Language Barriers, Internal Migration, and Labor Markets in General Equilibrium*

Amrutha Manjunath†
University of Nottingham

July 30, 2026

Abstract

I study how language barriers impact aggregate and distributional economic outcomes using rich microdata from India applied to a quantitative spatial general equilibrium framework. I first document three empirical facts, which are strongest for unskilled workers: first, workers migrate less often to locations where they face high language barriers; second, migrants with high language barriers are employed less often in speaking-intensive occupations; and third, migrants with high language barriers get a wage premium. To rationalize these facts, I then develop and estimate a static migration model in which heterogeneous workers sort across occupations and locations by skill and language, with wages accounting for worker selection and adjusting in general equilibrium. I show through the lens of the model how language barriers, by increasing worker sorting and selection, significantly obstruct internal migration and augment the skill premium; removing them would raise aggregate welfare by 10.9 percent. Finally, I calibrate costs of both program provision and learning languages to evaluate potential benefits of language programs for unskilled migrants. Using the calibrated model, I argue that welfare benefits of implementing language programs would outweigh costs.

Keywords: Language Barriers, Internal Migration, Spatial and Occupational Sorting, Selection, Skill Premium, Spatial General Equilibrium
JEL Codes: R13, R23, J61, J24, J31, O15

*First Draft: October 2024. I am indebted to my advisor Fernando Parro and mentors Enghin Atalay, Jonathan Eaton, Jingting Fan, Kala Krishna, and Jim Tybout for their advice and encouragement at every stage of research. I am grateful to Salomé Baslandze, Klaus Desmet, Simon Fuchs, Devaki Ghose, Hugo Lhuillier, Veronika Penciakova, Michael Peters, Joris Pinkse, Eric San Miguel, Bradley Setzler, Sebastian Sotelo, Stephen Yeaple, and Kai-Jie Wu and participants at workshops, conferences and seminars at Penn State, FRB Philadelphia, FRB Atlanta, UEA European Meeting 2024, UEA North American Meeting 2025, Nottingham, Essex, Birmingham, ISB Hyderabad, IIM Bangalore, Ashoka, RFBerlin, UCSD-GPS, among others, for their generous comments and suggestions. I thank the Data Dissemination Unit at the Office of the Registrar General of India for help with accessing confidential data and Hailey Gretz for excellent research assistance.

†School of Economics, University of Nottingham. Email: amrutha.manjunath@nottingham.ac.uk.

1 Introduction

The world is highly multilingual, with over 7,000 languages spoken today. In many countries, there is no *lingua franca*. In these countries, language barriers augment the cost of migration to a degree comparable to, and beyond, that of geographic barriers.[1] Migrants' decisions on where to move and in which occupations to work depend on the language(s) spoken in the destination. The implications of this manner of sorting may be consequential in labor markets that increasingly reward communication skills.[2] Further, as skilled workers are more likely to be proficient in languages spoken across labor markets (e.g., English, French), language barriers may contribute to the skill premium. This makes language a critical factor in understanding and mitigating economic disparities within multilingual countries.

Despite the prevalence of multilingual countries, the macroeconomic consequences of language barriers remain largely unexplored. In this paper, I ask: How do language barriers impact internal migration, the skill premium, and aggregate welfare?[3] How could language policy mitigate these effects? To answer these questions, I show both theoretically and empirically that language barriers impact labor market outcomes and significantly obstruct internal migration patterns. I proceed in three parts.

First, I document three empirical facts that relate language barriers to migration patterns and labor market outcomes of workers by skill. Second, to explain these facts, I develop a quantitative spatial general equilibrium model in which language is both a component of bilateral migration cost and a technological friction that varies with the speaking intensity of the destination occupation, and structurally estimate the model on Indian microdata. Third, I use the estimated model to perform two sets of counterfactual exercises: the first decomposes the aggregate, distributional, and spatial general equilibrium effects of language barriers, and the second weighs the welfare benefits of a language program against its implementation costs.

I begin by leveraging confidential, highly disaggregated district-to-district migration data from the Census of India. Specifically, this paper marks the first use of the 2011 wave of this restricted data, which consists of the universe of internal migrants and provides the joint distribution of their origins, destinations, and key demographic characteristics. To control for historical migration networks, I complement this with earlier restricted data from 2001. I further merge this with census data on languages spoken across districts and household surveys that provide information on workers' occupations and wages, among other demographic and labor market characteristics.

India's linguistic diversity, with over 270 languages spoken by significant populations, and

[1] For example, Kone et al. (2018) empirically demonstrate using a gravity specification that having a common language across districts in India is equivalent to reducing geographic barriers by 48 percent.

[2] Occupations that are intensive in communication skills have been shown to pay 6–11 percent higher wages across varying degrees of cognitive skills in the context of the United States (Deming, 2017).

[3] In this paper, welfare is defined as aggregate real income in the economy. This measure does not capture benefits from cultural diversity or multilingualism.

persistently low internal mobility provide an ideal setting for this research. Moreover, in India, proficiency in English varies substantially with education levels: it is widespread among college-educated workers but limited among those without college education (Azam et al., 2013). This provides additional context in which to examine the differential impact of language barriers on skilled and unskilled workers. I follow Fearon (2003) and measure language barriers by constructing a linguistic distance index between district pairs based on shared language branches and speaker populations.

Using the Indian microdata, I document three empirical facts that relate language, location, and labor market outcomes. In the first fact, I show that workers migrate less often to locations where they face high language barriers. Language may be operating through two distinct channels here: as a component of migration cost by making relocation more challenging, or as a determinant of labor market outcomes by changing workers' employment prospects and earnings potential. Further, this pattern is less pronounced for skilled (college) than unskilled (non-college) workers.

The next facts show how language barriers shape migrants' occupations and wages, and once again, how these differ between skilled and unskilled workers. In the second fact, I show that among migrant workers, those with high language barriers are employed less often in speaking-intensive occupations, consistent with diminished productivity in occupations that require communication.

In the third fact, I show that language barriers also generate a wage premium within comparable occupations, acting as a migration cost that screens out all but the most productive migrants. That is, only selected migrants who expect to earn higher wages choose to move despite these barriers. Crucially, both the occupation and wage patterns are strongly attenuated for college-educated migrants. This reflects how higher education in India provides proficiency in English, which reduces skilled workers' dependence on local languages.

Motivated by the empirical facts, I build a static quantitative spatial general equilibrium model of migration. In this framework, language is both a component of migration cost and a technological friction. Heterogeneous workers (skilled and unskilled) may face language barriers at potential destinations and choose between heterogeneous occupations (speaking and non-speaking) across locations. In addition, the production structure allows for complementarities by skill and occupation type. These features allow me to flexibly uncover the relationship of language to the labor market.

In the model, workers choose destinations and occupations based on their comparative advantage and subject to linguistic and geographic migration costs and amenities. This sorting mechanism explains why migrants are less likely to move to locations with high language barriers and are less frequently employed in speaking occupations at these locations (empirical facts 1 and 2). The model also accounts for selection effects. That is, since language barriers increase migration costs, only workers with higher potential wages choose to migrate to these destinations, explaining

the observed wage premium (empirical fact 3). The model incorporates labor demand considerations by allowing firms to differentiate between workers based on skill and language. This reflects varying linguistic and skill requirements across occupations that contribute to equilibrium wage adjustments.

The mechanisms in the model underscore the interplay between sorting and selection mechanisms in general equilibrium: removing language barriers can enhance productivity (increasing wages), and yet simultaneously eliminate the positive selection effect of language and potentially increase labor supply (decreasing wages). These countervailing forces render the net impact on wages theoretically ambiguous, making a structural model essential to determine the overall effects. The results suggest that the productivity-enhancing reallocation of workers dominates the downward pressure on wages.

I take the model to microdata from India and estimate novel elasticities of substitution between occupation and worker types. I find that, conditional on occupation and skill group, workers with and without language barriers are close substitutes in production, so the productivity consequences of language barriers operate through the estimated productivity terms rather than through imperfect substitutability. I also estimate the average productivity of worker types in speaking occupations. I find that unskilled workers who face language barriers suffer an efficiency penalty of roughly 40 percent in speaking occupations relative to non-speaking occupations, but this is not so for unskilled workers without language barriers or for skilled workers. Thus, the estimates corroborate my hypothesis that language barriers depress productivity for unskilled rather than skilled workers, and in speaking rather than non-speaking occupations.

Using the estimated model, I conduct two sets of counterfactual exercises. In the first set, I find that removing language barriers increases internal migration by 14.6 percentage points, decreases the skill premium by 3.8 percentage points, and increases aggregate welfare by 10.9 percent. I decompose these aggregate numbers into a migration cost channel and a productivity friction channel: the migration cost drives most of the welfare gain and essentially all of the migration response, while the productivity friction is the larger direct force compressing the skill premium. The joint gain exceeds the sum of the two channels in isolation, reflecting a complementarity between sorting workers into origin-destination pairs with high language barriers and removing the productivity penalty they face there.

The aggregate welfare gain from removing language barriers is comparable to, and somewhat larger than, the gain from reducing India's geographic frictions to a benchmark based on the United States (10.9 versus 8.4 percent), and has opposite-signed incidence on the skill distribution: removing language barriers compresses the skill premium while reducing geographic frictions slightly widens it. These aggregate numbers conceal substantial spatial heterogeneity: regional welfare gains concentrate in the linguistically distinct south and northeast, while the Hindi-speaking belt is flat to slightly negative.

In the second set of counterfactuals, I consider the cost-effectiveness of policies to reduce language barriers.[4] To implement this exercise, I extend the model by introducing language programs offered by the government to unskilled migrant workers facing language barriers. I incorporate two kinds of costs: the opportunity cost of learning languages and the cost of program provision. Unskilled migrants moving to regions with different languages weigh the benefit of participating in the labor market without language barriers against the opportunity cost of overcoming them. The government finances the cost of program provision by levying a uniform tax on all workers in the economy.

I first calibrate the costs of the program. I measure the opportunity cost of learning using information on time taken to overcome language barriers and wages forgone, where learning time increases with linguistic distance between origin and destination. Using the model and incorporating this cost, I predict the number of non-college workers by origin that would enroll in the program at each destination. Second, I calibrate the cost per student of program provision using budget reports from India's Central Institute of Indian Languages (CIIL), which runs language training programs across seven regional centers.

I then use the calibrated model to compute the welfare benefit of the program subject to these costs. Welfare under the policy is aggregate real income net of total program cost. The calibrated model implies a break-even cost of approximately $73 per hour of instruction. The cost per student computed from CIIL budget reports is approximately $5.2 per hour, well below this threshold, and the policy generates a welfare gain of 6.1 percent at this calibrated cost, approximately 56 percent of the gain from removing language barriers entirely. Thus, language programs would deliver aggregate welfare gains well above their calibrated costs.

Together, these findings establish language as a first-order spatial and labor market friction in multilingual economies undergoing structural transformation. The decomposition into a migration cost channel and a productivity friction channel clarifies how language shapes inequality and motivates language policy as a substantive component of inclusive development strategy.

In sum, this paper makes four contributions. First, I contribute new measurement. This paper marks the first use of the confidential 2011 district-to-district migration data from the Census of India, which I link to a bilateral linguistic distance index and to microdata on occupations and wages. Second, I document three new empirical facts that connect language barriers to internal migration, occupational sorting, and wage selection, and I show that each pattern is strongest for unskilled workers. Third, I develop and estimate the first quantitative spatial general equilibrium model that microfounds language as both a bilateral migration cost and an occupation-specific technological friction. The estimation delivers novel parameters, including the productivity penalty that language barriers impose in speaking-intensive occupations and the elasticity of substitution between

[4]This is motivated by language programs facilitated by the Federal Office for Migration and Refugees in Germany. These programs aim to help migrants achieve working proficiency in German for the purpose of labor market integration.

workers with and without barriers. Fourth, I quantify the aggregate stakes and the policy response. Language barriers suppress welfare by as much as India's geographic frictions benchmarked to United States levels; unlike geographic distance, they widen the skill premium. I also provide the first general equilibrium welfare evaluation of language training as a labor market policy, showing that benefits exceed calibrated costs by a wide margin.

Related Literature: This paper is related to several strands of literature. First, it contributes to the literature in economic geography and macroeconomic development that studies the aggregate and distributional implications of spatial frictions. Similar to Tombe and Zhu (2019), Fan (2019), and Allen et al. (2018), the model features multiple regions with costly labor mobility. While previous dynamic and static migration models have heavily emphasized geographic barriers (Kennan and Walker, 2011; Monras, 2018; Caliendo et al., 2019; Desmet and Parro, 2025), language has largely been ignored. In quantifying how spatial frictions limit the aggregate productivity gains of workers sorting by comparative advantage, this paper builds on Lagakos and Waugh (2013), Bryan and Morten (2019), Hsieh et al. (2019), Burstein et al. (2019), Burstein et al. (2020), and Bratsberg et al. (2023).

