	(0.210)	(0.453)	(0.230)
	0.126*** (0.030)	0.076** (0.032)	-0.067** (0.031)	-0.057* (0.032)	-0.069** (0.033)
Identification Test of Instruments:					
Kleibergen-Paap rk LM statistic	5.985	12.985	7.895	6.559	10.400
P-value	0.0032	0.001	0.000	0.0060	0.0012
Kleibergen-Paap rk Wald F statistic	10.026	13.265	11.421	16.473	23.920
10% maximal IV size	7.030	7.030	7.030	7.030	7.030
First Stage:					
Diff Dialect × Δ $\widehat{\text{Imports}}$	1.437*** (0.012)	1.436*** (0.206)	-1.772*** (0.204)	-1.401*** (0.088)	-1.150*** (0.078)
Δ $\widehat{\text{Imports}}$	-1.366*** (0.159)	-1.367*** (0.167)	-1.431*** (0.062)	-1.206*** (0.216)	-1.097*** (0.142)
Panel B:					
1998					
Counties Pair of Different Dialects					
Mean	0.454	0.443	0.413	0.388	0.418
Std. Dev.	0.323	0.306	0.264	0.252	0.268
Counties Pair of Same Dialect					
Mean	0.237	0.242	0.169	0.161	0.175
Std. Dev.	0.206	0.205	0.116	0.110	0.117
2007					
Counties Pair of Different Dialects					
Mean	0.523	0.470	0.397	0.381	0.402
Std. Dev.	0.357	0.303	0.278	0.274	0.283
Counties Pair of Same Dialect					
Mean	0.199	0.185	0.226	0.242	0.226
Std. Dev.	0.132	0.123	0.206	0.206	0.195
Observations	2,035	2,035	2,035	2,035	2,035
R-squared	0.460	0.447	0.400	0.388	0.370
Econ Controls	YES	YES	YES	YES	YES
Demo Controls	YES	YES	YES	YES	YES
Geo Controls	YES	YES	YES	YES	YES

Notes: Panel A displays the coefficients of trade penetration and its interaction with a different dialect dummy on the changes in the task index between 1998 and 2007. The classification of the task index follows Peri and Sparber (2009). Panel B reports the mean and standard deviation for county pairs from the same and different dialect regions in 1998 and 2007, respectively. The results reported are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

industry and is especially pronounced in relatively high-skilled industries (Acharya and Keller,2009). As a result, the growth of high-skilled industries should outpace the growth of low-skilled sectors, and counties with more skill-intensive industrial structures would assist this growth, which results in industrial agglomeration.

To examine whether the rising industrial similarity for the county pairs with high-skilled industrial structures is driven by the growth of high-skilled industries, we first decompose the manufacturing industries in the ASIF into those that are less skill-intensive and those that are more skill-intensive by using UNCTAD data on the skill and technology content of HS 6-digit industries as described in Basu(2011). [11] Since one ASIF industry can map to several SITC industries, we construct two measurements for high- and low-skilled industries. Our first measurement is less restrictive and claims one industry as low-skill-intensive (LSI) if it includes at least one SITC LSI industry. Our second measure is more restricted, and one ASIF industry that contains only LSI SITC industries is defined as LSI. For the more-skill-intensive classification, we include both medium- and high-skill-intensive industries.

We then compute the similarity index for low- and high-skill-intensive industries with the specification in Eq.(1) and run separate regressions for pairs of counties with different dialects and the same dialect. Table 4 displays the results: relative to LSI industries, the more-skill- intensive industries assimilate faster in the pairs of counties located in different dialect regions. We also plot the trend of the industrial similarities of low- and more-skill-intensive industries and separately by county pairs with the same and different dialects in Fig. A3. Importation increases the industrial similarity of high-skill-intensive industries for all pairs of counties. However, these industries grow faster in counties with different dialects. Our findings provide some suggestive evidence that the rising industrial similarity is driven by the growth of high- skill-intensive industries through technology spillovers from importation. Neighboring counties whose initial skills are higher facilitate such growth and thus develop more similar industrial structures.

[11] Based on the skill and technology content of goods, UNCTAD assigns each HS-6 category a basic skill intensity designation. These HS-6 industries are then mapped to SITC industries by using the HSSITC concordance from the Center for International Data at UC Davis. SITC 2-digit manufacturing industries (SITC codes 6, 7, and 8) are defined as low skill-intensive if they consist primarily of "Non-Fuel Primary Commodities", "Resource-Intensive Manufactures", and "Mineral Fuels". SITC 2-digit manufacturing industries that consist primarily of "Technology-Intensive Manufacturing" are instead designated as skill-intensive manufacturing industries.

Table 4
Import and Industrial Similarity by Skill Complexity

	Skill Complexity - 1		Skill Complexity - 2	
	Low	High	Low	High
	(1)	(2)	(3)	(4)
Diff Dialect × Δ Imports	-0.171***	1.614***	-0.030	0.025***
	(0.093)	(0.258)	(0.031)	(0.004)
Δ Imports	0.072	1.588	-0.039	0.037
	(1.229)	(1.357)	(0.038)	(0.042)
Identification Test of Instruments:				
Kleibergen-Paap rk LM statistic	7.496	6.866	7.315	10.765
P-value	0.0062	0.0088	0.0060	0.0010
Kleibergen-Paap rk Wald F statistic	48.542	34.389	247.505	36.157
10% maximal IV size	7.03	7.03	7.03	7.03
First Stage:				
Diff Dialect × $\Delta\ \widehat{\text{Imports}}$	-0.613*** (0.178)	0.634*** (0.135)	-0.256 (0.176)	1.893*** (0.257)
$\Delta\ \widehat{\text{Imports}}$	0.098 (0.840)	-1.032*** (0.275)	1.058*** (0.414)	-0.265 (0.176)
Observations	842	1,058	2,186	2,186
R-squared	0.026	0.012	0.039	0.067
Econ Controls	YES	YES	YES	YES
Demo Controls	YES	YES	YES	YES
Geo Controls	YES	YES	YES	YES

Notes: The ASIF manufacturing industries are first classified as either low or high skill intensive based on publicly available UNCTAD classifications described in Basu (2011). The first measurement is less restrictive and classifies an industry as low skill intensive if it includes at least one low-skill industry in SITC. The second measure is more restrictive, and an ASIF industry that contains only low-skill-intensive SITC industries is defined as a low-skilled industry. For the more-skill-intensive classification, we include both medium- and high-skill-intensive industries. The industrial similarity for low- or high-skill-intensive industries is calculated based on Equation 1. The regressions are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Technology spillover is responsible for the growth of high-technology industries (Dekle,2002; Cingano and Schivardi,2004). In our context, technology spillover occurs when the knowledge that was originally developed by one county (originating county) flows into its neighboring counties (recipient counties). When recipient counties exploit the technology of the originating county, they either apply the originating county's technology directly to similar product lines or combine it with other knowledge to create unique innovations. If this is the case, then the growth of high-skilled industries in counties with different dialects should be associated with an increasing trend in technology spillovers. As a result, two neighboring regions develop a more similar industrial structure.

