Market Access, Labor Mobility, and the Wage Skill Premium: New Evidence from Chinese Cities

Hongbing Li
Beijing University of Posts and Telecommunications
hongbin li@bupt.edu.cn

Hongbo Cai
Beijing Normal University
hongbocai@bnu.edu.cn

Suparna Chakraborty[+]
University of San Francisco
schakraborty2@usfca.edu

[+] Corresponding Author: Dept. of Economics, University of San Francisco, 2130 Fulton Street, San Francisco, CA 94117. Phone: (415) 422-4715;

Market Access, Labor Mobility, and the Wage Skill Premium: New Evidence from Chinese Cities

Abstract

As labor market reforms in developing economies cause economic regions to open up, providing greater market access to businesses, what happens to labor mobility and the consequent wage skill premium? Who gains – skilled or unskilled workers? Using data collected from a survey of 331 Chinese cities, this paper empirically analyzes the impact of increased market access on the wage skill premium. Using the theoretical framework of new economic geography and the tools of spatial externality combined with quantile regression techniques, we find that there exists an inverted U-shaped relationship between market access and the wage skill premium. An initial increase in market access initially attracts high-skilled labor with promises of an increased return that increases the wage skill premium till a threshold, beyond which it starts to decline. The conduit of this relationship is the impact that market access has on returns to education. The observed pattern is robust to alternative tests.

Key words: Market Access, Labor mobility, Wage Skill Premium, New Economic Geography, Spatial Externality

JEL Classification Code: F16, J24, J31, J61

1. Introduction

A causal observation of developing countries since 1980s tells us that the high-skilled workers' wages have witnessed a much larger increase as compared to the wages of low-skilled workers resulting in an *expansion* of the wage skill premium – a puzzling observation given the tenets of the new classical economics. According to the standard Stolper-Samuelson theorem of international trade, given that developing countries have a relative abundance of low skill labor, an increase in trade would imply an increase in demand for output produced by low skilled workers where the comparative advantages lie. This, in turn, would increase the wages of the low skill labor relative to the high skill labor narrowing the wage gap.

This theoretical proposition, however, is in contrast to the data from developing economies, especially China – the focus of our study. China is particularly well suited as our test case as empirical research has unambiguously shown its strong export orientation motivated by availability of cheap, unskilled labor (Fernald et. al., 1999; Ahearne et. al., 2006). According to the tenets of the Stolper-Samuelson theorem, we should have seen an increased return to unskilled labor and a narrowing of the wage gap as China became one of the world's largest exporter of labor intensive goods. Yet, according to the China Household Income Project (CHIP) survey, urban high-skilled worker's hourly wage in China increased from 4.26 Renminbi (RMB) in 1995 to 16.74RMB in 2007, and urban low-skilled worker's hourly wage increased from 3.09RMB in 1995 to 10.92RMB in 2007. Thus, for every 1RMB earned by the low skilled worker, the high skilled worker earned 1.38RMB in 1995 but 1.51RMB in 2005, suggesting a *widening* of the Chinese wage gap and intensification of wage inequality[1] and bringing into question the success of a sustainable development plan for the Chinese economy (Zhou et al., 2012)[2].

To explain this puzzle, the general equilibrium analysis of new economic geography

[1] To calculate the wage skill premium, we use the CHIP database and partition the skill levels by labors' education level. Workers who have a college degree and above are regarded as high-skilled labors. Workers who have a high school diploma and below are regarded as low-skilled labors.

[2] The negative effect of trade, or offshoring on labor through a decline in labor demand of medium and low skilled labor has also been documented by Foster-McGregor, Poeschl and Stehrer (2016)

(hereinafter, NEG) has considered alternative explanations like monopolistic competition, transportation costs and increasing returns to scale (Krugman, 1991; Fujita et al., 1999)[3]. NEG also allows introduction of spatial factors which can provide a reasonable explanation of the spatial concentration of economic activities, the return to education and income inequality attributable to regional geography.

In our study, we apply a core concept of NEG——market access (henceforth, MA), and define the workers' nominal wage in the NEG wage equation as the function of the local MA. In addition, we attempt to go a step further and analyze the transmission channel of this relationship through returns to education.

In the traditional NEG model, market access captures the interaction among returns to scale, spatial distribution of market demand and trade cost, which, in turn, determines the relationship between the final consumption and enterprise production (Krugman, 1991). In general, NEG theorizes that the regions which are closer to the economic epicenter would also experience higher returns to education and higher workers' wage (Fujita, 1999; Fingletion, 2006). Higher market access reduces the transportation cost of attracting enterprises to a centralized location that increases the wages in this region. Secondly, the enterprises' spatial agglomeration pushed by market access has prominent technology spillovers, promoting local labor productivity. For example, Campos and Dabusinskas (2009), studying the labor market in Estonia, find that returns to alternative occupation is a core reason for labor mobility amongst Estonian female labor force. Thus, taking market competition into consideration, the enterprises would pay a higher factor price, especially to attract qualified employees by high wages creating a "space lock" of wage advantage in economic centers (Ottaviano & Pinelli, 2006). In this context, NEG has been successfully applied by Redding and Venables (2004), Hanson (2005) and Head and Mayer (2006) to ascertain the positive impact of market access on nominal wages. For example, Redding and Venables (2004) show that a significant proportion of cross-

[3] In the existing literature, the scholars used different alternatives to measure the core market concepts of new economic geography. Hanson (2005), Head & Mayer (2004, 2006), uses market potential while Fujita et al. (1999), Redding & Venables (2004), Combes et al. (2008), Hering & Poncet (2010), Kamal et al. (2012), use market access. In our work, we use the description of Hering & Poncet (2010) modified to Chinese data.

country variation in per capita income could be attributed to geography. Head and Mayer (2006) document the importance of education while studying the link between geography and wage differentials.

Early research on Chinese skilled wage gap has so far concentrated on foreign direct investment (FDI), technology change and labor market institutions based on the works of Feenstra et al. (1996) and Acemoglu (1998). For example, Li and Xu (2008) finds that FDI can increase skilled workers' wage while Shao et al. (2010) suggests that the increase of industry export intensity may lead to a skill-based technological transformation in that industry. In a related study, Zhou et al. (2012) estimates the influence of trade, technology and institution on the wage premium of both skilled and unskilled workers and finds that trade and technology enhancements enlarge the wage gap, but employment protection policies of the political parties in action decrease the wage premium, so the ultimate effect is ambiguous[4].

One shortcoming of the above research is the absence of accounting for or analyzing any worker or enterprise characteristics that has been emphasized by labor economists as an important determinant of skill premium (Mincer, 1974). For example, Troske (1999) shows that employer size has a large and positive impact on employee wages. This becomes particularly important in the case of Chinese labor markets, which are often characterized by segmentation and discrimination affecting frictionless labor mobility, a basic assumption of traditional trade models. If we ignore these characteristics, the final estimation is likely to be biased.

Secondly, empirical research using the NEG framework has mostly relied on macro level data. It is only recently that some scholars have started analyzing micro level data in the NEG framework to analyze the effect of market access on average wages of a particular sector. The effect of market access on the wage skill premium is still a new area of research which has gained traction since the work of Hering and Poncet (2010) who used Chinese Household Income Project (CHIP) 1995 data to study the impact of market

[4] Other studies attempting to explain the wage gap include Liu et al. (2007), Liu and Yin (2008), Fan and Zhang (2009), and Xu (2012).

access on wage skill premium. Kamal et al. (2012) focused on average wages rather than the skill premium. Using CHIP1995 and CHIP2002, they found that during 1995 and 2002 market access is an important determinant of average wages, and both skilled and unskilled workers can benefit from higher market access.

We build our analysis from the baseline analysis of Hering and Poncet (2010). Updating the data to include the period post World Trade Organization (WTO) membership of China in 2002 when there were many labor market reforms to adhere to the international standards, and expanding the original study limited to 56 cities to all 331 cities in China, we apply the core NEG framework to study the relationship between market access and wage skill premium in China. To this end, we delve deeper and expand the basic NEG framework to allow for individual as well as enterprise level characteristics, and document their impact on the wage-gap.

Our work adds to the currently literature in two ways – one methodological and the second more conceptual in nature.

From a methodological standpoint, we create an updated index of market access by accessing provincial input-output matrix of China and using this matrix to calculate market access potential of every city in China- our current market access potential calculation encompasses 331 Chinese cities from urban metropolis like Shanghai to relatively smaller cities like Langfang in the Hebei province of China. In addition, we overcome the sample selection bias and endogeneity issues plaguing earlier research by incorporating a Heckman sample selection methodology with instrumental variables in our quantile regression analysis.

From a more conceptual standpoint, earlier studies have concentrated on supply and demand interactions to find the effect of market access on the labor market. In our study, we look at the issue of human capital and skill more carefully by looking at the issue of wage gap from the perspectives of returns to education by asking a simple question: does market access influence returns to education? The answer is “yes”. Not only does market access influence returns to education but more importantly, the interaction between market access and returns to education is an important conduit through which market access

influences the wage skill premium[5]. Our empirical analysis shows an *inverted U-shape relationship* between market access and wage skill premium: before reaching a critical mass, attempts to promote market access may widen the wage skill premium through increasing returns to education up to a threshold, after which the skill premium narrows.