Furthermore, a growing subset of this literature examines how these labor market frictions shape spatial and skill-based inequality, particularly in India (Desmet et al., 2015; Fan et al., 2023; Ghose, 2026). Relative to this work, I extend the lens of comparative advantage to a third dimension: language. By microfounding language as both a bilateral movement friction and an occupation-specific technological penalty, I study not only its aggregate macroeconomic costs, but also how it drives heterogeneous outcomes for skilled versus unskilled workers against the backdrop of service-led growth. This approach contrasts with Wang (2026), who treats linguistic distance as an ex-post correlate of spatial parameters in Indonesia; this paper instead isolates the structural communication friction underlying India's internal labor markets.

Second, this paper contributes to empirical work studying the role of language in trade, migration, and the labor market. Melitz (2008) introduced language to empirical gravity models of international trade, and Melitz and Toubal (2014) further show, using linguistic proximity measures, that the aggregate impact of linguistic factors on trade is twice that implied by a common-language gravity control. Desmet et al. (2012) construct linguistic-distance measures from Ethnologue trees and link them to political and redistributive outcomes. Adserà and Pytliková (2015) document the empirical significance of language in shaping international migration patterns. Dustmann and Fabbri (2003) study the importance of language proficiency of migrants to the UK in shaping their labor market outcomes, and Chiswick and Miller (2015) survey the economics of language in international migration. Peri and Sparber (2009) attribute the sorting of comparably skilled migrants away from communication-intensive tasks to imperfect substitutability between migrants and natives in partial equilibrium.

In the context of internal migration in India, Kone et al. (2018) and Imbert and Papp (2020)

provide evidence on how migration flows empirically relate to both distance and linguistic differences, and Chaudhary et al. (2025) document the long-run persistence of these barriers by combining colonial 1901 and contemporary 2001 district-pair data in a gravity framework. This paper embeds the linguistic-distance measure of Fearon (2003) in a structural framework that quantifies the labor market and welfare consequences of language barriers in general equilibrium.

Third, this paper builds on the literature studying language skills as a form of human capital investment. For example, Bleakley and Chin (2004) and Adda et al. (2022) show that language proficiency represents a crucial form of location-specific human capital that affects earnings and labor market outcomes. In the Indian context, Azam et al. (2013) show that returns to English language skills vary significantly across regions and occupations. This paper extends the location-specific human capital view of language by allowing the productivity penalty to vary with the speaking intensity of the destination occupation, so that the same linguistic distance is more binding in occupations that require communication.

The rest of this paper is structured as follows. Section 2 describes the data, Section 3 presents the three empirical facts, Section 4 develops the model, Section 5 takes it to the data, Section 6 conducts the counterfactual exercises, and Section 7 concludes.

2 Data

A significant challenge for empirical analysis of the impact of language barriers on internal migration and labor market outcomes is the paucity of comprehensive and disaggregated data. To address this, I obtain district-to-district internal migration data from the Census of India (2001, 2011) and combine this with data on language and labor market outcomes. I further create a concordance between occupations in the Indian household surveys and the Occupational Information Network (O*NET) to obtain a ranking of speaking intensity over occupations.

2.1 Data on Migration

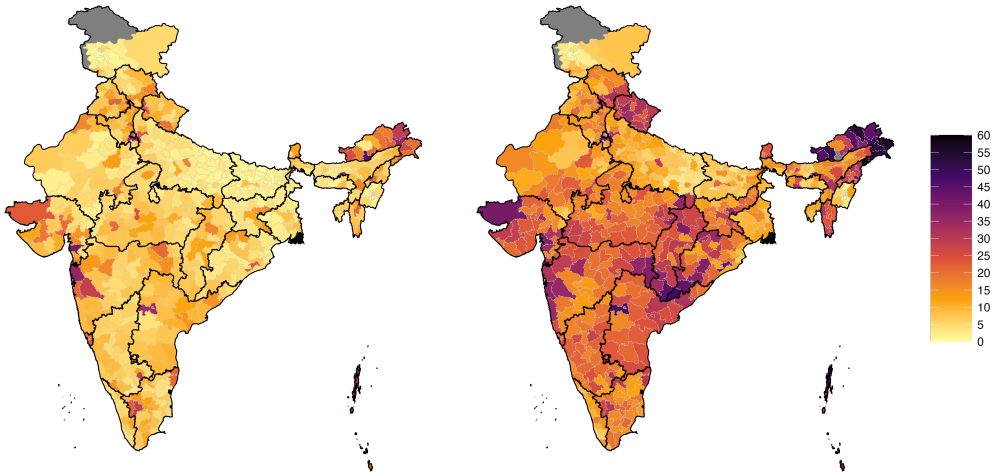
The publicly available data in the Census of India (2001, 2011) contains information for each destination district on migrants' education level, age, reason for migration, and duration of stay. However, analyzing the effects of language barriers on migration decisions requires identifying the district of origin jointly with these variables. Because this is not possible in the public release, I obtained confidential, highly disaggregated district-to-district migration data from the Office of the Registrar General of India.

This paper marks the first use of the 2011 wave of this data, which records migration flows jointly by education level, age, reasons for migration, and duration of stay. I rely primarily on this

novel 2011 cross-section to establish the main empirical facts.[5] In addition, to control for historical bilateral migration networks in robustness checks, I harmonize the earlier 2001 wave with the 2011 data to construct a separate district-level panel.

Since the focus of this paper is the interaction of language barriers with labor market outcomes, I focus my analysis on migration for work or business. These migrants make up about a tenth of all internal migrants and are the most likely to move across state lines, where they are most exposed to language barriers. And yet, work migrants largely move within their own states, either within their own district or to another district in the same state. In 2001, 65.9 percent of all migrant workers moved within their own state. In 2011, this was even higher, at 73.4 percent. This pattern of intra-state moves being more common than inter-state moves holds for both non-college (73.1 percent) and college migrant workers (75.7 percent) in 2011. Table 8 in Appendix A.1 reports the composition of internal migrants by all reasons, and Table 9 reports the full geographic decomposition of work migrants.

Figure 1: Share of Inter-District Migrant Workers in Population in 2011



(a) Inter-District Migration: Non-College (%) **(b) Inter-District Migration: College (%)**

Notes: Data from the district-to-district migration tables from Census 2011. The maps display the share of workers in each district who migrated from other districts with separate visualizations for college and non-college educated workers. For each district, the percentages shown are calculated by dividing the number of non-college and college inter-district migrants by the total non-college and college worker population, respectively, in the destination district. Districts are shaded on a gradient from light yellow (low migration rates around 0 percent) to deep purple (high migration rates around 60 percent), with darker colors indicating higher shares of migrant workers.

These aggregate statistics mask considerable spatial heterogeneity. Figure 1 maps the inter-district non-college and college migrant workers as a share of, respectively, the destination district population of non-college and college workers in 2011. For non-college workers, inter-district

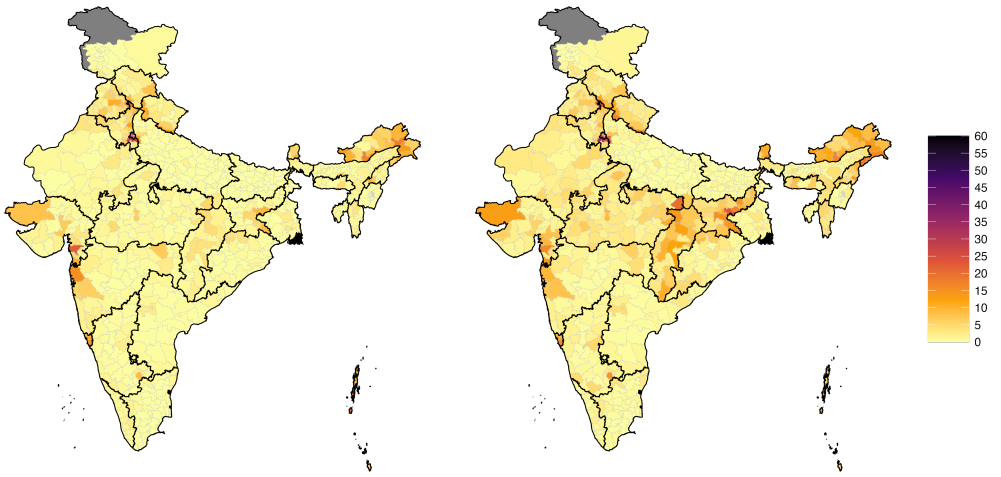
[5]Research papers on internal migration using 2001 district-to-district data are also relatively scarce. Recent papers that have used the 2001 data to study migration frictions in India are Kone et al. (2018), Rai (2023), and Ghose (2026).

migration remains relatively modest across most of India, with most districts showing rates below 15 percent. However, there are some notable exceptions, particularly in select districts of Northeast India and scattered urban centers, where migration rates reach higher levels.

On the other hand, college workers demonstrate higher mobility, with many districts showing migration rates of 20-30 percent, compared to the generally lower rates for non-college workers. This heightened mobility is especially pronounced in major urban and economic centers, districts in Northeast India, clusters of districts in central India, and select coastal regions. This contrast suggests that educational attainment significantly influences labor mobility in India, with college education appearing to substantially reduce barriers to inter-district movement.

Figure 2 maps the inter-state non-college and college migrant workers as a share of, respectively, the destination district population of non-college and college workers. These maps show that college-educated workers demonstrate higher inter-state mobility compared to their non-college counterparts. This difference is particularly pronounced in major urban and industrial centers, suggesting that higher education is associated with greater geographic mobility across state boundaries. The regions around Delhi NCR, coastal Maharashtra, and parts of Gujarat stand out with deeper orange hues in the college-educated workers' map, indicating these areas attract a substantial share of educated inter-state migrants relative to their local educated workforce.

Figure 2: Share of Inter-State Migrant Workers in Population in 2011



(a) Inter-State Migration: Non-College (%) **(b) Inter-State Migration: College (%)**

Notes: Data from the district-to-district migration tables from Census 2011. The maps display the share of workers in each district who came from other states, based on India's 2011 Census data. For each district, the percentages shown are calculated by dividing the number of non-college and college inter-state migrants by the total non-college and college worker population, respectively, in the destination district. This calculation yields the percentage of workers in each district who came from other states, computed separately for non-college and college workers. The color gradient, ranging from light yellow to darker orange, represents migration intensities from near 0 percent to up to 60 percent of the local workforce.

These patterns suggest a complex interplay between education, economic opportunities, and

geographic mobility in India. Urban and industrial centers appear to act as stronger magnets for inter-state migrants, particularly those with college education, while rural districts generally show lower migration rates across both education groups. This distribution likely reflects the concentration of formal sector jobs and higher-skill opportunities in urban areas, as well as the greater ability of educated workers to overcome the barriers associated with inter-state migration, including language.

2.2 Data on Language

India's linguistic diversity is vast, with over 1600 languages in daily use and 22 officially recognized in the Constitution (Government of India, 2017). This diversity is geographically structured, with distinct languages predominant within specific provinces. The States Reorganisation Act of 1956 reinforced this structure by redrawing colonial province boundaries to better align with linguistic regions. Consequently, inter-state migration often involves crossing linguistic boundaries, with workers entering labor markets that use a different lingua franca from their origin state.

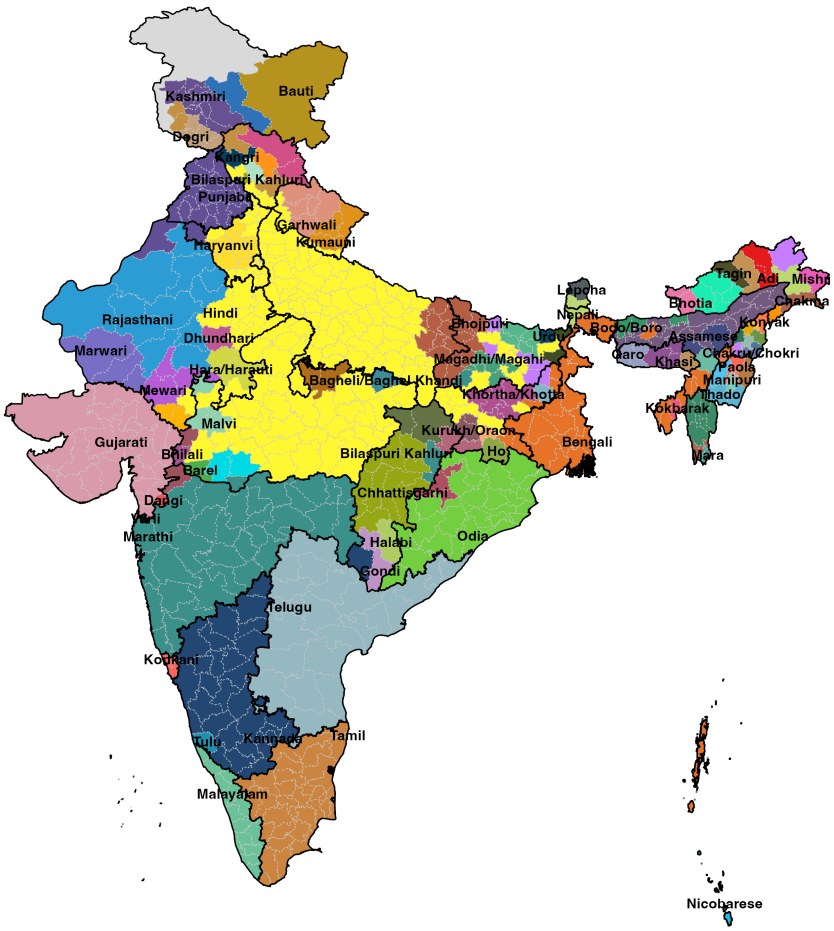
I obtain data on the distribution of languages in India from the Census of India (2001, 2011). This publicly available data contains information on the number of speakers of over 270 languages spoken by at least 10,000 people as a first language in every district. Figure 3 shows the major languages spoken across districts of India as recorded in the 2011 Census. An important feature of this data is that it accounts for heterogeneity within broad language categories such as Hindi, which is perceived as ubiquitous in the northern states. While the 2011 Census classifies 57 percent of the population as Hindi speakers, this classification subsumes over 56 dialects (including Bhojpuri, Magadhi, Rajasthani, and Chhattisgarhi), substantially overstating its ubiquity. Accounting for this heterogeneity, Hindi is spoken by 28 percent of the population.