Following Ellison et al.(2010), our technology flows from one county to another are measured based on the information flows of research and development (R&D). We detail the construction of technology spillovers across counties in Appendix B. The first three columns of Table 5 report the estimates for the changes in knowledge spillovers that concern cultural division and imports. The IV estimation indicates larger volumes of knowledge flow between paired counties if they are located in different dialect regions. The next three columns report the estimates for changes in R&D expenses. Counties from different dialect regions indeed invest more in R&D. Notably, both the knowledge spillovers and R&D expenses are measured in dollars and are reported in logarithmic form.

6. Heterogeneous Effects

The impact of trade liberalization on the industrial similarity between two adjacent counties has been the focus of our research. In this part, we categorize cities based on their trade exposure, talent levels, government interconnectedness, and the varied effects that they have.

6.1. Different Exposure to Trade

We divide cities into high and low groups based on their import values in 1998 to study the heterogeneous effect for regions with varied levels of trade exposure. The numbers over (below) the $50th percentile are classified as high (low) exposure. In the first two columns of Table 6, we show the coefficients for high-exposure and low-exposure locations. There is a clear distinction between regions with high trade exposure and those with low trade exposure, and adjoining counties with two distinct languages acquire a more comparable industrial structure when their trade exposure is strong.

Table 5
Dialect and Trade Impact on Knowledge Spillovers and R/D Expenses

	(1) Knowledge Spillover	(2)	(3)	(4) R/D Expenses	(5)	(6)
Diff Dialect × Δ Imports	0.637** (0.280)	0.456* (0.265)	0.480* (0.280)	0.015** (0.007)	0.014* (0.007)	0.012* (0.007)
Δ Imports	0.868*** (0.078)	0.888*** (0.074)	0.870*** (0.095)	0.009** (0.003)	0.010*** (0.004)	0.008** (0.004)
Identification Test of Instruments:						
Kleibergen-Paap rk LM statistic	9.260	9.230	9.120	6.084	6.530	6.518
P-value	0.0023	0.0024	0.0025	0.0106	0.0106	0.0107
Kleibergen-Paap rk Wald F statistic	54.163	55.208	51.104	60.644	63.365	62.710
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030
First Stage:						
Diff Dialect × Δ $\widehat{\text{Imports}}$	1.690*** (0.223)	1.690*** (0.224)	1.684*** (0.224)	1.171*** (0.101)	1.172*** (0.100)	1.173*** (0.101)
Δ $\widehat{\text{Imports}}$	0.497*** (0.230)	0.498*** (0.229)	0.492*** (0.232)	0.004 (0.007)	0.003 (0.006)	0.005 (0.007)
Observations	2,275	2,275	2,275	2,279	2,279	2,279
R-squared	0.318	0.330	0.332	0.019	0.027	0.035
Econ Controls	YES	YES	YES	YES	YES	YES
Demo Controls	NO	YES	YES	NO	YES	YES
Geo Controls	NO	NO	YES	NO	NO	YES

Notes: Both knowledge spillovers and R&D are in logarithms and measured in dollars. Knowledge spillovers from one county to another includes the information from the technology-flow matrix constructed by Scherer (2007). The detailed procedure for constructing the measure can be found in the technical notes in Appendix B. The regressions are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** p < 0.01, ** p < 0.05, * p < 0.1.

Table 6
Heterogeneity Analysis

	(1) Trade Exposure	(2)	(3) Skill Abundance	(4)	(5) Political Regulation	(6)
	High	Low	High	Low	Capital	Non-Capital
Diff Dialect × Δ Imports	0.227** (0.108)	-1.264 (0.786)	0.572** (0.227)	0.070 (0.211)	0.295*** (0.105)	1.940* (1.052)
Δ Imports	0.135*** (0.036)	1.798 (2.070)	0.099** (0.049)	0.196 (0.261)	0.198*** (0.025)	3.506*** (1.055)
Identification Test of Instruments:						
Kleibergen-Paap rk LM statistic	12.874	11.083	9.244	6.060	10.616	15.166
P-value	0.0003	0.0009	0.0060	0.0102	0.0011	0.0001
Kleibergen-Paap rk Wald F statistic	21.256	18.329	23.117	80.973	56.119	28.792
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030
First Stage: Diff Dialect × Δ $\widehat{\text{Imports}}$	1.384*** (0.026)	1.066*** (0.024)	1.409*** (0.020)	1.419*** (0.088)	1.159*** (0.230)	1.426*** (0.010)
Δ $\widehat{\text{Imports}}$	0.196 (0.205)	0.435*** (0.178)	1.619*** (0.242)	1.206*** (0.140)	1.185*** (0.132)	1.215*** (0.162)
Observations	1,046	980	684	537	85	1,951
R-squared	0.723	0.141	0.514	0.107	0.123	0.373

Notes: The first two columns report the estimates for cities with different degrees of trade intensity. High (low) exposure is defined as trade values in 1998 they lie above (below) the 50th percentile. The next two columns display the estimates for cities with different skill intensities. High- (low-) skilled cities are defined by whether the educational attainment of the labor force is placed in the top (bottom) 30th percentile. The last two columns present the results for capital and non-capital cities. The results reported are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** p<0.01, ** p<0.05, * p<0.1.

6.2. Different Skills Abundance

Despite the fact that our regressions account for labor force skill, it is still fascinating to see whether regions with a higher skilled workforce integrate faster than regions with a lower skilled workforce. This adds to the evidence for our theory and indicates that the process of knowledge spillovers differs when regions vary by different levels of ability due to their industrial systems. Regions with greater skill levels benefit more from knowledge spillover through imports than regions with lower skill levels. As a result, we divide the county pairings into two categories based on the labor force's skill intensity. The average level of education, which is just the average education for two paired counties, is used to determine the skill intensity.[12] The top (bottom) one-third of the distribution represents the high (low) skill group defined by schooling. The results are presented in the middle two columns of Table 6. The industrial similarity rate is actually higher in cities with a less-trained labor force.

6.3. Capital/Non-capital Cities

Industrial similarity may be influenced by the extent to which local government intervenes in the local market. Despite the fact that the share of SOEs in the local market is regulated, their relationship with the government may differ depending on whether the city is a capital city. On the one hand, when the provincial government develops its economic strategy, the investment policy may benefit the capital city by attracting more investment and enhancing industrial agglomeration. On the other hand, the capital's industrial structure may be more regulated by the government, which reduces the likelihood of industrial resemblance. The last two columns of Table 6 provide the results of industrial similarity based on capital city status between 1998 and 2007. In non-capital cities, the similarity grows faster than in capital cities during the pre-trade period, which demonstrates that greater government intervention slows the process of industrial absorption. Our findings corroborate the evidence presented in Bai et al.(2004).