The rest of the paper is organized as follows. In the next section, we describe our model, variables and the data used for the empirical analysis in greater detail. Section 3 presents the empirical analysis. Section 4 presents the results of our robustness checks and Section 5 concludes.

2. Model, Variables and Data

2.1 Model Setup

NEG tells us that technological externality can lead to substantial technology spillover by industry agglomeration in regions with a high level of market access. Working through an increased demand of high-skilled workers implying higher returns to education, it widens the wage skill premium between high-skilled workers and low-skilled workers. These higher wages have the potential to attract more workers to regions with increased market access, exhibiting urban migration phenomenon. On the flip side, skill-premium and discriminatory immigration policies in different regions raises the threshold and cost of labor migration, dampening the impact.

These opposing forces have two entirely different effects on the wage skill premium in high market access (region of final migration) as opposed to low market access areas (origin of migration): on the one hand, along with the decrease of high-skilled workers in regions of origin, the enterprises have to replace high-skilled workers by low-skilled workers, and raise the wage of low-skilled workers, thereby narrowing the wage skill premium. The impact on the region of final migration is the opposite. Increased demand for higher skilled labor necessitates a wage premium to encourage an agglomeration of

[5] The higher returns to education following a comprehensive economic reform is also documented by Campos and Jolliffee (2007) who document, using micro-level data from Hungary, that the large beneficiaries of economic reform are college or university educated labor and those involved in the service industries.

high-skilled workers in regions with increased market access and this contributes to an increase of the wage skill premium.

Our baseline model is common to Fujita et al. (1999), Redding and Venables (2004), Hering and Poncet (2010). In addition to these baseline models, in order to investigate the role of returns to education as a conduit and allowing for any nonlinearity in the relationship, we respectively introduce the interaction of a binary variable indicating skill and a measure of market access, and the interaction of a binary variable of skill and a measure of market access squared (to capture the non-linearities of this relationship) to give us our amended baseline regression, summarized as:

$$\begin{aligned}\ln w_{ic} = {} & \alpha_0 + \alpha_1 \ln MA_c + \alpha_2 \ln MA_c \times skill + \alpha_3 \ln MA_c^2 \times skill + \alpha_2 \ln gdp + \alpha_5 \ln capital \\ & + \alpha_6 marriage + \alpha_7 age + \alpha_8 age^2 + \alpha_9 gender + \alpha_{10} edu + \alpha_{11} SOE + \mu_{ic} + \varepsilon_{ic}\end{aligned} \quad (1)$$

where lnw_{ic} stands for hourly wages of labor i in city c. Our major variables of interest are the log of market access index of a city (MA_c) and its interactions with the skill dummy variable ($lnMA \times skill$), the interaction of its square and skill dummy variable ($lnMA^2 \times skill$), per capita real GDP ($lngdp$) and quantity urban human capital ($lncapital$) at a particular location. In our baseline specification, the skill dummy takes a value 1 if the labor has a college degree or above and is classified as high-skill, and is 0 otherwise (in our robustness checks, we test for alternative skill cut-off levels). In addition to the above variables, we also include individual characteristics to control for heterogeneous labor characteristics that can potentially impact our results. These labor characteristics include marital status ($marriage$, 1 for married, 0 for not), age (age), square of age (age^2), gender ($gender$, 1 for male, 0 for female), education level (edu), as well as ownership pattern of the enterprises (SOE,1 for state-owned, 0 for not state-owned). μ_{ic} captures all other controls, including whether the city under consideration is a provincial capital, or a port city, as well as industry type for the enterprises being studied and individual occupation of the labor included in the final sample (whether they are blue or white collar workers and nature of their primary job). The subscripts i, c respectively denotes worker and city. ε_{ic} is random disturbance term.

2.2 Theoretical Framework

2.2.1 Estimation of the Trade Equation

Following the two-region trade model of Redding and Venables (2004), we first express the total trade between region *r* and region *j* as

$$n_r p_r X_{rj} = n_r p_r^{1-\delta} T_{rj}^{1-\delta} G_j^{\delta-1} E_j \tag{2}$$

where E_j is the final consumption of manufacturing output in region *j*. X_{rj} is region *r*'s effective demand of output produced in region *j*. p_r is the ex-factory price and T_{rj} is a measure of iceberg transportation cost. G_j is the general price index of manufacturing products. σ is the measure of elasticity of substitution among differentiated products, with an assumption that σ>1.

The above equation has three critical components: (1) $n_r p_r^{1-\delta}$: this component measures the market supply capacity of export region *r*. (2) $G_j^{\delta-1} E_j$: this component measures the market demand potential of region *j*. (3) $T_{rj}^{1-\delta}$: this component measures the transportation or iceberg cost between two regions.

Using the above relationship, we can express the trade gravity equation in the NEG framework as a function of the trade flow $Trade_{rj}$, market supply capacity, market demand potential and transportation cost between region *r* and region *j*[6]:

$$\ln(Trade_{rj} = n_r p_r X_{rj}) = FX_r + \ln \phi_{rj} + FM_j \tag{3}$$

where FX_r and FM_j are binary variables of regions *r* and *j* which are intended to capture supply capacity and demand potential in different regions (these are referred to as the Exporter Dummy and the Importer Dummy variable). $ln\phi_{rj}$ captures barriers to free trade between the two regions – the iceberg costs. To capture trade barriers ϕ_{rj}, we include five alternative components: the log distance between two provinces ($lndist_{rr}$), indicator of an existing domestic trading relationship between provinces ($provincial_{ij}$),

[6] To derive equation (3) from equation (2), note that the trade flow $Trade_{rj}$, market supply capacity, market demand potential and transportation cost between region *r* and region *j* constitute a trade gravity equation in the traditional NEG framework. By taking natural logarithm of equation (2) on both sides and with some rearrangement of terms, we can get equation (3).

indicator of an existing international trading relationship between domestic provinces and foreign trade partners ($export_{ij}$), indicator of an existing trading relationship between foreign trade partners ($foreign_{ij}$) and indicator of an existing domestic trading relationship within foreign trade partners ($intranational_{ij}$). This renders a modified gravity equation of the form:

$$\ln(Trade_{rj}) = \alpha FX_r + \beta FM_j + \delta \ln dist_{rj} + \varphi export_{rj} + \chi foreign_{rj} + \vartheta provincial_{rj} + \xi intranational_{rj} + \psi contig_{rj} + \varepsilon_{rj} \quad (4)$$

where $lndist_{rj}$ measure the geographical distance between region r and region j. The interior distance between two cities can be calculated using the nonlinear distance measure given by the formula $(2/3\sqrt{area_c/\pi})^{\delta}$, where $area_c$ is the area of the downtown epicenter of the central city of a region (the regression coefficient of this variable would give us the marginal effect of trade within a province) and δ is the distance factor in the trade equation. It needs to be pointed out that in equation (4), $export_{ij}$, $foreign_{ij}$, $provincial_{ij}$ *and* $intranational_{ij}$ are all indicator variables of whether a trading relationship exists or not. The indicator takes a value 1 if the trading relationship exists and 0 if not. $Contig_{rj}$ is the dummy variable indicating a common border between two regions[7]. ε_{rj} is random disturbance term.

Data Sources

The trade flow data used in this paper is collected from different sources:

(1) The data of trade flow among domestic provinces in China comes from Chinese extended regional input-output table in 2002, which covers 30 provinces and 42 sectors.

(2) The trade flow data between domestic provinces and trade partners abroad comes from Chinese Customs. According to the Customs Statistics, provinces, on average, have 221 trade partner countries (or regions). We manually identify the exact trade partners for each province and incorporate it in our regressions.

(3) The data of international trade comes from the inter-state trade flow provided by DOTS (Direction of Trade Statistics) database from IMF.

[7] The role of a common border in enhancing trade in now an empirically established factor since first introduced in the works of Frankel and Romer (1999).

(4) There is no available data of trade flow within trade partners abroad, so following Hering and Poncet (2010), we replace it by the difference between gross domestic output and gross export, which comes from the WDI (World Development Indicators) database from World Bank.

To maintain uniformity in our analysis, the trade flow data within each province is expressed as the difference of provincial gross output and gross export and dispatch among provinces, where the data is collected from the China Statistical Yearbook of that year. We convert our RMB to US dollar (to get a standardized unit of measurement) using the average annual exchange rate.

Our next intensive data is the data on domestic distances between each city-pair in our database. Data of domestic distance is calculated based on *1:40,000,000* terrain database mapping provided by the National Fundamental Geographic Information System, which in turn is calculated in Euclidean straight line by ArcGIS. As to the domestic distances between domestic provinces and trade partners abroad, they are calculated by ArcGIS based on the geographical coordinates of the capital cities of each region.

Table 1 summarizes our results for equation (4). The estimated regression coefficient of distance δ is -0.96, suggesting a statistically significant (at 1% level or better) negative relationship between distance between two locales and the amount they trade consistent with the predictions of a standard gravity model and previous findings of Frankel and Romer (1999), Redding and Venables (2004) and Hering and Poncet (2010).

2.2.2 Calculation of Market Access

Next, we measure the relative market access into the city as compared to the province as a whole in which it is located based on the share of a city's GDP in the provincial GDP. Our construction of the market access measure, one of the major focus of our study is based on the methodology of Hering and Poncet (2010). We first construct the market access index at a provincial level and then split it by cities in a particular province.

