Linguistic Distance, Internal Migration and Welfare: Evidence from Indonesia*

Yao Wang†
Syracuse University

November 2021

Click here for the latest version

Abstract

This paper quantifies the effects of cultural barriers on internal migration and welfare by exploiting rich ethnolinguistic data in Indonesia and a spatial equilibrium framework. I estimate internal migration gravity using linguistic distances as a proxy for cultural barriers and instrument for current linguistic differences using data from the 1930 colonial census. I find inverted U-shaped effects of linguistic distance on migration. Longer linguistic distance encourages migration for linguistically close location pairs, but the pattern inverts as linguistic distance grows. The effects are more prominent for unskilled and older populations. To quantify welfare and distributional implications of linguistic barriers, I further develop a quantitative spatial model with heterogeneous skill groups, incorporating linguistic distance as migration and trade barriers. I find that a simulated reduction in linguistic distances by extrapolating the historical migration trend generates smaller welfare gains but improve equity more than a similar reduction in geographic barriers.

JEL Classifications: J61, N95, R12, R13, R23, Z13

Keywords: internal migration, economic geography, culture

*I am extremely grateful to Alex Rothenberg, Devashish Mitra, and Stuart Rosenthal for their guidance and support. I thank Gary Engelhardt, Alfonso Flores-Lagunes, John Yinger, Mengxiao Liu, Kristy Buzard, Giuseppe Germinario, Rachel Jarrold-Grapes, Samuel Saltmarsh, Maeve Moloney, Yanmin Yang, Yimin Yi, Tianyun Zhu, and seminar participants at Syracuse Graduate Student Workshop. All errors remain my own.

†Ph.D. candidate, Syracuse University. 110 Eggers Hall, Syracuse, NY 13244-1020. Email: ywang119@syr.edu.

1 Introduction

Migration is an important equilibrium device for equalizing income differences both within and across countries. However, the large regional and sectoral disparities in the developing world suggest the existence of substantial barriers to internal migration (Lagakos 2020; Gollin et al. 2014; Bryan and Morten 2019). While the existing literature has examined many factors contributing to those frictions,[1] the role of culture has been understudied, especially for internal migration in a less-developed country context.[2] As an integral part of people's lives, culture and language can deeply influence people's economic decisions, including migration, and therefore profoundly shape internal economic geography. Examining this issue in less-developed countries raises special policy interests. On the one hand, the effects of culture can be more salient in developing countries as many of them feature diverse ethnic backgrounds and weak nation building originating from colonial history. On the other hand, the cultural environment in those countries evolves rapidly as urbanization proceeds at an astonishing pace.

How do cultural barriers affect internal migration patterns? Answering this question is empirically challenging because culture is a multidimensional concept that is hard to measure. Moreover, inter-regional cultural connections are substantially influenced by migration patterns, which poses threats to identifying causal effects. If cultural frictions do affect migration, what are the welfare gains of reducing these barriers, and how are they shared between skilled and unskilled workers? When individuals relocate to more desirable locations as cultural barriers are reduced, other welfare components determined by labor and goods markets will adjust accordingly. The lower capacities to cope with cultural differences of less-educated individuals imply that they might benefit more. To account for all aspects of welfare gains from reducing cultural barriers, a quantitative framework is needed.

This paper quantifies the effects of cultural barriers on internal migration and welfare by exploiting rich ethnolinguistic data and a spatial equilibrium framework in the context of Indonesia. With more than 1,300 ethnic groups, each following different traditions and speaking their own native language, Indonesia is a fascinating setting to study cultural barriers and migration.[3] The paper starts by examining the reduced form effects of cultural barriers on inter-regional migration patterns. The rich ethnolinguistic data in Indonesia allows the construction of linguistic distance between locations to proxy cultural barriers. By merging each ethnicity to its corresponding native language in the *Ethnologue*, which categorizes languages into language trees based on various linguistic characteristics, I first construct the linguistic distance between each pair of ethnicities defined as the lack of shared language branches, following Fearon (2003) and Esteban et al. (2012). I then create the linguistic distances between locations by weighting the previous linguistic distances with current ethnic composition shares in each location. Combining this linguistic distance measure and inter-regional migration data, I estimate the migration elasticity to linguistic distance using a gravity specification (Yotov et al., 2016) for the overall Indonesian population and separately for the skilled and unskilled individuals.

[1] People in remote rural areas may have limited information about the benefits of moving to a city (Baseler 2019; Porcher 2019; Farré and Fasani 2013); migration could be risky and potential migrants cannot access formal insurance markets (Bryan et al. 2014; Munshi and Rosenzweig 2016; Morten 2019); in the absence of property rights, people may be locked in locations where they are born to protect their properties (De Janvry et al. 2015; Chen 2017; Gottlieb and Grobovsek 2019).

[2] Cultural barriers are more extensively studied as frictions for international trade and migration (see Adsera and Pytlikova (2015) for a review).

[3] This is according to the self-identified ethnic information in the 2010 Census.

I attempt to address the potential endogeneity of linguistic distance to migration. The endogeneity first comes from reverse causality —— linguistic distance is based on contemporary regional ethnic composition, which is potentially determined by current migration patterns. Moreover, there might be omitted variables that make locations simultaneously attractive to migrants and also culturally diverse. I tackle endogeneity by using historical ethnic composition data recorded in the 1930 colonial census to instrument for the contemporary one. This 1930 linguistic distance reflects predetermined cultural connections that are less contaminated by migration, as it was collected when the economy was predominantly agricultural and with relatively limited labor market integration.[4] It is also unlikely to be determined by current unobservable socio-economic determinants of migration patterns.

I find that while a modest level of linguistic distance is conducive to migration, linguistic distance has detrimental impacts on migration when it crosses a threshold. This inverted U-shaped effect reflects the trade-off between benefits and costs of living in a different cultural environment.[5] Linguistic distance increases migration costs due to the extra efforts required to overcome informal and informational barriers. However, a modest linguistic difference could be an amenity since people enjoy cooperating with people and consuming goods from different cultures. This inverted U-shaped effect is robust to using different linguistic distance measures, using different samples, and focusing on different types of migration (recent or lifetime). I also find significant asymmetric effects for people with different education levels or at different ages. Less-educated and older populations are generally more affected by linguistic differences, consistent with the literature (Bauernschuster et al., 2014).

In the second part of the paper, I construct a quantitative spatial model to understand the welfare and distributional implications of reducing linguistic distance. The model is based on Allen and Arkolakis (2018), where locations are connected through costly trade and migration. Motivated by heterogeneous migration elasticities across skill groups, I additionally incorporate multiple skill groups into the model. The skill division enters the model through two mechanisms. On the labor supply side, unskilled migration is more susceptible to cultural barriers; therefore, their migration propensity is more affected by reducing linguistic distance. On the labor demand side, skilled and unskilled workers are complementary in local production, and locations differ in their demand for skills. Whether reducing migration barriers benefits a certain skill group depends on whether those workers are connected to where demand for their skill is high.

To take the model to the data, I combine estimates of the migration elasticity for skilled and unskilled workers with wage and population data to estimate the skill intensity of production for each location. Next, I calibrate the trade gravity and other parameters from the literature and use the model structure to recover the unobserved exogenous local fundamentals. I find significant positive correlations between the recovered amenities and observed amenities using external data, which lends credibility to the model and parameter choices.

In the last part of the paper, I use the quantitative spatial model to evaluate the welfare implications of reducing linguistic distances. As people migrate to different locations, ethnic composition evolves, and linguistic distance changes. Inspired by this, as a baseline, I quantify the welfare implications of a

[4] In 1930, Indonesia was still under the rule of Dutch colonists and rapid industrialization and structural transformation had not taken off.

[5] This is different from findings of Falck et al. (2012) in Germany and Li (2018) in China, where they find detrimental effects, but consistent with Krieger et al. (2018)'s findings for international migration.

6.2% reduction in linguistic distances by extrapolating the migration trend from 1995 to 2000.[6] Reducing linguistic distances by 6.2% increases overall welfare by 0.13%. Moreover, it substantially reduces skill inequality by 0.85%, which is much larger in magnitude than a similar reduction in geographic barriers. With this reduction in linguistic frictions, people get access to the most productive or desirable places, such as Jakarta, Bandung, Bali, East Kalimantan, and Riau. Among those who benefit, the unskilled, whose migration is more constrained by cultural barriers and who are more likely to initially locate in linguistically isolated places, gain more when linguistic distance is reduced.

To demonstrate the potential mechanisms of the welfare effects of cultural barriers, in the second set of counterfactuals, instead of focusing on reducing linguistic barriers everywhere, I look at the welfare effects of reducing the linguistic distance by half between every pair of locations. By correlating the predicted welfare changes with observables or recovered locational characteristics, I show that the welfare implications of reductions in linguistic distance largely depend on where the reductions are. First of all, reducing linguistic barriers to economically significant cities can induce more welfare gains. Second, in locations that are linguistically accessible,[7] a reduction of linguistic distance can increase migration barriers. This is because within modest level of cultural disparity, people benefit more than lose from cultural heterogeneity, as demonstrated in the reduced form migration gravity estimation. As for the distributional effect, who gains more depends on local demand in the affected locations. Skilled (unskilled) workers gain more if reducing linguistic distances connects them to locations with high (low) skill intensity.

Lastly, I run another set of counterfactuals to draw more insights into the implications of cultural barriers. Because language and culture evolve slowly, one can expect that linguistic frictions may persistently form the economy. This is in stark contrast with its rapidly changing economic environment, as Indonesia actively participates in globalization and enacts local policies to promote economic growth. To examine how linguistic frictions interact with the rapidly transforming economic environment, I impose a 5% productivity increase in each location and calculate the local employment elasticity which measures how responsive local employment is to this local shock. The key insight is that the local employment elasticity is an endogenous variable depending on the location's linkages to other locations. One such linkage is linguistic distance. I find that places with an intermediate level of linguistic distance can attract more employment in response to positive productivity shocks. This implies that cultural connections between locations can impact the effectiveness of location-based policies and, therefore, should be considered in policy design.

With rich ethnic diversity and a long history of colonization, Indonesia has a number of policies to reduce ethnic and religious barriers and promote nation building, demonstrated by its national motto, "Unity in Diversity." For example, it has adopted a lingua franca, *Bahasa Indonesia*, for education and formal communication, which turned out to be effective as almost everyone could speak it in 2010 (Bazzi et al., 2019).[8] The reduced form and counterfactual results show that cultural barriers significantly impact people's migration decisions and thus affect economic welfare despite policy efforts.

[6] Specifically, by comparing the linguistic distance calculated using observed ethnic composition in 1995 and 2000, I find an average fall in linguistic distances by 6.2% in 50 years with this trend.

[7] Those places with short linguistic distance to other places.

[8] In addition, the government relocated more than 4.8 million volunteer migrants from Java and Bali to outer islands through a large-scale resettlement project, known as *Transmigration*, to encourage inter-group contact.

This paper is related to two strands of the existing literature. First, it contributes to understanding how culture affects migration and economic development (Collier 2017; Ginsburgh and Weber 2020). Although the importance of culture in international migration has been widely established (Adsera and Pytlikova, 2015), similar patterns within countries have been less explored. The few studies on this issue in a within-country context are in developed countries (Falck et al. 2012; Bauernschuster et al. 2014). Only recently have some studies begun using new datasets to examine the role cultural factors play in labor market integration in the developing world.[9] The closest study to this paper which also studies internal migration in a developing country is Li (2018), who focuses on a different but highly correlated distance measure, genetic distance, and finds it has significant and adverse effects on internal migration in China. Compared to the previous literature, I combine both contemporary and historical linguistic distance measures to identify the effects of cultural distance on internal migration. I focus on Indonesia, a distinct setting with extreme diversity and a long history of nation building. Moreover, I further incorporate the reduced-form results into a quantitative spatial model to quantify the general equilibrium implications of linguistic distance and conduct counterfactual analysis, which, to the best of my knowledge, has not been done in the literature. Second, this paper is also closely related to the recent literature using the spatial equilibrium model to study labor misallocation (Bryan and Morten 2019; Tombe and Zhu 2019; Fan 2019; Khanna et al. 2021; Monte et al. 2018). While many factors that affect labor allocation have been studied (for example, international trade, pollution, and commuting accessibility), this paper emphasizes the importance of cultural barriers. Compared to Bryan and Morten (2019), which also focuses on the migration frictions in Indonesia, I discover a unique inverted U-shaped effect of linguistic distance utilizing historical census data. I also additionally incorporate heterogeneous skill groups and goods trade to examine the welfare and equity implication of migration frictions.

2 Background

2.1 Migration and Regional Inequality in Indonesia

Indonesia is the fourth most populous country in the world, with more than 243 million people in 2010. As urbanization is proceeding at an astonishing pace, inter-regional migration has become an essential feature of the Indonesian economy. In 2010, around 10% of Indonesians lived in a different district from where she was born. Around 5% live in a different district from where she lived five years ago. However, empirical evidence shows that substantial migration frictions remain in Indonesia. Bryan and Morten (2019) find that migrants in Indonesia need to be compensated with a 39% increase in earnings compared to non-migrants, while Americans only require a 15% increase in incomes.

The lack of labor mobility exacerbates regional inequality. In 1983, the regional GDP per capita of the wealthiest district (Central Java) was almost 23 times the GDP per capita of the poorest district (South Bengkulu). The Indonesian government has implemented many placed-based policies to promote more balanced regional development, including regional development programs, building transportation infrastructure, and constructing elementary schools.[10] However, the disparity of regional development

[9] As an example of a related literature that focuses on goods trade, Fenske and Kala (2021) digitize the 1901 Census and historical price data of India, and detect significant negative correlations between linguistic distance and market integration in India.
[10] See Rothenberg and Temenggung (2019) for a review.

has become even more salient over time. Appendix Figure A.1 presents a histogram showing that the regional GDP per capita for the richest district (Central Java)'s GDP per capita is 88 times the poorest district (Lembata, an island in the Lesser Sunda Islands) in 2010.[11]

2.2 Language and Ethnicity in Indonesia

Indonesia is also an incredibly diverse country with more than 1,300 self-identified ethnicities scattered across more than 6,000 islands. Using the 2010 Census, Bazzi et al. (2019) show that the ethnic fractionalization index — the probability that two residents belong to different ethnicities — is around 0.81 in Indonesia. Each ethnicity speaks its own native language, and follows distinct customs and traditions. The differences among ethnicities remain partly due to Indonesia's archipelagic geography and lack of shared history. For most of history, different regions of Indonesia have been under the rule of several different independent kingdoms. It was not until the Dutch colonial period that the shared identity of Indonesia was formed. The lack of national identity limits inter-group interactions. Even after independence, many isolated outer islands residents have never met with people from other parts of the country and have no idea of what "Indonesia" means. Elizabeth Pisani mentions her failure to find a complete national map in Sumba — "In Sumba, the nation didn't exist" (Pisani, 2014).

The cautious attitude toward collecting ethnicity statistics demonstrates the extent to which the Indonesian government is sensitive to ethnicity issues. During the Old Order Era (1945-1967) and the New Order Era (1967-1998), the Indonesian government collected no data on ethnicity, believing that codifying ethnicity's diversity would hurt political stability. It was not until 2000 when the country had gone through an extensive democratization reform that the government started collecting information about ethnicity composition in Census.[12]

The Indonesian government also implements other policies to promote nation building and intergroup interactions. In 1928, the Youth Oath —– "One Land, One Nation, One Language" (*Satu Nusa, Satu Bangsa, Satu Bahasa*) was declared, and *Bahasa Indonesia*, a modified version of Malay, was promoted as the national language.[13] Subsequently, it is required to use Bahasa Indonesian school and official occasions. After independence, "Unity in Diversity" (*Bhinneka Tuggal Ika*) has become the state motto. In order to further simulate intergroup contact and relocate the population from densely populated Java and Bali to outer islands, the Indonesian government continued and expanded a large-scale resettlement program —– *Transmigration* — which the Dutch colonial government initiated. Through this project, landless households from Java, Bali, and Madura volunteered and were supported to migrate to less populated areas in Kalimantan, Sumatra, Sulawesi, Maluku, and Papua.

Despite the efforts to promote the integration of ethnic groups, ethnic cleavages remain. Although everyone can speak *Bahasa Indonesia*, only less than 20 percent use it as their daily language at home (Bazzi et al., 2019). As for the *Transmigration* program, the results are controversial. While Bazzi et al. (2019) find it has positive integration effects in diverse and vibrant locations, *Transmigration* is also

[11] Oil revenue is excluded in the calculation of gross domestic regional product.

[12] The 2010 Census continued the inclusion of the question about ethnicity and it. also include a question about people's daily language.

[13] Accirding to Ananta et al. (2014a), Bahasa Indonesia has its roots in Malay dialect spoken in the coastal areas of the Malacca Strait and the Riau islands, and it actually has been used as a langua franca in trade before 1928. People who speak Malay and Batawi can undertand it.

blamed for causing communal clashes between native settlers and newcomers, especially for those who are religiously antagonistic.[14]

3 Data

Spatial Units The primary spatial unit used in this analysis is the district (*Kabupaten* for the rural district and *Kota* for the urban district), which is the second-tier administrative unit in Indonesia. There were 416 districts in Indonesia in 2010, with an average size of around 6,000 square kilometers and a mean population of around 425,000.[15] Considering many large metro areas, like Jakarta and Surabaya, contain multiple districts, I dissolve districts within the same metro area into one district, following the delineation in Civelli et al. (2021) (see Appendix B.1 for defining of the spatial distribution of metro areas).

Contemporary Linguistic Distance To proxy for ethnic and cultural disparities, I construct distance between languages using the sixteenth edition of the *Ethnologue* database, following Esteban et al. (2012) and Fearon (2003).[16] In *Ethnologue*, each language is categorized by a language tree with a maximum number of 15 branches based on various characteristics, including lexicon, syntax, phonology, and grammar. The distance between two languages is approximated by the lack of shared language branches. Formally, the distance between language m and n is defined as,

$$\tau_{mn} = 1 - \left[\frac{branch_{mn}}{max(branch_m, branch_n)} \right]^{\kappa}$$

where $branch_{mn}$ is the number of shared language tree branches between the two languages, and $max(branch_m, branch_n)$ is the maximum possible number of common branches. κ determines the level of linguistic similarity is emphasized.[17] Following Esteban et al. (2012), I set $\kappa = 0.05$ as the baseline and set $\kappa = 0.5$ and $\kappa \to \infty$ for robustness checks. Around 56% of the 761 language pairs in Indonesia do not share any branches.

Then I use the local ethnic composition data from the Indonesia Census to calculate the linguistic distance between locations.[18] The linguistic distance between district i and j is the population-weighted

[14] In 1999, conflict broke out between the native Dayaks and Malays and the transmigrant Madurese. Two years later, the Dayaks and Madurese clashed again, resulting in thousands of deaths and many Madurese being displaced (van Klinken, 2003).

[15] This is almost equivalent to the US county in area, but larger in population. The average *kabupaten* population is almost double the population for the average county.

[16] There are several different ways to calculate the distance between languages, including Lexicostatistical Distance, Levenshtein distances, analyzing sound and so on. Ginsburgh and Weber (2020) has thorough discussion of those measure.

[17] A low value of κ separates languages that have very few branches in common from the rest. As κ increases, slight differences acquire greater salience while the larger differences play a less than proportional role. In the extreme case when $\kappa \to \infty$, the smallest difference is identified as an absolute difference, indistinguishable from deeper linguistic cleavages. In this case, the constructed linguistic distance collapses to the (opposite of) commonly-employed ethnolinguistic fractionalization (ELF) index (Alesina et al., 2003).

[18] To further construct linguistic distance between districts, it is necessary to match the ethnicity to language. I follow Bazzi et al. (2019)'s mapping of each ethnicity in the 2000 and 2010 Census to its corresponding language in *Ethnologue*. The 2010 Census records both ethnicity and people's native language. So they use the individual-level information on the home language available in the 2010 Census to define the native language for each ethnic group. In the 2000 Census, only ethnicity is recorded, so they merge with the 2010 Census using the ethnicity information to map to its language use.

sum of the distance between languages.

$$Ldist_{ij} = \sum_m \sum_n w_m^i w_n^j \tau_{mn} \tag{1}$$

where w_m^i is the 1995 population share of ethnicity speaking language m living in district i, and w_n^j is the population share of ethnicity speaking language n living in district j. Intuitively, this measure is the expected linguistic distance between two randomly selected individuals from different districts.

Historical Linguistic Distance To tackle the potential endogeneity of above contemporary linguistic distance with respect to migration, I obtained the ethnic composition from the 1930 Dutch East Indies Census (*Volkstelling, 1930*) for constructing historical linguistic distance. In 1930, when a unified national language had not been widely promoted nationwide, and people in different regions spoke their own native languages, Dutch colonists conducted the first full-count census that records ethnicity information.[19] The 1930 Dutch East Indies Census is the earliest complete record of ethnic composition in Indonesia one can get. After that, the post-colonization Indonesian government did not collect the ethnicity information until 2000 due to their efforts of enforcing homogeneity over Indonesia's population (Auwalin, 2020). To obtain general makeup information of the population in various parts of Dutch Indies, all populations were asked simple questions about their gender, civil conditions, age, the means of subsistence, literacy, ethnic group (referred to as "race" then), birthplace, and physical disabilities.[20]

The 1930 Dutch Indies Census adopted "social criteria" that uses language spoken, customs, and habits to distinguish ethnicity (van Klinken, 2003). The Census also documents the difficulties of accurately classifying different ethnicities then. On one hand, different locations have inconsistent names for the same ethnic group. On the other hand, it was often impossible to make a classification where the margins of the various groups overlapped as in the case of the Malays and Dayaks in Borneo, the South-Sumatrans and the Torajas in Celebes.

Despite these obstacles, the 1930 Census captures the most important ethnicities then with the assistance of many experts. Javanese is the most populous ethnic group, which accounts for 47% of the total inhabitants. The other major ethnic groups include Sundanese, Madurese, Minangkabau, and Buginese. Those ethnic groups do not distribute evenly spatially. The densely populated Java is almost ethnically homogeneous - nearly three-quarters of the native population was Javanese, while the remaining population almost entirely consists of Sundanese, Madurese, and Batavians. In contrast, outer islands show much more salient diversity.[21]

There are two inconsistencies between 1930 Colonial Census and the 2000 Census. The first is in definition of ethnicity. The 1930 Census is more general in ethnicity definition than the 2000 Census. While there are in total more than 1,300 ethnicities defined in the 2000 Census, in 1930 there are only 137 ethnic groups documented. Appendix Table A.1 shows the 15 major ethnicities defined in 1930 Census and their corresponding ethnicities in 2010 Census definition. The other one is the changes in administrative

[19] In central Kalimantan, Maluku and some other small islands, where interviews were hard to conduct, the data are estimated.