7. Conclusion

For a national or regional economy, industrial agglomeration has been regarded as a source of long-term economic growth. Economists have extensively researched the economic factors that contribute to the process of industrial similarity. In the present study, we investigate the dynamic effect of culture on industrial structure similarity throughout China's trade liberalization process by focusing on the relationship between cultural division and industrial similarity. Our research adds to the large body of literature by featuring cultural characteristics. Our study also contributes to understanding the relationship among trade, the regional industrial structure, and culture. Our findings shed new light on the cultural impact in forming the industrial structure and determining that industrial distribution has received little academic attention.

Acknowledgments

We would like to thank the editor Hongbin Li, two anonymous referees, Yonghong Zhou, Le Wang, Ruohan Wu and participants at various seminars for helpful comments and suggestions, and Zhou Zhiying for the excellent research assistance. Yuyun Liu acknowledges the grants of National Natural Science Funds of China (Grant NO.71803198), humanities and social science Fund of Ministry of education of China (Grant NO.18YJC790107). All errors are our own.

Appendix A. Tables and Figures

Table A1, Table A2, Table A3, Table A4, Table A5, Table A6, Table A7, Table A8, Table A9, Table A10, Table A11, Fig. A1, Fig. A2 and Fig. A3.

[12] Since trade could accelerate economic inequality across regions, workers might migrate to places with better economic opportunity, thereby changing the structure of the labor force. To avoid such endogeneity, the skill classification is only based on education in 1998.

Table A1
Summary Statistics of All Variables

	Same Dialect(1998)			*Different Dialects(1998)*		
	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs
Panel A - Industrial Similarity Measurement Similarity01	0.22	0.25	1757	0.17	0.26	411
Similarity02	0.17	0.17	1757	0.14	0.20	411
Whether same Lead Industry	0.61	0.49	1757	0.46	0.50	411
Whether same Industry	0.91	0.28	1757	0.70	0.46	411
Panel B - County Characteristics County i Employment Number	15827.08	41709.22	1757	6547.b27	12942.37	411
Enterprises Number	50.39	118.82	1757	25.88	44.83	411
Light Intensity	3.00	5.66	1757	1.37	2.80	411
Share of Population under 15 years	1.00	0.00	1757	1.00	0.00	411
Average Educational Attainment	5.96	1.12	1757	5.56	1.56	411
Share of Illiterate Population	25.12	11.44	1757	30.18	17.60	411
Elevation	783.87	745.63	1757	1444.95	1411.14	411
Soil Suitability	0.41	0.37	1757	0.30	0.32	411
County j Employment Number	7800.13	12545.15	1757	4460.27	10569.74	411
Enterprises Number	33.91	54.23	1757	20.06	42.41	411
Light Intensity	1.91	3.18	1757	0.93	1.99	411
Share of Population under 15 years	1.00	0.00	1757	1.00	0.00	411
Average Educational Attainment	5.68	1.03	1757	5.28	1.58	411
Share of Illiterate Population	27.39	12.21	1757	32.32	19.03	411
Elevation(j)	803.32	701.22	1757	1448.99	1393.58	411
Soil Suitability	0.43	0.37	1757	0.32	0.31	411
Panel C - Exposure to Trade						
Changes in Imports	5037.807		7006.08		411	
Changes in Input Imports from High-Income Countries	4148.365		5284.805		411	
Changes in Exports	9908.067		8721.631		411	

Notes: Panel A and Panel B details the dependent variables and independent variables used in the estimation. Values in Panel A are reported at the level of pairs of adjacent counties while values in Panel B are at county level. Similarity01 is the primary measure of industrial similarity used in Imb (2004). Similarity02 takes the sum of the minimum values of any two industries for a pair of counties, as constructed in Finger and Kreinin (1979). Panel C shows changes in trade exposure at the city level, including imports, inputs from high-income nations, and exports.

Table A2
Results of Exports

	(1)	(2)	(3)	(4)	(5)
Diff Dialect × Δ Exports	0.001	0.001	0.001	0.001	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Δ Exports	0.003	0.003	0.005* (0.003)	0.004	0.003
	(0.003)	(0.003)		(0.003)	(0.003)
Identification Test of Instruments:					
Kleibergen-Paap rk LM statistic	36.772	36.171	36.704	36.121	40.695
P-value	0.0000	0.0000	0.0000	0.0000	0.0000
Kleibergen-Paap rk Wald F statistic	237.054	142.486	144.473	137.711	106.436
10% maximal IV size	7.030	7.030	7.030	7.030	7.030
First Stage: Diff Dialect × Δ $\widehat{\text{Exports}}$	0.722*** (0.023)	0.695*** (0.042)	0.686*** (0.041)	0.684*** (0.044)	0.655*** (0.043)
Δ $\widehat{\text{Exports}}$	0.029	0.033	0.032	0.032	0.017
	(0.035)	(0.022)	(0.022)	(0.022)	(0.013)
Observations	2,052	2,036	2,036	2,036	2,052
R-squared	0.001	0.006	0.022	0.025	0.003
Econ Controls	NO	YES	YES	YES	YES
Demo Controls	NO	NO	YES	YES	YES
Geo Controls	NO	NO	NO	YES	YES
Province-Time Trend	NO	NO	NO	NO	YES

Notes: The values of exports are measured in $100,000. No weights are used in the re- gressions in the odd columns, and the total geographic areas of paired counties are used as weights in the regressions in the even columns. Standard errors are clustered at the prefecture city-level. *** p < 0.01, ** p < 0.05, * p < 0.1.

Table A3
Balance Test

	Coefficients	Standard Error
Employment Share*i*	0.240	0.784
Employment Share*j*	0.184	1.059
Firm per Population*i*	-0.001	-0.610
Firm per Population*j*	0.000	0.581
Light Intensity*i*	0.000	0.892
Light Intensity*j*	0.000	0.595
Share of Illiterate Population*i*	0.034	1.237
Share of Illiterate Population*j*	-0.023	-2.149
Average Educational Attainment*i*	1.147	0.143
Average Educational Attainment*j*	-2.877	-0.296
Share of Population under 15*i*	0.112	0.097
Share of Population under 15*j*	0.431	0.308
Elevation*i*	-0.239	-0.178
Elevation*j*	-0.095	-0.080
Soil Suitability*i*	-1.345	-0.943
Soil Suitability*j*	0.178	0.085

Notes: This table reports coefficients from regressing industry- specific weighted averages of prefecture characteristics on industry shocks and year fixed effects. Standard errors are clustered at 3-digit CIC codes. Regressions are weighted by average industry exposure.
The sample includes 298 industry × period observations.