First, following Redding and Venables (2004) and Hering and Poncet (2010), the MA of region *r* can be expressed as:

$$MA_r = \sum_j \phi r_j G_j^{\delta-1} E_j = \sum \phi_{rj} m_j \ ,$$

where m_j represents market capacity. Next, the individual demand potential of city c can be summarized as:

$$m_c = G_c^{\delta-1}E_c = (y_c / y_j)m_j = (y_c / y_j)G_j^{\delta-1}E_j = y_c / y_j \exp(FM_j) \quad (5)$$

where y_c and y_j are respectively the GDP of city c and the GDP of the province and m_j represents market capacity. Next, combining with equation (2), the city level measure of market access is further subdivided into four parts to better capture different dimensions of what "access" means –the internal city access – this would be related to local transportation and other amenities within the city (MA_{cc}), ease of access to other cities in the same province (MA_{cp}), ease of access to other provinces in the same country (MA_{cn}) and ease of access to trade partners abroad(MA_{cf}), which renders our final market access estimator of the city, MA_c:

$$MA_c = \phi_c G_c^{\delta-1}E_c + \sum_{k\in province} \phi_{ck} \frac{y_k}{\sum y_k} G_J^{\delta-1}E_j + \sum_{j\in China} \phi_{cj} G_j^{\delta-1}E_j + \sum_{j\in ROW} \phi_{cj} G_j^{\delta-1}E_j$$

where expanding each term, we get:

$$MA_c = dist_{cc}^{\delta}(y_c / y_j)\exp(FM_j) + \sum_{n\in province} dist_{cn}^{\delta} \frac{y_n}{\sum y_n} \exp(FM_j) + \sum_{j\in china} dist_{cj}^{\delta} \exp(\vartheta)\exp(FM_j) + \sum_{j\in foreign} dist_{cj}^{\delta} \exp(\varphi + \psi contig_{rj})\exp(FM_j) \quad (6)$$

Equation (6) is the fundamental equation we use to calculate the overall market access measure of a city. Using the 2002 Chinese regional extended input-output table, we have data on 30 provinces including 331 cities. Our calculation and empirical study covers all 331 cities.

Table 2 and Map 1 provides the results of our market access calculation based on equation (6) above. In Table 2, we list the top 10 and bottom 10 Chinese cities in terms of market access, and the map provides a visual representation of market access of all 331 cities across China[8]. We find that the mean value of market access of 331 cities is 8.8617, with the top five cities in China in terms of market access being the usual suspects, mostly the big metropolis - Shanghai (31.6038), Foshan (23.0575), Shenzhen (20.8949), Suzhou (18.2443) and Tianjin (17.8722). In contrast, the places that rank among the last five in

[8] Interested readers can get the full table of market access from the authors, which is not included in the main manuscript for brevity.

terms of ease of market access are Yili (3.0893), Hetian (3.0676), Akesu (3.0649), Kezilesukeerkezi autonomous prefecture (Kezhou for short, 2.8774) and Kashi (2.8702), which are relatively remote.

In general, we detect a large degree of heterogeneity in market access amongst different regions as expected. Ease of market access in the eastern and southern coastal regions is higher, while it is much lower in the northwest provinces especially Xinjiang, which perhaps is attributable to the local geography, the former being planes with easy access to seas and rivers and the latter being more mountainous and uneven terrain. From the view of market access decomposition, the domestic and internal market access plays a dominant role in market access composition of large cities such as Shanghai and Shenzhen. This is partly explained by the fact that within one country (or region), a close supporting network exists inside the manufacturing sector that often relies on its own core group for production – the vertical integration channel. Additionally, large cities being endowed with large population, often have a brisk market demand which often creates production networks as well as a diverse pool of product sets that often solely or in a large proportion exist to cater to the neighborhood demand, and the dependence on overseas market is weakened.

2.3 Variable Descriptions & Final Sample

In this section, we describe some of the additional control variables we use in our regression analysis, and our final regression sample. The major explanatory and additional control variables can be subdivided into two groups: (a) Variables controlling for city characteristics that we call City Characteristic Variables and (b) individual enterprise and labor characteristics that we call Individual Characteristic Variables.

City Characteristic Variables: The traditional NEG model considers the increasing returns to scale, the availability of non-human factor endowments and the externality of human capital as three channels through which the wage skill premium is affected by enterprise agglomeration (Hanson, 2003). Generally, the regions more open to trade and those with a stronger economy have the potential to attract more enterprises and qualified personnel. On the one hand, economies of scale and knowledge spillover can contribute

to the increase of labor productivity and wages. Conversely, the arrival of an abundance of high-skilled workers may intensify the competition in local labor market which in turn might push wages down[9]. We use per capita GDP and the level of human capital in a city to reflect the degree of agglomeration of productive factors in each city, where city level human capital, following Kamal et al (2012), is expressed as the mean level of the individual years of education of the population in the city sampling data. The data on GDP comes from statistical yearbooks of each city. The number of individual educational years comes directly from the population sampling surveys.

Individual Characteristic variables: Taking the individual characteristic from the wage equation (Mincer, 1974) into consideration, we choose marital status, age[10], gender, educational years, ownership type of working organization, occupation, and location to study their effects on individual wage. Here education level plays an important part. We classify education into seven groups: no education, primary school, junior high school, senior high school, junior college, undergraduate degree, and graduate degree and above. Taking junior college as the minimum threshold for skill, we regard individuals with a junior college level and above degree as high-skilled workers. In additional robustness tests, we use alternative education threshold like undergraduate degree and high school senior certificates and find generally consistent results. In addition, we control for individual gender, marital status, ownership type of the working organization (whether state owned enterprise, SOE, or private owned), occupation type, and location.

According to the code of province and city provided by population sampling surveys, we can match city characteristic variables to the individual in order to build up the final sample that is used in empirical analysis. Chinese individual data comes from sampling survey of 1% nationwide population in 2005. It uses the methodology of multi-stage stratified cluster probability proportion, classifying the entire country unit as the general, each province as the sub-general, and each survey community as the final sample unit, to

[9] It is important to note that since the difference of non-human factor endowments appear mainly at a more aggregated province and sector level (Hering & Poncet, 2010), our research by city requires no deliberate control.

[10] We also include square of age to account for any nonlinear effects of age on wage-gap. Previous studies show that wage shows a trend of increasing with age to a threshold and then declining –almost like the inverted U.

obtain a random sample of 13 million people (1.31% of national population) in 31 provinces. Consistent with previous research, we use 20% stochastic sampling, where filtered individuals are associated with non-agricultural "Hukou" and authorized job qualification.

Since we were not able to calculate the market access data for Tibet where data access is rather limited, we drop the population data of Tibet from our final sample. This gives us a final sample of 304,537 observations with 97, 966 skilled workers comprising 32.17% of the final sample and 206,571 unskilled workers that make up the remaining 67.83% of the final sample.

Table 3A reports the summary statistics of the primary explanatory variables of our model, both for the overall sample as well as skill levels. **Table 3B** displays the results of a *t-test* to detect any statistical differences in the characteristics of the two groups.

We detect a number of interesting characteristics that significantly differ between different skill cohorts. We find that high-skilled workers are younger (mean age 35.35 as compared to 38.09 for the low skilled workers), have a higher level of education as expected, and are also less likely to be married than a low skill worker (87.6% of the low skilled workers in our sample are married, as compared to 81.2% of high skilled workers). These numbers suggest a younger and a different demographics for high skilled workers (perhaps career oriented younger population) as compared to the low skilled workers in our sample. What is also interesting to note is that the proportion of high-skilled workers employed in state-owned enterprises (74.7%) is statistically significantly higher than low skilled workers (39.9%). Coming to the dependent variable of our model, the logarithm of the hourly wage, we find that it is significantly higher for the high-skilled workers ($2.12) as compared to the low-skilled workers ($1.44), as expected.

More informative for our purposes is **Figure 2** which is visual representation of the density distribution of the logarithm of wage level of high skilled and low skilled workers. The distribution of logarithm of wages of the high-skilled workers is positioned to the right of the low-skilled workers wage when superimposed on the same graph. This reflects that high-skilled workers are at middle and high level of the income distribution, while low-skilled workers are mostly at low and medium levels, suggesting potential existence

of wage skill premium in our sample.

The main focus of our analysis is a link between market access and wage skill premium. To inform us on the potential of any relationship between the two, in **Figure 3**, we provide a visual representation of the relation between log of market access and the log of real hourly wage rate at the 95% confidence band. The figure suggests a strong positive association between the two at 95% confidence interval. Finally, to detect any potential correlation between our explanatory variables, we present the results of Spearman correlation coefficients in **Table 3C**. We detect some collinearity between our explanatory variables, but in most cases it ranges from 30%-40%, which is lower than the rule of thumb of 70%-80% for considerable correlation resulting in multicollinearity issues (Angrist and Pischke, 2009).

3. Empirical Analysis

3.1: Potential Pitfalls and Regression Design

At this point, before we begin discussing the details of our empirical setup, it is useful to discuss in further detail some of the endogeneity as well as sample selection issues that could potentially affect the regression analysis. In any study of labor wages or wage differentials, one has to be aware of potential sample selection problems, *a la* Heckman. As pointed out by previous research, as well as noted by Hering and Poncet (2010), it is possible that the sample respondents who report their hourly wage are a self-selected group. In order to avoid biased ordinary least squares estimation, one therefore needs to also analyze the baseline using Heckman correction to test the robustness of initial conclusions.