Using data on the number of speakers of over 270 languages as mother tongue in each district and linguistic trees from Eberhard et al. (2024), I follow Fearon (2003), Spolaore and Wacziarg (2009), and Desmet et al. (2017) and construct a measure of language barriers for any region-pair o, d,

$$\text{Linguistic Distance}_{o,d} = \sum_m \sum_n (s_{o,m} \times s_{d,n} \times \text{dist. between languages}_{m,n}), \tag{1}$$

where $s_{o,m}$, $s_{d,n}$ denote the share of speakers for any language-pair m, n and the distance between them is inversely related to how many branches they have in common in linguistic trees. This index can be interpreted as the population-weighted likelihood of an average person in origin o being *unable* to communicate with an average person in destination d. A value of 0 indicates absolute convergence in languages, and a value of 1 indicates absolute divergence. This measure captures both direct and indirect communication barriers between regions, and accounts for the possibility that speakers of related languages may achieve some degree of mutual intelligibility.

Figure 3: Major Languages in Districts of India in 2011



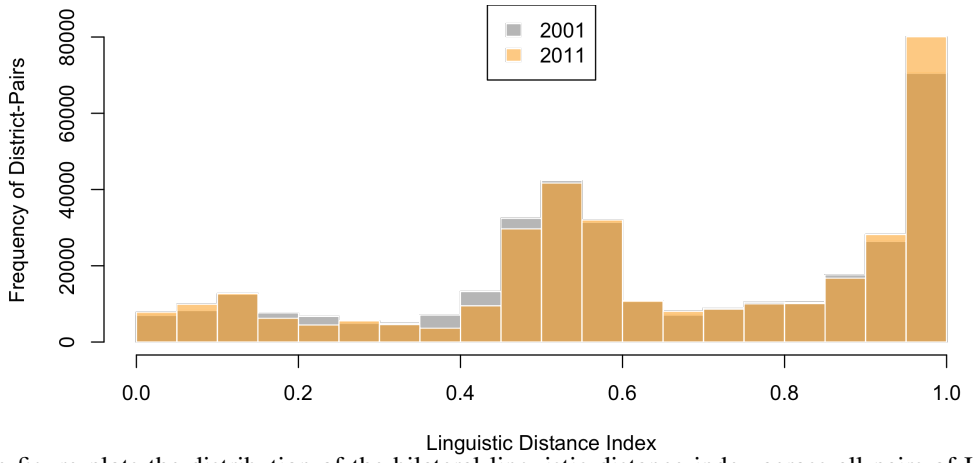
Notes: Data on languages spoken in each district from the Census of India in 2011. Each district is marked by a different color, corresponding to the language spoken at home by the largest number of people. These languages belong to five distinct language families: Indo-European, Dravidian, Austroasiatic, Tibeto-Burman, and Semito-Hamitic. Many of them are written in distinct scripts. While some language families are geographically concentrated (like Dravidian languages in the South), there is considerable linguistic variation even within regions.

Figure 4 plots the histograms of the linguistic distance index between every pair of Indian districts in 2001 and 2011. While some district pairs share similar language compositions (indices near 0), a significant portion of pairs have high linguistic dissimilarity, highlighting the considerable linguistic diversity across regions. The correlation coefficient of 0.92 between the two years indicates that these linguistic differences are highly persistent over time.

This index has been used in prior literature to study gravity relationships in international migration, international trade patterns, and market integration to show how language differences affect bilateral migration and trade flows and economic exchange between regions (Melitz, 2008; Fenske and Kala, 2021). Alternative measures of language barriers have been used in the literature, such as shared communication probability indices (Kone et al., 2018) and degrees of difference from Hindi

based on cognates and grammar (Shastry, 2012). These have been found to be highly correlated with this measure.

Figure 4: Histogram Plots of Linguistic Distance in 2001 and 2011



Notes: The figure plots the distribution of the bilateral linguistic distance index across all pairs of Indian districts, shown separately for the 2001 and 2011 Census waves. For a district pair (o,d), the index is constructed as in Equation 1: a population-weighted sum over all language pairs of the product of the mother-tongue speaker shares in each district and the linguistic distance between those languages. Speaker shares are computed from the more than 270 languages spoken by at least 10,000 people as a first language in each district, as recorded in the Census of India (Census of India, 2001, 2011). The distance between two languages is inversely related to the number of branches they share in the linguistic trees from Ethnologue (Eberhard et al., 2024), so that languages with more common ancestry are treated as closer. The index ranges from 0, when two districts share identical language compositions, to 1, when they are maximally dissimilar, and can be read as the population-weighted probability that a randomly chosen person in o cannot communicate with a randomly chosen person in d. Origin-origin pairs, in which a district is paired with itself, are excluded.

This linguistic landscape may significantly impact labor market outcomes, but less so for college-educated workers. English, though not indigenous, has become important in higher education and professional settings. Its importance varies across regions and sectors, often serving as a bridge in multilingual work environments. Workers proficient in English may have access to broader employment opportunities, particularly in sectors with international connections or requiring cross-state collaboration.

However, English is far from a lingua franca in India. The 2011 Census reported that 0.02 percent of the population spoke English as their first language, while 10.6 percent of the population spoke English as a first, second, or third language. This figure obscures substantial spatial heterogeneity, as the share of English speakers ranged from 2.3 percent in the state of Chhattisgarh to 41.8 percent in the state of Goa.

Unlike local languages, English is not tied to a regional linguistic homeland. Although its prevalence varies across space, who speaks it is determined far more by education than by origin. According to the India Human Development Survey (Desai et al., 2018), among college workers,

45 percent of natives speak English, while this share is lower for within-state migrants (27 percent) and out-of-state migrants (38 percent). Among non-college workers, English-speaking ability is rare across all groups. That is, only 2 percent of natives, 1 percent of within-state migrants, and virtually none of the out-of-state migrants speak English. These patterns suggest that English language skills are strongly correlated with education level (Azam et al., 2013) and play a potentially important role in inter-state labor mobility, particularly for college workers.

2.3 Data on Labor Market Outcomes

I use the monthly income tables of the Consumer Pyramids Household Survey (CPHS) of the Centre for Monitoring Indian Economy (2019) to obtain information on individuals' state of origin, occupations, wages, education level, and other demographic variables. The state of origin allows me to identify individual migrants that moved to the destination district from outside the state and link them to language barriers. I restrict the sample period to be from September 2017 to December 2019, as direct information on occupations was not collected prior to that. I match occupations to those in the O*NET database based on their descriptions to assign a composite 'speaking score' to each occupation.

Table 1: Occupations Ranked by O*NET Speaking Score

	Occupation	Composite Speaking Score
Speaking	Lawyers	91.5
	General Practitioners	78.0
	School Teachers	77.3
	Nurses	77.3
	Call Centre Workers	76.6
Non-Speaking	Electricians	60.3
	Software Developers	53.0
	Carpenters	53.0
	Agricultural Labourers	50.0
	Truck, Lorry Drivers	48.5

Notes: Data sourced from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with all 213 occupations in the full sample mapped to the O*NET database. The composite speaking score is an equal-weighted average of four O*NET measures (0–100): Speaking Importance, Active Listening Importance, Oral Comprehension Importance, and Oral Expression Importance. The occupation-weighted median score (≈68.9) is used to divide occupations into speaking and non-speaking categories.

Table 1 presents a sample of occupations categorized by their speaking intensity based on this score. Professional occupations like lawyers and general practitioners score at least 78, while manual and technical jobs like carpenters and truck drivers score below 55. Notably, the correlation

between skill and speaking intensity is not perfect: high-skill occupations like software developers have relatively low scores (53.0), while service jobs like call centre workers score above 76. The occupation-weighted median score is used to divide occupations into speaking and non-speaking categories. Appendix A.2 plots the full employment-weighted distribution of these scores in Figure 10, where the occupation-weighted cutoff separates a high-scoring mass of service and white-collar occupations from a lower-scoring mass of manual and agricultural workers.

Table 2: Wages and Employment in Speaking vs. Non-Speaking Occupations

	Wage Ratio		Employment Share	
	College	Non-College	College	Non-College
Speaking	1.29	1.20	0.09	0.25
Non-Speaking			0.04	0.61

Notes: Data sourced from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, restricted to workers in the 213 O*NET-matched occupations (the same sample as Figure 10). Occupations are split at the occupation-weighted median composite speaking score (≈68.9), the cutoff used in Fact 2. Wages are deflated using the monthly state-level consumer price index from the Reserve Bank of India. The wage ratio is the population-weighted mean real wage in speaking relative to non-speaking occupations, among workers with positive wages. The employment share is each speaking × education cell's share of all matched workers, weighted by the CPHS sampling weight; shares may not sum to one due to rounding.

Table 2 presents wage and employment patterns across speaking and non-speaking occupations, broken down by education level. Speaking occupations command higher wages regardless of education. That is, college workers earn a 29 percent premium and non-college workers earn a 20 percent premium in speaking relative to non-speaking occupations. The workforce composition shows that about 13 percent of workers have a college degree (9 percent in speaking and 4 percent in non-speaking occupations), while the remaining 87 percent are non-college workers.

The data also reveals that speaking occupations employ about 34 percent of all workers, while non-speaking occupations account for the remaining 66 percent of employment. Among college workers, employment leans toward speaking occupations (9 versus 4 percent of total employment), whereas non-college workers are overwhelmingly concentrated in non-speaking occupations (61 versus 25 percent of total employment).

3 Empirical Facts

In this section, I present three empirical facts about the relationship between language barriers, location choices, and labor market outcomes of internal migrant workers in India. Facts 1 through 3 document three distinct patterns on migration, occupations, and wages. Each of these patterns is systematically attenuated for college relative to non-college workers.

Fact 1 (Migration): *Workers migrate less often to locations with high language barriers, but this is attenuated for college relative to non-college.*

I use the 2011 Census, which records district-to-district migration by reason and education level, to measure both migration flows and language barriers. This allows me to focus on economic migrants, rather than those that migrate for other reasons, such as education or marriage. I measure bilateral language barriers using the linguistic distance index defined in Section 2.2, which ranges from 0 (no linguistic distance) to 1 (maximum linguistic distance).[6] For this analysis, I define a binary indicator for high language barriers, $\mathbb{1}\{\text{Barrier}_{o,d}\} \equiv \mathbb{1}\{\text{Linguistic Distance}_{o,d} \geq \text{Median}\}$, which corresponds to origin-destination pairs with linguistic distance above the full off-diagonal median. I define college-educated workers as those holding a graduate degree or higher, or a diploma certificate; all others are classified as non-college. I measure bilateral geographic barriers by the geodesic distance between geographic centers of districts o and d.

Then, I estimate the following Poisson Pseudo Maximum Likelihood (PPML) specification,

$$N_{o,d,e,g} = \exp\Big(\ln(P_{o,e,g}) + \beta_0 + \beta_1 \mathbb{1}\{\text{Barrier}_{o,d}\} + \beta_2 \mathbb{1}\{\text{College}\} + \beta_3 \mathbb{1}\{\text{Barrier}_{o,d}\} \cdot \mathbb{1}\{\text{College}\} + \beta_4 \ln(\text{Geo. Dist.}_{o,d}) + \gamma_o + \gamma_d + \gamma_g\Big) \cdot \varepsilon_{o,d,e,g}, \tag{2}$$

where $N_{o,d,e,g}$ is the number of workers at origin o with education $e \in \{\text{College}, \text{ Non-College}\}$ and gender g that migrate to destination d. Origin population $P_{o,e,g}$ enters as an offset (coefficient constrained to unity), so the regression explains the migration rate rather than the level.[7] I add origin and destination fixed effects, γ_o and γ_d, to control for time-invariant unobserved heterogeneity across districts, and gender fixed effects γ_g to account for systematic differences in migration patterns by gender.