Table A4
Influential industries – Rotemberg Weights

CIC	Industry Name	Alpha1	Beta1	Cumulative Weight
366	Electronics and Electronic Equipment	2.77	-0.01	0.22
133	Vegetable Oil Processing	1.34	0.00	0.33
231	Coking Industry	1.34	0.03	0.43
171	Cotton and Chemical fiber industry	0.83	0.00	0.50
134	Steel-making Industry	0.77	-0.01	0.56

Table A5
Robustness Checks – SOE/FDI Reforms

	(1) Baseline	(2) SOE	(3) FDI	(4) SOE & FDI
Diff Dialect × Δ Imports	0.200**	0.213	0.212	0.213
	(0.083)	(0.081)	(0.081)	(0.081)
Δ Imports	0.210**	0.209***	0.208*	0.209***
	(0.034)	(0.030)	(0.029)	(0.030)
Δ SOE Share		-0.011		-0.011
		(0.079)		(0.080)
FDI Firms			-0.000	-0.000
			(0.000)	(0.000)
Identification Test of Instruments:				
Kleibergen-Paap rk LM statistic	31.115	14.049	14.050	14.049
P-value	0.0000	0.0002	0.0002	0.0002
Kleibergen-Paap rk Wald F statistic	60.857	44.957	46.010	45.093
10% maximal IV size	7.030	7.030	7.030	7.030
First Stage:				
Diff Dialect x Δ $\widehat{\text{Imports}}$	1.407***	1.152***	1.393***	1.395***
	(0.307)	(0.162)	(0.033)	(0.032)
Δ $\widehat{\text{Imports}}$	0.311***	0.302***	0.301***	0.302***
	(0.108)	(0.112)	(0.111)	(0.114)
Observations	2,052	2,052	1,828	1,828
R-squared	0.325	0.325	0.416	0.417
Econ Controls	YES	YES	YES	YES
Demo Controls	YES	YES	YES	YES
Geo Controls	YES	YES	YES	YES

Notes: The results reported are weighted by the total geographic areas of paired counties. Δ*SOEShare* measures the changes in the share of State-Owned Enterprises (SOE) in all domestic firms between 1998 and 2007. FDI firms takes the logarithm of the number of FDI firms in 1998. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A6
Robustness Checks – Alternative Measurements of Similarity

	(1) SC2	(2)	(2) SC3	(4)	(5) SC4	(5)
Diff Dialect × Δ Imports	0.106*** (0.024)	0.153** (0.062)	0.187*** (0.019)	0.563*** (0.189)	0.588** (0.258)	0.858*** (0.180)
Δ Imports	0.140*** (0.044)	0.092 (0.056)	0.202*** (0.011)	0.340*** (0.114)	0.199*** (0.013)	0.325*** (0.103)
Identification Test of Instruments:						
Kleibergen-Paap rk LM statistic	11.718	16.386	9.667	14.773	9.667	14.773
P-value	0.0002	0.0000	0.0019	0.0001	0.0019	0.0001
Kleibergen-Paap rk Wald F statistic	44.034	67.122	35.439	63.840	35.439	63.840
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030
First Stage: Diff Dialect × Δ $\widehat{\text{Imports}}$	1.162*** (0.121)	1.100*** (0.066)	1.148*** (0.132)	1.093*** (0.066)	1.147*** (0.132)	1.094*** (0.065)
Δ $\widehat{\text{Imports}}$	1.445*** (0.025)	1.201*** (0.100)	1.433*** (0.008)	1.173*** (0.106)	1.432*** (0.007)	1.174*** (0.116)
Observations	2,457	2,283	2,269	2,253	2,269	2,253
R-squared	0.243	0.037	0.041	0.076	0.024	0.127
Econ Controls	NO	Yes	NO	Yes	NO	Yes
Demo Controls	NO	Yes	NO	Yes	NO	Yes
Geo Controls	NO	Yes	NO	Yes	NO	Yes

Notes: The second measurement of similarity follows Finger and Kreinin (1979), which takes the sum of the minimum values of any two industries for a pair of counties. The third and fourth measurements are two dummies. If two counties have the same industry with the largest share, *SC*3 takes value one. If two counties have at least one common industry, the value of *SC*4 equals one. The results reported are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A7
Robustness Checks – Measuring Trade by Tariff Reduction

	(1)	(2)	(3)	(4)
Diff Dialect × Δ Tariff	-0.022* (0.014)	-0.024** (0.013)	-0.033** (0.012)	-0.026*** (0.009)
Δ Tariff	-0.025 (0.025)	-0.024 (0.024)	-0.027** (0.013)	-0.019** (0.009)
Observations	2,052	2,052	2,052	2,052
R-squared	0.093	0.105	0.268	0.272
Econ Controls	NO	YES	YES	YES
Demo Controls	NO	NO	YES	YES
Geo Controls	NO	NO	NO	YES
For a 10% decrease in tariff reduces the similarity gap between counties of same and different dialects by	15.9%	16.1%	19.4%	17.5%

Notes: Changes in exposure to trade are measured by the reduction in tariffs between 1998 and 2007 following the measurement of Lu and Yu (2015). The regressions are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A8
Robustness Checks – Alternative Trade Periods

VARIABLES	(1) OLS	(2) Reduced	(3) IV	(4) OLS	(5) Reduced	(6) IV
Diff Dialect × Δ	0.015		0.011	0.005		0.004
	(0.019)		(0.018)	(0.004)		(0.003)
Δ Imports	0.006		0.006	0.007		0.006
IV(Diff Dialect × Δ Imports)		0.012			0.003	
		(0.021)			(0.003)	
IV(Δ Imports)		0.006			0.006	
		(0.006)			(0.005)	
Identification Test of Instruments:	(0.006)		(0.006)	(0.005)		(0.005)
Kleibergen-Paap rk LM statistic			19.166			20.181
P-value			0.0000			0.0000
Kleibergen-Paap rk Wald F statistic			414.929			181.633
10% maximal IV size			7.030			7.030
First Stage:						
Diff Dialect $\times \Delta \widehat{\text{Imports}}$			1.155*** (0.046)			1.019*** (0.003)
$\Delta \widehat{\text{Imports}}$			-0.004* (0.002)			-0.305*** (0.083)
Observations	1,006	1,006	1,006	1,006	1,006	1,006
R-squared	0.132	0.131	0.132	0.132	0.131	0.132
Econ Controls	YES	YES	YES	YES	YES	YES
Demo Controls	YES	YES	YES	YES	YES	YES
Geo Controls	YES	YES	YES	YES	YES	YES

Notes: The results reported are weighted by the total geographic areas of paired counties. The trade measurement in columns (1) to (3) is calculated based on the time period between 1990–1997, while trade measurement in columns (4) to (6) is calculated based on 2007–2016. The regressions are weighted by the total geographic areas of paired counties. Standard errors are clustered at the prefecture city-level. *** p < 0.01, ** p < 0.05, * p < 0.1.