In addition to the well-known problem of sample selection that plagues studies of labor wages, our baseline model also has potential endogeneity issues. Head and Mayer (2006) suggest that market access and the wage skill premium interactions can go both ways. Greater market access brings in highly skilled workforce who command a high wage, but at the same time, greater wage opportunities can attract labor, expanding output and in turn, improving the market access potential of a city or region. The existence of

such endogeneity can potentially lead to biased and inconsistent estimates if not controlled for. To solve this problem, Head and Mayer (2006), propose using the *centrality index* of each city as an instrumental variable in lieu of market access.

Centrality index $C_i = \ln \sum_{j \neq i} d_{ij}^{-1}$ is measured as the logarithm of the sum of the reciprocal of the geographic distance between every pair of cities in a sample.

Calculating the centrality index above, we find that centrality index and market access are significantly correlated at the 1% level or better. As argued by Rodrick (2004) and Head and Mayer (2006), the exclusion restriction is also met when we use centrality index, as geographical location that is the fundamental determinant of centrality index, is an exogenous variable uninfluenced by other economic factors.

We start our empirical analysis by reporting the baseline results conducted using a standard OLS. However, acknowledging that our sample might suffer from potential sample selection bias due to non-reporting of wage information by individuals in the survey, we also apply Heckman selection (Heckman, 1979) to test for robustness of our analysis. In additional robustness tests, we correct for endogeneity problems as well. Finally, in order to get more rigorous empirical results correcting for the twin problems of sample selection and endogeneity, following Wooldridge (2002), we modify the Heckman methodology to incorporate an instrumental variables approach and check the robustness of our baseline results.

3.2 The Benchmark Regression

We first report the results of OLS and Heckman selection applied to Equation (1), our benchmark specification. Our benchmark findings highlight two important trends. **Table 4 Columns 1 - 5** show that market access, in general, has a significantly positive effect on wages, with a 1% increase in market access potential leading to a 2.46% - 16.53% increase in wage level, depending on the regression column. What is more striking is that the regression coefficient of the interaction term, "ln*MA*skill*" is *significantly positive*, suggesting that skilled workers, on an average, gain *an additional* 1.62% - 16.66% increase in wages for every 1% increase in market access, suggesting a widening of the wage-gap. Given this evidence, when we introduce the interaction of the squared of

market access and dummy variable of skilled workers in the second column (lnMA^2*$skill$) to capture any potential non-linearity in the relationship, we find the coefficient to be *significantly negative* indicating declining returns to market access suggesting an inverted U-shaped relationship. The third, the fourth, and the fifth columns include additional controls like log of per capita GDP level of cities, log of human capital in a city, as well as indicators of capital city, port city, individual sector and individual occupation. We find that even with introduction of these additional controls, our baseline findings are robust.

Next, we use Heckman sample selection model (Heckman, 1979) to correct the sample selection bias, where the key is to introduce labor into the model using the form:

$$S_i^* = Z_i\gamma + \nu_i$$

$$S^* i \geq 0,$$

Observed wage (S=1), Non observed wage (S=0) (7)

where Z is a vector of observable individual characteristics affecting individual wage, and γ and ν are respectively the coefficient and the random disturbance term. We use the above relationship to amend the baseline wage equation to:

$$\begin{aligned}\ln w_{ic} = {} & \alpha_0 + \alpha_1 \ln MA_c + \alpha_2 \ln MA_c \times skill + \alpha_3 \ln MA_c^2 \times skill + \alpha_2 \ln gdp + \alpha_5 \ln capital \\ & + \alpha_6 marriage + \alpha_7 age + \alpha_8 age^2 + \alpha_9 gender + \alpha_{10} edu + \alpha_{11} SOE + \mu_{ic} + \lambda_{ic} + \varepsilon_{ic}\end{aligned} \quad (8)$$

In the equation (8), λ is inverse mills ratio of the individual *i*, and υ is random disturbance term. The **last column in Table 4** reports the Heckman two-stage estimation results. The coefficient of inverse Mills ratio is significant at 1% level, indicating the existence of sample selection bias, suggesting suitability of application of the Heckman methodology **(**Wooldridge, 2002，2006) in our context.

After controlling the variables above and considering sample selection bias, the primary results under OLS estimation still go through when we adopt the Heckman methodology. A 1% increase in market access now leads to a 9.21% increase in wage levels, and *an additional* 11.94% increase in wages for the high skilled group (coefficient of the interaction term, "lnMA*$skill$" is *significantly positive*). The inverted U relationship between market access and wages is still robust (the coefficient on lnMA^2*$skill$ is *significantly negative*).

Based on the mechanisms of the theory discussed in the previous sections, what our

findings so far suggest is that with increased market access, enterprise agglomeration can intensify the competition in the product market, and lead to an increase in demand for high-skilled workers, which in turn raises the returns to education and widens the wage gap. This process generates a migration of high skill workers to high market access areas. High-skilled workers self-select to move to high market access areas attracted by the high wages, crowding out the low skilled labor. This process eventually leads to an excess supply of high skilled workers in high market access areas. Therefore, after a critical migration threshold, we see the balance tilting with an overabundance of high skill workers and a relative scarcity of low skill workers, which helps to increase the low-skilled workers' relative wage, after the critical threshold and the wage skill premium starts to narrow. This generates the inverted U-shaped pattern of returns to market access that we observe in our data.

In terms of the other control variables, our findings are consistent with existing literature (Hering and Poncet). Per capita GDP and human capital of a city has positive effects on individual wages, which suggests a positive association between returns to workers and level of development of the region they operate in. As for individual characteristics, the more educational years, the higher wage. Married workers have higher wages than unmarried workers[11]. Male workers' wages are generally higher than the female's indicating potential of a gender bias. At the same time, age and age^2 is respectively associated with the wage positively and negatively, which illustrate an inverted U relationship between age and wage. Finally, our findings suggest that in the state-owned enterprises in China, workers on average, earn about 14% - 21% more wages than the private sector.

3.3 Subsample regressions controlling for Endogeneity and Sample Selection

Armed with our benchmark findings, we now investigate if there is any heterogeneity in the relationship between market access and wages based on regional disparities in the

[11] Note that one might argue that married workers might belong to a higher age cohort on average than unmarried workers. However, our results on the marital coefficient holds even after we control for age and age squared, which would capture the impact of age differential on wages.

level of market access. The argument here is that in regions with a comparatively higher level of initial market access, subsequent further opening up of markets might not result in the same wage outcome as in more closed regions that are now experiencing an increased market access. To investigate this, we classify the regions in our sample into two groups – initial high market access areas and the initial low market access areas based on the initial mean market access level of the cities in 1995. Our aim is to test if the inverted U-shaped relationship between increased market access and wages that we observe in our overall sample comes from initially higher market access regions when they further improve their market access, or comparatively lower initial market access regions when they improve their market access – this gives us further insight into the policy implications of improved market access based on initial regional disparity.

In addition, we present the results correcting for endogeneity (two stage least squares) and well as endogeneity and sample selection jointly (Heckman methodology with instrumental variables). **Table 5** reports the results of two stage least squares correcting for endogeneity (Columns 1 – 3) and the results of Instrumental Variables estimation[12] using the Heckman approach (Columns 4 – 6) correcting for endogeneity as well as sample selection. Note that to keep the regression consistent across the overall and grouped samples, we do not include market access variable by itself in the regression on the overall sample in Table 5, instead just focusing on the interaction of market access and skill dummy, but our results are robust when we include it by itself as well in additional checks on the overall sample.

Before discussing our findings, it is important to note the results of our Kleibergen-Paaprk LM and the Kleibergen-Paaprk Wald F-test. The null hypothesis of Kleibergen-Paaprk LM test is the under-identification of the instrumental variable (IV). If one rejects the null hypothesis, then IV is reasonable. The null hypothesis of Kleibergen-Paaprk Wald F test is that IV has a weak identification. Again, if we reject the null hypothesis, then we can conclude that IV is reasonable. Our *p*-values as well as critical values of Stock-Yogo (2005) test indicate that we can significantly reject the null hypothesis in both cases,

[12] The instrumental variables in Table 6 have passed the under-identification test and weak identification tests as discussed above, the complete results of which can be obtained from the authors.

making a case for use of our IV.

Coming to the findings, our results show that even after controlling for endogeneity and sample selection, the inverted U-shaped relationship between market access and wage still holds for the overall sample with the coefficient on the interaction term "ln*MA*skill*" being significantly positive as before and the coefficient on the interaction term, "lnMA^2**skill*" being significantly negative.

However, we find a divergent trend between regions with high versus low initial market access. In regions with low levels of initial market access, the relationship between market access and the wage gap still demonstrates the inverted U-shape as found in the overall sample. However, in regions with comparatively high levels of initial market access, the relationship demonstrates exactly the opposite trend - a regular U-shape! (coefficient of the interaction term, "ln*MA*skill*" is *significantly negative* and the coefficient on "*ln*MA^2**skill*" is *significantly positive*).