Table 3 presents the results. Column (1) establishes that language barriers are strongly associated with lower migration flows. Moving to a high-barrier destination is associated with 97 percent lower flows ($\exp(-3.361) - 1$, $p < 0.001$). This is a large magnitude, but it may partly reflect geographic distance. Column (2) controls for $\ln(\text{Geo. Dist.}_{o,d})$, and the language barrier remains large and highly significant: migration to high-barrier destinations is still 73 percent lower, even after

[6] Mother tongue is assigned at birth and does not change with migration, so language barriers are predetermined with respect to the migration decision. The remaining concern is reverse causality: migration from o to d mechanically reduces $\text{Linguistic Distance}_{o,d}$ by increasing the share of o's languages at d. This bias attenuates estimates toward zero, and is negligible for most pairs given that the average out-migration rate to any given destination is 0.01 percent of the origin's population. Estimates are therefore conservative lower bounds on the true barrier effect. Chaudhary et al. (2025) use an alternative measure and find consistent results for both colonial and contemporary India.

[7] The offset is not redundant with origin fixed effects in PPML. Each origin contributes observations across education groups and genders sharing a single origin intercept; without the offset, the barrier–college interaction would partly reflect compositional differences in origin populations rather than differential sensitivity to language barriers.

accounting for geographic distance. While linguistic and geographic distance are correlated, language boundaries are geographically abrupt. That is, neighboring districts can belong to entirely different linguistic regions, so the two measures capture distinct dimensions of bilateral barriers.

The preferred specification in Column (3) adds the interaction with college education. The language barrier for college workers is 25 percent smaller than for non-college workers ($0.335/1.333$, $p = 0.013$). In this specification, migration to high-barrier destinations is 74 percent lower for non-college workers but only 63 percent lower for college workers, a substantial gap that persists after controlling for geographic distance.

In summary, language barriers are a powerful predictor of lower migration even after controlling for distance, and college workers appear significantly better equipped to overcome them. Appendix B.1.1 shows the same pattern holds when using the continuous linguistic distance index in place of the binary barrier indicator. Appendix B.1.2 shows that adding controls for historical migration patterns more than doubles the attenuation ratio for college workers, which becomes highly significant ($p < 0.001$), confirming that language is a specific friction differentially associated with migration by education rather than a proxy for general mobility differences.

Table 3: Language and Migration of College vs. Non-College (Fact 1)

	Dep. Var.: $N_{o,d,e,g}$		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$	-3.361***	-1.292***	-1.333***
	(0.188)	(0.147)	(0.140)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}\}$			0.335**
			(0.134)
$\mathbb{1}\{\text{College}\}$			0.547***
			(0.062)
$\ln(\text{Geo. Dist.}_{o,d})$		-1.999***	-1.999***
		(0.122)	(0.122)
Origin District FE	Yes	Yes	Yes
Destination District FE	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes
Observations	1,352,568	1,352,568	1,352,568
Pseudo R-squared	0.628	0.858	0.861

Notes: The gravity model is estimated using the PPML estimator on the universe of bilateral economic migration flows, with fixed-effects for origin district, destination district, gender, and offset by 2011 origin district population. Standard errors are clustered two-way at origin and destination districts. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Fact 2 (Occupations): *Among migrant workers, those with high language barriers choose occupations that are intensive in speaking less often, but this is attenuated for college relative to*

non-college.

To study how language barriers might shape occupational sorting, I construct a composite index of speaking intensity by linking occupations in CPHS to those in the O*NET database. I build a comprehensive concordance between CPHS occupations and specific O*NET occupation codes, matching 213 CPHS occupations. The remaining 5 unmatched categories are non-workers who are not making active occupation choices, and the sample is restricted to labor market participants. The matched occupations each have a composite speaking index, which is the equal-weighted average of four O*NET Importance scores (Speaking, Active Listening, Oral Comprehension, and Oral Expression), each normalized to a 0–100 scale.[8]

I classify a worker as employed in a speaking-intensive occupation if their occupation's composite score exceeds the occupation-weighted median (≈68.9 on a 0–100 scale). I estimate the following linear probability model through least squares,

$$\begin{aligned}\Pr\left(\mathbb{1}\{\text{Speaking Occ.}_{i,o,d,j,t}\}=1 \mid \text{Covariates}_{i,o,d,j,t}\right) &= \beta_0+\beta_1\mathbb{1}\{\text{Barrier}_{o,d}\}\\ &+\beta_2\mathbb{1}\{\text{College}_{i,o,d,j,t}\}+\beta_3\mathbb{1}\{\text{Barrier}_{o,d}\}\times\mathbb{1}\{\text{College}_{i,o,d,j,t}\}\\ &+\beta_4\ln(\text{Geo. Dist.}_{o,d})+\mu\text{Controls}_{i,o,d,j,t}+\gamma_{o,t}+\gamma_{d,t},\end{aligned} \tag{3}$$

where $\text{Controls}_{i,o,d,j,t}$ include gender, age, caste, and lagged wages. I add origin-state × month fixed effects $\gamma_{o,t}$ and destination-district × month fixed effects $\gamma_{d,t}$ to control for time-varying push and pull factors.

Table 4 presents results for migrant workers across three specifications, and progressively builds up controls. All columns include both sets of fixed effects and control for log geographic distance, lagged log real wages, gender, age, and caste. As a benchmark, among native workers the college coefficient is 0.320 ($p < 0.01$). That is, college-educated natives are 32 percentage points more likely to work in speaking-intensive occupations, reflecting the strong educational gradient in occupational sorting in India.

Column (1) establishes the migrant baseline: college-educated migrants are 24 percentage points more likely to be in speaking-intensive occupations, somewhat below the native gradient. Column (2) adds the language barrier indicator, which carries a coefficient of -0.035 ($p < 0.05$): migrants facing high language barriers are substantially less likely to work in speaking-intensive occupations, even after conditioning on geography, ability (via lagged wages), and individual characteristics.

[8]Importance scores are drawn from the O*NET Skills database (Speaking and Active Listening) and the O*NET Abilities database (Oral Comprehension and Oral Expression). Eighty-seven percent of matched occupations have all four components; the composite averages over available components. The index is intuitive: lawyers (91.5), general practitioners (78.0), and school teachers (77.3) rank near the top; truck drivers (48.5), agricultural labourers (50.0), and carpenters (53.0) rank near the bottom. Table 1 presents selected examples.

The preferred specification in Column (3) adds the interaction with college education. The coefficient on the language barrier indicator is -0.049 ($p < 0.01$): non-college migrant workers facing a high language barrier are 4.9 percentage points less likely to work in speaking-intensive occupations. The interaction coefficient of $+0.057$ ($p < 0.01$) indicates that this gap is essentially eliminated for college workers, whose net effect is $-0.049 + 0.057 = +0.008$, effectively zero.

In summary, high language barriers are associated with a lower likelihood of working in speaking-intensive occupations for all migrants, but the gap is concentrated among non-college workers (4.9 percentage points), while for college workers it is essentially eliminated (a net effect of $+0.008$). This pattern is consistent with a model in which non-college workers who migrate across linguistic boundaries sort into occupations that do not require local language fluency (roles such as agricultural labor, construction, and manufacturing), where communication demands are minimal and productivity depends primarily on physical skills.

The attenuated gap for college workers is consistent with college education providing access to English as a lingua franca and to professional and managerial roles where English serves as the working language, substantially reducing the need to rely on local language proficiency. Appendix B.2 confirms these patterns hold under an alternative population-weighted speaking-occupation threshold (54.6), with state-to-state measures of linguistic distance, and when using the continuous composite speaking score as the dependent variable.

Fact 3 (Wages): *Among non-college migrant workers, those with high language barriers receive higher wages; this pattern is significantly attenuated for college workers.*

To examine how language barriers relate to migrant wages, I use the same CPHS panel and estimate a log wage equation. I estimate the following specification,

$$\begin{aligned} \ln(\text{Wages}_{i,o,d,j,t}) = \beta_0 &+ \beta_1 \mathbb{1}\{\text{Barrier}_{o,d}\} + \beta_2 \mathbb{1}\{\text{College}_{i,o,d,j,t}\} \\ &+ \beta_3 \mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\} + \beta_4 \ln(\text{Geo. Dist.}_{o,d}) \\ &+ \mu \text{Controls}_{i,o,d,j,t} + \gamma_{j,t} + \gamma_{o,t} + \gamma_{d,t} + \varepsilon_{i,o,d,j,t}, \end{aligned} \tag{4}$$

where $\text{Controls}_{i,o,d,j,t}$ are gender, age, and caste. I add origin-state $\times$ month fixed effects $\gamma_{o,t}$ and destination-district $\times$ month fixed effects $\gamma_{d,t}$ to absorb time-varying push and pull factors, and occupation $\times$ month fixed effects $\gamma_{j,t}$ to absorb differences in average pay across occupations within each period.

Table 5 presents results for migrant workers across three specifications, building up controls progressively, mirroring the structure of Table 4. The occupation $\times$ month fixed effect is central to this analysis. By absorbing variation across occupations within each period, it ensures that the barrier coefficient reflects within-occupation wage differences rather than mechanical differences

Table 4: Language and Occupations of College vs. Non-College (Fact 2)

	Dep. Var.: $\mathbb{1}\{\text{Speaking Occ.}_{i,o,d,j,t}\}$		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		-0.035**	-0.049***
		(0.015)	(0.016)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$			0.057***
			(0.014)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.243***	0.243***	0.235***
	(0.009)	(0.009)	(0.010)
$\ln(\text{Geo. Dist.}_{o,d})$	0.012*	0.017**	0.014**
	(0.007)	(0.007)	(0.007)
$\ln(\text{Wages}_{i,o,d,j,t-1})$	0.090***	0.090***	0.090***
	(0.007)	(0.007)	(0.007)
Origin State × Month FE	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes
Observations	96,014	96,014	96,014
Adj. R-squared	0.184	0.184	0.184

Notes: The linear probability model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, lagged log real wage, and fixed-effects for origin state × month, destination district × month. The dependent variable equals one for occupations whose composite O*NET speaking score exceeds the occupation-weighted median (≈68.9). Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in average pay or even speaking intensity between occupations. As a benchmark, the college wage premium among native workers is approximately 23 percent ($p < 0.01$).

Column (1) establishes the migrant baseline: college-educated migrants earn approximately 30 percent more than non-college migrants, and geographic distance carries a positive premium of about 2.5 percent, consistent with distance-based positive selection (Bryan and Morten, 2019). Column (2) adds the language barrier indicator, which enters with a coefficient of $+0.070$ ($p < 0.01$): conditional on occupation, migrants facing a high language barrier earn about 7 percent higher wages. This is consistent with positive selection, more productive workers being the ones who cross high-barrier corridors.

The preferred specification in Column (3) adds the interaction with college education. The barrier coefficient is $+0.082$ ($p < 0.01$): non-college migrant workers facing a high language barrier earn approximately 8 percent higher wages. This is consistent with language barriers acting as a strong selection mechanism for non-college workers. That is, only those with sufficiently high productivity find it worthwhile to migrate across linguistic boundaries, and these positively selected

Table 5: Language and Wages of College vs. Non-College (Fact 3)

	Dep. Var.: $\ln(\text{Wage}_{i,o,d,j,t})$		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		0.070***	0.082***
		(0.018)	(0.018)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$			-0.054***
			(0.015)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.262***	0.262***	0.269***
	(0.011)	(0.011)	(0.011)
$\ln(\text{Geo. Dist.}_{o,d})$	0.025***	0.016***	0.019***
	(0.005)	(0.006)	(0.006)
Occupation × Month FE	Yes	Yes	Yes
Origin State × Month FE	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes
Observations	101,207	101,207	101,207
Adj. R-squared	0.671	0.671	0.671

Notes: The linear regression model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, and fixed-effects for occupation × month, origin state × month, destination district × month. Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

workers command higher wages at their destination.

The interaction of -0.054 ($p < 0.01$) indicates a substantially smaller premium for college workers, whose net wage premium from barriers is approximately 3 percent, roughly one-third the non-college premium. The weaker selection gradient for college workers is consistent with the patterns documented in Facts 1 and 2. Since college workers are better able to navigate language barriers through English proficiency and access to occupations less reliant on local language, the barrier is less binding. Thus, crossing it requires less positive selection on unobserved productivity.

Geographic distance carries a premium of about 2 percent ($p < 0.01$), consistent with distance-based positive selection operating alongside language-based selection. Appendix B.3.2 shows the wage patterns are robust to measuring linguistic distance at either the state-district or state-state level, and Appendix B.3.1 confirms the pattern using the continuous linguistic distance index.

Summary: Taken together, the three empirical facts reveal that language barriers are associated with lower migration flows, sorting away from speaking-intensive occupations, and stronger positive wage selection. College education substantially attenuates each margin.

For migration (Fact 1), the barrier gap is 25 percent smaller for college workers, who are relatively more likely to cross linguistic boundaries. For occupational sorting (Fact 2), college education reduces the speaking-occupation penalty from 4.9 percentage points to essentially zero (a

net effect of $+0.008$). Non-college migrants facing high language barriers sort heavily away from speaking-intensive roles, while college migrants largely maintain access to such occupations. For wages (Fact 3), the positive selection premium narrows from 8 to 3 percent, a 66 percent attenuation. The large premium for non-college workers is consistent with stronger positive selection at linguistic boundaries, while the weaker premium for college workers suggests the barrier is less binding.