[20] The great majority of the population was enumerated twice in Java, but only enumerated once in outer islands. For some groups in Java, additional questions about dwelling conditions, religion and education are asked.

[21] In 1930, Dutch already finished its territory expansion which later form the territory of the current Republic of Indonesia. Maluku and Papua are dropped from the analysis due to the difficulty of merging ethnic group to the corresponding language in *Ethnologue* and the inconsistent quality of 1930 Colonial Census ethnicity data.

border as shown in Appendix Figure A.3. I correspond the border in 1930 to its most geographically approximate 2000 border manually. Most of the borders conincide perfectly to their modern counterparts. However, new district divisions have been created in the process of dencentralization. This inconsistency is concentrated for outer islands (See Appendix Section B.2 for more details). To accommodate the 1930 and 1995 linguistic distance to the same spatial units, I simply assume the recorded 1930 population was equally distributed within each administrative unit in 1930.[22]

Using the 1930 Dutch East Indies Census, the linguistic distance between district i and j are defined as follows,

$$Ldist1930_{ij} = \sum_m \sum_n w_m^i w_n^j \delta_{mn}$$

where w_m^i is the population share of ethnicity m living in district i, and w_n^j is the population share of ethnicity n living in district j, and δ_{mn} is an indicator indicator that equals one if m and n are the same ethnicity and 0 otherwise. Since the 1930 Dutch East Indies Census adapted a more general ethnicity definition than the one in the 2000 Census, it is difficult to precisely match each ethnicity defined in 1930 to its ISO codes. So instead of using τ_{mn} as in the contemporary linguistic distance definition (1), I treat each ethnic group defined in the 1930 Census as completely different linguistically, so this linguistic distance collapses to the extreme case when $\kappa \to \infty$ in linguistic distance definition. This approximation is arguably reasonable since more than half of the languages spoken in Indonesia do not share any common language tree branches.

Figure 1 shows the bin-scatter plot of the contemporary and historical linguistic distance.[23] Specifically, I group the historical linguistic distance into equal-sized bins, compute the mean of both contemporary and historical linguistic distances within each bin, and create a scatterplot of these data points. The graph shows a high correlation. A percentage increase in logged historical linguistic distance is correlated with a 0.79% increase in contemporary linguistic distance. Panel (A) and (B) of Figure A.4 show the calculated linguistic distance to Jakarta (colored in black) for 1995 and 1930, respectively. Two linguistic distance measures show consistent patterns - Jakarta is linguistically similar to the geographically nearby area but most different from Kalimantan and North Sulawesi. The high correlation between those two linguistic distance measures also demonstrates the slow evolution of regional ethnic composition.

Internal Migration Flow The data for internal migration flows is the 2010 Indonesia Population Census. This data records the current location, location five years ago, and birthplace (to district level) for all individuals in Indonesia. An individual is identified as a migrant if her location five years ago differ from her current location. Migration flows are aggregated from the individual-level data to the location-pair level.[24] In all analyses, I restrict the sample to individuals between 15 to 65 years old. Among 158 million individuals in the analysis, around 8 million (5%) changed their location from 2005 to 2010.

Regional Wage I obtain the regional wage data from 2009 and 2010 waves from the Indonesia National

[22] This induces measurement errors, but as shows in Figure 1, the partly extrapolated historical linguistic distance measure is highly correlated with the contemporary linguistic distance. I also replicate the main results with all variables consolidated to the 1930 spatial units as a robustness check.

[23] Bin-scatter plots more clearly show the correlation with large number of observations than scatter plots.

[24] This averages out observable and observed indivudual characteristics.

Labour Survey (SAKERNAS, *Survei Angkatan Kerja Nasional*). SAKERNAS is a national labor survey dataset representative at district level (for 2009 and 2010 waves), which records workers demographics including age, gender, education and current location, and rich labor market information including working hours per week, working experience, and monthly salary.

4 What Does Linguistic Distance Capture?

Linguistic distance is a summarized measure capturing an array of cultural dissimilarities. It is also highly correlated with some deep-root relatedness like genetic distance. In this section, I show empirical evidence for the different sets of practices and beliefs that linguistic distance captures.

Cultural Practice and Belief People speaking different languages may also follow different traditions and hold different beliefs. To demonstrate the cultural component of linguistic distance, I utilize the traditional practice data of ethnic groups, which is mainly based on *Ethnographic Atlas*, from Ashraf et al. (2020). This dataset contains the customs practices, including marriage, gender difference, community organization, and traditional economy of each ethnicity. With the data, I examine the extent to which cross-ethnicity linguistic distance is correlated to their similarity in traditional practices.

Table 1 shows the correlation of linguistic distance and similarity of custom practice between ethnic groups. Each observation in this set of regressions is an ethnicity-ethnicity pair. The dependent variables are indicators equal to one when the two groups share the same traditional practice. The significantly negative coefficients of all indicators for similarity provide strong evidence that it is more likely for linguistically close ethnicities to follows similar cultural practices.

Religious Dissimilarity Although Indonesia has been well known as a Muslim-majority country (Muslin took up to 87.5% of the total population in 2020), there is considerable regional variation in religious composition. While 28 out of 33 provinces are Muslim-majority (with Muslim share greater than 50%), East Nusa Tenggara is a Catholic-majority province, West Papua, Papua, and North Sulawesi were Protestant-majority, and the majority in Bali is Hindu. Moreover, religion is highly correlated with ethnicity. For example, Javanese and Sundanese are almost exclusively Muslim, while 95.2% of Balinese mostly embrace Hindu. To illustrate the correlation between linguistic and religious distances, I calculate the religious dissimilarity between district i and j as follow,

$$ReligionDist_{ij} = \sum_m \sum_n p_m^i p_n^j Same_{mn}$$

where p_m^i is the population share of religion m living in district i, p_n^j is the population share of religion n living in district j, and $Same_{mn}$ is an indicator equals one if m and n are the same religion, zero otherwise.

The last row of Table 1 show that the religious distance positively correlates with linguistic distance, with a correlation of 79%. Despite the high correlation, a simple regression of linguistic distance on religious distance shows that religious distance only explains 16% of the variation of linguistic distance, which implies that linguistic distance provides a rich measure of cultural disparities not limited to reli-

gious dissimilarity.[25]

5 Gravity Equation Estimation

5.1 Theoretical Prediction: How Does Linguistic Distance Affect Migration?

How linguistic distance affects migration is determined by the interplay between the benefits and costs of living in a culturally different environment. On the one hand, linguistic distance can have adverse effects on migration. First of all, there is direct communication cost related to speaking different languages. Although there is a lingua franca, *Bahasa Indonesia*, that nearly everyone can speak in Indonesia, the native language is still preferred on many informal occasions (Bazzi et al., 2019). Second, linguistic differences can directly affect skill transferability. By examining agricultural migrants engaged in the "Transmigration" program, Bazzi et al. (2016) find linguistic similarity important for their social interactions with natives and occupational adjustments. It is reasonable to hypothesize that language skills' effects are even more prominent for migrants in other sectors that involve more interactions. Third, heterogeneity can generate disamenities by inducing disarray and mistrust and disrupting cooperation and social stability (Ashraf and Galor, 2013). Lastly, linguistic distance can generate other informal barriers related to other aspects of culture, including traditional practices, values, and religion, as discussed in Section 3. People moving to more culturally distant places need more efforts to adapt to the destination's cultural environment (Falck et al., 2012). In addition, the literature shows that particular populations, for example, the less educated, are more susceptive to the costs of cultural barriers due to their lower capacities for acquiring language skills and adapting to a different cultural environment (Bauernschuster et al., 2014).

On the other hand, cultural disparities can bring benefits. First, heterogeneity can boost productivity if specialized tasks are complementary. Second, besides productivity, the diversity of available consumption goods and services is considered one of the attractive features of cities in urban literature (Ottaviano and Peri 2006; Trax et al. 2015; Glaeser et al. 2001).[26] Lastly, some people may prefer to live in a different cultural environment because of their pronounced inter-cultural interests or desire to take adventure (Krieger and Lange, 2010).

To sum up, the effects of linguistic distance on migration could be negative or positive, depending on the comparison of several different mechanisms. The literature further predicts that the beneficial effects of cultural distance should dominate at low levels and the detrimental effects should prevail at higher ones. On one hand, literature examining diversity on economic development shows that there are diminishing marginal returns to both heterogeneity and homogeneity (Ashraf and Galor, 2013). On the other hand, the benefits may only occur within a certain range of linguistic distance since it is necessary to comprehend and connect to other cultures for receiving those benefits (Krieger et al., 2018).

[25] Linguistic distance also capture other factors. According to the *dual inheritance theory* in social anthropology, linguistic distance is also highly correlated with genetic distance, which measures a more fundamental biological disparity between populations (Creanza et al., 2017). Although linguistic distance is persistent, migration induced by large-scale historical events or climate change can significantly alter linguistic distance. Also , since geographic barriers deterred migration and social interaction, they play an important role in linguistic distance. This is evident in Section 3, linguistic distance and geographic distance are highly correlated.

[26] This is also related to "love of variety" in preferences which has been the building block of many trade and spatial development model.

5.2 Fact 1: The Inverted U-shaped Effects of Linguistic Distance on Migration

In light of the above hypothesis of the conflicting impacts of linguistic distance on migration benefits and costs, the following gravity equation, extensively used in the literature in international trade (Anderson and Van Wincoop, 2003), is adopted to examine the influence of linguistic distance on migration propensity. This specification is based on a discrete location choice model that I will derive formally in Section 6,

$$\pi_{ij} = \exp\left\{\delta_i + \delta_j + \beta_l Ldist_{ij} + \beta_{l2} Ldist_{ij}^2 + \beta_g Gdist_{ij} + \beta_h Home_{ij} + \gamma X_{ij}\right\} + \epsilon_{ij} \tag{2}$$

where π_{ij} is the migration probability from i to j; $Ldist_{ij}$ refers to linguistic distance between i and j; $Gdist_{ij}$ refers to the logarithm of geographic distance, defined as the great circle distance between the district centroids. $Home_{ij}$ is an indicator that equals one if $i = j$, i.e.the destination is the same as origins, which captures that migrants are biased to live in their hometowns (Diamond, 2016). Henceforth, I refer to it as home bias. X_{ij} is a vector of origin-destination pairwise characteristics to control for other forces that could affect migration flows. It includes a dummy variable which equals one if the migration flow is across islands, the agroclimatic similarity index, and absolute differences in altitude and latitude (of district centroids) that captures the potential non-linear effects of geographic distances.[27] δ_i and δ_j are the origin and destination fixed effects, respectively. Including origin and destination fixed effects removes unobserved heterogeneity in sending and receiving districts. It also removes the multilateral resistance summarizing the attractiveness of all destinations relative to an origin since migration between a specific pair of locations is also influenced by the appeal of other locations.

Moreover, to tackle the endogeneity of the linguistic distance, in the baseline specification, I instrument the contemporary $Ldist_{ij}$ with historical linguistic distance constructed using the 1930 colonial Census discussed in Section 3. In 1930, Indonesia was still ruled by the Dutch colonists and remained agricultural. The main push and pull forces that motivate internal migration back then are quite different from those nowadays. So cultural distance constructed based on population in 1930 should not be affected by current migration trends and contemporaneous political and economic situations. The historical linguistic distance can strongly predict the contemporary one, as shown by the high correlation between the two measures in Figure 1. In addition, the first-stage F statistic is substantial, as shown in Columns 1 and 4 of Appendix Table A.2.

In all empirical estimations of migration gravity, I follow the recommended practice in empirical gravity literature (see Yotov et al. (2016) for a review). I use migration share $\hat{\pi}_{ij} = M_{ij}/L_i^0$ to approximate migration probability π_{ij}, where M_{ij} is migration flows between the source and destination location i and j and L_i^0 is the population of the origin location. I estimate all specifications using Poisson Pseudo-Maximum Likelihood (PPML) to correct heteroskedasticity of ϵ_{ij} and to get unbiased estimation even with a large share of zero migration flows (30% percent of the total migration flows are zeros) (Silva and Tenreyro, 2006). For PPML regression with instrument, I adapt a control function method and bootstrap

[27] I borrow Bazzi et al. (2016)'s measure of agroclimatic similarity. They use data from the Harmonized World Soil Database (HWSD) and other sources and construct a measure of the agroclimatic similarity between locations based on a set of geographic characteristics, including elevation, slope, ruggedness, altitude, distance to rivers and the sea coast, rainfall, temperature, and soil texture, drainage, acidity, and carbon content. The great circle distance, differences in latitudes and longitudes and the cross islands dummy are constructed using the administrative map of Indonesia in GIS software.

the standard errors following Lin and Wooldridge (2019) and Wooldridge (2015) (See Appendix C for detailed discussion of the estimation). I rescale linguistic distance so that it has a standard deviation of one, so the coefficients can be read as the effects of increasing the linguistic distance by one standard deviation. Lastly, to address the spatial correlation, the standard errors in all regressions are two-way clustered by origin and destination district (Cameron et al., 2011).

Table 2 presents the results of estimating specification (2). Column 1 shows that the unconditional correlation between linguistic disparity and migration probability is large and significantly negative. Column 2 further includes a quadratic term to test nonlinearity (I discuss the specification choices in detecting nonlinearity in Appendix C.2). Consistent with the prediction of the proposed hypothesis, linguistic distance is conducive to migration when it is modest, while it is detrimental to migration when it crosses a certain threshold. However, this can be confounded by many other determinants of migration that are also correlated with linguistic distance. Column 3 shows the result of adding other pairwise controls. After purging out the effects of other controls, both coefficients of the linear and quadratic terms remain statistically significant at the 1 percent level, but the magnitude of the linear term coefficient increases. In contrast, the magnitude of the quadratic term is reduced.

In the baseline specification in Column 4, I further instrument contemporary linguistic distance with the historical linguistic distance to tackle the potential endogeneity. As for the interpretation of the conditional effects of linguistic distance, the estimated linear and quadratic coefficients imply that half of a standard deviation increase on linguistic distance increases migration propensity by 10.7% for locations that are linguistically proximate (at the 5th quantile of linguistic distance distribution).[28] In contrast, half of a standard deviation increase in linguistic distance decreases migration propensity by 46% (at the 90th quantile of linguistic distance distribution). For location pairs with median linguistic distance, a half standard deviation increase in linguistic distance leads to a 39% reduction of migration propensity (the calculation is based on the last panel of Table 2).[29] Comparing to Column 3, the estimated coefficients of linguistic distance in the instrumented regression remain relatively stable in magnitude, suggesting that the potential endogeneity is not a significant concern conditional on controlling other pairwise determinants of migration propensity. The coefficients of linguistic distance are well-identified due to the sufficiently high reported Kleibergen-Papp F Statistic.

The identification of causal effects of linguistic distance on migration relies on the validity of historical instrument. First of all, according to the 1930 Colonial Census, only around 5% of the population resided in a diffrent location than their birthplace, and since the data did not differentiate if individuals were temporarily or permanently lived in a location, this overestimated the migration rate. This relative low level of migration rate, comparing to the 10% lifetime migration rate in 2010, suggests that the historical linguistic distance is less contaminated by migration and mainly captures predetermined cultural barriers related to exogenous historical and geographic factors. However, since the historical migration rate is non-zero, the historical linguistic distance can still affect current migration by previous migration network. To further assess the results are driven by endogeneity, I exclude from the sample location pairs involving the largest historical migration flows, where the endogeneity problem is likely to be most acute. In addition, I run the regression in Column 4 controlling for previous migration net-

[28] I choose a half of standard deviation increase to make no out of sample prediction.
[29] The effect size is calculated as $(e^{0.5\beta} - 1) * 100$

work following Beine et al. (2011) (See details in Appendix C.3). The effects of linguistic distance remains robust.

For visualization, I plot the marginal effects at each level of linguistic distance in Appendix Figure A.6.[30] For location pairs that are linguistically adjacent (below 1.62, which is the 8th percentile of the linguistic distance distribution), increasing linguistic encourages migration (marginal effects are positive). However, the effect flips (marginal effects turn negative) beyond 1.62.

Column 3 and 4 also reveals the significance of other controls. The coefficients of geographic distance are significantly negative. The magnitude is consistent with the estimation of Kone et al. (2018) in India, Tombe and Zhu (2019) in China, Falck et al. (2012) in Germany and Allen and Arkolakis (2018) in the US. It is larger than Bryan and Morten (2019)'s estimate in Indonesia potentially because we use different data and focus on different types of migration.[31] The estimated geographic distance effects on migration are also slightly larger than the comparable effects on goods trade in magnitude (Yotov et al., 2016), suggesting that it is harder to move people than goods across locations. The results also show that inter-island travel creates barriers for migration, although the effects are not significant after controlling other barriers. Interestingly, while the absolute longitude difference deters migration, the absolute latitude difference encourages migration, suggesting that people are more willing to migrate in the south-north direction but less so in the east-west direction.[32] Agroclimatic similarity increases migration significantly, which is consistent with Bazzi et al. (2016).[33] The coefficients of home bias $Home_{ij}$ are large, positive, and significant across all specifications, implysing that people have a significant preference to stay in their hometown. According to Column 4, people are almost ten times more likely to stay in their hometown than migrate to another location. This effect is also more notable than its counterpart in goods trade gravity by Yotov et al. (2016), which implies people are more attached to their hometown than goods.

The inverted U-shaped effect of linguistic distance on migration probability contrasts to the existing literature in other contexts (Bauernschuster et al. (2014) and Falck et al. (2012) in Germany; Li (2018) in China), where they find a significantly negative effect. However, the inverted U-shaped result is consistent with Krieger et al. (2018)'s findings for international migration.[34]

Robustness The inverted U-shaped result remains stable for several sensitivity checks. First, I split the sample by the 25th percentile of linguistic distance and run the PPML regressions with IV with the split samples. As shown in Appendix Table A.5, the linguistic distance effects on migration are positive, although not significant, below the 25 percentile and negative above the 25 percentile. Second, OLS and 2SLS estimations of the migration gravity equations show similar results with PPML estimation (see Appendix Table A.6). Third, the result is robust to creating linguistic distance using different κ (see

[30] The magional effect is calculated by plug in the value of linguistic distance into the derivatives of the linear and quadratic term.

[31] We use different data - while they use SUSENAS survey data, I use the full-count Census data, so the sample in this analysis covers larger part of the population in the country, including some extremely remote parts. I focus on recent migration (within 5 years) while they look at lifetime migration. Also, in their empirical gravity equation in the reduced form section, they do not control for the effects of other barriers than geographic distance.

[32] This pattern contradicts Diamond (1998)'s argument that differences in latitude are significant barriers to the transfer of technological innovations. While he looks at the universe of all countries, my analysis focuses on a more micro geography unit, i.e., Indonesia, and the most populous islands in Indonesia, Java, and Sumatra, have a predominantly East-West axis.

[33] While their study focuses on agricultural migrants, this analysis additionally includes non-agricultural migrant (86% of the total migrants). But the positive effect of agroclimatic similarity is still significant.

[34] They uses a slightly different explaining variable — genetic distance.

Appendix table A.7). Fourth, the results are not driven by outliers or migration to Jakarta (see Appendix A.8). Fifth, the results are robust to measuring migration using lifetime migration status as opposed to the recent ones (see Appendix table A.8). Lastly, the nonlinear effect remains significant using the colonial divisions of administrative units (see Appendix Table A.9).

5.3 Fact 2: Less Educated Individuals are More Affected by Linguistic Barriers

In this section, I run gravity equation (2) separately for people with different skill levels to test if they are differently sensitive to migration barriers using the baseline specification (2). In this analysis, skilled workers are defined as those with education levels equal to or higher than college, and unskilled workers are defined as those with education lower than college.

In Table 3, Column 1 reports the gravity equation estimation for skilled workers while Column 2 reports the results for unskilled workers. The inverted U-shaped effects of linguistic distance on both skilled and unskilled migration are consistent with the overall gravity estimation in Table 2. More importantly, skilled migration is less sensitive to linguistic distance than the unskilled since the coefficients of both the linear term and quadratic term are smaller in magnitude for the skilled. The conducive effect of linguistic distance is larger and persists in a larger range for unskilled when linguistic distance is low. However, when linguistic distance is larger, unskilled migration is also more deterred by linguistic distance than skilled migration. For the demonstrative purpose, I plot the marginal effects for skilled and unskilled workers at each level of linguistic distance in Panel A in Figure 2. It confirms that for both skilled and unskilled, the marginal effects of linguistic distance on migrations are positive for modest linguistic distance, and the effects decrease to negative as distance increases. Moreover, the magnitude of the size of the effect is more prominent for unskilled, which suggests they are more sensitive to linguistic distance when making migration decisions.

The lower capacities to overcome linguistic barriers for unskilled, suggested by the above results, raise concerns that linguistic distance have significant effects on selection on migrants, which is essential to forming human capital and promoting economic development. A location already isolated from other labor markets is further disadvantaged if it has more troubles attracting skilled labor. Following the migration selection literature (Borjas 1987; Chiquiar and Hanson 2005; Krieger et al. 2018), I measure the migrant selection using the skill composition of migrants (relative to the skill composition of the population in origin), $\ln M_{ij}^s/M_{ij}^u - \ln L_i^{s0}/L_i^{u0}$. Formally, it can be directly derived by log-transforming and taking differences of the gravity equations 18 for skilled and unskilled,

$$\ln \pi_{ij}^s - \ln \pi_{ij}^u = \delta_i - \xi_l Ldist_{ij} - \xi_{l2} Ldist_{ij}^2 - \xi_g Gdist_{ij} - (\eta\gamma^s - \eta\gamma^u)X_{ij} + \delta_j + \epsilon_{ij} \tag{3}$$

where the dependent variable is the log difference of skilled and unskilled migration propensities, further written as migration selection $\left(\ln M_{ij}^s/M_{ij}^u - \ln L_i^{s0}/L_i^{u0}\right)$ using the definition of π_{ij}^e. Migrants are positively (negatively) selected if migrant selection measure $\left(\ln M_{ij}^s/M_{ij}^u - \ln L_i^{s0}/L_i^{u0}\right)$ is positive (negative) as the migrants are disproportionally skilled (unskilled). By controlling origin fixed effects, destination fixed effects, and geographic distance, the effect of linguistic distance on migration selection ξ_l and ξ_{l2} can be identified.

Column 3 in Table 3 reports the estimation results for Equation (3), which are consistent with results

in Columns 1 and 2. Migration is negatively selected when locations are linguistically connected but positively selected for culturally remote location pairs. For demonstration, Panel B in Figure 2 plots the marginal effects of linguistic distance on migrant selection at each level of linguistic distance. The point estimates of marginal effect are positive for linguistic distance above 2.85 (88% of the sample), which implies linguistic distance does create additional barriers for unskilled when linguistic distance is sufficiently high. While for the other 12% linguistically close location pairs, the marginal effects are negative, suggesting linguistic distance is less of a problem when linguistic distance is modest.