One potential explanation for the above observation lies in the existence of spatial externality and the unbalanced regional distribution of returns to education, which might cause a migration of factors of production from regions with low market access to regions with high levels of market access. In such a scenario, one can imagine existing firms in regions with low market access provide competitive benefits, including higher wages to the skilled labor in order to attract or retain them when market access improves to meet the increased demand, which will lead to a widening of the wage gap, at least in the short run. In contrast, in regions with initial high levels of market access, the relative abundance of the existing skilled labor and the resulting competition due to new migrants from the low market access areas in the short run with improved mobility, will narrow the skilled wage gap, which matches our observed pattern. However, this decline would also reduce the attractiveness of these high market access areas for skilled labor migration as initial low market access areas become more competitive with increased market access, and therefore, after a certain critical threshold, there would be a reversal of pattern and skilled worker wages would rise to maintain the competitive edge of these regions and we would see the U-shaped pattern of the wage-skill premium.

While our study captures the short run effects, it would be worthwhile to put it in

context of some studies that shed light on the long run impact of migration due to market access. Marshall (1890) propounded the theory that enterprises adjacent to each other lead to technology spillover. If so, then one can argue that in the long run, regional agglomeration of labor and technology can increase regional labor productivity through technology spillover (Rahman & Fujita, 1990; Combes, 2000). One potential impact of such changes in regions with high market access would be skilled-biased technological change that would be observed due to the agglomeration caused by market access which would in turn encourage such technology spillovers. This in turn, would lead to an increase in the relative demand for high-skilled workers, raise returns to education, and widen the skilled wage gap. These long term differences in returns to education would encourage further migration of high-skilled workers to regions with high levels of market access (Xing et al., 2013) causing a shortfall of such labor in low market access regions. The potential cost of attracting and retaining the high skilled workers in regions with low levels of market access can be prohibitive and if this is a deterrent enough, many low skilled workers in regions with low market access would move to jobs requiring higher skills, narrowing the wage gap and explaining the urban migration flows recently witnessed in China.

4. Robustness Analysis

In the previous sections, while we have conducted our analysis using several methodological alternatives – Ordinary Least Squares, standard Heckman selection procedure to control for potential sample selection, two stage least squares to control for potential endogeneity as well as Heckman procedure taking into account endogeneity of variables, we kept our main variables – the measure of market access and the educational cut-off to determine skill levels, constant under separate methodological alternatives. In this section, we repeat our analysis by re- characterizing the indicators of the market access and the employee skills in order to examine whether alternative variable measurements change our baseline conclusions.

We begin by acknowledging that there is a long standing argument amongst experts about the empirical benchmark that allows division of high and low skills, especially the

use of post-secondary education as the cutoff of skill levels. In our benchmark regressions, we had used college education (junior college) as a threshold for skill. In this segment, we relax the assumption of a college degree and instead consider secondary education (graduation from high school) as a benchmark for skill levels.

Next, we define an alternative measure of market access. The Chinese input-output data table only exists in 2002 and 2007 (every five years) due to the special nature of its design. In our baseline regressions, we use the interprovincial input-output data table in 2002 when estimating distance coefficient values (δ). In this segment, we calculate the distance coefficient based on the latter period data and re-estimate our baseline regressions. In each case, we present the results based on regressions run on the overall sample as well as the high and low initial market access subgroups.

Table 6 columns 1 – 3 present the results with alternative skill-level cutoff while retaining the original measure of market access. In columns 4 – 6, we retain the original skill-level cutoff but use the alternative market access measure based on 2007 Chinese input-output matrix, and finally, in columns 7 – 9, we use the alternative skill-level cutoff as well as alternative measure of market access jointly to check robustness of our benchmark findings. We find that across all regressions conducted on the full sample, increased market access has a significantly positive association with wages, and the inverted U-shaped relationship between market access and skill level is robust.

One still finds a divergence in experience of the high market access and the low market access areas. While the low market access areas exhibit the inverted U-shaped pattern suggesting an initial expansion followed by a contraction of the wage skill premium, the high market access areas exhibit an exactly opposite pattern, even with an alternative measurement of skill as well as market access.

5. Conclusions and Policy Implications

Reconstructing the market access index on 331 Chinese cities, and using the theoretical foundations of the new economic geography model, this paper takes a closer look at the relation between wage-skill premium and market access.

Modifying the standard regression techniques to control for potential sample selection issues associated with observed wage and a potential endogeneity between wages and the market access variables, we find a robust inverted U-shaped relationship

market access and wage gap, suggesting a potential widening of the wage gap as a region or a city improves its market access potential, which declines after a critical threshold is reached. However, this relationship is not universal, and is primarily observed for low market access areas as they open up their markets. For areas with already a high level of market access, the relationship is exactly the opposite. Our findings suggest a role of urban migration patterns in explaining wage skill premium observed in China. High market access areas have the potential to attract skilled labor, and low market access areas, as they open up, enter into a bid to attract skill with promises of high wage in order to retain or attract skilled labor, which initially leads to an increase in the wage gap. In contrast, high market access areas already have an existing pool of skilled labor, so further migration works to drive down the skill premium and narrow the wage gap. Thus, in the low market areas we notice what we term a "demand pull" wage gap created by firm competition for skilled labor, while in areas of high market access, further improvement in market access only serve to reduce the wage gap in a "supply push" environment driven by excess supply of skilled labor. Each phenomenon continues till a critical migration threshold is reached beyond which there is a reversal of the trend. Thus, our findings suggest a potential explanation of the urban migration patterns and existing wage-gap witnessed in China.

References

[1] Acemoglu, D. (1998) Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality. Quarterly Journal of Economics, 113(4): 1055-1089.

[2] Ahearne A G, Fernald J G, Loungani P, et al. Flying Geese or Sitting Ducks: China's Impact on the Trading Fortunes of Other Asian Economies，SSRN Electronic Journal, 2006.

[3] Angrist, J. D., Pischke, J. S., & Pischke, J. S. (2009). Mostly harmless econometrics: an empiricist's companion (Vol. 1). Princeton: Princeton university press.

[4] Campos, N.F. and D. Jolliffe (2005) Does market liberalisation reduce gender discrimination? Econometric evidence from Hungary, 1986-1998, Labour

Economics,12(1): 1-22.

[5] Campos, N.F. and A. Dabusinskas (2009) So Many Rocket Scientists, So Few Marketing Clerks: Estimating the Effects of Economic Reform on Occupational Mobility in Estonia, IZA Working Paper 3886.

[6] Combes, P. (2000) Economic Structure and Local Growth: France, 1984–1993, Journal of Urban Economics, 47(3): 329-355.

[7] Campos, N. F., & Dabušinskas, A. (2009). So many rocket scientists, so few marketing clerks: Estimating the effects of economic reform on occupational mobility in Estonia. European Journal of Political Economy, 25(2), 261-275.

[8] Fan, J. and Zhang, Y. (2009) Economic Geography and Regional Wage Gap, Economic Research, 2009, 8: 73-84.

[9] Fernald J, Edison H, Loungani P. Was China the first domino? Assessing links between China and other Asian economies，Journal of International Money & Finance, 1999, 18(4):515-535.

[10] Feenstra, R. C., & Hanson, G. H. (1996). Globalization, outsourcing, and wage inequality (No. w5424). National Bureau of Economic Research.

[11] Frankel, J. and D. Romer (1999) Does Trade Cause Growth? American Economic Review, 89(3): 379-399

[12] Foster-McGregor N., J. Poeschl and R. Stehrer. (2016) Offshoring and the Elasticity of Labour Demand, Open Economies Review, 27(3): 515-540.

[13] Fujita, M., Krugman, P., and Venables, J. (1999), The Spatial Economy: Cities, Regions and International Trade, Cambridge, Mass: MIT Press

[14] Fingleton, B., & López‐Bazo, E. (2006). Empirical growth models with spatial effects. Papers in regional science, 85(2), 177-198.

[15] Frankel, J. A., & Romer, D. (1999). Does trade cause growth? American economic review, 379-399.

[16] Hanson, G. H. (2003), Firms, Workers and the Geographic Concentration of Economic Activity, in Clark, G. M., Feldman, M. G. (Eds.) The Oxford Handbook of Economic Geography, New York: Oxford University Press: 477-494.

[17] Hanson, G.H. (2005), Market Potential, Increasing Returns, and Geographic

Concentration, Journal of International Economics, 67: 1-24.

[18] Head, K. and Mayer, T. (2006), Regional Wage and Employment Responses to Market Potential in the EU, Regional Science and Urban Economics, 36: 573-594.

[19] Heckman, J. (2005), Sample Selection Bias as a Specification Error, Econometrica, 47: 153-162.

[20] Heckman, J. J. (1979). Statistical models for discrete panel data. Department of Economics and Graduate School of Business, University of Chicago.

[21] Hering, L. and Poncet, S. (2010), Market Access Impact on Individual Wages: Evidence from China, The Review of Economics and Statistics, 92(1): 145-159.

[22] Kamal, F., Lovely, E., and Ouyang, P. (2012), Does deeper integration enhance spatial advantages? Market access and wage growth in China, International Review of Economics and Finance, 2012, Vol.23, pp.59-74.

[23] Krugman P. (1991), Increasing Returns and Economic Geography, Journal of Political Economy, Vol.99, pp.483-499.

[24] Liu, X., He, X., and Yin, X., Market Access and Regional Wage Gap: Empirical Research Based on the Panel Data of Chinese Prefecture-level City, Management World, 2007, Vol.9, pp.48-55.