Two features of this pattern are worth noting. First, the attenuation is largest on the occupational margin and smallest on the migration margin. This is consistent with college education being most effective at neutralizing language barriers in day-to-day work interactions, where English as a lingua franca substitutes directly for local language fluency, and somewhat less effective at reducing the fixed costs of crossing into a linguistically distant region. Second, while college education essentially eliminates the speaking-occupation penalty, on the migration and wage margins the attenuation is partial but never complete: even college workers face meaningful language penalties when deciding where to move and what they earn. Local languages remain important even in professional settings, as English alone does not fully replace them. This is consistent with English proficiency reducing but not eliminating the salience of local language competence in Indian labor markets.

These patterns point to language barriers operating both as migration costs and as occupation-specific productivity shifters. A spatial equilibrium model must account for this systematic heterogeneity by education to accurately characterize how language barriers shape aggregate and distributional economic outcomes. In the following section, I show how these facts can be rationalized in such a model with multiple locations, multiple occupations, and workers that are heterogeneous in productivity, education, and language.

4 A Quantitative Spatial Model

In this section, I develop a static Roy-Fréchet quantitative spatial general equilibrium model of migration in the presence of language barriers. First, I describe preferences and derive sorting and selection equations. Second, I explain the production structure, where heterogeneous occupations (speaking and non-speaking) in each region use heterogeneous labor (skilled and unskilled, and further, with and without language barriers) as imperfect substitutes in a nested constant elasticity of substitution (CES) structure. Third, I impose market clearing conditions and define general equilibrium.

Building on standard spatial models, I incorporate language barriers as both a productivity effect and a movement friction. In particular, I model language as a technological friction that affects the average productivity of workers, and more so in speaking-intensive occupations. That is, in any region and conditional on their education, workers with language barriers are less productive on

average, and more so, in occupations that are speaking-intensive. I also model language as a component of migration cost. This is to capture how language barriers, beyond affecting workers' productivity, can impede their decision to move to regions where they cannot speak the language. Finally, I allow each occupation to imperfectly substitute between workers with and without language barriers, conditional on their skill. Appendix C.1 contains detailed derivations of the main equations of the model.

4.1 Environment

Let there be R regions in the economy, which represent the districts of India. I denote the origin region by o and the destination region by d[9]. Workers may be skilled or unskilled. This is defined by their education e, that is, by whether they have a college degree (c) or not (nc). I assume that each region o is endowed with $N_{o,c}$ units of skilled workers and $N_{o,nc}$ units of unskilled workers.

At any potential destination d, a worker may face a language barrier (b) or not (nb). I formally define language barrier as a mapping, $\mathscr{L} : \{(o,d)\} \longrightarrow \{b, nb\}$, that specifies whether a worker born in o faces a barrier at d (or not), depending on the language at the destination relative to the origin. Notice that since language barrier is a function of both origin and destination, whether a worker faces a language barrier in the labor market is determined in equilibrium.

Further, let there be $J = 2$ types of occupations in each region, defined by whether they are speaking-intensive (e.g., salesperson, professor, waiter) or not (e.g., computer programmer, construction worker, cable operator). This is to capture that language at the destination may be more important while working in some occupations than others.

A firm in each region uses a nested CES production structure with three layers: the outer nest aggregates over speaking and non-speaking occupations, the middle nest aggregates over skilled and unskilled workers, and finally, the inner nest aggregates over skilled and unskilled workers with or without language barriers. This form allows the degree of substitution between labor inputs to be flexible. Workers with or without language barriers are imperfect substitutes within each occupation and skill group.

I assume that workers that differ by skill and anticipate facing different language barriers in equilibrium make idiosyncratic productivity draws for each occupation $j \in J$ and destination $d \in R$ from a Fréchet distribution. This form allows the mean productivity of skilled and unskilled workers to be different, and moreover, to depend on whether they face a language barrier at the destination, and whether occupations are speaking-intensive. Formally, the distribution is,

[9]While the theoretical model is written generally for locations R, data constraints in the CPHS require restricting the origin dimension o to states for the structural estimation. However, the destination dimension d is maintained at the more granular district level. This preserves the district-level geography for all equilibrium labor markets, prices, wages, and amenities, while origin-destination frictions $\tau_{o,d}$ map state-to-district flows.

$$z^i_{j,e,\mathscr{L}(o,d)} \sim^{i.i.d.} G_{j,e,\mathscr{L}(o,d)}(z^i) = \exp(-A_{j,e,\mathscr{L}(o,d)}(z^i)^{-\theta}), \tag{5}$$

where the location parameter, $A_{j,e,\mathscr{L}(o,d)}$, governs absolute advantage. It determines the average productivity of workers with education e with language barrier $\mathscr{L}(o,d)$ in occupation j. The dispersion parameter, θ, reflects the importance of comparative advantage. As θ decreases, there is a greater difference between the productivity draws of workers with different education and language barriers in speaking vs. non-speaking occupations.

This specification does not impose any restrictions on the relative magnitudes of the mean productivity draws for different groups. But we may expect, for instance, that the draws for unskilled workers have a lower mean if they face a language barrier, and that the mean is even lower when the occupation is speaking-intensive, that is, $A_{spk,nc,b} < A_{spk,nc,nb}$.

I assume that the location parameters for college workers do not depend on language barrier status. Specifically, $A_{spk,c,b} = A_{spk,c,nb} \equiv A_{spk,c}$ and $A_{nonspk,c,b} = A_{nonspk,c,nb} \equiv A_{nonspk,c}$. This is motivated by Fact 2, which shows that the sorting of workers across occupations in response to language barriers is effectively eliminated for college workers. College workers still face language barriers as a component of migration cost, and college workers with and without language barriers remain imperfect substitutes in production.

In contrast, the location parameters for non-college workers retain their dependence on barrier status. The location parameter vector therefore takes 6 values, 2 for college and 4 for non-college. For compactness, I continue to write $A_{j,e,\mathscr{L}(o,d)}$, with the understanding that it does not depend on $\mathscr{L}(o,d)$ when $e = c$.

4.2 Preferences

I assume that workers' preferences are linear in consumption and amenities, which are discounted by geographic and language barriers at the destination relative to their origin. Workers supply one unit of labor inelastically. They take wages per unit of productivity, $w_{d,j,e,\mathscr{L}}$, as given and maximize their utility subject to a budget constraint. They spend their entire income on the consumption good, which is priced at P.

Thus, the indirect utility of worker i with education e from origin $o \in R$ who works in occupation $j \in J$ at destination $d \in R$ is given by

$$U^i_{o,d,j,e} = \underbrace{(1-\tau^G_{o,d})^{\eta^G}(1-\tau^L_{o,d})^{\eta^L_e}\ \alpha_d\ (w_{d,j,e,\mathscr{L}}/P)}_{\equiv\ \Lambda_{o,d,j,e}}\ \underbrace{z^i_{j,e,\mathscr{L}(o,d)}}_{\text{Fréchet draw}}, \tag{6}$$

where $\tau^G_{o,d}$ is the geographic barrier, $\tau^L_{o,d}$ is the language barrier, η^G and η^L_e are parameters governing the sensitivity of utility to geographic and language migration frictions, respectively, and α_d are region-specific amenities. I allow η^L_e to differ across skill groups $e \in \{c, nc\}$ to capture that

language barriers may impede college and non-college workers differently. Worker i chooses (d, j) s.t. $U^i_{o,d,j,e} \geq U^i_{o,d',j',e}\ \forall (d', j') \neq (d, j)$.

From the properties of the Fréchet distribution, the probability that workers from origin o with education e choose occupation j and destination d is given by

$$\pi_{o,d,j,e} = \frac{A_{j,e,\mathscr{L}(o,d)}\ (\Lambda_{o,d,j,e})^{\theta}}{(\Phi_{o,e})^{\theta}}, \tag{7}$$

where $(\Phi_{o,e})^{\theta} \equiv \sum_{(d',j')} A_{j',e,\mathscr{L}(o,d')}\ (\Lambda_{o,d',j',e})^{\theta}$. This is the key sorting equation, which captures the idea that workers sort based on relative productivities, income, migration costs, and amenities. From the law of large numbers, it follows that the number of workers from origin o with education e who choose occupation j at destination d is $N_{o,d,j,e} = \pi_{o,d,j,e} N_{o,e}$. This is consistent with Facts 1 and 2.

Further exploiting properties of the Fréchet distribution, the average productivity of a worker from origin o with education e in occupation j at destination d is given by

$$E\left[z^i_{j,e,\mathscr{L}(o,d)} \mid U^i_{o,d,j,e} \geq U^i_{o,d',j',e}\ \ \forall (d', j') \neq (d, j)\right] = \tilde{\Gamma}(\pi_{o,d,j,e}/A_{j,e,\mathscr{L}(o,d)})^{-1/\theta}, \tag{8}$$

where $\tilde{\Gamma} \equiv \Gamma(1 - 1/\theta)$ and $\Gamma(\cdot)$ is the gamma function. The expression for the average wage of a worker from origin o with education e in occupation j at destination d,

$$\bar{w}_{o,d,j,e} = \tilde{\Gamma}(\pi_{o,d,j,e}/A_{j,e,\mathscr{L}(o,d)})^{-1/\theta} w_{d,j,e,\mathscr{L}} = \frac{\tilde{\Gamma}\ \Phi_{o,e}\ P}{(1 - \tau^G_{o,d})^{\eta^G} (1 - \tau^L_{o,d})^{\eta^L_e} \alpha_d}, \tag{9}$$

embeds this selection mechanism. The first equality scales the average productivity from equation 8 by the wage per efficiency unit. It shows that, holding the latter fixed, a smaller share $\pi_{o,d,j,e}$ means the marginal worker is drawn from higher in the productivity distribution.

The second equality shows that in equilibrium this selection on productivity is exactly offset by the wage per efficiency unit $w_{d,j,e,\mathscr{L}}$, so that average wages do not vary across occupations. Instead, for a given origin and education, average wages are higher at destinations with greater migration frictions and lower amenities. In particular, workers that face a higher linguistic barrier $\tau^L_{o,d}$ to a destination are more positively selected and earn higher average wages within any occupation. This selection effect is attenuated for college workers through η^L_e. This within-occupation, barrier-driven wage premium is the model's counterpart to Fact 3.

4.3 Regional Production

In each region $d \in R$, a representative firm uses a triple-nested CES production function to produce y_d units of a good, which has price p_d. This nested structure allows for varying degrees of

substitutability between different types of labor inputs. The three nests partition workers along the dimensions of occupation, education, and language barrier status, with a separate elasticity governing substitution within each grouping. The outer nest combines labor inputs from speaking and non-speaking occupations in the region using a CES function with elasticity of substitution κ and share parameter $\phi_{d,j}$,

$$y_d = \Big[\sum_{j \in J} (\phi_{d,j})^{\frac{1}{\kappa}} (\ell_{d,j})^{\frac{\kappa-1}{\kappa}} \Big]^{\frac{\kappa}{\kappa-1}},$$

where college and non-college labor in each occupation are combined in the middle nest using a CES function with elasticity of substitution ρ,

$$\ell_{d,j} = \Big[(\ell_{d,j,c})^{\frac{\rho-1}{\rho}} + (\ell_{d,j,nc})^{\frac{\rho-1}{\rho}} \Big]^{\frac{\rho}{\rho-1}},$$

where college and non-college workers with and without language barriers in each occupation are combined in the inner nest using a CES function with elasticity of substitution ν,

$$\ell_{d,j,e} = \Big[(\ell_{d,j,e,b})^{\frac{\nu-1}{\nu}} + (\ell_{d,j,e,nb})^{\frac{\nu-1}{\nu}} \Big]^{\frac{\nu}{\nu-1}}, \text{ where } e \in \{c, nc\}.$$

The ordering of the nests reflects the hierarchical nature of production decisions and substitution patterns in labor markets. I place the speaking versus non-speaking occupational distinction in the outer nest because it represents the most fundamental technological choice firms face: how to allocate production between occupations that require different speaking intensities. The college versus non-college distinction appears in the middle nest since skill-based substitution occurs within each occupational category, with potentially different patterns between speaking and non-speaking tasks. Finally, language barriers enter the innermost nest because they represent a friction that modifies worker productivity within skill groups, rather than a fundamental technological distinction.

The inner nest governs the substitutability between workers with and without language barriers through elasticity ν, which is assumed to be common across occupations and skill groups. Importantly, if this elasticity turns out to be large, the inner nest collapses and the structure reduces to a more conventional production function that distinguishes only between skilled and unskilled labor, nesting this model within the broader literature on education-based labor market sorting.