Table A9
Robustness Checks – Removing Most Concentrated Industries

	(1) Baseline	(2) SC_{R1}	(3) SC_{R2}	(4) SC_{R3}	(5) SC_{R4}	(6) SC_{R5}
Diff Dialect × Δ Imports	0.186**	0.182**	0.175*	0.174*	0.176*	0.170**
	(0.079)	(0.089)	(0.091)	(0.091)	(0.092)	(0.086)
Δ Imports	0.205***	0.154***	0.158***	0.157***	0.154***	0.154***
	(0.030)	(0.038)	(0.037)	(0.037)	(0.038)	(0.036)
Identification Test of Instruments: Kleibergen-Paap rk LM statistic	14.116	16.079	16.013	16.013	16.013	16.078
P-value	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
Kleibergen-Paap rk Wald F statistic	47.016	41.365	41.167	41.167	41.167	41.179
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030
First Stage: Diff Dialect $\times \Delta \widehat{\text{Imports}}$	1.113***	1.078***	1.079***	1.078***	1.080***	1.078***
	(0.150)	(0.108)	(0.109)	(0.109)	(0.110)	(0.109)
$\Delta \widehat{\text{Imports}}$	0.013	0.332***	0.331***	0.331***	0.332***	0.330***
	(0.011)	(0.113)	(0.114)	(0.114)	(0.113)	(0.114)
Observations	2,036	1,831	1,827	1,827	1,826	1,821
R-squared	0.034	0.038	0.036	0.036	0.037	0.035
Econ Controls	YES	YES	YES	YES	YES	YES
Demo Controls	YES	YES	YES	YES	YES	YES
Geo Controls	YES	YES	YES	YES	YES	YES

Notes: According to the computed Rotemberg weights, The top five industries with the highest concen- trations of imports include Electronics and Electronic Equipment Industry, Vegetable Oil Processing Industry, Coking Industry, Cotton and Chemical fiber Industry, and Steel-making Industry. After gradually deleting them, the similarity indices of SC_{R1} - SC_{R2} are recalculated.

Table A10
Robustness Checks – Intermediate Input Import Shock

	(1) SC1	(2)	(3) SC2	(4)	(5) SC3	(6)	(7) SC4	(8)
Diff Dialect × Δ Imports	0.120***	0.113***	0.101***	0.097***	0.080***	0.124***	0.111***	0.108***
	(0.008)	(0.017)	(0.008)	(0.016)	(0.024)	(0.038)	(0.012)	(0.027)
Δ Imports	0.116***	0.106***	0.108***	0.093***	0.125***	0.141***	0.124***	0.110***
	(0.004)	(0.015)	(0.004)	(0.013)	(0.007)	(0.026)	(0.007)	(0.020)
Identification Test of Instruments: Kleibergen-Paap rk LM statistic	8.798	8.080	9.041	8.040	8.798	8.080	8.798	8.080
P-value	0.0030	0.0045	0.0026	0.0046	0.0030	0.0045	0.0030	0.0045
Kleibergen-Paap rk Wald F statistic	324.082	240.651	346.488	242.672	324.082	240.651	324.082	240.651
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030	7.030	7.030
First Stage:								
Diff Dialect × Δ $\widehat{\text{Imports}}$	1.211***	1.123***	1.213***	1.228***	1.211***	1.230***	1.211***	1.229***
	(0.046)	(0.058)	(0.045)	(0.058)	(0.046)	(0.058)	(0.046)	(0.058)
Δ $\widehat{\text{Imports}}$	1.428***	1.376***	1.420***	1.380***	1.418***	1.377***	1.428***	1.376***
	(0.048)	(0.019)	(0.010)	(0.018)	(0.048)	(0.0189)	(0.048)	(0.028)
Observations	1,263	1,247	1,426	1,303	1,263	1,247	1,263	1,247
R-squared Econ Controls	0.448	0.476	0.326	0.570	0.179	0.244	0.276	0.329
Demo Controls Geo Controls	NO NO NO	YES YES YES	NO NO NO	YES YES YES	NO NO NO	YES YES YES	NO NO NO	YES YES YES

Notes: The results reported are weighted by the total geographic areas of paired counties. SC1-SC4 are defined and computed in the same way they were previously. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A11
Robustness Checks – Excluding Firms Founded after 2001

	(1) SC1	(2)	(3) SC2	(4)	(5) SC3	(6)	(7) SC4	(8)
Diff Dialect × Δ Imports	0.209**	0.182**	0.168	0.274***	0.022	0.028*	0.393	0.445*
	(0.103)	(0.089)	(0.107)	(0.097)	(0.159)	(0.010)	(0.296)	(0.267)
Δ Imports	0.234***	0.155***	0.044	0.066**	0.043	0.011	0.072	0.066
	(0.037)	(0.038)	(0.041)	(0.033)	(0.041)	(0.097)	(0.054)	(0.058)
Identification Test of Instruments: Kleibergen-Paap rk LM statistic	14.314	16.079	14.088	16.518	14.088	16.518	14.088	6.518
P-value	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001
Kleibergen-Paap rk Wald F statistic	41.408	41.365	79.756	62.710	79.756	62.710	79.756	62.710
10% maximal IV size	7.030	7.030	7.030	7.030	7.030	7.030	7.030	7.030
First Stage:								
Diff Dialect × Δ $\widehat{\text{Imports}}$	1.078***	1.386***	2.042***	1.684***	2.042***	1.684***	2.042***	1.684***
	(0.108)	(0.035)	(0.193)	(0.226)	(0.193)	(0.226)	(0.193)	(0.193)
Δ $\widehat{\text{Imports}}$	0.011	0.331***	1.433***	1.173***	0.792***	0.492***	0.792***	1.173***
	(0.012)	(0.114)	(0.008)	(0.232)	(0.202)	(0.232)	(0.202)	(0.101)
Observations	1,847	1,831	2,444	2,279	2,444	2,279	2,444	2,279
R-squared Econ Controls	0.012	0.038	0.002	0.090	0.001	0.016	0.007	0.036
Demo Controls Geo Controls	NO NO NO	YES YES YES	NO NO NO	YES YES YES	NO NO NO	YES YES YES	NO NO NO	YES YES YES

Notes: The results reported are weighted by the total geographic areas of paired counties. SC1-SC4 are defined and computed in the same way they were previously. Standard errors are clustered at the prefecture city-level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