To understand the effect size, I present the marginal effects at the 1st, 5th, 50th, 90th percentile of the linguistic distance distribution in Table 4. For location pairs that are exceptionally culturally similar (at the first percentile of linguistic distance distribution), increasing linguistic distance by a half standard deviation increases migration propensity by 25% for skilled workers and 64% for unskilled workers. For location pairs at the fifth percentile, as larger linguistic distance still increases migration propensity for unskilled workers, the effect is insignificant for skilled workers. For location pairs that are particularly culturally remote (at 90th percentile), a half standard deviation increase in linguistic distance deters skilled migration by 39% and unskilled migration by 62%.

For robustness check, I estimate separate gravity equations for more defined education groups, i.e., no education, primary school, junior high, senior high, college, and above. As shown in Appendix Table A.10, the results are robust to different skill splits as the higher educated population are monotonically more immune to linguistic distance.

5.4 Fact 3: The Negative Effect of Linguistic Barriers Intensifies with Age

Benefiting from the advancement of information and transportation technologies and promoting cultural diversity, Indonesians are increasingly mobile, venturing into a broader labor market that is more culturally distant from their own ethnic identity. To examine if migrants are less sensitive to linguistic distance, in this section, I look at how migration at different ages is differently affected by linguistic barriers. Nevertheless, instead of looking at recent migration (within five years) that I focus on in the previous analysis, I look at migrations that happen throughout people's lifetime to partly capture migration that happens in different periods.[35]

I run separate migration gravity equations analogous to Eq.(2) separately for each age group (results shown in Appendix Table A.10). In Figure 3, I plot the marginal effects at the 5th percentile and at the 75 percentile for different age groups to demonstrate age differences in the elasticity of migration to linguistic distance. The marginal effects (and confidence interval) at the 5th percentile are red, while the marginal effects at the 75th percentile are blue. Each dot is for a specific age group indicated in the x-axis. The marginal effects at the 5th percentile for all age groups are positive, but there is no significant difference across age groups. However, marginal effects at the 75th percentile are negative for all age groups. While the differences across groups are not statistically significant, the magnitudes suggest that older cohorts seem to be more impacted.

[35] Lifetime migration is identified if individual's birthplace is different from their current location.

6 A Quantitative Spatial Model

The estimation of migration gravity in the previous section demonstrates significant partial equilibrium effects of cultural barriers on internal migration. However, it is not sufficient for us to learn about the overall and distributional welfare implications of linguistic distance. Linguistic distance leads to inefficient labor allocation by inducing migration barriers, creating welfare losses through wages, prices, and amenities. Moreover, the unskilled might be hurt more as their migration is more constrained by cultural barriers. This section develops a quantitative model based on Allen and Arkolakis (2018) to account for a rich set of mechanisms that linguistic distance can affect welfare. To quantify the potential distributional effect, I extend the model to multiple skill groups following the recent economic geography literature (Allen and Donaldson 2020; Farrokhi and Jinkins 2019; Khanna et al. 2021; Fan 2019).

6.1 Setup

There are total N locations indexed by i for origin and j for destination in the country. Each location $i, j \in N$ differs in its amenity and productivity. Locations are connected through costly trade and migration. Skilled and unskilled workers decide where to locate based on the utility they can obtain in each location. Representative firms in each location produce dinstinct varieties using skilled and unskilled labor, which is the Armington assumption (Armington, 1969). Each location differs in its demand for skill depending on its endowment and technology. For example, large cities like Jakarta and Surabaya are better equipped with desirable amenities for service industries, and therefore, their local production relies more on skilled workers. In equilibrium, local wages and prices adjust to clear goods and labor markets.

6.2 Migration Flows

There are $\bar{L}$ workers in total who differ in level of skill $e \in \{s, u\}$, where s stands for skilled and u for unskilled workers. Workers are born in location i (origin), and choose the location to locate j (destination). Each location is endowed with initial (exogenous) labor stock $L_i^0 = L_i^{0s} + L_i^{0u}$. After workers choose their location, the local population is the aggregate of skilled and unskilled population, $L_i = L_i^s + L_i^u$.[36] Workers value local consumption goods and amenity and choose location to reside with the most desirable bundle of wage, price and amenity, discounted by the migration cost incurred when moving out of their birthplace. By utility optimization, the indirect utility of worker i with education $e \in \{s, u\}$ who is born in district i and moves to district j is,

$$V_{ij}^e(i) = \frac{w_j^e \epsilon_{ij}^e(i)}{P_j D_{ij}^e} u_j^e \tag{4}$$

where w_j^e is the nominal wage, and P_j is the price index in location j. u_j^e is the local amenity in location j, which is further decomposed as,

$$u_j^e = \bar{u}_j^e L_j^\beta \tag{5}$$

[36] This is still static setup that assume individual only make their location choice once.

where $\bar{u}_j^e$ is an exogenous part that is predetermined (i.e. soil quality, distance to coast), and L_j^β is an endogenous part that is determined by local population L_j and the elasticity β, which captures the congestion forces (i.e. traffics, housing prices).[37] The exogenous part $\bar{u}_j^e$ is different for education level e. For example, skilled workers may have a stronger preference to cleaner environment than unskilled (Khanna et al., 2021).

D_{ij}^e is the migration cost of moving from location i to j. Skilled and unskilled workers are differently sensitive to various migration barriers, $D_{ij}^s \neq D_{ij}^u$. $\epsilon_{ij}^e(i)$ is worker i's idiosyncratic preference draw that follows a Frechet distribution $F(x) = \exp\{-x^{-\eta}\}$, with shape parameter η that governs the dispersion of the draw. As η decreases, workers' preference are more dispersed over space, implying that they are more attached to certain places due to there preference so that they are less responsive to changes in other components of welfare, i.e. wage, price and amenity.

Each worker chooses the location that offer her the highest indirect utility. By the properties of the Frechet distribution, the probability that worker with education e that born in i moves to j is,

$$\pi_{ij}^e = \frac{(u_j^e w_j^e / P_j D_{ij}^e)^\eta}{\sum_{j'}(u_{j'}^e w_{j'}^e / P_{j'} D_{ij'}^e)^\eta} \tag{6}$$

Eq.(6) shows that individuals are more likely to migrate to locations with higher wage and lower price. They also prefer location that are easy to access relative to all other locations, which is captured by the multilateral-resistance term $\Phi_i^e = \left[\sum_{j'}(u_j^{e'} w_{j'}^e / P_{j'} D_{ij'}^e)^\eta\right]^{\frac{1}{\eta}}$ summarizing the appeal of all migration options for location i.

The *labor supply* in location j for education group e can be expressed as follow,

$$\tilde{L}_j^e = \sum_{o \in N} \pi_{ij}^e L_i^{e0} \tag{7}$$

By the property of Frechet distribution, the average welfare of workers living in location i is,

$$\mathbb{E}\left[\max_j V_{ij}^e(i)\right] = \Gamma(1 - \frac{1}{\eta})\Phi_i^e \tag{8}$$

where $\Gamma(\cdot)$ is the Gamma function, and $\Gamma(1 - \frac{1}{\eta})$ is a constant.[38]

6.3 Demand for goods

Workers have uniform constant elasticity of substitution (CES) preference over varieties of consumption goods produced in each location with the elasticity of substitution σ. Skilled and unskilled workers in i face the same price, and substitute goods produced in different location the same way. However, they earn different wages w_j^e. By maximization of utility constrained by income, and aggregation of the demand for goods of all local population, the total expenditure of population in i on the variety

[37] This in contrast to Diamond (2016) and Tsivanidis (2019) where they assume endogenous amenities depend on demographic composition across skill groups rather than total population. Since the total population and the skilled share of population is highly correlated, those two assumptions deliver similar conclusion.

[38] Allen and Arkolakis (2014) shows that this model is isomorphic to a model with amenity spillover, that is, amenity $\bar{u}_j^e = \tilde{u}_j^e {L_j}^\lambda$ is a function of population. They also show the isomorphism to Redding (2016) with certain budget share spend on housing, the supply of which is inelastic.

produced in location j is,

$$X_{ij} = \tau_{ij}^{1-\sigma} p_i^{1-\sigma} P_j^{\sigma-1} Y_j \tag{9}$$

where $Y_j = w_j^s L_j^s + w_j^u L_j^u$ is the total income in location j, and $P_j = \left(\sum_{i' \in N} p_{i'j}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}$ is the Dixit-Stigliz price index that summarizes consumer access to varieties in location j. p_{ij} is the unit price of variety produced in j for i, which is the factory-door price p_i discounted by iceberg trade cost $p_{ij} = \tau_{ij} p_i$.

6.4 Production

Each location i is assumed to produce a differentiated good y_i using skilled labor L_i^s and unskilled labor L_i^u. There are many firms producing homogeneous goods and the market is perfect competitive. The production function of a representative firm in location i is,

$$y_i = A_i \left(\theta_i^s L_i^{s\,\frac{\rho-1}{\rho}} + \theta_i^u L_i^{u\,\frac{\rho-1}{\rho}}\right)^{\frac{\rho}{\rho-1}} \tag{10}$$

where θ_i^e is the factor share of skill e and $\theta_i^s + \theta_i^u = 1$. When θ_i^s is higher, skilled workers weights more in district i's production, i.e. district i is more skill intensive. Henceforth I refer θ_i^s as skill intensity in location i. ρ is the elasticity of substitution of skilled and unskilled labor.[39] A_i is the overall productivity in location i,

$$A_i = \bar{A}_i L_i^{\alpha} \tag{11}$$

which consists of an intrinsic part $\bar{A}_i$ that is predetermined by exogenous characteristics, such as soil quality, distance to coast and etc., and an endogenous part that is determined by local population with elasticity α which captures the agglomeration externalities such as labor market pooling, input sharing, and knowledge spillovers (Rosenthal and Strange, 2004).

Profit maximization implies the following labor demand for skilled and unskilled labor,

$$p_i A_i \left(\theta_i^s L_i^{s\,\frac{\rho-1}{\rho}} + \theta_i^u L_i^{u\,\frac{\rho-1}{\rho}}\right)^{\frac{1}{\rho-1}} \theta_i^s L_i^{s\,-\frac{1}{\rho}} = w_i^s \tag{12}$$

$$p_i A_i \left(\theta_i^s L_i^{s\,\frac{\rho-1}{\rho}} + \theta_i^u L_i^{u\,\frac{\rho-1}{\rho}}\right)^{\frac{1}{\rho-1}} \theta_i^u L_i^{u\,-\frac{1}{\rho}} = w_u^u \tag{13}$$

Eq.(12) shows that wage for skilled worker is not only determined by the skilled population, which captures the classic downward sloping labor demand curve, but it is also associated with total local population by agglomeration and the unskilled population by skill complementarity.

By perfect competition, the unit price of y_i is,

$$p_i = \frac{1}{A_i}\left[(\theta_i^s)^{\rho}(w_i^s)^{1-\rho} + (\theta_i^u)^{\rho}(w_i^u)^{1-\rho}\right]^{\frac{1}{1-\rho}} \tag{14}$$

which suggests that the production cost for each location depends on local productivity, labor costs for skilled and unskilled labor and skill intensity.

[39] Here I abstract from capital as this framework is isomorphic to the case that capital are mobile and rents are equal across space Allen and Arkolakis (2014).

Plug Eq.(14) into the gravity Eq.(9), the goods gravity equation is,

$$X_{ij} = \tau_{ij}^{1-\sigma} \left\{ \frac{1}{A_i} \left[(\theta_i^s)^\rho (w_i^s)^{1-\rho} + (\theta_i^u)^\rho (w_i^u)^{1-\rho} \right]^{\frac{1}{1-\rho}} \right\}^{1-\sigma} P_j^{\sigma-1} Y_j \tag{15}$$

6.5 Equilibrium

Given trade cost τ_{ij}, migration cost M_{ij}, local fundamental productivity $\bar{A}_i$ and amenities $\bar{u}_i^e$, initial population L_i^{0e}, preference parameter $\{\sigma, \eta\}$, production technology parameters $\{\theta_i, \rho\}$, an equilibrium is defined over a series of endogenous parameters $\{w_i^e, L_i^e, \omega_i^e, \phi_i^e, p_i, P_i\}$ such that,[40]

- **Labor Market Clearing:** Labor demand stated in Eq.(12) and Eq.(13) equals labor supply stated in Eq.(7) for each location i and each education group e.
- **Goods Market Clearing:** For each location i, the total payment to labor equals total sales, and the total expenditure equals to total payments to goods.
- **Closed Model:** The model is closed in the sense that the total population is constant.[41]

6.6 Existence and Uniqueness of the Equilibrium

The sufficient conditions for the existence and uniqueness of the equilibrium are provided formally by Allen and Donaldson (2020). First of all, the existence of equilibria with positive populations is always guaranteed as long as the local fundamentals, migration, and trade frictions are positive. Second, the uniqueness of the equilibrium depends on the relative strength of agglomeration and congestion forces. When agglomeration forces are stronger than dispersion forces, there could be multiple equilibria. In this case, strong black hole equilibria with all economic activity concentrated in one point arises. In the model developed above, congestion forces include the negative effects of population on amenities, β, and dispersion of idiosyncratic preferences, η, which creates dispersion of economic activities away from locations with large populations. Agglomeration forces are governed by the positive effect of population on productivity, α. Given the parameter values I use in the model (discussed in the next section), I expect a unique equilibrium. In addition to the comparison of congestion and agglomeration forces, the product differentiation assumption implies that people have incentives to move to less populated locations in order to provide labor for the global demand of the local good, which can be seen as another dispersion force that further guarantees the uniqueness of the equilibrium.

7 Take the Model to the Data

This section describes how I estimate and calibrate the structural parameters and recover unobserved local amenity and productivity.

[40] See Appendix D.2 for the formal expressions.

[41] The initial population of a location should equal to the total out-flows of population to all location (include stayers), and the total population should equal to the in-migration from all locations (include stayers).

Calibrated Parameters Table 5 summarizes the calibrated parameters from related literature. For agglomeration parameter α, I calibrate it to 0.085 following the discussion by Bryan and Morten (2019). The recent literature shows a higher estimation of agglomeration forces in developing countries than in developed countries (Chauvin et al., 2017).[42] For Armington elasticity of substitution σ, I follow the related literature and set it to 6 (Simonovska and Waugh 2014; Feenstra et al. 2018). For dispersion parameters of preference distribution, I set $\eta = 1.6$ following Tombe and Zhu (2019) and Bryan and Morten (2019).[43] I calibrate the trade cost of goods following the international trade literature (Allen and Donaldson 2020; Yotov et al. 2016; Head and Mayer 2014; Gurevich et al. 2021; Egger and Lassmann 2012).

$$\tau_{ij}^{1-\sigma} = -\ln Gdist_{ij} - 0.5 \ln Ldist_{ij} \tag{16}$$

Estimation of Skill Intensity for Each Location I recover θ_j^e for each location by observed wage and population using the following property of CES production function,

$$\frac{w_j^s L_j^s}{w_j^s L_j^s + w_j^u L_j^u} = \frac{\theta_j^s L_j^{s\,\frac{\rho-1}{\rho}}}{\theta_j^s L_j^{s\,\frac{\rho-1}{\rho}} + (1-\theta_j^s) L_j^{u\,\frac{\rho-1}{\rho}}} \tag{17}$$

Figure 4 shows the distribution of the estimated local skill intensity θ_j^e. Major cities, for example, Yogyakarta, Jakarta and Surabaya are among the most skill-intensive locations.

Estimation of Internal Migration Cost With further parameterization the migration costs,

$$D_{ij} = exp\left\{\beta_l Ldist_{ij} + \beta_{l2} Ldist_{ij}^2 + \beta_g Gdist_{ij} + \beta_h Home_{ij} + \gamma X_{ij}\right\} \tag{18}$$

Eq.(6) can be expressed as follows,

$$\pi_{ij}^e = \exp\{\underbrace{\ln(\eta \tilde{w}_j^e) + ln(\eta u_j^e)}_{\text{destination FE}} + \underbrace{\ln \sum_{d'} (\alpha_{d'} w_{d'}^e / P_{d'} D_{od'}^e)^\eta}_{\text{origin FE}} - \eta\beta_g Gdist_{ij} - \eta\beta_l Ldist_{ij} - \eta\beta_{l2} Ldist_{ij}^2 - Home_{ij} - \eta\gamma X_{ij} + \epsilon_{ij}\} \tag{19}$$

where $\tilde{w}_j^e = w_j^e / P_j$ is the real wage in location j. This is exactly the gravity specification I use in Section 5. The results have been shown in Column (2) and (4) of Table 3 in Section 5.

Recovering Locational Fundamentals Combining the estimated and calibrated parameters and observed wage and employment data, I follow Allen and Arkolakis (2018) to recover location fundamentals ($\bar{u}_i^e$ and $\bar{A}_i$) that rationalize the observed data as a model equilibrium by the following procedures:

[42] The estimation is around 0.01 to 0.02 in developed world (Rosenthal and Strange, 2004)

[43] Fan (2019) estimate η separately for the skilled and skilled, and find a slightly larger η for skilled workers. Ideally I can use the empirical gravity equation combined with appropriate instrument to estimate η for different skilled group.

Step 1: Invert model to get price and welfare composite With the knowledge of the paramater vectors $\{T_{ij}, M_{ij}, Y_i, L_i^s, L_i^u, L_i^{s0}, L_i^{u0}\}$, where $T_{ij} = \tau_{ij}^{1-\sigma}$ and $M_{ij}^e = (D_{ij}^e)^{-\eta}$, I first invert a unique (to-scale) set of price or welfare composites $\{p_i^{\sigma-1}, P_i^{\sigma-1}, (\omega_i^s)^\eta, (\Phi_i^s)^\eta, (\omega_i^u)^\eta, (\Phi_i^u)^\eta\}$ that are consistent to the equilibrium conditions in Section 6.5.[44]

Appendix Figure A.9 shows the inverted local welfare $(\omega_i^u)^\eta$ and $(\omega_i^s)^\eta$ and local price composite $p_i^{\sigma-1}$. Java has higher welfare composite and lower price composite than outer islands. Major cities, such as Jakarta and Surabaya in Java, Makassar in Sulawesi, Balikpapan in Kalimantan also show higher welfare composite and lower price composite. Intuitively, these are places where there are lots of varieties produced.

Step 2: Get $\bar{u}_i^e$ and $\bar{A}_i$ Using Calibrated Parameter Values With calibration of $\{\sigma, \alpha, \eta^s, \eta^u\}$ from the literature, I can further recover $\bar{A}_i$ and α_i^e from the definition of the inverted price and welfare composites as follows,

$$\ln(p_i^{\sigma-1}) = (\sigma-1)\ln\left[(\theta_i^s)^\rho (w_i^s)^{1-\rho} + (\theta_i^u)^\rho (w_i^u)^{1-\rho}\right]^{\frac{1}{1-\rho}} - \alpha(\sigma-1)\ln L_i - (\sigma-1)\ln \bar{A}_i \tag{20}$$

$$\ln(\omega_i^s)^\eta = \eta \ln w_i^s - \eta \ln P_i + \eta\beta \ln L_i + \eta \ln \bar{u}_i^s \tag{21}$$

$$\ln(\omega_i^u)^\eta = \eta \ln w_i^u - \eta \ln P_i + \eta\beta \ln L_i + \eta \ln \bar{u}_i^u \tag{22}$$

where $\{p_i^{\sigma-1}, (\omega_i^s)^\eta, P_i\}$ are recovered from the inversion in the first step, and $\{w_i^e, L_i\}$ are observed. Then $\{\bar{A}_i, \bar{u}_i^s, \bar{u}_i^u\}$ are residuals. Appendix Figure A.10 shows the inverted local fundamentals $\bar{u}_i^e$ and $\bar{A}_i$.

To check the validity of the model, I correlate the recovered local amenity from the model with the observed local amenity from external data, PODES 2011, that are not used in the previous estimation.[45] Table 6 shows the correlation between the inverted $\bar{u}_i^e$ and observed amenity, including retail amenities, transportation, education and medical facilities. The estimated amenity is higher in locations with better accessibility to retail, education, and medical facilities and higher-quality road conditions.

8 Counterfactuals

This section uses the quantitative spatial model developed in the last section to quantify the productivity and welfare implications of reducing linguistic distance.[46] Linguistic distance, as a component of migration and trade frictions, restricts the opportunity sets available to workers and firms, creating welfare loss. Since unskilled workers are more susceptible to linguistic migration barriers, it is harder for them to access productive and desirable places, so linguistic distance also intensifies inequality. However, linguistic distance can be altered. On the one hand, it can be altered by policies such as promoting the national language, *Bahasa Indonesia*. On the other hand, it naturally evolves with the dynamics of demographics and migration. As Indonesia rapidly urbanizes, people from diverse ethnic backgrounds

[44] See Appendix D.2 for formal expressions.

[45] PODES (Village potential Census) is a census of Indonesian villages conducted approximately every three years by BPS (Indonesias statistical agency). It collects detailed information from community informants about community characteristics, such as demographics, geography, as well as social and economic infrastructure.

[46] The definition of welfare and productivity is specified in Appendix D.5

migrate to economic opportunities, changing the linguistic distance between locations by changing local ethnic composition. In the first counterfactual in this section, I quantify the welfare implications of reductions in linguistic distances by extrapolating the historical trend of demographic changes. In the second set of counterfactuals, I impose a 50% reduction in linguistic distance for each pair of locations to demonstrate the mechanisms. In the last set of counterfactuals, I keep linguistic distances constant and calculate the local employment elasticity of each location to illustrate how linguistic barriers can determine the responsiveness of the local economy to changing economic environment.

There are two caveats in this analysis. First, I only consider the economic welfare determined by real wage, amenity, and migration cost and leave out other social aspects of welfare, such as nation building and cultural preservation. Second, I assume changes in linguistic distances only affect migration and trade frictions directly in this analysis.

8.1 Overall Reduction in Linguistic Distances

From 1995 to 2000, linguistic distance has decreased by 0.6% on average due to changes in ethnic composition. With this rate, linguistic distances would decrease by 6.2% in 50 years.[47] In this counterfactual, I assume a nationwide reduction in linguistic distances equal to 6.2% for demonstration.

Overall Impact Column 1 in Table 7 presents the impacts of a common 6.2% reduction in the linguistic distance on aggregate welfare, GDP, and inequality. The reduction of linguistic distances induces a 0.13% welfare gain (row 1) and boosts production by 0.29% (row 2). Besides the overall effects, it substantially reduces the skill inequality by 0.86%. However, it also intensifies the regional inequality, defined as the welfare gap across locations, by 0.28% (see Appendix D.5 for details of the calculations).