The regional firm takes wages and prices as given and maximizes profits,

$$\max_{\{\ell_{d,j,e,\mathscr{L}}\}} \; p_d y_d \; - \sum_{j \in J} \sum_{e \in \{c,nc\}} \sum_{\mathscr{L} \in \{b,nb\}} w_{d,j,e,\mathscr{L}} \, \ell_{d,j,e,\mathscr{L}}.$$

The first-order condition for each $\ell_{d,j,e,\mathscr{L}}$ implies the labor demand system

$$\ell_{d,j,e,\mathscr{L}} = \left(\frac{MP_{d,j} \cdot MP_{d,j,e}}{w_{d,j,e,\mathscr{L}}} \right)^{\nu}, \tag{10}$$

where

$$MP_{d,j} \;=\; p_d \, (y_d)^{1/\kappa} (\phi_{d,j})^{1/\kappa} (\ell_{d,j})^{\frac{1}{\rho} - \frac{1}{\kappa}}, \qquad MP_{d,j,e} \;=\; (\ell_{d,j,e})^{\frac{1}{\nu} - \frac{1}{\rho}}.$$

I assume that markets are perfectly competitive, so prices are equal to marginal costs (equivalently, unit costs under constant returns to scale),

$$p_d \;=\; \Big[\sum_{j \in J} \phi_{d,j} \, (w_{d,j})^{1-\kappa} \Big]^{\frac{1}{1-\kappa}},$$

where

$$w_{d,j} \;=\; \Big[(w_{d,j,c})^{1-\rho} + (w_{d,j,nc})^{1-\rho} \Big]^{\frac{1}{1-\rho}},$$

and

$$w_{d,j,e} \;=\; \Big[(w_{d,j,e,b})^{1-\nu} + (w_{d,j,e,nb})^{1-\nu} \Big]^{\frac{1}{1-\nu}}, \qquad \text{where } e \in \{c, nc\}.$$

4.4 Aggregate Production

I assume that an aggregate firm combines the regional goods into a single consumption good, Y, with price P, using a CES function,

$$Y = \Big[\sum_{d \in R} (y_d)^{\frac{\sigma-1}{\sigma}} \Big]^{\frac{\sigma}{\sigma-1}},$$

where σ is the elasticity of substitution between regional goods and determines the degree to which price differences across regions affect the overall price index. That is, higher values of σ indicate greater substitutability between regional goods, and hence less sensitivity of the aggregate price index to regional price variations. The firm takes prices as given and maximizes profit, which is given by $PY - \sum_{d \in R} p_d y_d$. The first-order condition determines regional good demand,

$$y_d = (p_d / P)^{-\sigma} Y, \tag{11}$$

where the consumption good is assumed to be traded costlessly across regions and there is no international trade. This implies that all regions face the same price index, which is defined by $\Big[\sum_{d \in R} (p_d)^{1-\sigma} \Big]^{\frac{1}{1-\sigma}}$. Under perfect competition, the price P equals the marginal cost, which, in this case, is identical to the expression for the price index. I assume that the consumption good is the numeraire, i.e., $P = 1$.

4.5 Market Clearing Conditions and Equilibrium

The labor market clearing condition equates labor supply in region d and occupation j from workers with education e and language barrier $\mathscr{L}(o,d)$ with demand from regional firms,

$$\sum_{o:\mathscr{L}(o,d)=\mathscr{L}} \underbrace{\tilde{\Gamma}\left(\pi_{o,d,j,e}/A_{j,e,\mathscr{L}(o,d)}\right)^{-\frac{1}{\theta}}}_{\text{avg. productivity}} \underbrace{\pi_{o,d,j,e}\, N_{o,e}}_{\text{no. of workers}} = \ell_{d,j,e,\mathscr{L}}, \tag{12}$$

where the average productivity (corresponding to equation 8) is multiplied by the number of workers to compute the labor supply in terms of efficiency units. The regional goods market clearing condition equates the supply of the regional good in region d with demand from the aggregate firm,

$$y_d = (p_d/P)^{-\sigma} Y. \tag{13}$$

The aggregate goods market clearing condition equates the supply of the consumption good from the aggregate firm with the total demand from workers,

$$Y = \sum_{d\in R} \sum_{j\in J} \sum_{e\in\{c,nc\}} \sum_{\mathscr{L}\in\{b,nb\}} (w_{d,j,e,\mathscr{L}}\, \ell_{d,j,e,\mathscr{L}})/P. \tag{14}$$

The equilibrium in this model is defined as follows. Given fundamentals, $\Theta \equiv \{A_{j,e,\mathscr{L}(o,d)},\ \theta,\ \sigma,\ \phi_{d,j},\ \kappa,\ \rho,\ \nu,\ N_{o,e},\ \tau^G_{o,d},\ \tau^L_{o,d},\ \eta^G,\ \eta^L_e,\ \alpha_d\}_{o,d\in R,\ j\in J,\ e\in\{c,nc\},\ \mathscr{L}(o,d)\in\{b,nb\}}$, an equilibrium is a set of wages and prices, $\{w_{d,j,e,\mathscr{L}}, p_d, P\}_{d\in R, j\in J, e\in\{c,nc\}, \mathscr{L}\in\{b,nb\}}$, that solves the optimization problems of consumers, regional firms, and the aggregate firm (corresponding to equations 6, 10, and 11), as well as market clearing conditions for labor, regional goods, and the aggregate good (corresponding to equations 12, 13, and 14). Appendix C.2 describes the algorithm used to solve for this equilibrium.

5 Taking the Model to the Data

In this section, I outline the procedure to take the model to the data and discipline its exogenous structural parameters: $(\eta^G, \eta^L_e, A_{j,e,\mathscr{L}(o,d)}, \kappa, \rho, \nu)$.[10] These parameters include the spatial migration elasticities governing geographic and linguistic distance, the technological productivity frictions, and the nested CES production elasticities. Since the spatial sorting of workers and local labor demand are intricately linked in general equilibrium, these parameters cannot be recovered through isolated reduced-form regressions. Instead, I jointly identify and estimate the full parameter vector using a unified Generalized Method of Moments (GMM) framework with instrumental

[10] See Table 17 in Appendix D for a summary of all model parameters and variation in the data from which they are identified.

variables (IV).

I proceed in five steps. First, I define the geographic and temporal dimensions of the empirical economy. Second, I explain the identification strategy, establishing the theoretical moment conditions and shift-share IVs used to isolate exogenous spatial and labor-demand variation. Third, I describe the joint computational algorithm and the bootstrap procedure used for inference. Fourth, I present the structural estimates. Finally, I validate the estimated model's fit to the observed data.

5.1 Empirical Setting and Data Aggregation

I use the CPHS to construct migration flows from origin states ($R_o = 27$) to destination districts of India ($R_d = 522$), disaggregated by occupations and skill. While the theoretical model defines origin and destination regions generically, in this empirical implementation I map origins to the state level to reflect the reliable origin tracking in the CPHS, while retaining finer district-level precision for destinations.

I then link this data to the bilateral geodesic distance and the linguistic distance index, which map to $\tau^G_{o,d}$ and $\tau^L_{o,d}$, respectively, in the model. I aggregate the monthly CPHS waves into four-month periods and use the window September 2017 to December 2019 for the main estimation, which yields $T = 7$ periods, indexed $t = 1, \ldots, 7$. Each period is treated as an independent cross-section of the static model, with repeated periods serving to expand the variation available for identification.

5.2 Identification

5.2.1 Identification of Migration Frictions, Productivity Shifters, and Amenities

The model features two bilateral migration frictions: geographic distance ($\tau^G_{o,d}$), governed by the elasticity η^G, and linguistic distance ($\tau^L_{o,d}$), governed by the skill-specific elasticity η^L_e. A central empirical challenge is distinguishing the role of language as a bilateral migration friction from language as a local technological productivity friction ($A_{j,e,b}$). While substituting the observed average wage selection equation into the sorting equation analytically cancels $A_{j,e,b}$, doing so eliminates the ability to structurally identify the language productivity friction.

Instead, identification requires matching the model's unsubstituted sorting probabilities to the observed bilateral flows. To take the model's deterministic equilibrium to the data, I assume that observed flows $N_{o,d,e,t}$ correspond to the theoretical gravity prediction multiplied by an empirical error term, $\zeta_{o,d,e,t}$. This econometric residual captures measurement error in the migration data as well as unobserved, time-varying idiosyncratic bilateral frictions not captured by the static model. The population conditional expectation is:

$$E\left[\zeta_{o,d,e,t} \mid \tau^G_{o,d}, \tau^L_{o,d}, \mathbf{X}_t\right] = 1$$

This moment condition pins down the parameters by exploiting their distinct dimensionality. The migration elasticities (η^G and η^L_e) operate at the bilateral corridor level and are mathematically identified by the orthogonality of this empirical error term to the exogenous bilateral distance data. In contrast, the productivity frictions ($A_{j,e,b}$) operate at the local labor market level and are identified by matching the model's predicted worker sorting across specific occupations and education groups to the observed local allocations. Finally, destination amenities (α_d) are disciplined by the destination-time fixed effects within $\mathbf{X}_t$ that capture aggregate local pull factors.

Evaluating these sorting probabilities relies on unobserved, structurally consistent efficiency wages. Thus, this spatial block cannot be estimated in isolation and must be jointly identified alongside the nested CES labor demand moments.

5.2.2 Identification of Production Elasticities

The production elasticities (κ, ρ, and ν) are identified from the nested labor demand system. To empirically identify the parameters governing these nests, I introduce time variation across the seven four-month periods.

Taking the log time-difference of the relative CES first-order conditions isolates an econometric residual. Since the theoretical model is static and deterministic, this residual conceptually captures any unobserved, time-varying local relative demand shocks (such as high-frequency fluctuations in the occupation share parameters $\phi_{d,j}$) and measurement error in the data. For the outer nest, the empirical residual is defined as:

$$\Delta_t \varepsilon_{d,j,t} = \Delta_t \ln\left(\frac{w_{d,j,t}}{w_{d,j',t}}\right) + \frac{1}{\kappa}\Delta_t \ln\left(\frac{\ell_{d,j,t}}{\ell_{d,j',t}}\right)$$

Since local relative wages and relative labor quantities adjust simultaneously in equilibrium, robust estimation via ordinary least squares is not possible. To achieve causal identification, I construct nested time-varying shift-share instruments. Specifically, I interact predetermined pre-sample origin shares with leave-out, occupation-specific origin-supply growth, $\Delta_t \ln N_{o,-d,j,t}$, to isolate exogenous supply-driven local labor adjustments. The outer-nest instrument $Z^{\kappa}_{d,t}$ contrasts the speaking and non-speaking occupation legs, each formed from destination d's pre-sample origin composition and the growth of origin labor supply at *other* destinations.

Identification relies on the exclusion restriction that these supply-driven shift-share instruments are orthogonal to the unobserved local relative demand shocks, yielding the population moment condition:

$$E\left[Z^{\kappa}_{d,t} \cdot \Delta_t \varepsilon_{d,j,t}\right] = 0.$$

I apply an equivalent identification strategy to the middle and inner nests to separately identify ρ and ν. The full stacked system of moment conditions is collected in Appendix D.2, and the structural residuals and instruments for these lower nests are derived in Appendices D.4–D.7. The middle-nest instrument uses by-education leave-out supply growth $\Delta_t \ln N_{o,-d,j,e,t}$, while the inner-nest instrument weights the common occupation-j supply growth by the difference between the barrier and no-barrier pre-sample origin shares.

5.3 Estimation Procedure and Inference

To estimate the full parameter vector, I replace the population moment conditions defined above with their sample analogs. The exact construction of the stacked sample moments and the GMM objective function are detailed in Appendix D.2. Since the empirical moments depend on unobserved, structurally consistent efficiency wages, the parameters cannot be estimated via separate, reduced-form regressions. Instead, I stack the sample moments and estimate the parameters simultaneously using a joint GMM framework.

The estimation proceeds via a nested iterative algorithm. In the inner loop, for a given guess of the parameter vector, I solve the model's equilibrium conditions to recover the implied efficiency wages and destination price indices that support the observed spatial allocations. In the outer loop, the GMM objective function evaluates the joint fit of the stacked moments. For the spatial block, the outer loop minimizes the PPML objective corresponding to the bilateral sorting flows. Concurrently, for the production block, the structural CES residuals are evaluated via a Two-Stage Least Squares (2SLS) projection, where the time-differenced endogenous relative labor quantities are projected onto the time-varying shift-share instruments. The algorithm continuously updates the parameter vector until the joint objective function is minimized. To validate the identifying variation required for the short-run production elasticities, I report the first-stage F-statistics and associated weak-instrument diagnostics for each of the nested labor demand equations.

Finally, I address statistical inference. Since the bilateral structure of the spatial gravity moments naturally induces correlation across observations, and since the CPHS tracks workers across repeated panel waves, standard independent-observation inference would be biased downwards. I therefore compute standard errors using a multiplier (Bayesian) cluster bootstrap at the destination-district level. In each replication, every one of the 522 destination districts draws a random exponential weight, the weights enter the stacked moment conditions, and the full joint structural model is re-estimated. Clustering at the destination level makes inference robust to arbitrary correlation within a destination across origin corridors, occupations, education groups, and time periods, which subsumes serial correlation within any migration corridor into that destination. The multiplier form has a further advantage over resampling with replacement: every district remains in the sample in every replication, so the general equilibrium is well defined for each draw and thin corridors are never mechanically emptied. Since inference is conducted through this full structural

bootstrap rather than through asymptotic 2SLS formulas, the reported standard errors natively incorporate the finite-sample sampling distribution of the shift-share instruments.

5.4 Estimation Results

The structural estimates appear in Table 6. They confirm the central hypothesis of the paper: language barriers depress productivity for unskilled rather than skilled workers, and in speaking rather than non-speaking occupations.