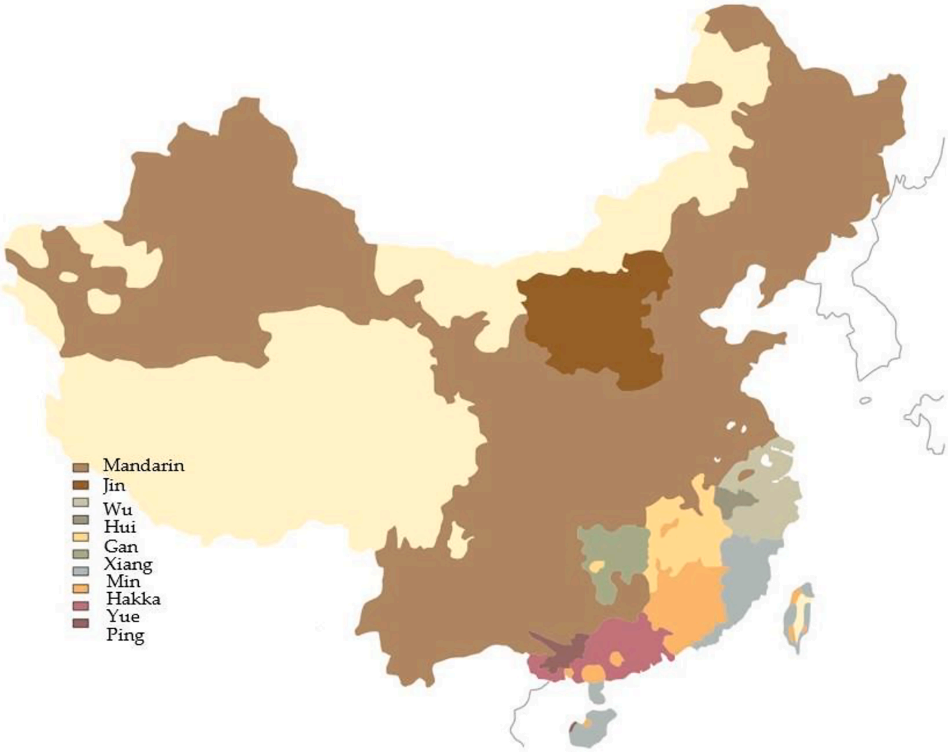

Fig. A1. Atlas of Chinese Dialects

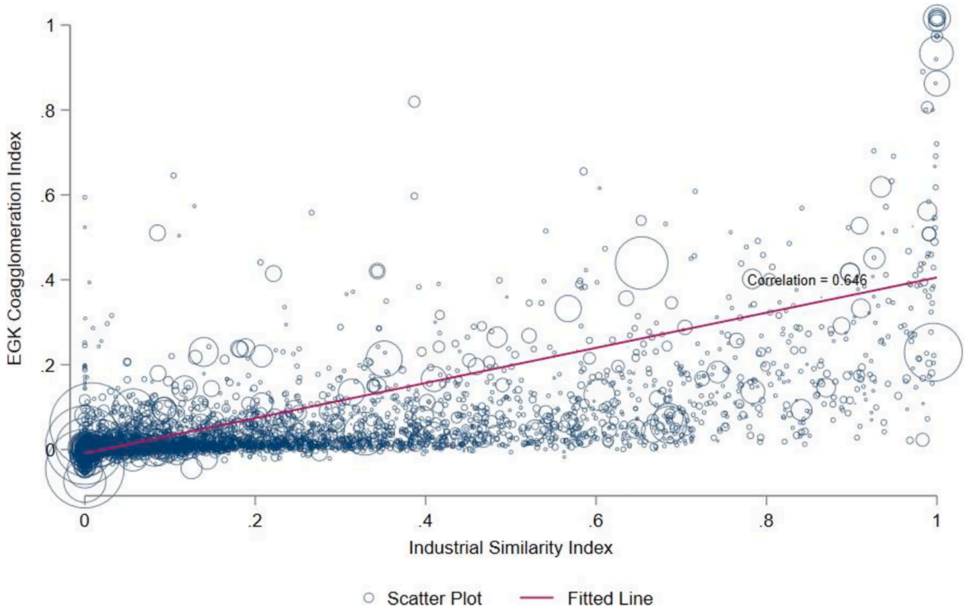

Fig. A2. Comparison between EGK Coagglomeration Index and Industrial Similarity Index

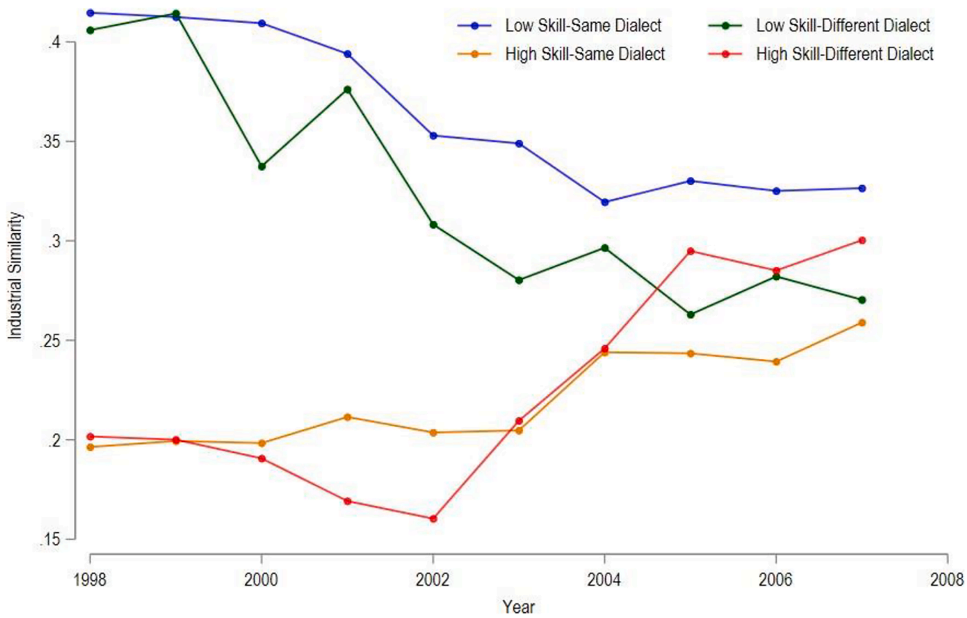

Fig. A3. Similarity of High- and Low-Skilled Industries between 1998 and 2007

Appendix B. Technical Notes

B.1. Construction of Instruments of Trade

In particular, Feenstra, Ma, and Xu (2019) show that in the situation of monopolistic competition, a constant elasticity of substitution (CES)-based utility model can estimate China's imports of good j from nation e as follows:

$$lnM_{jt}^{China,e} = \gamma_{jt} + \gamma_1 \ln\left(\tau_{jt}^{China,e}\right) + \gamma_2\left(\sum_{k\neq China} M_{jt}^{k,e}\right) + \gamma_3 \ln\left(\sum_{k\neq China} \frac{M_{j0}^{k,e}}{\sum_{k\neq China} M_{j0}^{k,e}}\left(\tau_{jt}^{k,e}\right)^{\sigma-1}\right) + \beta_4 d^{China,e} + \varepsilon_{jt}^{e} \tag{5}$$