To get a sense of the magnitudes of those effects, I conduct another counterfactual where the geographic barriers are reduced by a comparable 6.2%. Such reduction in geographic barriers (physical distances) can be achieved by policies like building transportation infrastructure, which has been widely used to promote labor market integration. While welfare gain from linguistic distance reduction is about 60% of the welfare gain from the equally 6.2% decrease in geographic barriers, reducing linguistic distances and geographic distances induces similar GDP increases. Interestingly, reducing linguistic barriers has a much more significant impact on skill inequality - it substantially decreases skill inequality by 0.86% compared to the 0.003% decrease by reducing geographic barriers. This is because linguistic distances play a more significant role in driving the difference in mobility between skills than geographic distance, as shown in the gravity estimates in Table 3.[48]

Relocation of Labor Figure 5 maps the changes in population (normalized) and skill composition in each district when linguistic distances are reduced, where locations experiencing increases are colored in blue and locations experiencing decreases are colored in red. Panel A shows that central Java, the most linguistically connected to other locations, largely losses population. With the reduction in linguistic distances, they get less attractive since people prefer modest level of cultural heterogeneity to pure ho-

[47] I put people back in their 1995 district and calculate the linguistic distance based on ethnic composition in 1995. Appendix Figure A.8 shows the distribution of percentage change in linguistic distance for all location pairs. More than 87% of location pairs experience drop in linguistic distance.

[48] This is also consistent with Bauernschuster et al. (2014)' s finding in Germany, where they argue that it is easier for more educated to cross language barriers and adapt to different cultural environment.

mogeneity, which is demonstrated empirically in the reduced form Section C. As migration barriers are reduced in previously isolated areas, people get access to the most productive or desirable places, such as Jakarta, Bandung, Bali, East Kalimantan, and Riau, so there are large migrant inflows in those places. Panel B further shows that places that gain the most population also experience the largest decreases in skill composition, implying that the migrant inflows are driven by unskilled workers.

Mechanisms What determines the aggregate and distributional welfare effects of the reduction in linguistic distance? To help demonstrate this, I decompose the welfare gain, starting with a simplified version of the model and adding its ingredients stepwise to isolate the corresponding impacts. The results are shown in Table 8. Column 1 shows the aggregate welfare changes and Column 2 shows the changes in skill inequality by the reduction in linguistic distances.

Row 1 presents welfare gains using the framework of Allen and Arkolakis (2018) with no skill divide of labor, and does not assume non-monotonic effects of linguistic distance on migration. I re-estimate a migration gravity similar to specification (2) without the quadratic term of linguistic distance for the whole population. The result is shown in column 2 of Appendix Table A.13. Linguistic distance has an overall detrimental effect on migration with a magnitude much smaller than the median marginal effects (-0.99) estimated using the non-monotonic specification (Panel B of Table 2). Appendix Figure A.11 plots the expectation of migration propensity (conditional on all control variables) at each level of linguistic distance using monotonic and non-monotonic specifications. The slope of the expected migration propensity from the monotonic specification, though negative, is much flatter than the slope of the negative section of the one from non-monotonic specification.

Row 2 shows results still considering the basic framework of Allen and Arkolakis (2018) but assuming non-monotonic effects of linguistic distance. Comparing Row 2 and 1, ignoring the non-monotonicity substantially underestimates the welfare gain from the reduction of linguistic distance by more than a half because although the overall effect of linguistic distance estimated in Appendix Table A.13 is negative, it substantially underestimates the magnitude of the detrimental effect, especially for linguistically remote locations.

Row 3 shows results using a model with skill complementarity in production and where both skill groups suffer the same migration cost.[49] Comparing Row 3 and Row 2, the welfare gain decreases from 0.20% predicted by the simplest model without skill division to 0.19% predicted by a model with skill complementarity. In a model with skill complementarity, welfare is determined by efficiency in the allocation of overall labor and in the allocation of skills. While reducing linguistic distance can allocate labor more efficiently, it may not necessary allocate both skilled and unskilled labor more efficiently simultaneously.

Row 4 shows the full model with both skill complementarity and skill differences in migration cost. Although all models predict that unskilled workers benefit more and inequality decreases, comparing Row 4 and Row 3, the magnitude of the inequality reduction more than double from 0.360% to 0.856% when the skill difference in migration elasticity is incorporated. Since skilled workers are more immune to migration costs, they are already more efficiently allocated before reducing linguistic distance. In contrast, unskilled workers bear a greater incidence of migration frictions. Therefore when migration cost is reduced, the unskilled benefit more than the skilled, and inequality further falls.

[49] I set the migration cost to skilled migration cost.

Since I allow linguistic distance to enter both migration and trade costs, I also consider a model where linguistic distance only enters migration barriers instead of trade costs in Row 5, which allows decomposition of the welfare gains that can be directly accounted to changes in migration barriers. The welfare is increased by 0.09%, and the inequality is reduced by 0.86%, suggesting that about 66% of the welfare gain and almost all equity improvement in the full model (Row 3) is accounted for by the reduction in migration barriers.

Further Reduction in Linguistic/Geographic Barriers This 6.2% is a lower bound of reduction in linguistic distances as it purely results from changes in ethnic composition by migration. However, other changes in the socio-economic environment in Indonesia also reduce cultural barriers. For example, the advancement of information technology and the widespread of social media has allowed people from diverse backgrounds to communicate more freely. In addition, as discussed in Section 2, the Indonesian government has put substantial efforts into promoting national language and nation building. In Appendix Figure A.12, I run counterfactuals with further reductions in linguistic distances and with reductions in geographic barriers for comparison. The results demonstrate sizable welfare gains from reducing linguistic distances, although with a magnitude of gain smaller than removing geographic barriers.

8.2 Pairwise Reduction in Linguistic Distances

Although it seems unrealistic to reduce cultural barriers nationwide, it is more empirically feasible to promote cultural exchange between certain locations. How do those pairwise reductions in linguistic distance affect welfare? What determines their welfare implications? To examine these questions, I conduct 27,028 counterfactuals in each of which I reduce the linguistic distance between a particular pair of locations by 50%. I find substantial heterogeneity in the predicted welfare changes of these local reductions in linguistic distances, which vary from welfare loss by 0.002%, by linking Denpasar to Belu, to welfare gain by almost 0.007%, by connecting Batam to Padang. To examine what determines the welfare implications of reducing linguistic distance, I regress the predicted aggregate welfare changes of the 27,028 pairwise counterfactuals on a range of observed or recovered locational variables,

$$\Delta W_{ij} = \beta_1 Ldist0_{ij} + \beta_2 Ldist0_{ij}^2 + X_{ij}\theta + \delta_i + \delta_j + \epsilon_{ij} \tag{23}$$

Where ΔW_{ij} is the percentage change in aggregate welfare from reducing linguistic distance between location i and j, $Ldist0_{ij}$ is the initial level of linguistic distance between location i and j, and X_{ij} is a vector of minimum of origin and destination characteristics, including initial population, exogenous local productivity, and exogenous local amenity. δ_i and δ_j are origin and destination fixed effects.

Table 9 shows the results. In column 1, I regress the aggregate welfare changes on the initial linguistic distance. Consistent with the inverted U-shaped gravity equation estimates, reducing linguistic distances for initially close locations hurts welfare as it deprives cultural heterogeneity that people value. Nevertheless, for linguistically remote locations, the larger the initial linguistic distance, the more welfare gain it generates. In Column 2, I include the geographic distance. Conditional on initial linguistic distance, reducing linguistic barriers for more geographically approximate location induces more welfare. In columns 3-5, I further include the minimum of local productivity, amenity, and initial population

between origin and destination by step. The aggregate welfare positively correlates with those characteristics of locations that directly experience the reduction in linguistic distance. In sum, the full regression result shown in column 5 delivers essential implications for policies to remove cultural barriers —— targeting different places generates heterogeneous welfare effects. On the one hand, promoting cultural exchange between linguistically remote and geographically close locations can generate more benefits. On the other hand, it is more efficient to target economically significant locations.

Appendix Table A.14 shows results of similar regressions for skill inequality. Besides relating the inequality effects to initial linguistic distance, local fundamentals, and population, in Column 5, I further examine the role of local skill intensity. Reducing linguistic distance to locations with higher demand for skills benefits skilled workers more, thus increasing inequality, suggesting that the inequality effects of reducing linguistic distance depend on which group is connected to locations that demand their specific skills the most.[50]

8.3 Linguistic Distance and Local Employment Elasticity

Although Indonesia has made extensive efforts to promote language homogeneity and intergroup interactions, linguistic disparities still form significant barriers to internal migration, as I have demonstrated in the previous sections. In contrast to the slow evolution of language and culture, the Indonesian economy is undergoing rapid transformation against globalization. How do these slowly-moving linguistic and cultural frictions affect the responsiveness of the local economy to the rapidly evolving economic environment? To help to understand this, following Monte et al. (2018), I compute 233 counterfactual exercises where I shock each location with a 5% productivity increase, holding productivity levels in other locations and other parameters constant. In each counterfactual, I calculate the local employment elasticity, denoted as E_i, for each location to measure how responsive local employment is to local demand shocks. The local employment elasticity is defined as follows,

$$E_i = \frac{\%\Delta L_i}{\%\Delta \bar{A}_i} \tag{24}$$

where L_i is the population and $\bar{A}_i$ is the exogenous productivity level in location i.[51] Local employment elasticity is of great policy interest since it determines the extent to which environmental changes (e.g., climate shocks) and placed-based policies (e.g., infrastructure construction and regional development programs) can affect the economy.[52] With a positive labor demand shock, a location with high local employment elasticity can quickly adjust by attracting more labor. In contrast, a place with a low local employment elasticity is more stagnant with less inflow of employment.

Panel A in Figure 6 shows the estimated kernel density for the distribution of the local employment elasticity, which is determined by general equilibrium, with respect to the productivity shock across these treated locations, with the confidence interval shown in gray shading. I plot the kernel density for both skilled (dash line) and unskilled employment (solid line). First, the mean of local employment elasticities is 0.13 for unskilled and 0.23 for skilled employment. Both are well below one due to migration

[50] This is consistent with Tsivanidis (2019)'s discussion about the inequality effects of public transit infrastructure in Bogota

[51] In this specific setting, $\%\Delta\bar{A}_i = 5$.

[52] Monte et al. (2018) has a nice discussion about the importance of local employment elasticity. They show that commuting openness of the local labor market increases local employment elasticities.

frictions — workers can not perfectly respond to the local demand shock. These local employment elasticities are also way below the estimation in Monte et al. (2018) where they assume free mobility. Second, skilled workers, who are less affected by a list of migration barriers, have a larger local employment elasticity, suggesting that skilled employment can adjust to economic shocks more effectively than unskilled employment. Third, there are significant variations of local employment elasticities across locations for skilled and unskilled workers.

Understanding the variations of local employment elasticities is important as it directly biases the evaluation of the true effects of any place-based productivity shocks. To detect the extent to which linguistic and geographic frictions can explain the variations, I construct two market access measures, for each location i,

$$MA_i = \sum_{j \neq i} L_j * Dist_{ij} \tag{25}$$

where L_j is the population in i, and $Dist_{ij}$ is the geographic or linguistic distance between i and j. Depending on the type of distance in use, I refer to the market access MA_i as geographic market access, denoted as GMA_i, or linguistic market access, denoted as LMA_i. This measure can be intuitively understood as how connective a location is to all other locations in terms of attracting labor. Appendix Figure A.7 shows the spatial distribution of these market access measures. Java, which is closer both geographically and culturally to other densely populated places, shows both higher geographic market access GMA_i and higher linguistic market access LMA_i.

Panel B in Figure 6 plots the calculated local employment elasticity against the geographic market access while assigning each dot (corresponding to a location) to blue, orange, and red if its linguistic market access is low, middle, or high, respectively.[53] In general, local employment elasticity increases with geographic market access. However, this tendency shows diverging patterns for locations with different linguistic market access. Linguistically accessible places (high linguistic market access), colored in red, are clear outliers with lower local employment elasticities on average. In contrast, locations with intermediate geographic and linguistic market accesses (colored in orange) have the highest local employment elasticities. This is also shown in Appendix Table A.15, which summarizes the average local elasticities for locations with different levels of geographic and linguistic market access.

This set of results demonstrate the role migration frictions, especially linguistic distance, play in determining the economic effects of local productivity shocks by local employment elasticities, which have further welfare consequences. For example, if the government considers providing a tax break for firms in one of the two locations: Central Java, the most linguistically accessible and has relatively low local employment elasticity; or North Sumatra, which has an intermediate linguistic market access and has relatively high local employment elasticity. Residents in Central Java will benefit more from the tax break since labor is relatively less mobile there, and local workers are able to capture more of the economic rent generated by the shock. However, other locations are less affected by this tax break since they have less access to it. In contrast, a tax break in North Sumatra benefits less to (initial) residents there but more to residents in other locations.[54] Therefore, it is crucial for policymakers to take cultural

[53] Low, middle and high LMA_i are divided by the 50th, 75th percentiles of the distribution of LMA_i .
[54] See Moretti (2010) for detailed discussions.

barriers into account in policy design.

9 Conclusion

This paper finds significant effects of cultural disparities on internal migration in Indonesia, one of the world's most populous and diverse countries. Cultural distance, proxied by linguistic distance, has inverted U-shaped effects on migration propensity, reflecting the trade-off between the benefits and costs of living in a different cultural environment. While linguistic distance deters migration flows between Jakarta and culturally distant Makarssa, it encourages migration between Jakarta and linguistically similar central Java. This pattern has been found in international migration (Krieger et al., 2018), and, to my knowledge, is documented for the first time in internal migration.

Building on the reduced-form results, I further quantify the welfare and distributional implications of reducing cultural barriers. First, I develop a quantitative spatial model incorporating asymmetric migration elasticities to cultural barriers of different skill groups. Second, I use the model to quantify the welfare effects of a simulated reduction in linguistic distances by extrapolating historical migration trends. I find a sizable overall welfare gain from this reduction in linguistic distance, as people from initially isolated areas get access to more desirable and productive places. I also find that unskilled workers benefit more than skilled workers, originating from the asymmetry in migration elasticities to cultural barriers between skills.

The economic consequences of cultural and ethnic disparities have been a widely concerned topic in the developing world. For example, Easterly and Levine (1997) finds that Africa's growth tragedy can be significantly attributed to its ethnic fractionalization. This paper contributes to understanding these economic consequences by demonstrating that cultural frictions create substantial labor misallocation. These distortion effects are significant even in Indonesia, with its long history of nation building policies. One can expect even more significant detrimental effects in countries with more intense ethnic tensions, such as Rwanda and Sudan, and other highly multi-linguistic contexts, like India and South Africa. Interestingly, the inverted U-shaped effects of cultural disparity found in this paper show an analogy to Ashraf and Galor (2013), where they find hump-shaped effects of genetic diversity, which is highly correlated to cultural diversity, on comparative economic development across countries. This paper points to the possibility of labor market integration being one of the mechanisms underlying their findings.

This paper only captures a general composite of cultural disparity. A more detailed decomposition of linguistic distance, combined with proper identification strategies, will offer more insights into disassembling the underlying mechanisms. The model in this paper is static, but linguistic distance is dynamically determined by migration. Future work would extend it to a dynamic setting. Moreover, I consider only one aspect of culture — inter-regional cultural barriers — and assume they directly affect migration and trade costs. However, as demonstrated in voluminous development and urban literature, other aspects of culture, such as cultural diversity, can profoundly impact local productivity and consumption amenities. A potential extension to this paper is to model both linguistic distance and cultural diversity and allow them to affect a broader range of variables, including migration and trade frictions, local productivity, and amenity.

References

ADSERA, A. AND M. PYTLIKOVA (2015): "The role of language in shaping international migration," *The Economic Journal*, 125, F49–F81.

ALESINA, A., A. DEVLEESCHAUWER, W. EASTERLY, S. KURLAT, AND R. WACZIARG (2003): "Fractionalization," *Journal of Economic growth*, 8, 155–194.

ALLEN, T. AND C. ARKOLAKIS (2014): "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 129, 1085–1140.

——— (2018): "Modern spatial economics: a primer," in *World Trade Evolution*, Routledge, 435–472.

ALLEN, T. AND D. DONALDSON (2020): "Persistence and Path Dependence in the Spatial Economy," Tech. rep., National Bureau of Economic Research.

ANANTA, A., E. N. ARIFIN, M. S. HASBULLAH, N. B. HANDAYANI, AND A. PRAMONO (2014a): *Demography of Indonesia's Ethnicity*, ISEAS–Yusof Ishak Institute.

——— (2014b): *THE NEW CLASSIFICATION: Uncovering Diversity*, ISEAS–Yusof Ishak Institute, 39–67.

ANDERSON, J. E. AND E. VAN WINCOOP (2003): "Gravity with gravitas: A solution to the border puzzle," *American economic review*, 93, 170–192.

ARMINGTON, P. S. (1969): "A theory of demand for products distinguished by place of production," *Staff Papers*, 16, 159–178.

ASHRAF, N., N. BAU, N. NUNN, AND A. VOENA (2020): "Bride price and female education," *Journal of Political Economy*, 128, 591–641.

ASHRAF, Q. AND O. GALOR (2013): "The'Out of Africa'hypothesis, human genetic diversity, and comparative economic development," *American Economic Review*, 103, 1–46.

AUWALIN, I. (2020): "Ethnic identity and internal migration decision in Indonesia," *Journal of Ethnic and Migration Studies*, 46, 2841–2861.

BARRIOS, S., L. BERTINELLI, AND E. STROBL (2006): "Climatic change and rural–urban migration: The case of sub-Saharan Africa," *Journal of Urban Economics*, 60, 357–371.

BASELER, T. (2019): "Hidden income and the perceived returns to migration: Experimental evidence from kenya," *Unpublished Working Paper, University of Rochester*.

BAUERNSCHUSTER, S., O. FALCK, S. HEBLICH, J. SUEDEKUM, AND A. LAMELI (2014): "Why are educated and risk-loving persons more mobile across regions?" *Journal of Economic Behavior & Organization*, 98, 56–69.

BAZZI, S., A. GADUH, A. D. ROTHENBERG, AND M. WONG (2016): "Skill transferability, migration, and development: Evidence from population resettlement in Indonesia," *American Economic Review*, 106, 2658–98.

——— (2019): "Unity in diversity? How intergroup contact can foster nation building," *American Economic Review*, 109, 3978–4025.

BEINE, M., F. DOCQUIER, AND Ç. ÖZDEN (2011): "Diasporas," *Journal of Development Economics*, 95, 30–41.

BORJAS, G. J. (1987): "Self-selection and the earnings of immigrants," *The American economic review*, 531–553.

BRYAN, G., S. CHOWDHURY, AND A. M. MOBARAK (2014): "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh," *Econometrica*, 82, 1671–1748.

BRYAN, G. AND M. MORTEN (2019): "The aggregate productivity effects of internal migration: Evidence from Indonesia," *Journal of Political Economy*, 127, 2229–2268.

BURCHFIELD, M., H. G. OVERMAN, D. PUGA, AND M. A. TURNER (2006): "Causes of sprawl: A portrait from space," *The Quarterly Journal of Economics*, 121, 587–633.

CAMERON, A. C., J. B. GELBACH, AND D. L. MILLER (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

CHAUVIN, J. P., E. GLAESER, Y. MA, AND K. TOBIO (2017): "What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States," *Journal of Urban Economics*, 98, 17–49.

CHEN, C. (2017): "Untitled land, occupational choice, and agricultural productivity," *American Economic Journal: Macroeconomics*, 9, 91–121.

CHIQUIAR, D. AND G. H. HANSON (2005): "International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States," *Journal of political Economy*, 113, 239–281.

CIVELLI, A., A. GADUH, A. D. ROTHENBERG, AND Y. WANG (2021): "Urban Sprawl and Social Capital: Evidence from Indonesian Cities," Working Paper.

COLLIER, P. (2017): "Culture, politics, and economic development," *Annual Review of Political Science*, 20, 111–125.

COMBES, P.-P. AND L. GOBILLON (2015): "The empirics of agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 5, 247–348.

CREANZA, N., O. KOLODNY, AND M. W. FELDMAN (2017): "Cultural evolutionary theory: How culture evolves and why it matters," *Proceedings of the National Academy of Sciences*, 114, 7782–7789.

DE JANVRY, A., K. EMERICK, M. GONZALEZ-NAVARRO, AND E. SADOULET (2015): "Delinking land rights from land use: Certification and migration in Mexico," *American Economic Review*, 105, 3125–49.

DIAMOND, J. M. (1998): *Guns, Germs, and Steel: the Fates of Human Societies*, New York: W. W. Norton & Co.

DIAMOND, R. (2016): "The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000," *American Economic Review*, 106, 479–524.

EASTERLY, W. AND R. LEVINE (1997): "Africa's growth tragedy: policies and ethnic divisions," *The quarterly journal of economics*, 1203–1250.

EGGER, P. H. AND A. LASSMANN (2012): "The language effect in international trade: A meta-analysis," *Economics Letters*, 116, 221–224.

ESTEBAN, J., L. MAYORAL, AND D. RAY (2012): "Ethnicity and conflict: An empirical study," *American Economic Review*, 102, 1310–42.

FALCK, O., S. HEBLICH, A. LAMELI, AND J. SÜDEKUM (2012): "Dialects, cultural identity, and economic exchange," *Journal of urban economics*, 72, 225–239.

FAN, J. (2019): "Internal geography, labor mobility, and the distributional impacts of trade," *American Economic Journal: Macroeconomics*, 11, 252–88.

FARRÉ, L. AND F. FASANI (2013): "Media exposure and internal migration—Evidence from Indonesia," *Journal of Development Economics*, 102, 48–61.

FARROKHI, F. AND D. JINKINS (2019): "Wage inequality and the location of cities," *Journal of Urban Economics*, 111, 76–92.

FEARON, J. D. (2003): "Ethnic and cultural diversity by country," *Journal of economic growth*, 8, 195–222.

FEENSTRA, R. C., P. LUCK, M. OBSTFELD, AND K. N. RUSS (2018): "In search of the Armington elasticity," *Review of Economics and Statistics*, 100, 135–150.

FENSKE, J. AND N. KALA (2021): "Linguistic Distance and Market Integration in India," *The Journal of Economic History*, 81, 1–39.

GINSBURGH, V. AND S. WEBER (2020): "The economics of language," *Journal of Economic Literature*, 58, 348–404.

GLAESER, E. L., J. KOLKO, AND A. SAIZ (2001): "Consumer city," *Journal of economic geography*, 1, 27–50.

GOLLIN, D., D. LAGAKOS, AND M. E. WAUGH (2014): "The agricultural productivity gap," *The Quarterly Journal of Economics*, 129, 939–993.