Panel B reports the technological productivity terms in mean-efficiency units, normalized to 1 in the non-speaking occupation within each skill group. The pattern in the point estimates is sharp. For non-college workers, a speaking occupation crossed under a language barrier carries an efficiency of 0.59, a penalty of roughly forty percent relative to the non-speaking baseline. The same occupation without a barrier instead earns a premium (1.13), and a non-speaking occupation under a barrier is essentially unaffected (0.99). The barrier therefore bites only to the extent that the occupation depends on language: language-intensive work done across a barrier is badly degraded, while language-light work is not. For college workers, the speaking occupation carries a large efficiency premium (1.88), reflecting educated workers' comparative advantage in communication-intensive tasks.

By construction, the college terms do not vary with barrier status, a restriction motivated by Fact 2, which shows that the occupational sorting of college workers is essentially unaffected by language barriers. This pattern is consistent with college education in India conferring English proficiency, which insulates skilled workers from local-language barriers. The qualitative ordering across cells is robust across specifications and starting values; the precise levels are less tightly pinned, and I treat them accordingly.

The production elasticities corroborate this account. For the outer nest, the elasticity of substitution between speaking and non-speaking occupations is $\hat{\kappa} = 3.2$, a novel estimate with no direct benchmark in the literature. Its magnitude implies that the two occupation groups are imperfect substitutes: more substitutable than college and non-college labor ($\hat{\rho} = 1.7$), but far from perfect substitutes. A waiter who interacts with customers and a chef who prepares the food both contribute to the restaurant's service, so neither occupation type can be dispensed with entirely, yet the occupation mix can adjust when relative wages change. For the middle nest, the elasticity of substitution between college and non-college workers is $\hat{\rho} = 1.7$, close to established benchmarks: Katz and Murphy (1992) estimate 1.4 for the United States and Khanna and Morales (2023) estimate 1.7 for India. The agreement with this developed- and developing-country evidence lends credibility to the estimation strategy and points to limited substitutability between skill types in production. Although the first-stage F-statistic of 7.8 indicates the shift-share instrument is only moderately strong by conventional linear standards, the close match of the point estimate to these established benchmarks provides economic validation that the instrument isolates the relevant relative-supply

Table 6: Structural Parameter Estimates

	Estimate
Panel A: Migration Costs	
Geographic, η^G	1.74^{***}
	(0.043)
Linguistic, College, η^L_c	0.39^{***}
	(0.056)
Linguistic, Non-College, η^L_{nc}	0.53^{***}
	(0.096)
Panel B: Technological Productivity, $A^{1/\theta}_{j,e,\mathscr{L}(o,d)}$ (mean efficiency)	
College	
Non-Speaking	$1.00^{\dagger}$
Speaking	1.88
Non-College	
Non-Speaking, No Barrier	$1.00^{\dagger}$
Non-Speaking, Barrier	0.99
Speaking, No Barrier	1.13
Speaking, Barrier	0.59
Panel C: Production Elasticities	
Outer Nest, κ	3.23^{***}
	(0.090)
	[11.1]
Middle Nest, ρ	1.71^{***}
	(0.033)
	[7.8]
Inner Nest, ν	41.2^{***}
	(0.228)
	[7.75]

Notes: This table reports structural parameters estimated jointly via GMM using shift-share IVs. Standard errors (in parentheses) are computed from a multiplier (Bayesian) cluster bootstrap over the 522 destination districts, re-estimating the full joint structural model in each replication. First-stage F-statistics for the shift-share instruments are reported in brackets. **Panel A** reports the geographic migration elasticity (η^G) and skill-specific linguistic migration elasticities (η^L_c, η^L_{nc}). **Panel B** reports technological productivity terms expressed in mean-efficiency units ($A^{1/\theta}_{j,e,\mathscr{L}}$, evaluated at $\theta = 3.5$). Since the mean of a Fréchet distribution scales with $A^{1/\theta}$, each entry is the average productivity of the cell's distribution of draws, relative to the baseline cell. Cells are defined by education, occupation speaking intensity, and barrier status, and A is estimated directly at this level. Consistent with Fact 2, college terms do not vary by barrier status. Cells marked † serve as the normalized baseline (set to 1) for each skill group; all other entries in Panel B are estimated relative to these baselines. **Panel C** reports the nested-CES production elasticities: outer (κ), middle (ρ), and inner (ν). As detailed in the text, the inner-nest instrument is only moderately strong, though the primary productivity consequences of language barriers operate through Panel B rather than ν. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

variation without severe finite-sample bias. Crucially, the alignment of my short-run estimates with these established long-run benchmarks validates their use in anchoring the language-policy counterfactuals in Section 6.

For the inner nest I estimate a single elasticity of substitution between workers with and without language barriers, common across occupations and skill groups, of $\hat{\nu} \approx 41$. As with the middle nest, the first-stage diagnostic ($F = 7.75$) indicates the instrument is only moderately strong, so the exact level is pinned down imprecisely, and I treat it with caution. The model's core mechanism is nonetheless insensitive to the point estimate. Whether the elasticity is 30, 40, or 50, the structural implication is identical: conditional on occupation and skill, barrier and no-barrier workers are close substitutes in production, and the productivity consequences of language barriers operate through the Fréchet location terms of Panel B, not through imperfect substitutability within the inner nest.

The GMM procedure also recovers the migration-cost elasticities: $\hat{\eta}^G \approx 1.74$ for geographic distance, with $\hat{\eta}^L_c \approx 0.39$ and $\hat{\eta}^L_{nc} \approx 0.53$ for linguistic distance among college and non-college workers. These levels are not directly comparable to the reduced-form gravity coefficients of Fact 1: the elasticities enter utility, so flows respond to the friction terms with elasticities $\theta\eta^G$ and $\theta\eta^L_e$, and the continuous index $\tau^L_{o,d}$ differs from Fact 1's binary barrier indicator. What the structural model adds is a separation of two channels that the reduced form bundles: a direct utility cost of the bilateral frictions, parameterized by η^G and η^L_e, and a labor-market channel operating through destination wages. Geographic distance has no productivity counterpart in the model, so its effect on flows operates entirely through the utility cost. Language does have a productivity counterpart, in the Panel B terms, where $\hat{A}^{1/\theta}_{spk,nc,b} \ll \hat{A}^{1/\theta}_{spk,nc,nb}$ for non-college workers in speaking-intensive occupations. The structural decomposition reattributes part of what the reduced form reads as a movement cost to this productivity wedge, leaving $\hat{\eta}^L_e$ to capture only the utility component.

Two comparisons with Fact 1 survive the change of scale. The ordering $\hat{\eta}^L_{nc} > \hat{\eta}^L_c$ matches the reduced-form finding that the barrier coefficient is larger for non-college workers, since the Fréchet dispersion parameter θ multiplies both linguistic elasticities equally. For the same reason their ratio is unit-free: $\hat{\eta}^L_c/\hat{\eta}^L_{nc} \approx 0.74$ implies a college utility cost 26 percent below the non-college cost, nearly identical to the 25 percent attenuation estimated in the flow data ($0.335/1.333$, Fact 1). The estimation does not impose this match.

That $\hat{\eta}^L_{nc}$ exceeds $\hat{\eta}^L_c$ indicates that, even after netting out the productivity channel, part of the linguistic friction operates outside the labor market. This is consistent with non-college migrants relying more heavily on local-language networks for housing, schooling, and public services at the destination. The linguistic elasticity levels carry wider, right-skewed bootstrap uncertainty than the production-side parameters. Their ordering, however, is unambiguous: $\hat{\eta}^G > \hat{\eta}^L_{nc} > \hat{\eta}^L_c$ holds in every valid bootstrap replication, and the 95 percent bootstrap confidence intervals (percentile

method) for the two linguistic elasticities do not overlap. Separating the utility and productivity channels is essential for the counterfactuals, because shutting down the migration-cost channel and equalizing the productivity terms trace out distinct general-equilibrium responses for aggregate output and the college wage premium.

5.5 Model Fit and Validation

I assess the model fit by comparing the worker allocations implied by the estimated model to their data counterparts. The natural target is the sorting margin that the PPML moments discipline: the distribution of workers across destinations, occupations, education groups, and barrier legs. Figure 11 in Appendix D.9 plots model against data flows at this level. The fit is close. Log flows line up along the forty-five-degree line, with a correlation of 88 percent at the destination-by-occupation-by-education level and 83 percent once the barrier leg is added. The model reproduces both the level of sorting and its composition across the speaking and non-speaking occupations and the barrier and no-barrier legs that the counterfactuals act on.

The model also fits well on dimensions the moments do not target, though naturally less tightly than on those they do. The full bilateral origin-to-destination migration matrix is not a direct estimation target: the sorting moments discipline the destination-side allocation of each origin's workers, not the complete origin-by-destination gravity matrix. Since the parsimonious structural model does not absorb idiosyncratic bilateral noise into exact structural residuals, the model and data correlate less strongly on the full bilateral matrix (about 60 percent), with the largest discrepancies on thin corridors carrying few migrants. This is the expected pattern for a model estimated on the sorting margin, and it does not affect the destination-level quantities that drive the counterfactual exercises.

6 Counterfactuals

In this section, I use the estimated model to quantify the aggregate impact of language barriers on internal migration, the skill premium, and welfare, and to document the spatial heterogeneity behind these aggregates. I then benchmark these effects against the geographic distance frictions that also shape India's spatial economy.

6.1 Quantifying the Effect of Language Barriers on Internal Migration, the Skill Premium, and Welfare

In the model, language barriers enter in two ways, and each has distinct economic implications. The first is the impact on the worker's choice problem: language barriers are a component of

bilateral migration costs that discourage workers from moving across language lines, generating a friction in the spatial allocation of labor. The second is the impact on the worker's productivity: language enters as a technological friction that lowers the average productivity of workers who face barriers, and more so in speaking-intensive occupations. This generates a wedge between barrier and non-barrier workers that varies systematically with occupation. In the single-ν specification this wedge is carried entirely by the non-college Fréchet locations $A_{j,nc,b}$: college productivity is pooled across barrier status, so for college workers the barrier operates only through the migration-cost channel.

Beyond these are general equilibrium forces and selection effects that shape how each component of language barriers maps into aggregate outcomes. The general equilibrium forces operate through wage repricing from the inner CES nest, skill-premium repricing through the middle nest with elasticity ρ, and regional price adjustments through the outer CES nest. The selection effects operate separately on the worker side: as more workers crowd into a (d, j, e, b) cell, the marginal arrival draws from further down the left tail of the Fréchet distribution, so per-worker productivity falls even before any wage adjusts.

The aggregate impact of language barriers is the joint effect of language both as a component of migration cost and as a technological friction, in the presence of general equilibrium forces and selection effects; the relative importance of each is an empirical question. To answer it, I conduct three counterfactual exercises that shut down each channel separately and then both jointly. I conduct a fourth exercise that lowers geographic migration costs to a benchmark level to gauge how the effects of language barriers compare to this benchmark spatial friction. Each exercise is solved as a new general equilibrium with all other parameters held at baseline. The aggregate results are summarized in Table 7.

Counterfactual 1: Removing Language Migration Costs. In counterfactual 1, I shut down language as a component of bilateral migration costs by setting $\tau^L_{o,d} = 0$ for all $(o,d) \in R \times R$, while holding the technological friction $A_{j,e,b}$ and all other model fundamentals at baseline. The exercise increases internal migration, as measured by the share of workers crossing a state border, by 12.5 percentage points and welfare, as measured by aggregate real income Y, by 8.9 percent. This is the dominant channel: language suppresses aggregate output primarily by blocking mobility. Because $\tau^L_{o,d}$ enters bilateral migration costs directly, removing it reallocates labor across space toward its best-match districts and occupations, and that reallocation is where most of the efficiency gain lives — the migration-cost channel moves the entire non-college distribution, not just a narrow margin. Part of the per-worker gain is nonetheless absorbed by selection: reallocating workers into a (d, j, e, b) cell raises $\pi_{o,d,j,e}$ and pulls down the marginal arrival's Fréchet draw.

The aggregate skill premium barely moves (-0.61 percentage points). Removing $\tau^L_{o,d}$ does shift relative supply at individual destinations, since inflows tilt toward non-college workers, but these local shifts largely net out across space, and the channel leaves the barrier productivity wedge

itself untouched; compressing the premium at barrier corridors is the role of the technological-friction channel. Among migrants, however, the premium *widens* (+7.78 percentage points), because removing the cost admits a more positively selected pool of college movers into high-skill-intensive destinations.

Counterfactual 2: Removing Language as a Technological Friction. In counterfactual 2, I shut down the technological friction by setting $A_{j,e,b} = A_{j,e,nb}$ for all (j,e), while holding $\tau^L_{o,d}$ and all other model fundamentals at baseline. Because college productivity is already barrier-pooled, this exercise operates entirely on non-college workers, raising the productivity of those who face barriers. It increases welfare by 0.9 percent and internal migration by 0.7 percentage points — an order of magnitude smaller than the migration-cost channel of counterfactual 1. The migration response is small because $A_{j,nc,b}$ does not alter the bilateral migration wedge. Removing the technological penalty directly improves the productivity distribution barrier-corridor workers draw from, making those cells more attractive in the sorting choice, but this channel primarily operates on local occupational switching rather than overcoming the large spatial distances that suppress inter-state flows. The aggregate welfare gain is likewise small: the technological penalty is severe but *narrow*, precisely because barrier-corridor non-college workers have already sorted away from the speaking-intensive occupations where it bites (Fact 2), so few workers sit on the margin that equalizing A lifts.