[25] Marshall, A., Principles of Economics. London, Macmillan, 1890.

[26] Mincer, J., Schooling, Experience and Earnings, New York: Columbia University Press, 1974.

[27] Ottaviano, G., and Pinelli, D., Market Potential and Productivity: Evidence from Finnish Regions, Regional Science and Urban Economics, 2006, Vol.36(5), pp.636-657.

[28] Rahman, H. and Fujita, M. (1990), Product Variety, Marshallian Externalities, and City Sizes. Journal of Regional Science, 30(2):165–183.

[29] Redding, S., Venables, A., Economic Geography and International Inequality, Journal of International Economics, 2004, Vol. 62(1), pp.53-82.

[30] Rodrik, D., Subramanian, A. and Trebbi, F. (2004) 'Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development' in: Journal of Economic Growth, 9: 131-165.

[31] Robert C., Gordon, F., and Hanson, H., Globalization, Outsourcing, and Wage Inequality, NBER Working Paper, 1996, No. 5424.

[32] Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. Identification and inference for econometric models: Essays in honor of Thomas Rothenberg.

[33] Troske, K. R. (1999). Evidence on the employer size-wage premium from worker-establishment matched data. Review of Economics and Statistics, 81(1), 15-26.

[34] Wooldridge, J., Econometric Analysis of Cross Section and Panel Data, Cambridge, MA: MIT Press, 2002.

[35] Wooldridge, P. D. (2006). The changing composition of official reserves. BIS Quarterly Review, September.

[36] Xu, B., & Li, W. (2008). Trade, technology, and China's rising skill demand1. Economics of Transition, Vol.16(1), 59-84.

[37] Xu, D., Market Access and the Disparity of Regional Development: Empirical from China City Economy, World Economic Forum, 2012, Vol.1, pp.33-52.

[38] Zhao, Z., Shi, M., and Yang, J., Market Proximity, Supply Adjacent and Apace China Manufacturing Distribution –Analysis Based on Input Output Model among Chinese Provinces, Economics quarterly, 2012, Vol.3, pp.1059-1078

[39] Zhou, S., Yang, H., and Li, K., Trade, Technology, Institution and Chinese Industrial Sector Wage Premium, Chinese Economic issues, 2012, Vol.01, pp.22-31.

Table 1: Trade and distance: testing the Gravity Model

In this table, we estimate the trade equation and capture the relationship of trade with the distance variable $dist_{rj}$ controlling for trading relationships:

$$\ln(Trade_{rj}) = \alpha FX_r + \beta FM_j + \delta \ln dist_{rj} + \varphi export_{rj} + \chi foreign_{rj} + \vartheta provincial_{rj} + \xi intranational_{rj} + \psi contig_{rj} + \varepsilon_{rj}$$

where $export_{ij}$, $foreign_{ij}$, $provincial_{ij}$ *and* $intranational_{ij}$ are all indicator variables of whether a trading relationship exists or not. The indicator takes a value 1 if the trading relationship exists and 0 if not. $Contig_{rj}$ is the dummy variable indicating a common border between two regions. $Indist_{rj}$ measure the geographical distance between region *r* and region *j*. Distance is captured as the geographical distance between two cities at the center of the region and calculated using the nonlinear distance measure given by the formula $(2/3\sqrt{area_c/\pi})^{\delta}$

Dependent Variable：ln(Trade)	
Explanatory Variables	
lndist	-0.9561*** (-44.01)
countig	0.7950*** (6.78)
export	4.2551*** (60.75)
foreign	7.4044*** (40.5)
provincial	5.7980*** (26.63)
Intra-national	12.0735*** (43.7)
Constant	16.0269*** (46.88)
Fixed Effects:	Yes
Exporter fixed effects	Yes
Importer fixed effects	Yes
Observations	26, 414
Within R2	0.70

Note: ***: significance at the 1% level; **: significance at 5% level; *: significance at 10% level. The t-statistic are in the parenthesis.

Table 2: Market Access (MA) differences across regions（year：2002）

We calculate market access of each of the 331 Chinese cities in our sample using the measure:

$$MA_c = dist_{cc}^{\delta}(y_c / y_j)\exp(FM_j) + \sum_{n \in province} dist_{cn}^{\delta} \frac{y_n}{\sum y_n} \exp(FM_j)$$
$$+ \sum_{j \in china} dist_{cj}^{\delta} \exp(\vartheta)\exp(FM_j) + \sum_{j \in foreign} dist_{cj}^{\delta} \exp(\varphi + \psi contig_{rj})\exp(FM_j)$$

The table below provides the list of top 10 and bottom 10 Chinese cities in our sample in terms of market access.

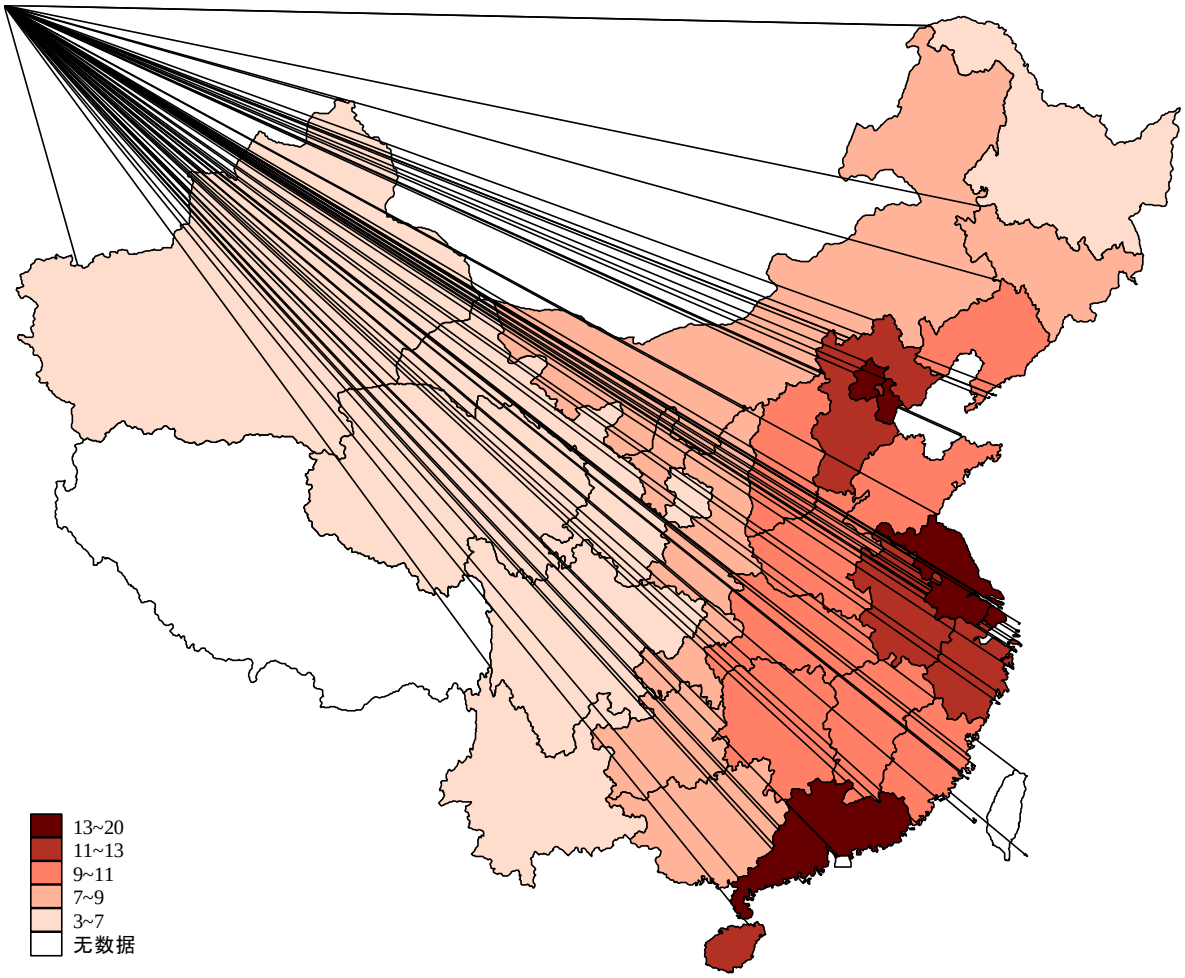
Top 10 cities in terms of market access				Bottom 10 cities in terms of market access			
City	**MA_{cc}+MA_{cp}+MA_{cn}**	**MA_{cf}**	**Total MA**	**City**	**MA_{cc}+MA_{cp}+MA_{cn}**	**MA_{cf}**	**Total MA**
Shanghai	30.4934	1.1105	31.6038	Aletai	2.6922	0.6755	3.3678
Foshan	22.3226	0.7349	23.0575	Shihezi	2.6808	0.6760	3.3568
Shenzheng	20.1622	0.7327	20.8949	Kelamayi	2.5696	0.6799	3.2495
Suzhou	17.4936	0.7507	18.2443	Tacheng	2.4347	0.6858	3.1206
Tianjin	16.9831	0.8891	17.8722	Boertala	2.4248	0.6868	3.1116
Wuhu	16.9058	0.7323	17.6381	Yili	2.4007	0.6886	3.0893
Guangzhou	16.2659	0.8299	17.0958	Hetian	2.3609	0.7067	3.0676
Beijing	16.2664	0.8096	17.0760	Akesu	2.3718	0.6931	3.0649
Dongguan	16.1736	0.7357	16.9093	Kezhou	2.1651	0.7123	2.8774
Tangshan	15.8562	0.8412	16.6975	Kashi	2.1563	0.7140	2.8702

Notes：Market access is measured under four different criteria: local transportation and other amenities within the city (MA_{cc}）, ease of access to other cities in the same province (MA_{cp}), ease of access to other provinces in the same country (MA_{cn}) and ease of access to trade partners abroad(MA_{cf}). The final column combines all four criteria to give the total market access potential of each city in our sample of 331 cities.