GOTTLIEB, C. AND J. GROBOVSEK (2019): "Communal Land and Agricultural Productivity," *Journal of Development Economics*.

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1984): "Pseudo maximum likelihood methods: Applications to Poisson models," *Econometrica: Journal of the Econometric Society*, 701–720.

GUREVICH, T., P. HERMAN, F. TOUBAL, AND Y. YOTOV (2021): "One nation, one language? domestic language diversity, trade and welfare," .

HEAD, K. AND T. MAYER (2014): "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of international economics*, Elsevier, vol. 4, 131–195.

KATZ, L. F. AND K. M. MURPHY (1992): "Changes in relative wages, 1963–1987: supply and demand factors," *The quarterly journal of economics*, 107, 35–78.

KHANNA, G., W. LIANG, A. M. MOBARAK, AND R. SONG (2021): "The productivity consequences of pollution-induced migration in China," *NBER Working Paper*.

KHANNA, G. AND N. MORALES (2017): "The IT Boom and Other Unintended Consequences of Chasing the American Dream," *Center for Global Development Working Paper*.

KLEEMANS, M. AND J. MAGRUDER (2018): "Labour market responses to immigration: Evidence from internal migration driven by weather shocks," *The Economic Journal*, 128, 2032–2065.

KONE, Z. L., M. Y. LIU, A. MATTOO, C. OZDEN, AND S. SHARMA (2018): "Internal borders and migration in India," *Journal of Economic Geography*, 18, 729–759.

KRIEGER, T. AND T. LANGE (2010): "Education policy and tax competition with imperfect student and labor mobility," *International Tax and Public Finance*, 17, 587–606.

KRIEGER, T., L. RENNER, AND J. RUHOSE (2018): "Long-term relatedness between countries and international migrant selection," *Journal of International Economics*, 113, 35–54.

LAGAKOS, D. (2020): "Urban-rural gaps in the developing world: Does internal migration offer opportunities?" *Journal of Economic perspectives*, 34, 174–92.

LI, N. (2018): "The Long-Term Consequences of Cultural Distance on Migration: Historical Evidence from China," *Australian Economic History Review*, 58, 2–35.

LIN, W. AND J. M. WOOLDRIDGE (2019): "Testing and correcting for endogeneity in nonlinear unobserved effects models," in *Panel Data Econometrics*, Elsevier, 21–43.

MONTE, F., S. J. REDDING, AND E. ROSSI-HANSBERG (2018): "Commuting, migration, and local employment elasticities," *American Economic Review*, 108, 3855–90.

MORETTI, E. (2010): "Local labor markets," Tech. rep., National Bureau of Economic Research.

MORTEN, M. (2019): "Temporary migration and endogenous risk sharing in village india," *Journal of Political Economy*, 127, 1–46.

MUNSHI, K. AND M. ROSENZWEIG (2016): "Networks and misallocation: Insurance, migration, and the rural-urban wage gap," *American Economic Review*, 106, 46–98.

OTTAVIANO, G. I. AND G. PERI (2006): "The economic value of cultural diversity: evidence from US cities," *Journal of Economic geography*, 6, 9–44.

PEPINSKY, T. B. (2016): "Colonial migration and the origins of governance: Theory and evidence from Java," *Comparative Political Studies*, 49, 1201–1237.

PISANI, E. (2014): *Indonesia, etc.: Exploring the improbable nation*, WW Norton & Company.

PORCHER, C. (2019): "Migration with costly information," *Princeton University Mimeo.*

REDDING, S. J. (2016): "Goods trade, factor mobility and welfare," *Journal of International Economics*, 101, 148–167.

ROSENTHAL, S. S. AND W. C. STRANGE (2004): "Evidence on the nature and sources of agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2119–2171.

ROTHENBERG, A. D. AND D. TEMENGGUNG (2019): "Place-Based Policies in Indonesia," .

SILVA, J. S. AND S. TENREYRO (2006): "The log of gravity," *The Review of Economics and statistics*, 88, 641–658.

SIMONOVSKA, I. AND M. E. WAUGH (2014): "The elasticity of trade: Estimates and evidence," *Journal of international Economics*, 92, 34–50.

TOMBE, T. AND X. ZHU (2019): "Trade, migration, and productivity: A quantitative analysis of china," *American Economic Review*, 109, 1843–72.

TRAX, M., S. BRUNOW, AND J. SUEDEKUM (2015): "Cultural diversity and plant-level productivity," *Regional Science and Urban Economics*, 53, 85–96.

TSIVANIDIS, N. (2019): "Evaluating the impact of urban transit infrastructure: evidence from Bogotá's TransMilenio," *Unpublished manuscript*.

VAN KLINKEN, G. (2003): "Ethnicity in Indonesia," in *Ethnicity in Asia*, ed. by C. Mackerras, RoutledgeCurzon, chap. 4, 64–87.

WOOLDRIDGE, J. M. (2015): "Control function methods in applied econometrics," *Journal of Human Resources*, 50, 420–445.

YOTOV, Y. V., R. PIERMARTINI, J.-A. MONTEIRO, AND M. LARCH (2016): *An advanced guide to trade policy analysis: The structural gravity model*, World Trade Organization Geneva.

Table 1: Correlation between Linguistic Distance and Customs

	Independent Var.: Linguistic Distance					
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:						
Similar Subsistence economy	-0.171***					
	(0.007)					
Similar Inheritance rule for land		-0.288***				
		(0.007)				
Similar Descent: patrilineal/matrilineal/bilateral			-0.247***			
			(0.007)			
Similar Post-marriage Residence: Patrilocal/Matrilocal/Neolocal				-0.078***		
				(0.007)		
Similar Settlement patterns					-0.204***	
					(0.007)	
Religious Distance						0.785***
						(0.004)
N of Ethnic Pairs	44,100	40,320	44,100	44,100	44,100	
N of Region Pairs						87,025
Adjusted R-squared	0.013	0.034	0.029	0.003	0.024	0.156

Notes: This table reports the correlation of linguistic distance and cultural practices. The dependent variables are indicators equal to one when the two ethnic groups share the same traditional practice. Subsistence economy refers to the main type of production for subsistence, including gathering, intensive agriculture, extensive agriculture and mixture of different type; Inheritance rule for land include matrilineal heirs, patrilineal heirs and children; Settlement pattern include dispersed homesteads, hamlets, semi-sedentary, villages/towns, and complex permanent. Robust standard errors, two-way clustered at the district level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 2: Linguistic Distance and Migration

	PPML			PPML with IV
	(1)	(2)	(3)	(4)
linguistic distance, 1995	-0.756***	0.600***	1.175***	0.766***
	(0.040)	(0.143)	(0.136)	(0.170)
linguistic distance, 1995, squared		-0.326***	-0.235***	-0.236***
		(0.035)	(0.033)	(0.035)
log(geographic distance)			-1.270***	-1.020***
			(0.053)	(0.083)
Cross Islands (1 0)			-0.136	-0.034
			(0.110)	(0.112)
Absolute difference in latitute			0.054***	0.030**
			(0.015)	(0.015)
Absolute difference in longtitute			-0.086***	-0.093***
			(0.012)	(0.012)
Agroclimatic Similarity			0.034***	0.027***
			(0.009)	(0.007)
Hometown Bias	6.785***	7.572***	1.708***	1.974***
	(0.115)	(0.122)	(0.216)	(0.389)
Optimal Linguistic Dist.		0.919	2.496	1.622
Median Distance		3.746	3.746	3.746
Mean Distance		3.493	3.493	3.493
N of region pairs	54,289	54,289	54,289	54,289
N Clusters	233	233	233	233
Pseudo R-squared	0.793	0.793	0.799	0.799
Kleibergen-Paap Wald Rank F Stat				60.425
Under Id. Test (KP Rank LM Stat)				41.359
p-Value				0.000
Marginal Effects for Linguistic Dist. at				
5th percentile				0.203**
				(0.085)
50th percentile				-0.999***
				(0.159)
90th percentile				-1.220***
				(0.195)
Origin Kabu FE	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2). Column 1-3 report PPML estimates. For Column 3, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 3: Cultural Distance and Migration Selection by Skill

	Migration Propensity		Migration Selection
	Skilled	Unskilled	
	(1)	(2)	(3)
linguistic distance, 1995	0.312**	0.816***	-0.679***
	(0.150)	(0.172)	(0.107)
linguistic distance, 1995, squared	-0.144***	-0.240***	0.119***
	(0.034)	(0.035)	(0.026)
log(distance)	-1.030***	-1.040***	0.527***
	(0.068)	(0.080)	(0.042)
Cross Islands (1 0)	-0.422***	-0.321***	0.023
	(0.079)	(0.121)	(0.045)
Absolute difference in latitude	0.047***	0.051***	-0.043***
	(0.013)	(0.015)	(0.005)
Absolute difference in longitude	-0.034***	-0.084***	0.030***
	(0.009)	(0.014)	(0.005)
Agroclimatic Similarity	0.035***	0.023***	0.003
	(0.007)	(0.007)	(0.003)
Hometown Bias	0.950***	2.059***	1.091***
	(0.299)	(0.377)	(0.170)
Optimal Linguistic Dist.	1.082	1.696	2.850
Median Distance	3.746	3.746	3.746
Mean Distance	3.493	3.493	3.493
N of region pairs	54,289	54,289	54,289
N Clusters	233	233	233
Pseudo R-squared	0.778	0.801	
Adjusted R-squared			0.277
Kleibergen-Paap Wald Rank *F* Stat	63.381	63.381	63.381
Under Id. Test (KP Rank LM Stat)	43.579	43.579	43.579
p-Value	0.000	0.000	0.000
Origin Kabu FE	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) by skill. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 4: Cultural Distance and Migration, by Skill, Marginal Effects by Quantiles

	Propensity			Selection
	Overall	Skilled	Unskilled	
	(1)	(2)	(3)	(4)
1th percentile	0.718***	0.283*	0.767***	-0.655***
	(0.176)	(0.147)	(0.177)	(0.102)
5th percentile	0.123	-0.081	0.161*	-0.355***
	(0.089)	(0.081)	(0.090)	(0.045)
50th percentile	-1.003***	-0.770***	-0.986***	0.214**
	(0.145)	(0.128)	(0.145)	(0.097)
90th percentile	-1.210***	-0.896***	-1.197***	0.318***
	(0.177)	(0.154)	(0.177)	(0.119)
Origin Kabu FE	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes

Notes: The table shows the marginal effects at each quantiles. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 5: Parameterization of the Model

Parameter	Baseline Value	Robustness Value Range	Literature
Agglomeration Parameters α	0.85	$[0.02, 0.85]$	Rosenthal and Strange 2004; Chauvin et al. 2017
Congestion Parameters β	-0.04	$[-0.05, 0]$	Combes and Gobillon 2015
Skill Elasticity of Substitution ρ	1.24	$[1.24, 1.8]$	Khanna et al. 2021; Khanna and Morales 2017; Katz and Murphy 1992
Armington Elasticity of Substitution σ	6	6	Simonovska and Waugh 2014; Feenstra et al. 2018
Dispersion Parameter η_s	1.6	$[1.7, 2]$	Tombe and Zhu 2019; Khanna et al. 2021 Fan 2019
Trade cost elasticity τ_{ij}^g	-1	-1	Head and Mayer 2014
Trade cost elasticity τ_{ij}^l	-0.5	-0.5	Egger and Lassmann 2012

Table 6: Inverted Amenity and Observed Amenities

	Hotels per 1,000 Residents (1)	Shopping Mall Accessibility (2)	Road Condition (3)	Primary/Kindergarten Edu. Accessibility (4)	Secondary/Junior Edu. Accessibility (5)	Medical Facilities Accessibility (6)
Inverted Amenity for Skilled	0.239**	0.128**	0.356***	0.231***	0.024	0.027
	(0.117)	(0.058)	(0.067)	(0.078)	(0.039)	(0.034)
Inverted Amenity for Unskilled	0.058	0.706***	0.067	0.418***	0.370***	0.259***
	(0.171)	(0.066)	(0.099)	(0.107)	(0.048)	(0.044)

Notes: This table shows the correlation between the recovered amenities and observed amenities. Each column shows result for a certain type of amenity. The observed amenity data is from PODES 2011. Standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 7: Aggregate and Distributional Impacts of Reducing Distance by 6.2%

	Reducing Linguistic Distance (1)	Reducing Geographic Distance (2)
Aggregate Welfare	0.134	0.240
Aggregate Production	0.293	0.271
Skill Inequality	-0.856	-0.003
Regional Inequality	0.284	-0.200

Notes: This table shows the percentage change in welfare, production and inequality from reducing linguistic distance (column 1) or geographic distance (column 2) by 6.2%. The equilibrium with the distance reduction is computed using the quantitative model developed in Section 6.

Table 8: Welfare Effects of Reducing Linguistic Distance by 6.2%: Decomposing the Channel

	Aggregate Welfare (1)	Inequality (2)
(1) No Skill divide, Monotonic Effect of Linguistic Distance	0.089	.
(2) No Skill divide, Non-monotonic Effect of Linguistic Distance	0.201	.
(3) Skill Complementarity, Same Migration cost, Non-monotonic Effect	0.189	-0.360
(4) (Baseline) Skill Complementarity, Different Migration cost, Non-monotonic Effect	0.134	-0.856
(5) No Linguistic barriers in Trade	0.089	-0.855

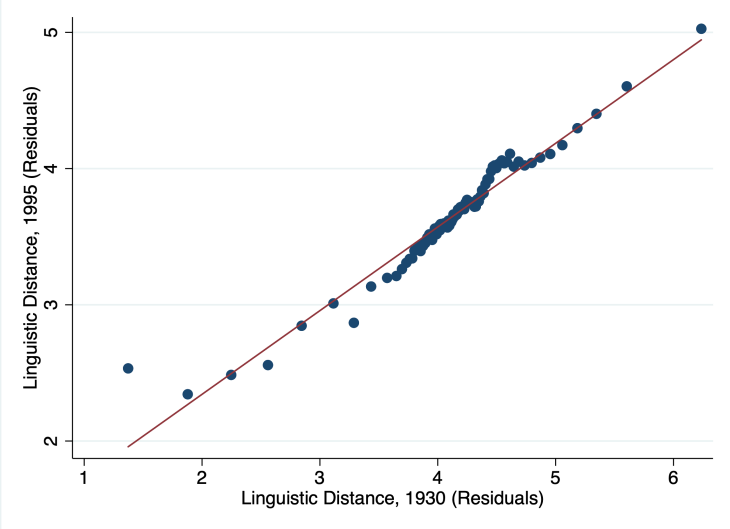
Notes: This Table shows the percentage change in welfare and inequality from reducing linguistic distance by 6.3% using different model. Row 1 considers a model without divide in skills, and assume monotonic effect of linguistic distance on migration, which is exactly the model in Allen and Arkolakis (2018). Row 2 still considers the basic setting in Allen and Arkolakis (2018) but assumes non-monotonic effects of lingusitic distance. Row 3 considers a model with skill complementarity in production but skilled and unskilled workers are faced with same migration costs. Row 4 considers the full model with skill complementarity in production and skilled and unskilled workers are faced with different migration costs. Row 5 considers a model where linguistic distance only enters migration costs but not trade costs.

Table 9: Determinants of Welfare impacts of Pairwise Reduction in Linguistic Distance

	(1)	(2)	(3)	(4)	(5)
Initial Linguistic Distance, 1995	-0.004***	0.001***	0.001***	0.001***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Initial Linguistic Distance, 1995, squared	0.001***	0.001***	0.001***	0.001***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log(Geographic Distance)		-0.008***	-0.008***	-0.008***	-0.008***
		(0.000)	(0.000)	(0.000)	(0.000)
Productivity Level (minimum)			0.001	0.000	0.002*
			(0.001)	(0.001)	(0.001)
Amenity Level (minimum)				0.008***	0.006***
				(0.001)	(0.001)
Log Population (minimum)					0.007***
					(0.000)
Parabola	2.218	-1.039	-1.088	-1.229	-3.922
N of region pairs	27,026	27,026	27,026	27,026	24,529
R-squared	0.312	0.399	0.399	0.403	0.411
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes

Notes: This table shows the correlation of the aggregate welfare changes (in percentage) with the observed or inverted location characteristics. For convenience, I rescale the dependent variables by multiplying 100. Robust standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

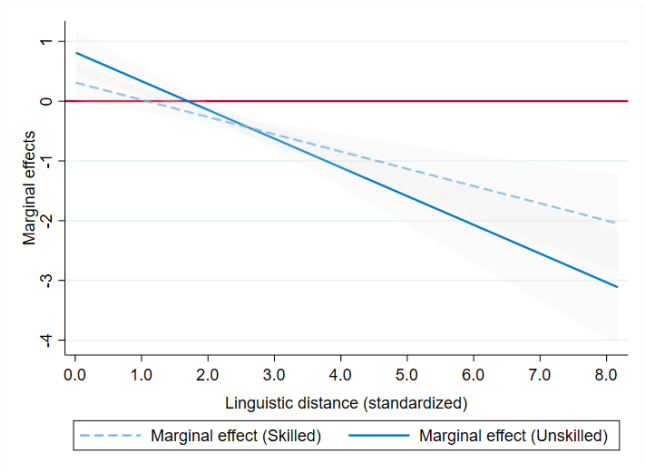
Figure 1: Correlation between linguistic distance 1995 and 1930



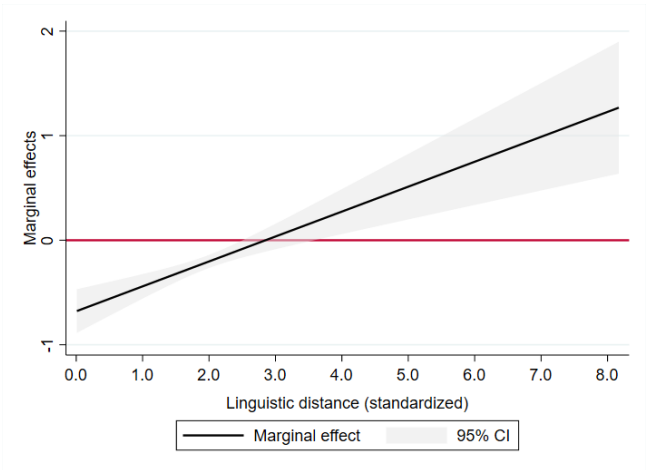
Notes: This figure shows the bin scatter plot of current linguistic distance and historical linguistic distance, purging out origin fixed effects, destination fixed effects and other pairwise controls.

Figure 2: Marginal Effects by Skill

(A) COMPARE SKILLED AND UNSKILLED

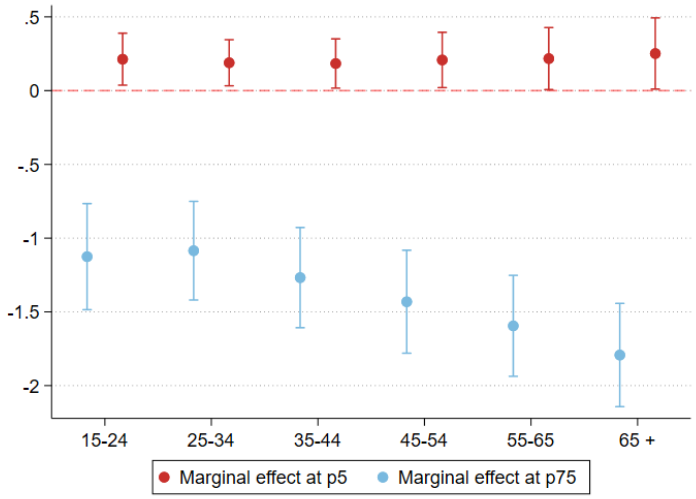


(B) SELECTION



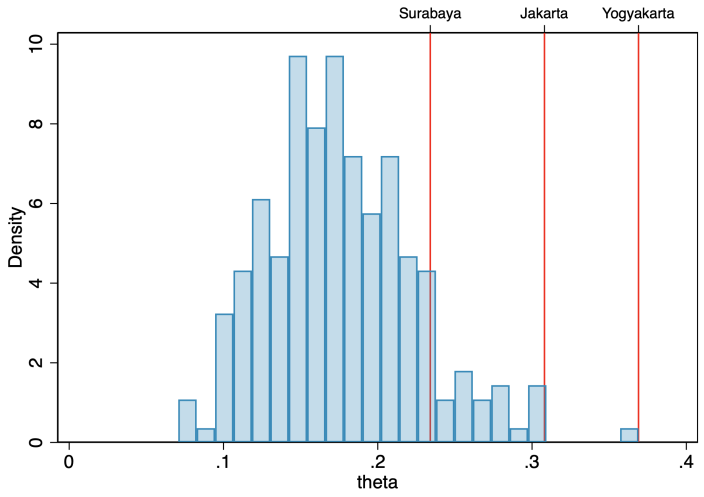
Notes: Panel A shows the marginal effects of linguistic distance on migration propensity for skilled (dash line) and unskilled (solid line), and panel B shows the marginal effect on migration selection against the level of linguistic distance. Panel A is obtained by evaluating the non-monotonic specification in column 1 (skilled) and 2 (unskilled) of Table 3 for each level of linguistic distance. The marginal effects are computed by $\beta_l + 2 * \beta_{ls} * Ldist$. Panel B is obtained using similar calculation using regression in column 3 of Table 3.

Figure 3: Linguistic Distance and Migration, by Age, Marginal Effects in 5th and 90th Percentile



Notes: This figure shows the marginal effects of linguistic distance on migration propensity for different age group. Each dot represents an estimated marginal effect of a certain age group, denoted in the x-axis. Blue dots (and corresponding confidence interval) are marginal effects at the 75th percentile of linguistic distance distribution and red dots (and corresponding confidence interval) are marginal effects at the 5th percentile of linguistic distance distribution. The marginal effects are obtained by evaluating the non-monotonic specification in Appendix Table A.11 and each dot corresponds to a column. The marginal effects are computed by $\beta_l + 2 * \beta_{ls} * Ldist$.