China's imports of good j from country e at time t, $M_{jt}^{China,e}$, are determined by the following five items. The term $\tau_{jt}^{China,e}$ corresponds to the import tariffs imposed by China on exporter e on good j at time t. The coefficient γ_1 is expected to be negative, and based on the structural model, $\gamma_1 = 1-\sigma$, where σ represents the elasticity of substitution. The term $\ln\sum_{k\neq c} M_{jt}^{k,e}$ corresponds to the total exports from e to all other countries excluding China, which is equivalent to country e's export demand from the rest of the world. Note that this term is similar to the instrument used in ADH (2013), where U.S. imports from China are instrumented by using other high-income countries' imports from China. The term that interacts with γ_3 corresponds to a measure of the average import tariffs that country e faces when exporting good j to the rest of the world, which differs from other countries' MFN tariffs if, for instance, e and partner country k are members of the same preferential trade agreement. In this case, the weights used rely on the trade flows from 1990 (base year). The term $d^{China,e}$ stands for the distance between China and country e. Notably, term γ_{jt} captures other changes in China's demand for good j over time. ε_{jt}^{e} is the error term, which captures export variety and the marginal costs of production from the rest of the world over time.

We estimate the above Eq.(5) by using the following specification:

$$lnM_{jt}^{China,e} = \alpha_{et} + \alpha_{ej} + \alpha_{jt} + \beta_1 \ln\left(\tau_{jt}^{China,e}\right) + \beta_2 \ln\left(\sum_{k\neq China} \frac{M_{jo}^{k,e}}{\sum_{k\neq China} M_{jo}^{k,e}} \left(\tau_{jt}^{k,e}\right)^{\sigma-1}\right) + \varepsilon_{jt}^{e} \tag{6}$$

where the set of country-time fixed effects (α_{et}) captures the distance between the U.S. and country e and other measures of trade costs that could change over time such as the formation of a free trade agreement between the U.S. and country e. In addition, the demand from the rest of the world for country e's product j at time t is absorbed by the country-time fixed effects. The set of country-product fixed effects, α_{ej}, controls for China's time-invariant demand factors for good j from country e. Finally, the set of product fixed effects that varies by year, α_{jt}, captures other changes in China's demand for good j over time.

Coefficient β_1 is expected to be negative, and based on the literature, we set $\sigma = 3$ in calculating the average import tariff, whose coefficient is β_2. Notably, this assumption is well within the range of the reported measures of the elasticity of substitution in the literature. We also consider setting parameter σ to values ranging from 1 to 7, and the results are robust to different values for this parameter. β_2 is expected to be positive since if country e faces higher average tariffs on good j from the rest of the world, then e will export more to China. Based on Eq.(6), we construct the predicted natural logarithm of China's imports from country e in good j at time t, $\ln \widehat{M_{jt}^{China,e}}$, and taking its exponential, we obtain China's predicted imports from country e in good j at time t, $\widehat{M_{jt}^{China,e}}$

In the second stage, we sum China's expected global imports at the two-digit CIC industry level, as represented by industry s, to obtain the following:

$$\widehat{M}_{st} = \sum_{j\in s}\sum_{e} \widehat{M_{jt}^{China,e}} \tag{7}$$

In the last step, we construct the instrument for China's local market i's import exposure by using the following:

$$\Delta\widehat{M}_{it} = \sum_{s} \frac{L_{is,t^0}}{L_{s,t^0}} \frac{\Delta\widehat{M}_{st}}{L_{i,t^0}} \tag{8}$$

where $\frac{L_{is,t0}}{L_{s,t0}}$ denotes the employment share of the local market in industry s of the initial period.

ΔMst is China's predicted change in imports from the world in industry s between 1998 and 2007.

B.2. Regional Coagglomeration Index for Counties i and j

Following the approach in Ellison, Glaeser, and Kerr (2010), the coagglomeration index is constructed:

$$\gamma_{ij} = \frac{\sum_n (s_{in} - x_n)(s_{jn} - x_n)}{1 - \sum x_n^2}$$

where n indexes industry. s_{in} is the share of industry n's employment in county i. x_n is the employment share in industry n in the country, which measures the aggregate size of industry n.

B.3. Explanation of the Construction of the Knowledge Spillover Index

For a county pair ij, we define $Tech\ Spill\ Over_{i\leftarrow j}$ as how R&D in county j benefits county i, or the direction of spillover in R&D is from j to i. We follow the same approach for $Tech\ Spill\ Over_{j\leftarrow i}$ to describe the spillover from i to j.

$$Tech\ spill\ over_{i\leftarrow j} = \sum_k W_k \times \sum_l RD_{j,l} \times \alpha_{k\leftarrow l}$$

where W_k measures the GDP weights of industry k in county i. $\alpha_{k\leftarrow l}$ is how much one dollar spent on R&D in industry l benefits R&D in industry k, i.e., the spillover coefficient. $RD_{j,l}$ is actual R&D in county j and industry l. $\sum_l RD_{j,l} \times \alpha_{k\leftarrow l}$ is the spillover effect from all industries in county j on industry k in county i. $\sum_k$ is the weight of industry k in county $i \times \sum_l RD_{j,l} \times \alpha_{k\leftarrow l}$ is the spillover effect from all industries in county j on all industries in county i.

Similarly, we can construct $T\ ech\ Spill\ Over_{j\leftarrow i}$, which indicates the benefit for county j:

$$Tech\ spill\ over_{j\leftarrow i} = \sum_k W_k \times \sum_l RD_{i,l} \times \alpha_{k\leftarrow l}$$

To compute $\alpha_{k\leftarrow l}$ or the R&D across industries, we follow Scherer (2007) and construct an R&D matrix. This matrix indicates how R&D in industry l flows to (is used in) another industry k.

This industry-level technology spillover illustrates how industry k receives inflows of R&D from other industries.