Figure 1: Market Access Map:
Visual representation of market access across China

We calculate market access of each of the 331 Chinese cities in our sample using the measure:

$$MA_c = dist_{cc}^{\delta}(y_c / y_j)\exp(FM_j) + \sum_{n \in province} dist_{cn}^{\delta} \frac{y_n}{\sum y_n} \exp(FM_j)$$
$$+ \sum_{j \in china} dist_{cj}^{\delta} \exp(\vartheta)\exp(FM_j) + \sum_{j \in foreign} dist_{cj}^{\delta} \exp(\varphi + \psi contig_{rj})\exp(FM_j)$$



MAP 1—Provincial Market Access: Source: Authors' calculations

Table 3A: Descriptive statistics of major variables overall and grouped by skill levels

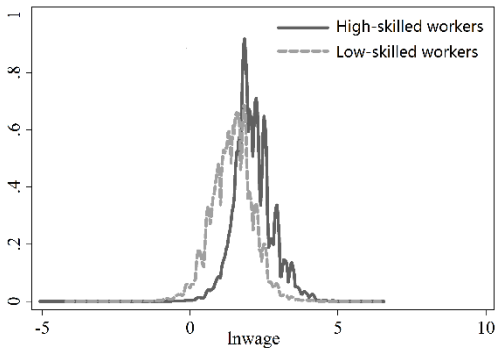
	N	Mean	Standard Deviation	Minimum	Maximum
City Level Characteristics					
lnMA	304, 537	2.1839	0.3877	1.0544	3.4533
ln*gdp*	304, 537	9.8015	0.7674	7.7816	11.3915
ln*capital*	304, 537	1.3564	0.0700	1.1026	1.4994
High-skilled (with Junior College degree or above) Individual Characteristics					
ln*wage*	97,966	2.1207	0.6483	-5.0752	6.5511
marriage	97,966	0.8122	0.3905	0	1
age	97,966	35.3512	8.7968	16	60
age^2	97,966	1327.0890	665.7610	256	3600
gender	97,966	0.5743	0.4944	0	1
edu	97,966	5.3964	0.5487	5	7
SOE	97,966	0.7471	0.4347	0	1
Low-skilled (with Senior high school degree or below) Individual Characteristics					
ln*wage*	206,571	1.4357	0.6938	-4.1997	6.4377
marriage	206,571	0.8760	0.3295	0	1
age	206,571	38.0888	9.5548	16	60
age^2	206,571	1542.0500	731.7304	256	3600
gender	206,571	0.5911	0.4916	0	1
edu	206,571	3.4066	0.6569	1	4
SOE	206,571	0.3986	0.4896	0	1

Note: We summarize the city characteristics over the entire sample comprising of 304,537 observations across 331 Chinese cities. However, the individual level characteristics are summarized over the two sub-groups – (a) high skilled labor force and (b) low-skilled labor force. High skill labor is defined as individuals with a junior college education level or above, and low skilled labor force comprises the remaining who have not completed a junior college education.

Table 3B: Differences in individual characteristics of high and low skilled population – a T-test

Variables	Low-skilled		High-skilled		Difference of Means (High – Low)
	Sample	*Mean*	*Sample*	*Mean*	
lnwage	206571	1.4357	97966	2.1207	-0.685***
marriage	206571	0.8760	97966	0.8122	0.122***
age	206571	38.0888	97966	35.3512	4.420***
age^2	206571	1542.0500	97966	1327.0890	347.336***
gender	206571	0.5911	97966	0.5743	-0.039***
edu	206571	3.4066	97966	5.3964	-2.053***
SOE	206571	0.3986	97966	0.7471	-0.345***

Note: ***: significance at the 1% level; **: significance at 5% level; *: significance at 10% level.



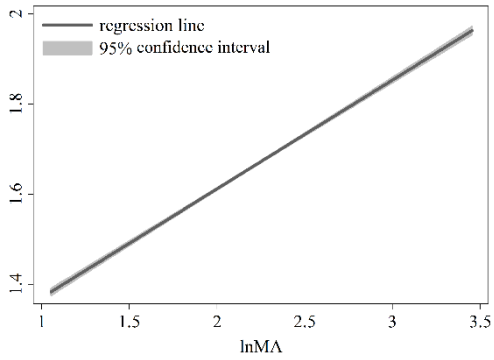


Figure 2: Density Distribution of Skilled *Lnwage* **Figure 3: The Correlation between *lnwage* and *lnMA***

Note: See previous tables and text for classification of skill levels and the calculation of the market access variables

Table 3C: The correlation coefficient matrix of major variables

	lnwage	**lnMA*skill**	**lngdp**	**lncapital**	**marriage**	**age**	**gender**	**edu**	**soe**
lnwage	1								
lnMA*skill	0.467***	1							
lngdp	0.385***	0.292***	1						
lncapital	0.344***	0.390***	0.613***	1					
marriage	-0.015***	-0.180***	-0.091***	-0.081***	1				
age	0.027***	-0.220***	-0.008***	-0.013***	0.663***	1			
gender	0.118***	0.039***	-0.013***	-0.005***	-0.028***	0.100***	1		
edu	0.502***	0.856***	0.151***	0.244***	-0.200***	-0.279***	0.051***	1	
soe	0.288***	0.282***	-0.044***	0.052***	0.178***	0.083***	0.111***	0.344***	1

Note：***, **, and * indicate it respectively passes significance test at the 1%, 5%, and 10% level.

Table 4: Benchmark regression results – skill premium and market access

We run the regression:

$$\ln w_{ic} = \alpha_0 + \alpha_1 \ln MA_c + \alpha_2 \ln MA_c \times skill + \alpha_3 \ln MA_c^2 \times skill + \alpha_4 \ln gdp + \alpha_5 \ln capital + \alpha_6 marriage + \alpha_7 age + \alpha_8 age^2 + \alpha_9 sex + \alpha_{10} edu + \alpha_{11} SOE + \mu_{ic} + \lambda_{ic} + \varepsilon_{ic}$$

where λ is inverse mills ratio of the individual *i* which is used in column (6). Columns 1 – 5 present the results of a standard OLS, while Heckman sample selection is taken into account in column 6.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Ordinary Least Square Estimation					**Heckman Selection**
ln*MA*	0.1653***	0.1880***	0.0246***	0.0460***	0.0873***	0.0921***
	(55.02)	(52.88)	(7.25)	(13.31)	(25.91)	(26.89)
ln*MA*skill*	0.0162***	0.1050***	0.1666***	0.1663***	0.1205***	0.1194***
	(8.48)	(13.68)	(23.56)	(23.56)	(17.62)	(17.49)
lnMA^2**skill*		-0.0362***	-0.0527***	-0.0519***	-0.0335***	-0.0332***
		(-11.94)	(-18.91)	(-18.65)	(-12.46)	(-12.36)
ln*gdp*			0.3482***	0.3173***	0.2559***	0.2548***
			(234.17)	(180.04)	(126.77)	(125.53)
ln*capital*				0.6435***	0.7552***	0.8408***
				(32.56)	(35.59)	(39.44)
marriage	-0.0896***	-0.0907***	0.0111***	0.0152***	0.0253***	0.0373***
	(-20.14)	(-20.38)	(2.69)	(3.70)	(6.37)	(9.30)
age	0.0270***	0.0271***	0.0234***	0.0232***	0.0223***	0.0184***
	(25.31)	(25.40)	(23.77)	(23.60)	(23.54)	(18.90)
age^2	-0.0002***	-0.0002***	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(-18.47)	(-18.57)	(-18.82)	(-18.78)	(-19.70)	(-14.78)
gender	0.1605***	0.1607***	0.1772***	0.1786***	0.1607***	0.1475***
	(67.72)	(67.80)	(81.18)	(81.97)	(73.56)	(64.90)
edu	0.3052***	0.3001***	0.2443***	0.2374***	0.1859***	0.1836***
	(164.51)	(157.80)	(138.25)	(133.60)	(105.27)	(101.81)
SOE	0.1448***	0.1428***	0.2090***	0.2047***	0.1572***	0.1429***
	(57.52)	(56.62)	(89.32)	(87.51)	(62.40)	(56.81)
capital control	NO	NO	NO	NO	YES	YES
port control	NO	NO	NO	NO	YES	YES
sector control	NO	NO	NO	NO	YES	YES
occupation control	NO	NO	NO	NO	YES	YES
inverse Mills ratio						-0.6083***
						(-18.91)
constant	-0.6825***	-0.7158***	-3.4707***	-3.9629***	-3.2255***	-3.2150***
	(-33.34)	(-34.66)	(-155.23)	(-147.01)	(-98.28)	(-97.07)
Observations	304537	304537	304537	304537	304537	295101
$\overline{R^2}$	0.2935	0.2938	0.4016	0.4037	0.4478	0.4245

Note: the t-statistic is presented in the parenthesis; ***, **, and * significance at the 1%, 5%, and 10% levels.