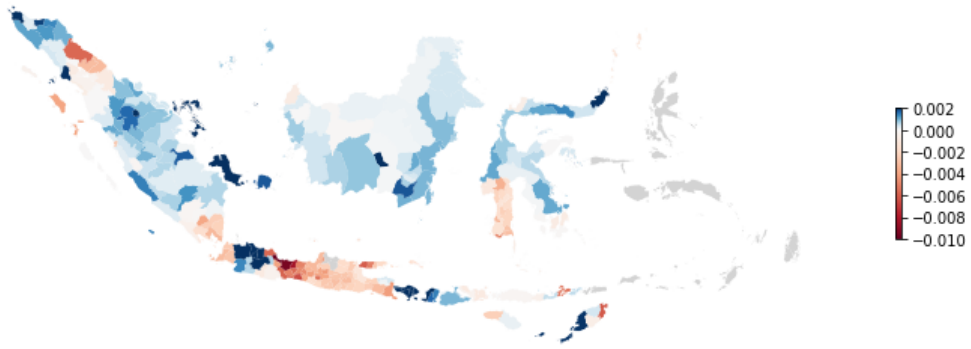
Figure 4: Estimated θ_j^s



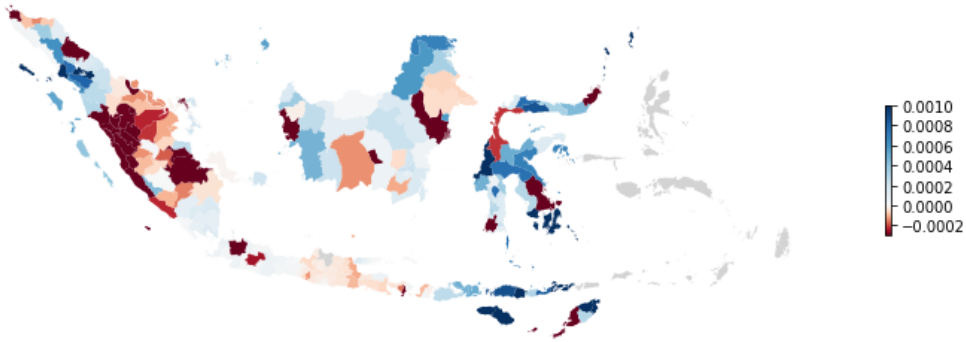
Notes: This figure shows the estimated skill intensity θ_j^s for each location using expression (17). Wage data is from SAKERNAS 2009 and 2010, and the population data is from the 2010 Indonesia Census.

Figure 5: Change in Spatial Distribution of Economy to a 6.2% Drop in Linguistic Distance

(A) POPULATION



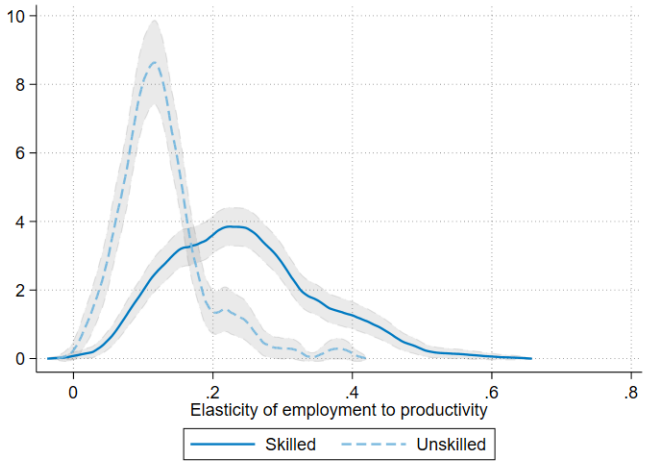
(B) SKILL COMPOSITION



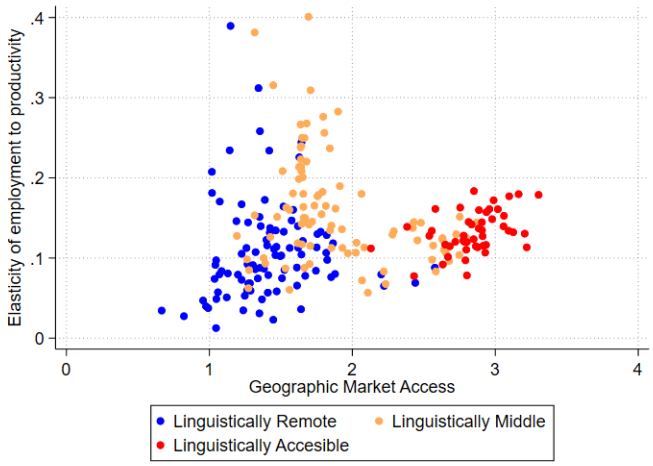
Notes: This figure maps the changes of population (panel A) and skill composition (panel B) when linguistic distance is dropped by 6.2%. Those locations in red undergo a reduction in population (skill share) while those in blue undergo increases, and the darker the color the larger the size of the changes.

Figure 6

(A) DISTRIBUTION OF EMPLOYMENT ELASTICITIES



(B) LOCAL EMPLOYMENT ELASTICITIES AND LINGUISTIC/GEOGRAPHIC MARKET ACCESS



Notes: Panel A shows kernel density of local employment elasticities for skilled (solid) and unskilled (dash), and Panel B plots the local employment elasticities against geographic market access, with each dot colored in different color according to linguistic market access.

A Appendix Tables and Figures

Table A.1: Crosswalk of Ethnicity definition between Census

Ethnicity 1930	Ethnicity 2010	Population Share 1930 (%)
Java	Java	48.2
Sunda	Sunda	14.0
Madura	Madura	0.07
From South Celebes	Baras, Bentong, Bingi, Duri, Kalumpang, Makassar, Rongkong, etc.	0.05
Minangkabau	Minangkabau	0.03
From Bali and Lombok	Bali Hindu, Bali Majapahit, Bali Aga	0.03
From Timor	Bima, Trunyan	0.03
From South Sumatra	Palembang, Daya, Enim, Gumai, Kikim	0.02
Malay	Melayu Asahan, Melayu Riau, Melayu Langkat, Melayu Banyu Asin, Melayu Lahat, etc.	0.02
Batak	Batak Angkola, Batak Simalungun, Batak Tapanuli, Batak Pakpak Dairi, Batak Karo, etc.	0.02
Batavian	Java, Malay, Sunda	0.02
From North Celebes	Bintauna, Bolaang Itang, Sangir, Talaud, etc.	0.01
Banda	Banda	0.01
From North Sumatra	Asahan, Dairi, Lubu, Nias, Siberut, Siladang, etc.	0.01
Dayak	Dayak Abai, Dayak Apalin, Dayak Apoyan, Dayak Badat, Dayak Bahau, Dayak Darok, etc.	0.01

Notes: This table shows the crosswalk between the 15 most populous ethnic groups in 1930 and the ethnic groups in 2010 Census. Some of the classification rules are based on Ananta et al. (2014b).

Table A.2: First Stage

	linear term			Quadratic term		
	$\kappa = 0.05$	$\kappa = 0.5$	$\kappa \to \infty$	$\kappa = 0.05$	$\kappa = 0.5$	$\kappa \to \infty$
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnic Share, 1930	0.366***	0.422***	0.468***	-0.444	-0.396	-0.084
	(0.087)	(0.102)	(0.096)	(0.425)	(0.521)	(0.488)
Ethnic Share, 1930, squared	0.038**	0.047**	0.034*	0.548***	0.691***	0.604***
	(0.018)	(0.021)	(0.020)	(0.092)	(0.112)	(0.104)
Cross Islands (1 0)	0.046	0.059	0.053	-0.364	-0.199	-0.211
	(0.058)	(0.067)	(0.061)	(0.296)	(0.360)	(0.309)
Agroclimatic Similarity	-0.005**	-0.006**	-0.006**	-0.040***	-0.045***	-0.052***
	(0.002)	(0.002)	(0.002)	(0.011)	(0.014)	(0.014)
d_lat	-0.005	-0.006	-0.010***	-0.044**	-0.073***	-0.106***
	(0.003)	(0.004)	(0.004)	(0.019)	(0.023)	(0.021)
d_long	0.009***	0.009***	0.005**	0.097***	0.108***	0.080***
	(0.002)	(0.002)	(0.002)	(0.013)	(0.015)	(0.013)
N	54,756	54,756	54,756	54,756	54,756	54,756
N Clusters	234	234	234	234	234	234
Adj. R^2	0.810	0.757	0.782	0.853	0.730	0.770
Adj. R^2 (Within)	0.601	0.618	0.621	0.487	0.530	0.541
Regression F-Stat	215.9	220.0	219.0	178.6	198.2	202.9
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows the first stage results of the IV regression. The dependent variable of all regressions is the current linguistic distance. Robust standard errors, two-way clustered at the district level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.3: Linguistic distance and migration, adding more polynomial terms

	(1)	(2)	(3)	(4)
linguistic distance, 1995	0.155***	1.312***	1.834***	1.246**
	(0.043)	(0.174)	(0.339)	(0.485)
linguistic distance, 1995, squared		-0.260***	-0.536***	-0.068
		(0.038)	(0.167)	(0.344)
linguistic distance, 1995, cubed			0.039	-0.090
			(0.024)	(0.086)
linguistic distance, 1995, 4th order				0.011*
				(0.007)
log(distance)	-1.207***	-1.222***	-1.221***	-1.220***
	(0.078)	(0.080)	(0.082)	(0.082)
Cross Islands (1 0)	-0.166	-0.144	-0.144	-0.142
	(0.143)	(0.130)	(0.129)	(0.130)
Agroclimatic Similarity	0.030***	0.033***	0.032***	0.032***
	(0.010)	(0.010)	(0.010)	(0.010)
Absolute difference in latitute	0.056**	0.063***	0.065***	0.064***
	(0.022)	(0.021)	(0.021)	(0.021)
Absolute difference in longtitute	-0.130***	-0.105***	-0.108***	-0.106***
	(0.017)	(0.017)	(0.017)	(0.017)
self	1.363***	2.037***	2.224***	2.082***
	(0.284)	(0.319)	(0.304)	(0.293)
N of region pairs	48,841	48,841	48,841	48,841
N Clusters	221	221	221	221
Pseudo R-squared	0.797	0.798	0.798	0.798
Origin Kabu FE	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) with more polynomial terms of linguistic distance. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.4: Cultural Distance and Migration, Robustness of Instrument

	Baseline	**Drop Significant Historical Migration Pairs**	**Control Migration Network**
	(1)	**(2)**	**(3)**
linguistic distance, 1995	0.766***	0.609***	0.707***
	(0.183)	(0.173)	(0.179)
linguistic distance, 1995, squared	-0.236***	-0.203***	-0.218***
	(0.040)	(0.038)	(0.039)
log(geographic distance)	-1.020***	-0.989***	-0.953***
	(0.083)	(0.093)	(0.085)
Cross Islands (1 0)	-0.034	0.087	-0.013
	(0.134)	(0.128)	(0.126)
Absolute difference in latitute	0.030	0.031	0.023
	(0.019)	(0.022)	(0.018)
Absolute difference in longtitute	-0.093***	-0.114***	-0.100***
	(0.016)	(0.021)	(0.015)
Agroclimatic Similarity	0.027***	0.050***	0.026***
	(0.010)	(0.013)	(0.010)
Hometown Bias	1.974***	1.818***	0.643
	(0.343)	(0.369)	(0.734)
Previous Migration Network			2.140**
			(0.987)
Optimal Linguistic Dist.	1.622	1.503	1.624
Median Distance	3.746	3.809	3.746
Mean Distance	3.493	3.577	3.493
N of region pairs	54,289	38,025	54,289
N Clusters	233	195	233
Pseudo R-squared	0.799	0.803	0.799
Kleibergen-Paap Wald Rank F Stat	84.277	62.030	82.318
Under Id. Test (KP Rank LM Stat)	47.821	33.548	47.181
p-Value	0.000	0.000	0.000
Origin Kabu FE	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes

Notes: This table demonstrates the robustness of the historical linguistic distance as instrument. In column 2, I drop the locations that were important migration exporters or importers in 1930. In column 3, I further control previous migration networks using migration propensity in 2000. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.5: Linguistic Distance and Migration, Split Sample

	Below	Above
	(1)	**(2)**
linguistic distance, 1995	0.100	-1.145***
	(0.101)	(0.162)
log(distance)	-0.971***	-1.344***
	(0.083)	(0.102)
Cross Islands (1 0)	0.102	-0.033
	(0.173)	(0.134)
Agroclimatic Similarity	0.037***	0.023*
	(0.013)	(0.012)
Absolute difference in latitute	0.037	0.060***
	(0.028)	(0.020)
Absolute difference in longtitute	-0.122***	-0.045***
	(0.021)	(0.014)
Hometown Bias	2.009***	-3.339***
	(0.289)	(0.663)
N of region pairs	13,556	40,950
N Clusters	217	233
Pseudo R-squared	0.768	0.815
Pairwise Controls	Yes	Yes
Origin Kabu FE	Yes	Yes
Destination Kabu FE	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) by slitting the sample by the 25th percentile of the linguistic distance. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.6: Linguistic Distance and Migration, Linear Regression

	Baseline	**OLS**	**2SLS**	**PPML**
	(1)	**(2)**	**(3)**	**(4)**
linguistic distance, 1995	0.766***	1.155***	1.403***	0.501***
	(0.183)	(0.116)	(0.156)	(0.182)
linguistic distance, 1995, squared	-0.236***	-0.245***	-0.322***	-0.217***
	(0.040)	(0.024)	(0.038)	(0.039)
log(geographic distance)	-1.020***	-1.763***	-1.696***	
	(0.083)	(0.049)	(0.054)	
Geographic distance				-0.005***
				(0.001)
Geographic distance, squared				0.000***
				(0.000)
Cross Islands (1 0)	-0.034	0.045	0.012	0.060
	(0.134)	(0.061)	(0.068)	(0.141)
Absolute difference in latitute	0.030	0.084***	0.079***	0.165***
	(0.019)	(0.007)	(0.007)	(0.050)
Absolute difference in longtitute	-0.093***	0.010*	0.011*	0.042
	(0.016)	(0.006)	(0.007)	(0.056)
Agroclimatic Similarity	0.027***	-0.003	-0.005	0.038***
	(0.010)	(0.004)	(0.004)	(0.010)
Hometown Bias	1.974***	0.129	0.425*	5.759***
	(0.343)	(0.226)	(0.235)	(0.216)
Optimal Linguistic Dist.	1.622	2.353	2.176	1.155
Median Distance	3.746	3.746	3.746	3.746
Mean Distance	3.493	3.493	3.493	3.493
N of region pairs	54,289	39,413	39,413	54,289
N Clusters	233	233	233	233
Pseudo R-squared	0.799			0.799
Kleibergen-Paap Wald Rank F Stat	84.277		84.277	96.743
Under Id. Test (KP Rank LM Stat)	47.821		47.821	52.035
p-Value	0.000		0.000	0.000
Origin Kabu FE	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2). For all regressions, I use the OLS instead of PPML. Robust standard errors, two-way clustered at the district level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.7: Cultural Distance and Migration, Different κ

	$\kappa = 0.05$	$\kappa = 0.5$	$\kappa \to \infty$
	(1)	**(2)**	**(3)**
linguistic distance, 1995	0.766***	1.037***	0.981***
	(0.183)	(0.148)	(0.156)
linguistic distance, 1995, squared	-0.236***	-0.259***	-0.245***
	(0.040)	(0.027)	(0.029)
log(distance)	-1.020***	-1.013***	-1.004***
	(0.083)	(0.085)	(0.084)
Cross Islands (1 0)	-0.034	-0.021	-0.005
	(0.134)	(0.129)	(0.126)
Absolute difference in latitute	0.030	0.027	0.019
	(0.019)	(0.019)	(0.019)
Absolute difference in longtitute	-0.093***	-0.089***	-0.096***
	(0.016)	(0.017)	(0.017)
Agroclimatic Similarity	0.027***	0.027***	0.024**
	(0.010)	(0.010)	(0.010)
Hometown Bias	1.974***	2.339***	2.232***
	(0.343)	(0.330)	(0.320)
Parabola	1.622	1.998	2.005
Median Distance	3.746	4.259	4.433
Mean Distance	3.493	3.888	4.002
N of region pairs	54,289	54,289	54,289
N Clusters	233	233	233
Pseudo R-squared	0.799	0.799	0.799
Adjusted R-squared			
Kleibergen-Paap Wald Rank F Stat	84.277	103.036	93.403
Under Id. Test (KP Rank LM Stat)	47.821	50.479	51.157
p-Value	0.000	0.000	0.000
Origin Kabu FE	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) using different linguistic distance measures based on different choices of κ. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.8: Cultural Distance and Migration, Different Sample

	Baseline	**Drop Linguistic Dist.** $> 95\%$	**Drop Migration Flow** $> 95\%$	**Drop Migration Flow** $= 0$	**2000 Recent Migration**	**2010 Lifetime Migration**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
linguistic distance, 1995	0.766***	1.095***	1.239***	0.724***	0.786***	0.814***
	(0.183)	(0.168)	(0.128)	(0.173)	(0.169)	(0.182)
linguistic distance, 1995, squared	-0.236***	-0.321***	-0.279***	-0.224***	-0.258***	-0.270***
	(0.040)	(0.035)	(0.027)	(0.038)	(0.040)	(0.040)
log(geographic distance)	-1.020***	-1.012***	-1.375***	-1.025***	-0.941***	-1.084***
	(0.083)	(0.086)	(0.072)	(0.082)	(0.077)	(0.078)
Cross Islands (1 0)	-0.034	-0.035	-0.044	0.008	-0.214	0.113
	(0.134)	(0.130)	(0.099)	(0.136)	(0.153)	(0.092)
Absolute difference in latitute	0.030	0.037**	0.070***	0.027	-0.017	0.054***
	(0.019)	(0.018)	(0.012)	(0.019)	(0.021)	(0.019)
Absolute difference in longtitute	-0.093***	-0.092***	-0.029**	-0.073***	-0.067***	-0.079***
	(0.016)	(0.017)	(0.011)	(0.016)	(0.013)	(0.014)
Agroclimatic Similarity	0.027***	0.026**	-0.001	0.025**	0.037***	0.030***
	(0.010)	(0.010)	(0.006)	(0.010)	(0.011)	(0.010)
Hometown Bias	1.974***	2.230***	.	1.939***	1.652***	-0.032
	(0.343)	(0.341)		(0.348)	(0.340)	(0.332)
Optimal Linguistic Dist.	1.622	1.708	2.218	1.619	1.523	1.507
Median Distance	3.746	3.714	3.769	3.607	3.746	3.746
Mean Distance	3.493	3.386	3.553	3.336	3.493	3.493
N of region pairs	54,289	51,572	51,571	39,413	54,056	54,289
N Clusters	233	231	233	233	232	233
Pseudo R-squared	0.799	0.798	0.110	0.789	0.776	0.677
Kleibergen-Paap Wald Rank F Stat	84.277	88.201	76.975	84.277	87.503	76.548
Under Id. Test (KP Rank LM Stat)	47.821	47.846	45.309	47.821	47.646	46.660
p-Value	0.000	0.000	0.000	0.000	0.000	0.000
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) using different samples. The title of each column describes the samples. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.9: Linguistic Distance and Migration, using administrative division in 1930

	PPML			PPML with IV
	(1)	(2)	(3)	(4)
linguistic distance, 1995	-0.595***	0.767	1.430***	1.450***
	(0.072)	(0.579)	(0.488)	(0.482)
linguistic distance, 1995, squared		-0.452**	-0.384**	-0.416**
		(0.209)	(0.168)	(0.168)
log(geographic distance)			-0.810***	-0.777***
			(0.128)	(0.135)
Cross Islands (1 0)			0.305	0.355
			(0.221)	(0.234)
Absolute difference in latitute			0.031	0.028
			(0.036)	(0.036)
Absolute difference in longtitute			-0.175***	-0.177***
			(0.027)	(0.027)
Agroclimatic Similarity			0.103***	0.103***
			(0.019)	(0.019)
Hometown Bias	7.392***	7.838***	3.174***	3.242***
	(0.166)	(0.212)	(0.511)	(0.530)
Optimal Linguistic Dist.		0.848	1.862	1.741
Median Distance		2.658	2.658	2.658
Mean Distance		2.319	2.319	2.319
N of region pairs	15,129	15,129	15,129	15,129
N Clusters	123	123	123	123
Pseudo R-squared	0.778	0.778	0.782	0.782
Kleibergen-Paap Wald Rank F Stat				2028.379
Under Id. Test (KP Rank LM Stat)				3202.179
p-Value				0.000

Notes: The regressions shows the results of estimation of gravity equation (2) collapsing all variables to the colonial definition of districts. For Column (3), I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.10: Cultural Distance and Migration Selection, by Education Group

	No School	Primary School	Middle School	High School	Higher Education
	(1)	(2)	(3)	(4)	(5)
linguistic distance, 1995	1.297***	1.023***	0.671***	0.616***	0.312**
	(0.158)	(0.163)	(0.150)	(0.226)	(0.153)
linguistic distance, 1995, squared	-0.318***	-0.260***	-0.195***	-0.208***	-0.144***
	(0.038)	(0.038)	(0.035)	(0.047)	(0.033)
log(distance)	-1.063***	-1.131***	-1.132***	-1.047***	-1.030***
	(0.083)	(0.088)	(0.079)	(0.082)	(0.062)
Cross Islands (1 0)	-0.285**	-0.265*	-0.340***	-0.421***	-0.422***
	(0.141)	(0.148)	(0.122)	(0.145)	(0.108)
Absolute difference in latitute	0.069***	0.083***	0.060***	0.016	0.047***
	(0.021)	(0.024)	(0.020)	(0.024)	(0.018)
Absolute difference in longtitute	-0.091***	-0.092***	-0.078***	-0.072***	-0.034***
	(0.020)	(0.019)	(0.017)	(0.017)	(0.011)
Agroclimatic Similarity	0.026***	0.021**	0.026***	0.026***	0.035***
	(0.009)	(0.010)	(0.008)	(0.010)	(0.010)
Hometown Bias	3.317***	2.491***	1.477***	1.231***	0.950***
	(0.348)	(0.366)	(0.322)	(0.363)	(0.263)
Parabola	2.036	1.969	1.718	1.478	1.082
Median Distance	3.746	3.746	3.746	3.746	3.746
Mean Distance	3.493	3.493	3.493	3.493	3.493
N of region pairs	54,289	54,289	54,289	54,289	54,289
N Clusters	233	233	233	233	233
Pseudo R-squared	0.825	0.817	0.798	0.766	0.778
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) for differen education groups. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.11: Cultural Distance and Migration Selection, by Age Group