References

Acharya, R.C., Keller, W., 2009. Technology transfer through imports. Canadian Journal of Economics/Revue canadienne d'économique 42 (4), 1411–1448.
Aker, J.C., Klein, M.W., O'Connell, S.A., Yang, M, 2014. Borders, ethnicity and trade. Journal of Development Economics 107, 1–16.
Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R., 2003. Fractionaliza- tion. Journal of Economic Growth 8 (2), 155–194.
Autor, H.D., Dorn, D., Hanson, G.H., 2013. Trade liberalization and markup dispersion: evidence from china's wto accession. American Economic Review 103 (6), 2121–2168.
Baccini, L., Impullitti, G., Malesky, E.J., 2019. Globalization and state capitalism: Assessing vietnam's accession to the wto. Journal of International Economics 119, 75–92.
Bai, C.-E., Du, Y.J., Tao, Z., Tong, S.Y., 2004. Local protectionism and regional specializa- tion: evidence from china's industries. Journal of International Economics 63 (2), 397–417.
Barr, A., Serra, D., 2010. Corruption and culture: An experimental analysis. Journal of Public economics 94 (11-12), 862–869.
Basu, S.R., 2011. Retooling trade policy in developing countries: Does technology intensity of exports matter for gdp per capita. Policy Issues in International Trade and Commodities 56.
Baxter, M., Kouparitsas, M.A., 2005. Determinants of business cycle comovement: a robust analysis. Journal of Monetary Economics 52 (1), 131–157.
Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in- differences estimates? The Quarterly journal of economics 119 (1), 249–275.
Bombardini, M., Li, B., 2020. Trade, pollution and mortality in china. Journal of Interna- tional Economics, 103321.
Borusyak, K., Hull, P., Jaravel, X., 2018. Technical report. National Bureau of Economic Research.
Brandt, L., Van Biesebroeck, J., Wang, L., Zhang, Y., 2017. Wto accession and performance of chinese manufacturing firms. American Economic Review 107 (9), 2784–2820.
Cingano, F., Schivardi, F., 2004. Identifying the sources of local productivity growth. Journal of the European Economic association 2 (4), 720–742.
Coe, D.T., Helpman, E., 1995. International r&d spillovers. European economic review 39 (5), 859–887.
Dekle, R., 2002. Industrial concentration and regional growth: evidence from the prefectures. Review of Economics and Statistics 84 (2), 310–315.
Desmet, K., Gomes, J.F., Ortuño-Ortín, I., 2020. The geography of linguistic diversity and the provision of public goods. Journal of Development Economics 143.
Ellison, G., Glaeser, E.L., 1999. The geographic concentration of industry: does natural advantage explain agglomeration? American Economic Review 89 (2), 311–316.
Ellison, G., Glaeser, E.L., Kerr, W.R., 2010. What causes industry agglomeration? evidence from coagglomeration patterns. American Economic Review 100 (3), 1195–1213.
Erbaugh, M.S., 1995. Southern chinese dialects as a medium for reconciliation within greater china. Language in Society 24 (1), 79–94.
Feenstra, R.C., Lipsey, R.E., Deng, H., Ma, A.C., Mo, H., 2005. Technical report. National Bureau of Economic Research.
Feenstra, R.C., Ma, H., Xu, Y., 2019. Us exports and employment. Journal of International Economics 120, 46–58.
Finger, J.M., Kreinin, M.E., 1979. A measure of 'export similarity' and its possible uses. The Economic Journal 89 (356), 905–912.
Frenken, K., Oort, F.V., Verburg, T., 2007. Related variety, unrelated variety and regional economic growth. Regional Studies 41 (5), 685–697.
Gang, F., Hope, N.C., 2013. The role of state-owned enterprises in the chinese economy. China-United States Exchange Foundation, US-China 2022, 1–21.
Gershman, B., 2020. Witchcraft beliefs as a cultural legacy of the atlantic slave trade: Evidence from two continents. European Economic Review 122, 103362.
Goldsmith-Pinkham, P., Sorkin, I., Swift, H., 2020. Bartik instruments: What, when, why, and how. American Economic Review 110 (8), 2586–2624.
Guiso, L., Sapienza, P., Zingales, L., 2009. Cultural biases in economic exchange? The quarterly journal of economics 124 (3), 1095–1131.
Imbert, C., Seror, M., Zhang, Y. and Zylberberg, Y. (2018), 'Migrants and firms: Evidence from china'.
Imbs, J., 2004. Trade, finance, specialization and synchronization. Review of Economics and Statistics 86 (3), 723–734.
Kim, S., 1995. Expansion of markets and the geographic distribution of economic activities: the trends in us regional manufacturing structure, 1860–1987. The Quarterly Journal of Economics 110 (4), 881–908.
Krugman, P., 1980. Scale economies, product differentiation, and the pattern of trade. The American Economic Review 70 (5), 950–959.
Krugman, P., 1991. Increasing returns and economic geography. Journal of Political Economy 99 (3), 483–499.
Krugman, P., Venables, A.J., 1995. Globalization and the inequality of nations. The quarterly journal of economics 110 (4), 857–880.
Li, R., Wurm, S., Baumann, T., Lee, M.W, 1988. Language atlas of China: parts I and II. Longman Group (Far East) Limited, Hong Kong.
Liu, Y., Jiao, Y., Xu, X., 2020. Promoting or preventing labor migration? revisiting the role of language. China Economic Review 60.
Liu, Y., Zhou, Z., Chen, Q., 2021. Linguistic identity monopoly in the spatial distribution of chinese industrial chain.
Long, C., Zhang, X., 2012. Patterns of china's industrialization: Concentration, specializa- tionand clustering. China Economic Review 23 (3), 593–612.
Lu, J., Tao, Z., 2009. Trends and determinants of china's industrial agglomeration. Journal of Urban Economics 65 (2), 167–181.
Lu, Y., Yu, L., 2015. Trade liberalization and markup dispersion: evidence from china's wto accession. American Economic Journal: Applied Economics 7 (4), 221–253.
Luttmer, E.F., Singhal, M., 2011. Culture, context, and the taste for redistribution. American Economic Journal: Economic Policy 3 (1), 157–179.
Melitz, J., Toubal, F., 2014. Native language, spoken language, translation and trade. Journal of International Economics 93 (2), 351–363.
Melitz, M.J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. econometrica 71 (6), 1695–1725.
Melitz, M.J., Ottaviano, G.I., 2008. Market size, trade, and productivity. The review of economic studies 75 (1), 295–316.
Michaels, G., Rauch, F., Redding, S.J., 2019. Task specialization in us cities from 1880 to 2000. Journal of the European Economic Association 17 (3), 754–798.
Nunn, N., Wantchekon, L., 2011. The slave trade and the origins of mistrust in africa. American Economic Review 101 (7), 3221–3252.
Peri, G., Sparber, C., 2009. Task specialization, immigration, and wages. American Economic Journal: Applied Economics 1 (3), 135–169.
Purcell, V., 1947. Chinese settlement in malacca. Journal of the Malayan Branch of the Royal Asiatic Society 20 (1), 115–125.

Rosenthal, S.S., Strange, W.C., 2003. Geography, industrial organizationand agglomera- tion. Review of Economics and Statistics 85 (2), 377–393.
Song, L., 2015. Routledge Handbook of the Chinese Economy. Routledge, chapter State and Non-State Enterprises in China's Economic Transition, New York.
Teso, E., 2019. The long-term effect of demographic shocks on the evolution of gender roles: Evidence from the transatlantic slave trade. Journal of the European Economic Association 17 (2), 497–534.
Xu, B.M.I., 1999. Dictionary of Chinese Dialect. Zhonghua Book Company.