Table 5: Grouped regression by region controlling for endogeneity and sample selection

We re-run the baseline regression with two-stage least squares methodology to control for endogeneity, as well as Heckman correction with instrumental variables to control for the twin issues of sample selection and endogeneity. Columns 1 to 3 present the results of two stage least squares and columns 4 to 6 present the results of a Heckman correction with instrumental variables, where λ is inverse mills ratio of the individual *i* . The results are presented for the overall sample as well as by sub-groups presented by market access – regions with initially high market access and regions with initially low market access.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	overall	**high-MA**	**low-MA**	**overall**	**high-MA**	**low-MA**
ln*MA*skill*	0.1981***	-4.9823***	0.1851***	0.2237***	-4.7095***	0.1966***
	(25.50)	(-12.51)	(8.10)	(28.66)	(-12.14)	(8.55)
lnMA^2**skill*	-0.0494***	1.0044***	-0.0461***	-0.0554***	0.9493***	-0.0477***
	(-26.63)	(12.60)	(-6.65)	(-29.72)	(12.23)	(-6.85)
ln*gdp*	0.2568***	0.3689***	0.2069***	0.2558***	0.3711***	0.2038***
	(123.99)	(60.81)	(78.34)	(123.34)	(61.31)	(76.91)
ln*capital*	0.6778***	-0.3648***	0.7649***	0.7514***	-0.2935***	0.8347***
	(30.22)	(-7.31)	(26.35)	(33.36)	(-5.90)	(28.61)
marriage	0.0249***	0.0120*	0.0370***	0.0359***	0.0204***	0.0486***
	(6.22)	(1.69)	(6.69)	(8.83)	(2.87)	(8.68)
age	0.0223***	0.0189***	0.0248***	0.0187***	0.0164***	0.0210***
	(22.83)	(11.73)	(17.80)	(18.65)	(9.99)	(14.64)
age^2	-0.0002***	-0.0002***	-0.0003***	-0.0002***	-0.0002***	-0.0002***
	(-19.06)	(-9.78)	(-14.65)	(-14.67)	(-7.95)	(-11.33)
gender	0.1593***	0.1634***	0.1652***	0.1475***	0.1532***	0.1538***
	(72.76)	(44.01)	(53.38)	(64.54)	(39.92)	(47.58)
edu	0.0293***	6.3626***	0.0271	0.0009	6.0252***	0.0101
	(3.64)	(12.86)	(1.47)	(0.11)	(12.50)	(0.54)
SOE	0.1589***	0.1003***	0.2014***	0.1448***	0.0927***	0.1826***
	(61.37)	(22.87)	(53.20)	(55.71)	(21.38)	(47.96)
inverse Mills ratio				-0.5303***	-0.3756***	-0.5506***
				(-13.73)	(-6.02)	(-10.62)
Kleibergen-Paap	30,000	579.863	30,000	26,000	575.506	29,000
rk LM test	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Kleibergen-Paap	100,000	315.567	390,000	100,000	313.499	37,000
rk Wald F	(7.03)	(7.03)	(7.03)	(7.03)	(7.03)	(7.03)
Constant	-3.4223***	-3.0267***	-3.0856***	-3.4183***	-3.0772***	-3.0511***
	(-97.03)	(-39.77)	(-66.53)	(-95.97)	(-40.68)	(-65.10)
Observations	304,537	149,585	154,952	295,101	146,203	148,898
Pseudo R^2	0.4469	0.1981	0.4552	0.4234	0.1865	0.4295

Notes: the t-statistic is presented in the parenthesis; ***, **, and * significance at the 1%, 5%, and 10% levels. In addition to the variables above, the regression also controls for capital, port, sector and occupation fixed effects.

Table 6: Robustness test with alternative measures of skill and market Access

We re-run the baseline regression using Heckman correction with instrumental variables to control for endogeneity in addition to sample selection issues. Columns 1-3 summarizes results under an alternative skill cut-off (high school education). Columns 4-6 summarizes results under an alternative definition of market access, and columns 7-9 brings the two together. Columns 1, 4 and 7 present results on the full sample, while columns 2, 5, and 8 focuses on the high market access subgroup, and columns 3, 6 and 9 focuses on the low market access subgroup.

	Robustness Test 1			**Robustness Test 2**			**Robustness Test 3**		
	Overall	**High MA**	**Low MA**	**Overall**	**High MA**	**Low MA**	**Overall**	**High MA**	**Low MA**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
lnMA	0.3256***			0.3451***			0.3442***		
	(54.18)			(59.19)			(59.25)		
ln*MA*skill*	0.4552***	-12.9869*	1.3156***	0.0152*	-4.7099***	0.1966***	0.1162***	-12.9883*	1.3157***
	(9.47)	(-1.68)	(7.73)	(1.95)	(-12.14)	(8.55)	(2.76)	(-1.68)	(7.73)
lnMA^2**skill*	-0.1552***	2.7137*	-0.3433***	-0.0123***	0.9494***	-0.0477***	-0.0757***	2.7140*	-0.3433***
	(-13.95)	(1.71)	(-6.51)	(-6.70)	(12.23)	(-6.85)	(-7.78)	(1.71)	(-6.51)
ln*gdp*	0.1891***	-0.0029	0.2169***	0.1814***	0.3710***	0.2038***	0.1806***	-0.0030	0.2169***
	(68.43)	(-0.01)	(66.47)	(75.99)	(61.32)	(76.92)	(67.61)	(-0.02)	(66.47)
ln*capital*	0.7349***	3.2900	0.6070***	0.8100***	-0.2929***	0.8346***	0.8111***	3.2907	0.6069***
	(28.84)	(1.60)	(16.81)	(36.28)	(-5.89)	(28.60)	(32.93)	(1.60)	(16.82)
marriage	0.0403***	0.0040	0.0586***	0.0406***	0.0204***	0.0486***	0.0385***	0.0040	0.0586***
	(9.87)	(0.08)	(8.61)	(10.08)	(2.87)	(8.68)	(9.56)	(0.08)	(8.61)
age	0.0209***	-0.0962	0.0246***	0.0198***	0.0164***	0.0210***	0.0193***	-0.0962	0.0246***
	(20.14)	(-1.38)	(14.28)	(19.90)	(9.99)	(14.64)	(18.84)	(-1.38)	(14.28)
age^2	-0.0002***	0.0012	-0.0003***	-0.0002***	-0.0002***	-0.0002***	-0.0002***	0.0012	-0.0003***
	(-16.56)	(1.41)	(-11.85)	(-16.05)	(-7.95)	(-11.33)	(-15.18)	(1.41)	(-11.85)
gender	0.1554***	-0.4235	0.1863***	0.1490***	0.1532***	0.1538***	0.1449***	-0.4236	0.1863***
	(56.01)	(-1.22)	(36.23)	(65.77)	(39.92)	(47.58)	(54.70)	(-1.22)	(36.23)

Table 6 continued:									
edu	0.1435***	5.0017*	-0.2073***	0.2454***	6.0257***	0.0101	0.2562***	5.0022*	-0.2073***
	(8.69)	(1.73)	(-4.68)	(29.60)	(12.50)	(0.54)	(17.65)	(1.73)	(-4.68)
SOE	0.1416***	0.8674*	0.1240***	0.1543***	0.0927***	0.1826***	0.1583***	0.8674*	0.1240***
	(39.01)	(1.89)	(16.13)	(59.58)	(21.38)	(47.96)	(46.67)	(1.89)	(16.13)
inverse Mills ratio	-0.3311***	-9.6297*	-0.0506	-0.4655***	-0.3756***	-0.5506***	-0.5028***	-9.6307*	-0.0505*
	(-7.12)	(-1.71)	(-0.63)	(-12.18)	(-6.02)	(-10.62)	(-11.32)	(-1.71)	(-0.63)
KleibergenPaap	1244.170	285.440	575.477	27000	29000	570.447	1509.920	29000	285.402
rk LM test	(0.00)	（0.00）	（0.00）	(0.00)	（0.00）	（0.00）	（0.00）	（0.00）	（0.00）
KleibergenPaap	742.067	149.662	313.476	100000	37000	360.446	1098.770	36000	149.641
rk Wald F	(7.03)	（7.03）	（7.03）	(7.03)	（7.03）	（7.03）	（7.03）	（7.03）	（7.03）
Constant	-3.3227***	-11.0565**	-2.0467***	-3.5212***	-3.0777***	-3.0509***	-2.9196***	-11.0578**	-2.0467***
	(-66.16)	(-2.23)	(-18.25)	(-99.55)	(-40.68)	(-65.08)	(-80.88)	(-2.23)	(-18.26)
Observations	295,101	146,203	148,898	295,101	146,203	148,898	295,101	146,203	148,898
Pseudo R^2	0.43	0.40	0.42	0.43	0.21	0.43	0.43	0.41	0.22

Note: the t-statistic is presented in the parenthesis; ***, **, and * significance at the 1%, 5%, and 10% levels. In addition to the variables above, the regression also controls for capital, port, sector and occupation fixed effects.