	15-24	25-34	35-44	45-54	55-64	Above 65
	(1)	(2)	(3)	(4)	(5)	(6)
linguistic distance, 1995	0.783***	0.732***	1.018***	1.194***	1.331***	1.328***
	(0.225)	(0.161)	(0.162)	(0.179)	(0.211)	(0.210)
linguistic distance, 1995, squared	-0.238***	-0.205***	-0.266***	-0.311***	-0.351***	-0.365***
	(0.050)	(0.038)	(0.038)	(0.041)	(0.047)	(0.046)
log(distance)	-1.023***	-1.029***	-0.942***	-0.899***	-0.877***	-0.866***
	(0.089)	(0.080)	(0.081)	(0.081)	(0.085)	(0.086)
Cross Islands (1 0)	-0.340**	-0.412***	-0.346***	-0.329***	-0.337***	-0.339**
	(0.144)	(0.106)	(0.111)	(0.115)	(0.130)	(0.156)
Absolute difference in latitute	0.045*	0.077***	0.084***	0.080***	0.075***	0.058**
	(0.023)	(0.018)	(0.019)	(0.020)	(0.023)	(0.025)
Absolute difference in longtitute	-0.100***	-0.072***	-0.082***	-0.096***	-0.100***	-0.099***
	(0.019)	(0.016)	(0.015)	(0.015)	(0.016)	(0.019)
Agroclimatic Similarity	0.026**	0.027***	0.025***	0.027***	0.025**	0.019*
	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.011)
Hometown Bias	1.640***	1.936***	3.227***	4.013***	4.424***	4.637***
	(0.370)	(0.304)	(0.313)	(0.321)	(0.348)	(0.351)
Parabola	1.646	1.786	1.913	1.920	1.897	1.820
Median Distance	3.746	3.746	3.746	3.746	3.746	3.746
Mean Distance	3.493	3.493	3.493	3.493	3.493	3.493
N of region pairs	54,289	54,289	54,289	54,289	54,289	54,289
N Clusters	233	233	233	233	233	233
Pseudo R-squared	0.772	0.791	0.814	0.826	0.829	0.832
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table shows the results of estimation of gravity equation (2) for differen age groups. For all regressions, I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.12: Estimating ρ using catchment area rainfall shock

	Dep. Var.: $\ln(w^s_{jt}/w^u_{jt})$
Ln(Skilled L/Unskilled L)	-0.809
	(1.057)
Own Rainfall Shock	-0.021
	(0.015)
N	1,223
Adjusted R-squared	0.330
Kleibergen-Paap Wald Rank F Stat	1.693
Under Id. Test (KP Rank LM Stat)	2.005
p-Value	0.157
Kabu FE	Yes
Year FE	Yes
Elasticity of Substitution ρ	1.24

Notes: The table shows the results of estimating ρ using relative labor demand equation (43) and using rainfall shock constructed in Eq.44 as instrument. Robust standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.13: Overall Effects of Linguistic Distance and Migration using a Monotonic Specification

	Baseline (1)	**Monotonic Specification (2)**
linguistic distance, 1995	0.766***	-0.225***
	(0.183)	(0.069)
linguistic distance, 1995, squared	-0.236***	
	(0.040)	
log(geographic distance)	-1.020***	-1.001***
	(0.083)	(0.077)
Cross Islands (1 0)	-0.034	-0.100
	(0.134)	(0.146)
Absolute difference in latitute	0.030	0.018
	(0.019)	(0.022)
Absolute difference in longtitute	-0.093***	-0.111***
	(0.016)	(0.015)
Agroclimatic Similarity	0.027***	0.024**
	(0.010)	(0.010)
Hometown Bias	1.974***	1.462***
	(0.343)	(0.285)
Optimal Linguistic Dist.	1.622	
Median Distance	3.746	
Mean Distance	3.493	
N of region pairs	54,289	54,289
N Clusters	233	233
Pseudo R-squared	0.799	0.799
Kleibergen-Paap Wald Rank *F* Stat	84.277	290.930
Under Id. Test (KP Rank LM Stat)	47.821	52.020
p-Value	0.000	0.000

Notes: The table campares the results of estimation of gravity equation (2) using a non-monotonic specification and a monotonic specification. I use the control function method to implement instrumented estimation in PPML (Wooldridge, 2015). Bootstrapped standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.14: Determinants of Inequality impacts of Pairwise Reduction in Linguistic Distance

	(1)	(2)	(3)	(4)	(5)
Initial Linguistic Distance, 1995	-0.812***	-0.813***	-0.837***	-0.838***	-0.858***
	(0.050)	(0.050)	(0.050)	(0.055)	(0.056)
Initial Linguistic Distance, 1995, squared	0.191***	0.191***	0.194***	0.196***	0.199***
	(0.010)	(0.010)	(0.010)	(0.012)	(0.012)
Productivity Level (minimum)		-0.013	0.096	0.065	-0.042
		(0.101)	(0.101)	(0.114)	(0.115)
Amenity Level (minimum)			-0.885***	-1.150***	-1.122***
			(0.082)	(0.100)	(0.099)
Log Population (minimum)				0.249***	0.173***
				(0.063)	(0.063)
Skill Intensity (minimum)					0.814***
					(0.118)
Parabola	2.131	2.132	2.159	2.139	2.151
N of region pairs	27,026	27,026	27,026	24,529	24,529
R-squared	0.281	0.281	0.284	0.293	0.294
Origin Kabu FE	Yes	Yes	Yes	Yes	Yes
Destination Kabu FE	Yes	Yes	Yes	Yes	Yes

Notes: This table shows the correlation of the skill inequality changes (in percentage) with the observed or inverted location characteristics. For convenience, I rescale the dependent variables by multiplying them by 100. Robust standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.15: Summary of Local Employment Elasticity

Color	Geographic MA	Linguistic MA	Average Local Employment Elasticity
Blue	Low	Low	0.103
Orange	Middle	Middle	0.150
Red	High	High	0.128

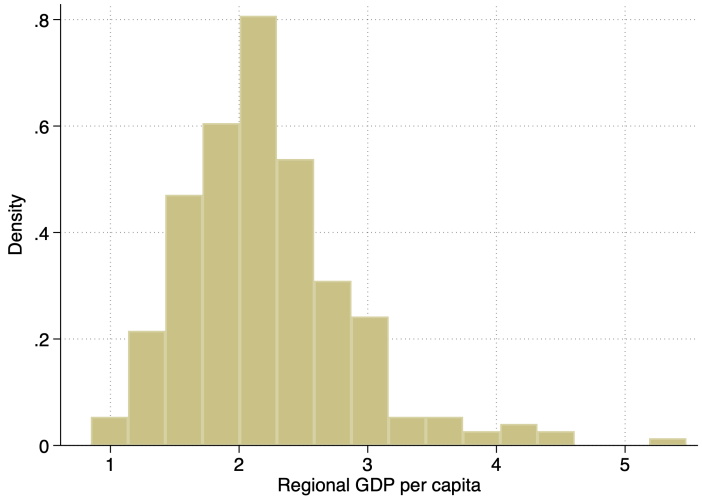
Notes: This table summarizes the average local employment elasticities for locations with different levels of linguistic and geographic market access.

Table A.16: Estimating η and σ

	Eq.20		Eq.21		Eq.22	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
lnw	-0.608***	3.526				
	(0.022)	(4.904)				
lnl	-0.535***	-0.534***				
	(0.008)	(0.127)				
lnws			1.350***	1.887		
			(0.251)	(3.626)		
lnwu					1.869***	1.679
					(0.213)	(2.138)
N	234	234	234	234	234	234
Adjusted R-squared	0.967	-2.404	0.210	0.228	0.359	0.420
Kleibergen-Paap Wald Rank F Stat		0.311		0.663		2.037
Implied σ		4.5				
Implied α		0.15				
Implied η_s				1.887		
Implied η_u						1.679

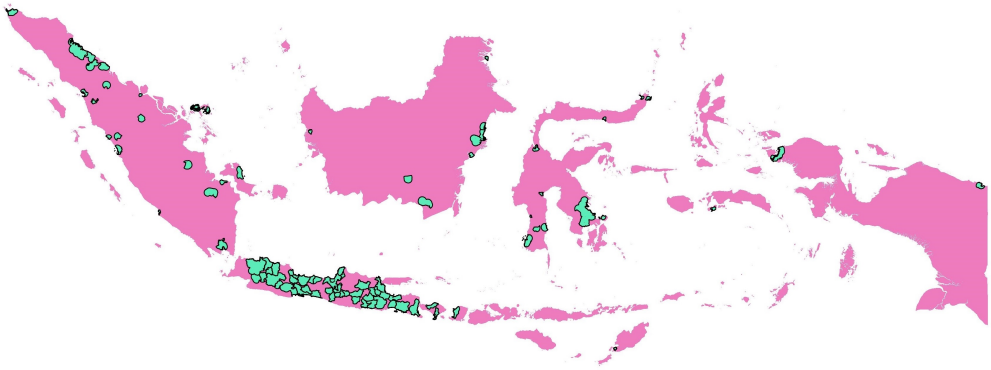
Notes: This table shows the results of estimating η and σ using the structural instruments following Allen and Arkolakis (2018). Robust standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Figure A.1: Regional GDP per capita



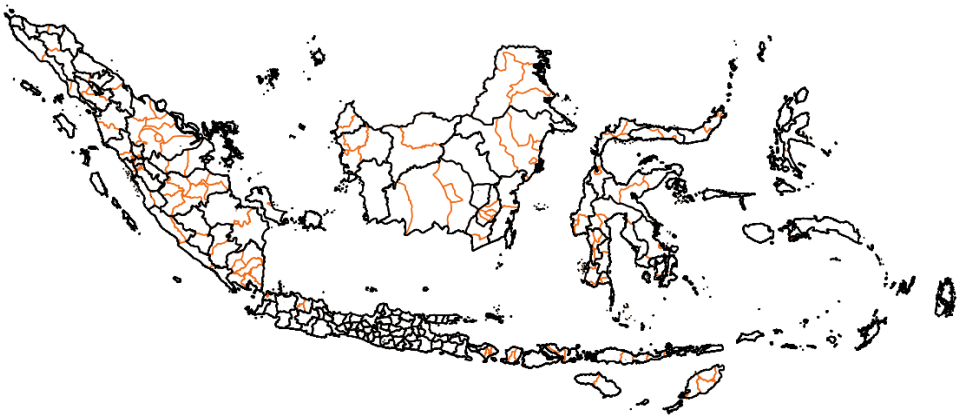
Notes: This figure shows the distribution of regional GDP per capita (IDR Million) of each *Kota* or *Kabupaten* in Indonesia in 2010. The data is from INDO-DAPOER. Oil revenue is excluded in the calculation.

Figure A.2: The metro area in Indonesia (Civelli et al., 2021)



Notes: This figure presents a map of urban areas in Indonesia, using Civelli et al. (2021)'s approach for delineating metro areas, which itself follows Burchfield et al. (2006). More details are included in Appednix B.1.

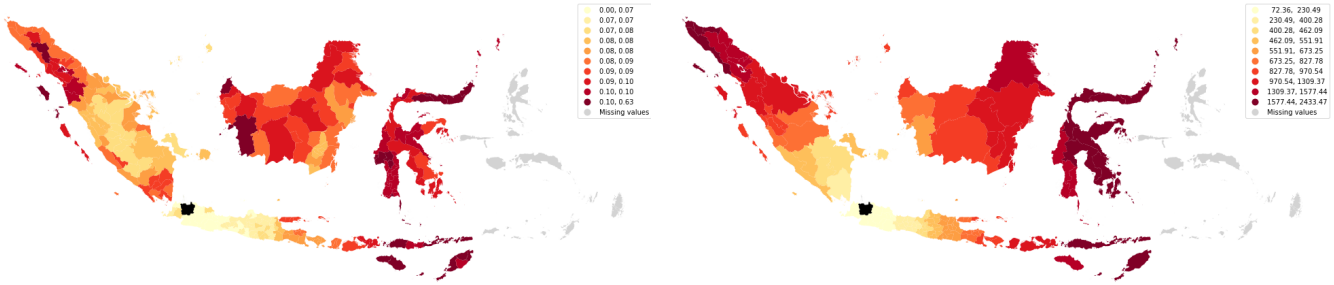
Figure A.3: Compare administrative border in 1930 and 2000



Notes: This figure compares the modern administrative border definded in 2000 (in orange) and colonial border defined in 1930 (in black).

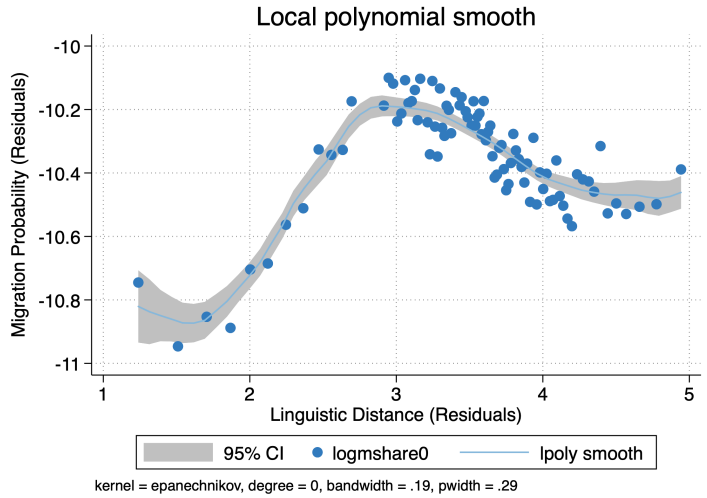
Figure A.4: Linguistic Distance to Jakarta

(A) LINGUISTIC DISTANCE, 1995 **(B)** LINGUISTIC DISTANCE, 1930



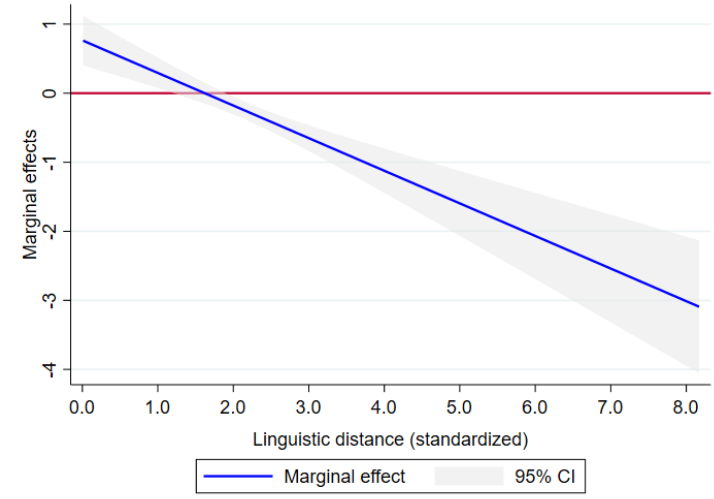
Notes: This figure maps the calculated linguistic distance to Jakarta (in black). Darker color signifies longer distance.

Figure A.5: Residual-residual Plots of Migration and Linguistic Distance



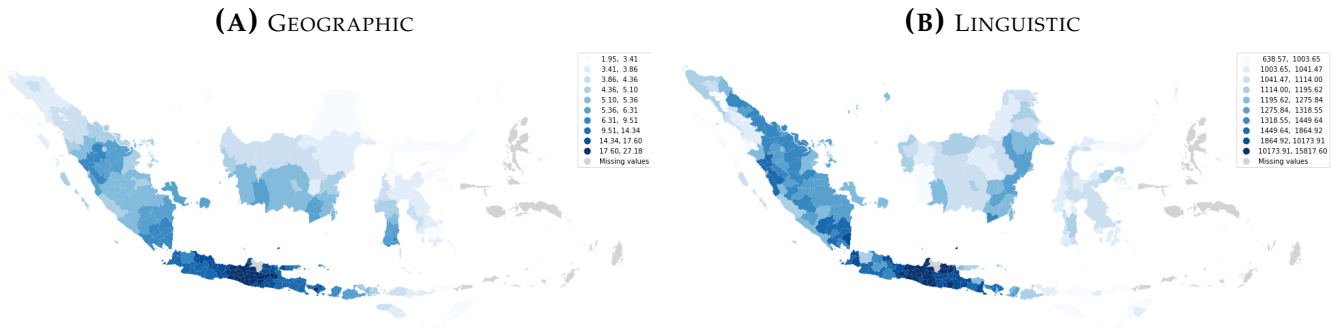
Notes: This figure shows the bin scatter plot of migration propensity and linguistic distance, purging out origin fixed effects, destination fixed effects and other pairwise controls. The blue line is the local polynomial smooth using Epanechnikov kernel, rule-of-thumb bandwidth and local cubic function. The shaded area is the confidence interval.

Figure A.6: Linguistic Distance and Migration, Marginal Effects



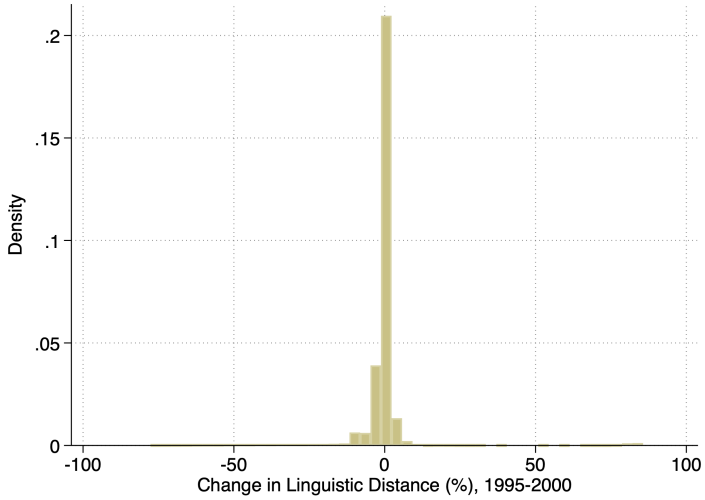
Notes: This figure shows the marginal effects of linguistic distance on migration propensity. It is obtained by evaluating the non-monotonic specification in column 4 of Table 2 for each level of linguistic distance. The marginal effects are computed by $\beta_l + 2 * \beta_{ls} * Ldist$.

Figure A.7: Market Access based on Geographic and Linguistic Distance



Notes: This figure maps the caculated geographic and linguistic market access using definition (25). Darker color signifies larger market access.

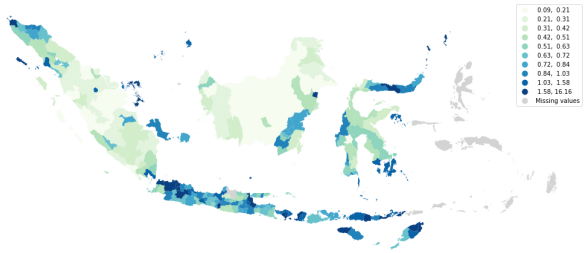
Figure A.8: Change in Linguistic Distance, 1995-2000



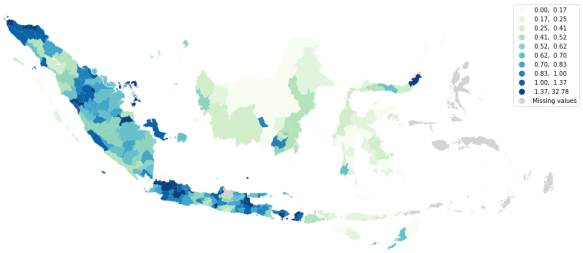
Notes: This figure shows the histogram of changes in linguistic distance from 1995 to 2000 using the Indonesia Census data.

Figure A.9: Inverted composite

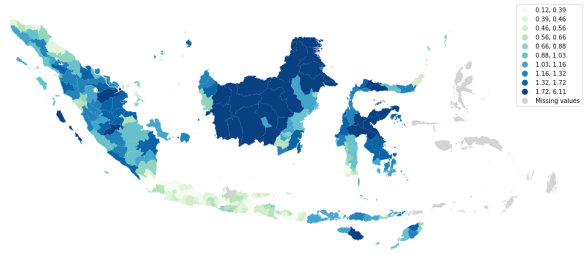
(A) LOCAL WELFARE FOR SKILLED $(\omega_i^s)^{\eta^s}$



(C) LOCAL WELFARE FOR UNSKILLED $(\omega_i^u)^{\eta^u}$



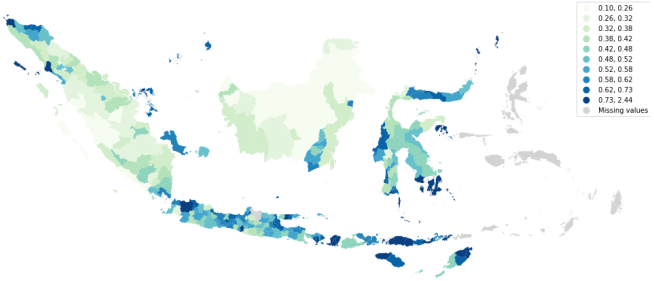
(B) LOCAL PRICE $p_i^{\sigma-1}$



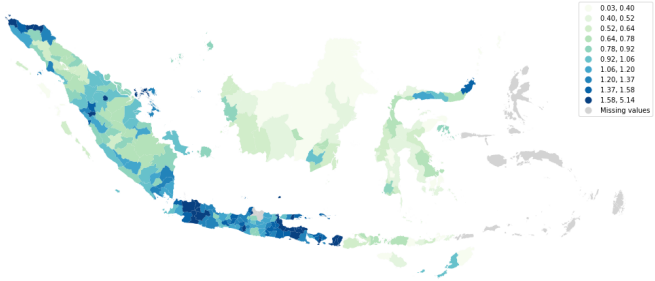
Notes: This figure maps the inverted welfare composites and price compostites using the equilibrium conditions Eq.(36)-Eq.(41).

Figure A.10: Inverted local fundamentals $\bar{u}_i^e$ and $\bar{A}_i$

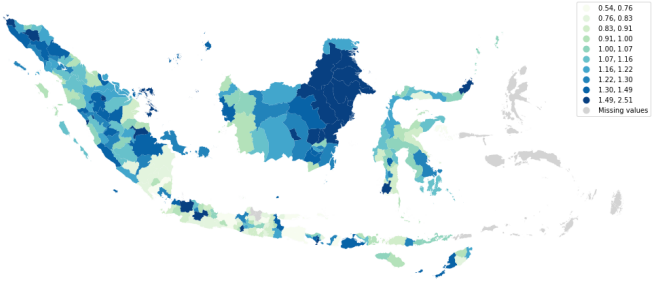
(A) EXOGENOUS AMENITY FOR SKILLED $\bar{u}_i^s$



(C) EXOGENOUS AMENITY FOR UNSKILLED $\bar{u}_i^u$

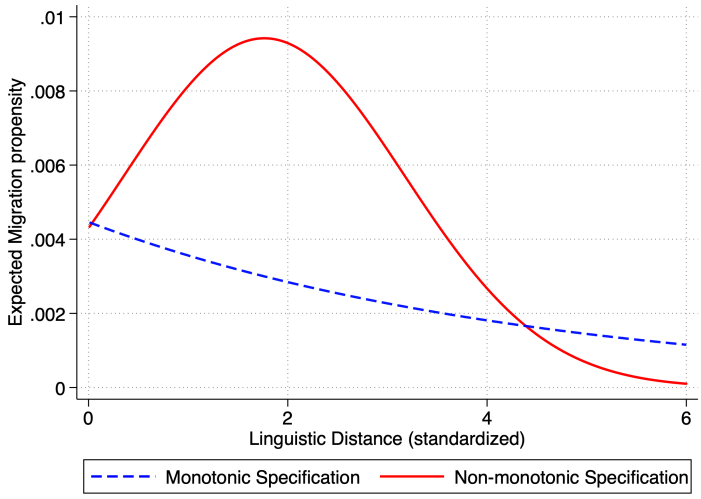


(B) EXOGENOUS PRODUCTIVITY $\bar{A}_i$



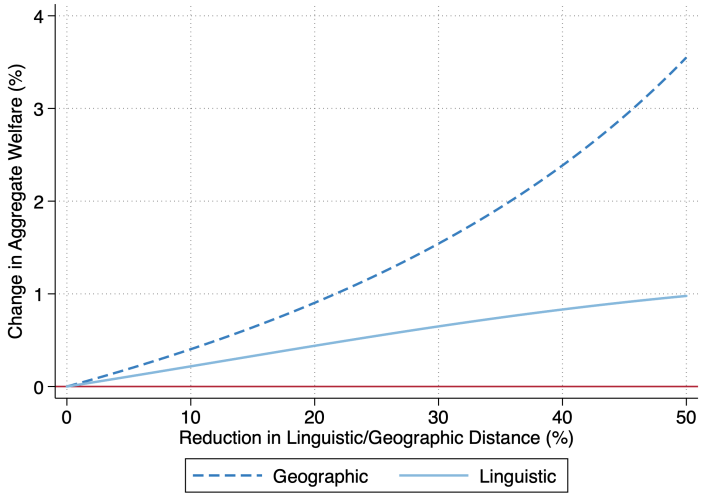
Notes: This figure maps the local fundamentals $\bar{u}_i^e$ and $\bar{A}_i$ using the definitions Eq.(20)-Eq.(22).

Figure A.11: Conditional Expectation of Migration Propensity Using Monotonic or Non-monotonic Specification



Notes: This figure plots the conditional expectation of migration propensity, i.e., $\mathbb{E}(M_{ij}|\mathbf{X}_{ij})$ by linguistic distance. To construct the figure, I use the estimates that condition on the full set of control variables from Columns (1) and (2) of Appendix Table A.13. Conditional expectation of the migration propensity is computed relative to its sample averages. The non-monotonic one is red and solid and monotonic one is blue and dash.

Figure A.12: Welfare Impacts of Further Reduction in Geograhic/Linguistic Distance



Notes: This figure plots the percentage changes in aggregate welfare against the percentage changes in linguistic distance (solid) or geographic distance (dash).

B Data Appendix

B.1 Definition of Metro Areas

I follow Civelli et al. (2021)'s definition of metro areas in Indonesia. They adopt a morphological approach to city definitions following Burchfield et al. (2006). They start by identifying a list of 83 urban regions in Indonesia by a population threshold of 100,000. Then, instead of using the administrative border, they identify the physical extent of urban areas based on patterns of high-density built-up areas, as measured with high-resolution satellite data.[58] They use this approach to identify 80 urban metropolitan areas as mapped in A.2. Half of these areas are located on the Inner Islands of Java and Bali, a quarter of them are on Sumatra, and the remaining quarter are in other parts of the Outer Islands. The largest urban areas identified are Jakarta, a megacity with a population size of more than 30 million, followed by Bandung, Surabaya, and Medan, with a population of 100,000 and 2 million.

B.2 More Details of Merging 1930 Colonial Census

Compared to the detailed over 1,000 ethnicity categories defined in 2000 and 2010 census, the Census 1930 has a much coarser definition of ethnicities. There are over 100 ethnic groups defined, but most of the variety comes from less-populated outer islands. The extent of details also depends on the islands. Since Javanese, Sundanese, and Madurese were dominant in population in Java, the Census only document very general groups of other ethnicities based on their homeland.[59] While in other islands, more detailed ethnicities are recorded. For example, ethnic groups in south Sumatra are further divided into ethnicities from Palembang, ethnicities from Lampung, ethnicities from Bengkulu, and others. In my analysis, in Java, I treat those defined as "From south Sumatra" as they belong to the same ethnic group. While in the outer islands, I treat those from Palembang and those from Lampung as different ethnic groups, although they both belong to the "From south Sumatra" group.

The spatial unit defined in the 1930 Census is also not thoroughly or perfectly consistent with the contemporary administrative division in Indonesia. As Java had been the economic and political center of the Dutch Indies, population data is collected at a finer level than in outer-islands. In contrast, the regional divisions are coarser in Sumatra and even coarser in other thinly-populated islands. The population and ethnicity share data are available at the third of five administrative levels in colonial Java, *Regentschaap* (regency) in Java, which mostly coincide with the district (*kabupaten/kota*) in 2000. I follow Pepinsky (2016)'s procedure to match the 1930 border with the contemporary administrative boundaries. Specifically, when new divisions have been created, I use the district boundaries which most closely match current boundaries. Where borders have been removed, I examine district boundaries again and use those which most closely match current boundaries.

C Reduced Form Analysis Appendix

C.1 PPML estimation

Taking log of the main gravity specification Eq.(2) deliver the following specification,

$$\ln \pi_{ij} = \delta_i + \delta_j + \beta_l Ldist_{ij} + \beta_{l2} Ldist_{ij}^2 + \beta_g Gdist_{ij} + \beta_h Home_{ij} + \gamma X_{ij} + \ln \epsilon_{ij} \tag{26}$$

This induces two problems: first, those zero migration flows are dropped by the nature of taking logarithm transformation, which induces selection bias. Second, the logarithm-transformed error term $\ln \epsilon_{ij}$ depends on

[58] They use Global Human Settlement Layer (GHSL), produced by the European Commission's Joint Research Centre (JRC). These data were created by applying machine learning techniques to 40 years of Landsat satellite imagery to measure the locations and intensity of human settlements, including buildings and physical infrastructure.

[59] Minor groups in Java are divided into "From south Sumatra", "From north Sumatra", "From south Celebes", "From Timor" and so on.

higher moments of ϵ_{ij}. If there is heteroskedasticity, the expectation of the error term is correlated to the independent variables since it includes its variance.

Silva and Tenreyro (2006) suggest using PPML (Pseudo Poisson Maximum Likelihood) estimator as a solution. For a model $y_i = exp(x_i\beta) + \epsilon_i$, PPML uses the set of first-order conditions,

$$\sum_{n=1}^{N}\{y_i - \exp(x_i\beta)\}x_i = 0$$

It does not require the data in fact distributed as Poisson (Gourieroux et al., 1984). As long as the explanatory variables are correctly specified, that is, $\mathbb{E}(y_i|x) = \exp(x_i\beta)$, the PPML estimator provides a consistent estimation of the nonlinear model in Eq.(2) by including the zero migration flows and effectively dealing with heteroskedasticity. In addition, the PPML estimator is consistent in the presence of fixed effects, while other nonlinear fixed effect model suffers from incidental parameters problem, which introduces bias. The PPML estimator is superior to other alternative count data models since it is consistent regardless of how the data are in fact distributed. Using PPML is also more convenient than using the Heckman selection model to deal with selection since it does not require one to know selection variables in advance, and it additionally deals with heteroskedasticity.

In the baseline specification, I instrument the contemporary linguistic distance, both for the linear and quadratic terms, with historical linguistic distance. I apply a control function method following Lin and Wooldridge (2019) and Wooldridge (2015).

Specifically, in the first stage, I run the following regression,

$$Ldist_{ij} = \delta_i + \delta_j + \beta_l Ldist1930_{ij} + \beta_g Gdist_{ij} + \beta_h Home_{ij} + \gamma X_{ij} + \varepsilon_{ij}$$

then I get the residuals $\hat{\varepsilon}_{ij}$ from the first-stage regression, and run the following regression using $\hat{\epsilon}_{ij}$ as the control function,

$$\pi_{ij} = \exp\left\{\delta_i + \delta_j + \beta_l Ldist_{ij} + \beta_{l2} Ldist_{ij}^2 + \beta_g Gdist_{ij} + \beta_h Home_{ij} + \gamma X_{ij}\right\} + \hat{\varepsilon}_{ij} + \epsilon_{ij}$$

and I bootstrap the standard errors.

C.2 The specification choice in detecting nonlinearity

I base my specification firstly on the theoretical predictions discussed in Section 5.1; secondly from the existing literature, especially for the choices and operation of pair-wise controls besides linguistic distance ((Krieger et al., 2018); (Yotov et al., 2016)); thirdly, directly from the data.

Appendix Figure A.5 shows the bin-scatter plots of migration probability (purging out geographic distance and origin and destination fixed effects) against the linguistic distance (purging out geographic distance and origin and destination fixed effects). For linguistic distance residuals that are relatively small, we can see that the residuals (migration probability that other controls cannot explain) increase with the linguistic distance residuals. When linguistic distance is larger, there is a negative relationship between the residual migration probability and linguistic distance.

I also conduct sensitivity analysis by adding more polynomials into the regressions. Appendix Table A.3 shows that adding a cubic term or a fourth-order term does not affect the results.

C.3 Robustness of Instrument

Considering the historical migration rate is not zero and historical can affect current migration through historical migration networks, it is possible that the historical linguistic distance cannot fully control the potential endogeneity. In this section, I conduct the following robustness checks.

First, according to the 1930 Colonial Census, in 1930, the major migrant exporting areas includes Batavia (Jakarta), Kedoe (Kebumen, Wonosobo, Temanggung, Purworejo and Magelang), and Kediri, and the major migrant importing areas includes Riau, Jambi and Lampung. Migrants to or from those locations accounted for more than 80% of the total migration in 1930. So I exclude those areas from the baseline regression in Column 4 of Table 3. The result is shown in Column 2 of Appendix Table A.4. The effects of linguistic distance fall just slightly, but it is mainly in line with the baseline results.

Second, the primary concern about the instrument is that the historical linguistic distance can affect the historical migration network and thus affect current migration patterns. So I further control migration networks using the previous (lifetime) migration propensity in 2000 as a proxy for migration networks following Beine et al. (2011). A caveat is that introducing migrant networks to the model is potentially endogenous and explains part of the effect of linguistic distance on migration. The result is shown in Column 3 of Appendix Table A.4. Again the coefficients of linguistic distance are slightly smaller in magnitudes but generally consistent with the baseline.

D Model Appendix

D.1 Derivation of Demand for Goods

Worker's utility is,

$$U_j^e = \left(\sum_i q_{ij}^e \frac{\sigma-1}{\sigma} \right)^{\frac{\sigma}{\sigma-1}} u_j^e \tag{27}$$

where σ is the degree of substitutability between products produced by different locations, q_{ij} is the quantity of variety produced in i and consumed in j. u_j^e is the local amenity.

and the budget constraint is,

$$\sum_{i \in N} p_{ij} q_{ij}^e = w_j^e \tag{28}$$

By maximization of utility (27) constrained by (28) for each worker, a worker of skill e living in j spend $q_{ij}^e = p_{ij}^{1-\sigma} P_j^{\sigma-1} w_j^e$ on variety produced by i. Aggregate it for all local population,

$$X_{ij} = p_{ij}(q_{ij}^s L_j^s + q_{ij}^u L_j^u) = p_{ij}^{1-\sigma} P_j^{\sigma-1} Y_j \tag{29}$$

D.2 Formal Definition of Equilibrium Conditions

The equilibrium can be formally expressed as follows,

$$w_i^s L_i^s + w_i^u L_i^u = \sum_{d \in N} X_{ij} = \sum_{d \in N} \tau_{ij}^{1-\sigma} \left\{ \frac{1}{A_i} \left[(\theta_i^s)^\rho (w_i^s)^{1-\rho} + (\theta_i^u)^\rho (w_i^u)^{1-\rho} \right]^{\frac{1}{1-\rho}} \right\}^{1-\sigma} P_j^{\sigma-1} Y_j \tag{30}$$

$$P_i = \left(\sum_{d \in N} {\tau_{ji}}^{1-\sigma} \left\{ \frac{1}{A_j} \left[(\theta_j^s)^\rho (w_j^s)^{1-\rho} + (\theta_j^u)^\rho (w_j^u)^{1-\rho} \right]^{\frac{1}{1-\rho}} \right\}^{1-\sigma} \right)^{\frac{1}{1-\sigma}} \tag{31}$$

$$L_i^s = \sum_{d \in N} \frac{(\bar{u}_i^s w_i^s / P_i D_{ji}^s)^\eta}{(\Phi_j^s)^\eta} L_j^{s0} \tag{32}$$

$$\Phi_i^s = \left(\sum_{d \in N} (\bar{u}_j^s w_j^s / P_j D_{ij}^s)^\eta \right)^{\frac{1}{\eta}} \tag{33}$$

$$L_i^u = \sum_{d \in N} \frac{(\bar{u}_i^u w_i^u / P_i D_{ji}^u)^\eta}{(\Phi_j^u)^\eta} L_j^{u0} \tag{34}$$

$$\Phi_i^u = \left(\sum_{d \in N} (\bar{u}_j^u w_j^u / P_j D_{ij}^u)^\eta \right)^{\frac{1}{\eta}} \tag{35}$$

Rewrite the equilibrium equation (30)-(35), and denote $p_i = \frac{1}{A_i} \left[(\theta_i^s)^\rho (w_i^s)^{1-\rho} + (\theta_i^u)^\rho (w_i^u)^{1-\rho} \right]^{\frac{1}{1-\rho}}$, and $\omega_i^e = \bar{u}_i^e w_i^e / P_i$ for $e \in \{s, u\}$, the following system is used to invert the price and welfare composites in Section 7.

$$p_i^{\sigma-1} = \sum_{j \in N} T_{ij} \left(\frac{Y_j}{Y_i} \right) P_j^{\sigma-1} \tag{36}$$

$$\left(P_i^{\sigma-1} \right)^{-1} = \sum_{j \in N} T_{ji} (p_j^{\sigma-1})^{-1} \tag{37}$$

$$[(\omega_i^s)^\eta]^{-1} = \sum_{j \in N} M_{ji}^s \left(\frac{L_j^{s0}}{L_i^s} \right) [(\Phi_j^s)^\eta]^{-1} \tag{38}$$

$$(\Phi_i^s)^\eta = \sum_{j \in N} M_{ij}^s (\omega_j^s)^\eta \tag{39}$$

$$[(\omega_i^u)^\eta]^{-1} = \sum_{j \in N} M_{ji}^u \left(\frac{L_j^{u0}}{L_i^u} \right) [(\Phi_j^u)^\eta]^{-1} \tag{40}$$

$$(\Phi_i^u)^\eta = \sum_{j \in N} M_{ij}^u (\omega_j^u)^\eta \tag{41}$$

D.3 Estimate parameters

D.3.1 Estimation of the Substitutability of Skilled and Unskilled Workers: ρ

By dividing skilled labor demand Eq.(12) and unskilled labor demand Eq.(13), I can derive relative labor demand curve as follows,

$$\ln \frac{w_j^s}{w_j^u} = \ln \frac{\theta_j^s}{\theta_j^u} - \frac{1}{\rho} \ln \frac{L_j^s}{L_j^u} \tag{42}$$

To get an unbiased estimate of ρ, I use rainfall shocks from migration sources as labor supply shifters following Kleemans and Magruder (2018). To do this, I first add a time subscript t to Eq.(42) and use labor and rainfall data from 2000 to 2010.

$$\ln \frac{w_{jt}^s}{w_{jt}^u} = -\frac{1}{\rho} \ln \frac{L_{jt}^s}{L_{jt}^u} + X_{jt-1} + \alpha_t + \alpha_j + \epsilon_{jt} \tag{43}$$

Then I construct the rainfall supply shock as follows,

$$RainShock_{jt} = \sum_{i \in C(j)} MigrantShare_{ij} Rainfall_{it} \tag{44}$$

where $C(j)$ is a set of locations that send migrants to location j, which I henceforth refer to as catchment area, and $MigrantShare_{ij}$ is the share of migrants from i in j in 1995-2000, which is predetermined compared to the rainfall and labor data (2000-2010). $Rainfall_{it-1}$ is the rainfall in location i in the previous year $t-1$, and I standardize it to have a mean of zero and a standard deviation of 1 following the literature. Literature has shown that climate shocks stimulate migration, and the effects are different for different population (Barrios et al. 2006; Kleemans and Magruder 2018). For X_{jt-1}, I also control location j's own rainfall in year $t-1$, since it directly affects the labor demand and labor supply in location j.

The exclusion restriction relies on the assumption that $RainShock_{jt}$, which is a weighted rainfall shock in migration sending locations for location j, only affects the wage premium $(\ln w_{jt}^s/w_{jt}^u)$ through relative labor supply $(\ln L_{jt}^s/L_{jt}^u)$ and does not affect the labor demand in j directly. Appendix Table A.12 shows the results. The estimated elasticity of substitution of 1.24 is comparable but smaller than 1.41 in the US (Katz and Murphy, 1992), 1.24 in China (Khanna et al., 2021), and 1.7 in India (Khanna and Morales, 2017).

D.3.2 Recovering agglomeration parameters α, Armington elasticity of substitution σ and dispersion parameter η^e

I follow Allen and Arkolakis (2018) and Allen and Donaldson (2020) to recover location fundamentals ($\bar{u}_i^e$ and $\bar{A}_i$), agglomeration Parameters α, Armington elasticity of substitution σ and dispersion parameter η^e. Here I assume

η^e is different for skilled and unskilled workers. To begin with, I calibrate an initial guess of the key parameters from the literature, as shown in Table 5.

Estimation of Dispersion Parameters: σ, η^e and α uisng the model structure After step 7, I can use Eq.(20), Eq.(21), and Eq.(22) to estimate σ, η^e, and α. However, the residuals in Eq.(20)-Eq.(22), $\ln \bar{A}_i$ and $\ln \bar{u}_i^e$, determine wage and population through general equilibrium, so the OLS estimates are biased. To recover unbiased estimates for σ, η^e, and α, I follow Allen and Arkolakis (2018) to get a set of instruments for endogenous wage, price and population. First, I regress the inverted $\bar{u}_i^e$ and $\bar{A}_i$ on geographic variables, and get predicted $\hat{\bar{u}}_i^{\,e}$ and $\hat{\bar{A}}_i$, which capture the part of exogenous amenities and productivity that is determined by observed geographic characteristics (distance to coast, distance to rivers, ruggedness, elevation, and soil quality. I also assume migration cost is sorely determined by geographic distance $\hat{M}_{ij}$. Then I plug back the predicted $\hat{\bar{u}}_i^{\,e}$, $\hat{\bar{A}}_i$ and $\hat{M}_{ij}$ into the model structure and attain a set of equilibrium $\hat{w}_i^{\,s}$, $\hat{w}_i^{\,u}$, $\hat{L}_i$, and $\hat{P}_i$ of a hypothetical world where local fundamentals and migration frictions are solely determined by observed geographic variables.

Then I use those hypothetical endogenous variables as instruments for the actual wage, price and population in estimating (20), (21), and (22). Table A.16 shows the estimation results of Eq.(20), Eq.(21), and Eq.(22) using these model-based instruments.

D.4 Algorithm for Counterfactuals

Now I know the exogenous variables set $\{A_i, \bar{u}_i^s, \bar{u}_i^u\}$, and parameters $\{\sigma, \alpha, \eta, \rho, \theta_s, \theta_u\}$. Suppose M_{ij}^s and M_{ij}^u change to $\hat{M}_{ij}^s$ and $\hat{M}_{ij}^u$. The following are the steps for updating variables $\{\hat{X}_i\}^t$, where t denotes an iteration. To start with, I take a first guess of $\{w_i^s, w_i^u, P_i\}$.

First, update counterfactual migration flows and population for both skilled and unskilled workers $\{L_i^e\}^1$, and welfare index $\{\Phi_i^e\}^1$ according to equation 32 to 34;

Second, calculate $\{Y_i\}^1$ from $\{L_i^e\}^1$ in frist step and initial guess of w_i^s, w_i^u, plug into Eq.(30) and get $\{p_i\}^1$.

Third, plug $\{p_i\}^1$ into Eq.31 and get $\{P_i\}^1$.

Lastly, update wages using labor equilibrium conditions in Eq.(12) and Eq.(13).

So far I have update all endogenous variables once. Now I calculate the different of $\{\hat{X}_i\}^1$ and $\{\hat{X}_i\}^0$. I repeat the iteration using the updated guess of $\{w_i^s, w_i^u, P_i\}$ until the difference is smaller than the tolerance level $(1e-9)$.

D.5 Measure Welfare, Production and Inequality

Welfare By the properties of the Frechet distribution, the expected regional welfare for location i is

$$\mathbb{E}(V_i) = \Gamma(1 - \frac{1}{\eta})\Phi_i$$

where $\Phi_i = \left[\sum_{j'} \left(\frac{\tilde{w_{j'}^{\,}} u_{j'}}{D_{ij'}}\right)^{\eta}\right]^{\frac{1}{\eta}}$ summarize the appeal of all migration options for location i. This welfare is the welfare before individual makes any migration choices (before the realization of idiosyncratic preference).[60]

The aggregate welfare is the sum of the expected regional welfare across all regions, weighted by initial population share.

Production For production, the real GDP is defined as nominal GDP deflated by the overall local price index,

$$\widetilde{Y}_i = \frac{Y_i}{P_i} = \frac{w_i^s L_i^s + w_i^u L_i^u}{P_i}$$

and the aggregate production is

[60] See Tombe and Zhu (2019) for derivation of the expression.

$$\widetilde{Y} = \sum_{i}^{N} S_i \widetilde{Y}_i$$

where S_i is the initial contribution of location i to national real GDP.

Regional Inequality To measure the inequality across regions, I use the Theil index defined. as follows,

$$T = \sum_{i=1}^{N} w_i V_i \ln V_i \tag{45}$$

where N is the set of all location, w_i is the population share of location i, and V_i is the expected welfare level in location i.

Skill Inequality I use the skilled welfare over unskilled welfare V_i^s / V_i^u to measure skill inequality. The overall skill inequality is the sum of the expected skill inequality across all regions, weighted by initial population share.

D.6 Further Reduction in Linguistic Distance and Geographic Distance

As mentioned in Section 8.1, 6.2% is a lower bound of reduction in linguistic distances as many policies can effectively reduce linguistic barriers. In this section, I estimate the welfare implications by a larger percentage reduction in linguistic barriers. For comparison, I also conduct a set of counterfactuals where I impose a similar percentage reduction in geographic barriers. Appendix Figure A.12 plots the aggregate welfare changes against the percentage reduction in distances, where the solid line for linguistic distance and dash line for geographic distance. Several interesting patterns emerge. On the one hand, reduction of linguistic distance generally generates smaller welfare gains than a similar percentage reduction in geographic barriers. On the other hand, as further reduction in geographic barriers steadily generates more welfare gains, the marginal benefits of reduction in linguistic barriers decreases.