Although its aggregate footprint is small, this is the channel that reshapes the skill distribution. The skill premium compresses by 1.49 percentage points, more than under the migration-cost channel. The friction removal is asymmetric by education: it raises only non-college barrier productivity, so non-college effective labor expands in barrier corridors while college is unchanged. This relative supply shock in fact pushes non-college efficiency wages *down* in the middle CES nest with elasticity ρ, but that general-equilibrium headwind is completely overwhelmed by the direct surge in non-college productivity. Observed average non-college wages therefore rise substantially, and it is this direct productivity effect, not CES repricing, that compresses the premium. Among migrants the premium likewise compresses slightly (-1.27 percentage points).

Counterfactual 3: Removing Both Language Channels. In counterfactual 3, I apply counterfactuals 1 and 2 jointly, setting $\tau^L_{o,d} = 0$ and $A_{j,e,b} = A_{j,e,nb}$. This corresponds to the full removal of language barriers from the estimated economy. The exercise increases internal migration by 14.6 percentage points, compresses the aggregate skill premium by 3.77 percentage points, and increases welfare by 10.9 percent.

Ex ante, the joint result is ambiguous on two dimensions. First, the two channels act on overlapping margins: removing $\tau^L_{o,d}$ moves workers into the same corridors where productivity rises under $A_{j,e,b} = A_{j,e,nb}$, so the levers could be substitutes or complements. Second, the two channels pull the migrant skill premium in opposite directions: removing $\tau^L_{o,d}$ widens it through positive se-

lection into high-skill-intensive destinations, while removing the technological friction compresses it at barrier corridors. The data alone cannot separate complementarity from substitution or sorting from selection; I rely on the structural model to disentangle them.

The welfare gain exceeds the sum of counterfactuals 1 and 2 in isolation (+9.8 percent), reflecting a complementarity between the two channels: the workers who newly arrive in barrier corridors when $\tau^L_{o,d}$ is removed are the same workers who become more productive when the technological friction is removed, and each channel raises the marginal product of the other. On the aggregate premium, the widespread friction-side compression easily dominates any localized widening from the migration-cost channel, so the overall premium compresses by 3.77 percentage points. Among migrants, however, the strong positive spatial selection prevails, widening their premium by 9.38 percentage points.

Reading the decomposition. The two channels have a clean division of labor. The *migration-cost* channel (τ^L) drives the aggregate output gain and essentially all of the migration response (+8.9 percent and +12.5 p.p. of the joint +10.9 percent and +14.6 p.p.). The *technological-friction* channel (A) drives the occupational reallocation and is the larger of the two direct forces compressing the aggregate skill premium (−1.49 percentage points versus −0.61 from the migration-cost channel), while contributing little to aggregate output on its own. The two direct effects sum to −2.10 of the joint −3.77 percentage points; the remainder reflects the same complementarity as for welfare, since the workers who newly cross barriers once τ^L is removed are those whose productivity penalty the A channel removes. Language suppresses output primarily by blocking mobility, and shapes the skill distribution through the productivity penalty and its interaction with mobility.

Counterfactual 4: Comparison with Geographic Migration Costs. Counterfactual 4 assesses how language barriers compare to the benchmark friction in the spatial economy: geographic distance. Removing geographic migration costs entirely ($\tau^G_{o,d} = 0$) raises welfare by 49.5 percent and migration by 52.3 percentage points — an implausibly large number that reflects how tightly distance binds India's low observed mobility, and one that overstates any feasible policy or comparison. Following Bryan and Morten (2019), I therefore benchmark against the United States, where internal language barriers are effectively zero and geographic mobility is high. Their estimated gravity distance elasticities imply that US flows decay $0.553/0.717 = 0.77$ times as steeply with distance as Indonesia's; I apply the same proportional reduction to India, scaling η^G by 0.77 so that India's distance-decay is made "as flat, relative to itself, as the US is relative to Indonesia." Because this is a ratio applied to the estimated elasticity, it does not require taking a stand on the level of η^G.

Reducing geographic frictions to this US benchmark raises welfare by 8.4 percent and migration by 7.4 percentage points. The aggregate welfare gain from removing language barriers (10.9 percent, counterfactual 3) is therefore *comparable to* — indeed somewhat larger than — the gain

from making India's geography as frictionless as the United States'. Language is a first-order spatial friction: not a second-order correction to geography, but a barrier of the same magnitude as the distance between a developing economy and a mobile advanced one. Combining the US geography benchmark with the removal of language migration costs ($\tau^L_{o,d} = 0$) raises welfare by 19.6 percent.

The two frictions also differ in incidence. Reducing geographic costs slightly *widens* the skill premium (+0.81 percentage points), while removing language barriers *compresses* it (−3.77 percentage points). Language barriers are not merely comparable to geographic frictions in the aggregate; they have a qualitatively different, and regressive, incidence across the skill distribution.

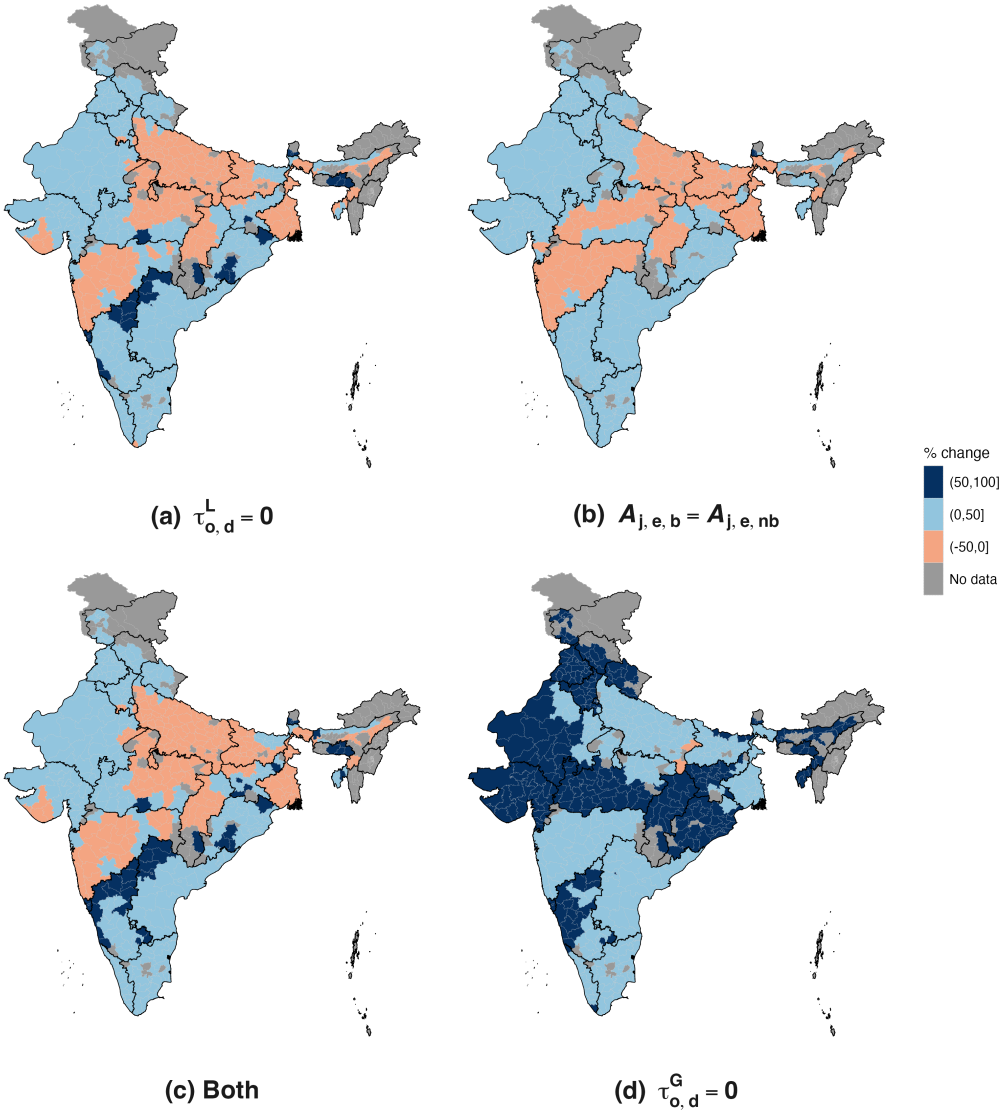
Table 7: Aggregate Effects of Removing Language and Geographic Barriers

	Internal Migration	Skill Premium (all)	Skill Premium (migrants)	Welfare (Y)
Baseline (levels)	23.1%	265.6%	255.7%	1464.1
CF1: $\tau^L_{o,d} = 0$	+12.5 p.p.	−0.61 p.p.	+7.78 p.p.	+8.9 percent
CF2: $A_{j,e,b} = A_{j,e,nb}$	+0.7 p.p.	−1.49 p.p.	−1.27 p.p.	+0.9 percent
CF3: Both	+14.6 p.p.	−3.77 p.p.	+9.38 p.p.	+10.9 percent
CF4: geography → US ($\eta^G \times 0.77$)	+7.4 p.p.	+0.81 p.p.	+0.51 p.p.	+8.4 percent
Full US benchmark (CF4 + $\tau^L = 0$)	+22.9 p.p.	−0.11 p.p.	+6.57 p.p.	+19.6 percent
$\tau^G_{o,d} = 0$ *(remove all geography)*	+52.3 p.p.	+2.51 p.p.	+5.98 p.p.	+49.5 percent

Notes: The baseline row reports levels; all other rows report changes relative to the estimated baseline equilibrium. Internal migration is the share of workers who cross a state border. The baseline migration share is the model-implied level under the calibrated amenities and lies above the observed inter-state migrant share; all counterfactual changes are read relative to this common model baseline. The skill premium is the percentage by which the population-weighted mean wage of college workers exceeds that of non-college workers, i.e., the wage ratio minus one, ×100; "all" is over all workers earning income and "migrants" is over cross-state migrants only. Changes in the skill premium are reported in percentage points of this premium. Welfare is aggregate real income Y. CF1 sets $\tau^L_{o,d} = 0$; CF2 removes the technological friction by equalizing $A_{j,e,b}$ with the no-barrier value (non-college only, since college productivity is barrier-pooled); CF3 applies both. CF4 is the geographic benchmark: rather than remove geography entirely, it reduces India's geographic distance-decay to approximately the US level by scaling η^G by the Bryan–Morten (2019) US/Indonesia gravity ratio $0.553/0.717 = 0.77$. The two indented rows report the full US benchmark (geography-to-US combined with the removal of language migration costs, $\tau^L_{o,d} = 0$; the technological friction remains at baseline) and, as an upper bound, the removal of geographic migration costs entirely.

The aggregate numbers in Table 7 conceal substantial spatial heterogeneity in the effects of removing language barriers. I depict this heterogeneity through four maps: the percentage change in regional welfare, the percentage-point change in the in-migration rate at each destination, the percentage change in efficiency units per worker at each destination, and the change in the skill premium at each destination.

Figure 5: Counterfactual Changes in Regional Welfare



Notes: Each panel maps the percentage change in regional welfare at the district level under one of four counterfactuals, relative to the estimated baseline. Regional welfare at district d is aggregate real income $p_d y_d$. Panel (a) sets $\tau^L_{o,d} = 0$ (counterfactual 1); panel (b) sets $A_{j,e,b} = A_{j,e,nb}$ (counterfactual 2); panel (c) applies both (counterfactual 3); panel (d) sets $\tau^G_{o,d} = 0$ (the remove-all-geography benchmark of Table 7). Blue indicates positive changes; red indicates negative changes. The color scale is percentile-clipped at the 95th percentile and is symmetric around zero. Districts shaded grey have no data.

Regional Welfare. Figure 5 maps the percentage change in regional welfare $p_d y_d$. Under counterfactuals 1 and 3, the gains concentrate in the Dravidian south and the northeast — Karnataka, Andhra Pradesh, Telangana, and Kerala, together with the northeastern states — while the Hindi belt is flat to slightly negative and West Bengal declines more sharply, as both cede workers to newly accessible destinations. Counterfactual 2 produces a flatter and smaller pattern, and the full removal of geographic costs in panel (d) is more diffuse, spreading large gains across northern and

central districts rather than concentrating them in the high-barrier south.

The structural precondition for these maps is the share of a district's out-of-state origins for which it has a language barrier. Southern India, much of the northeast, and parts of Odisha sit at 0.7 to 1.0 on this measure, while the Hindi belt sits at 0.2 to 0.4. This asymmetry is genealogical: the south is Dravidian-speaking, while most other origins are Indo-Aryan. Counterfactuals that operate on $\tau^L_{o,d}$ generate the largest reallocation precisely where this fraction is highest, which is why the language maps light up the south and the geographic map does not.

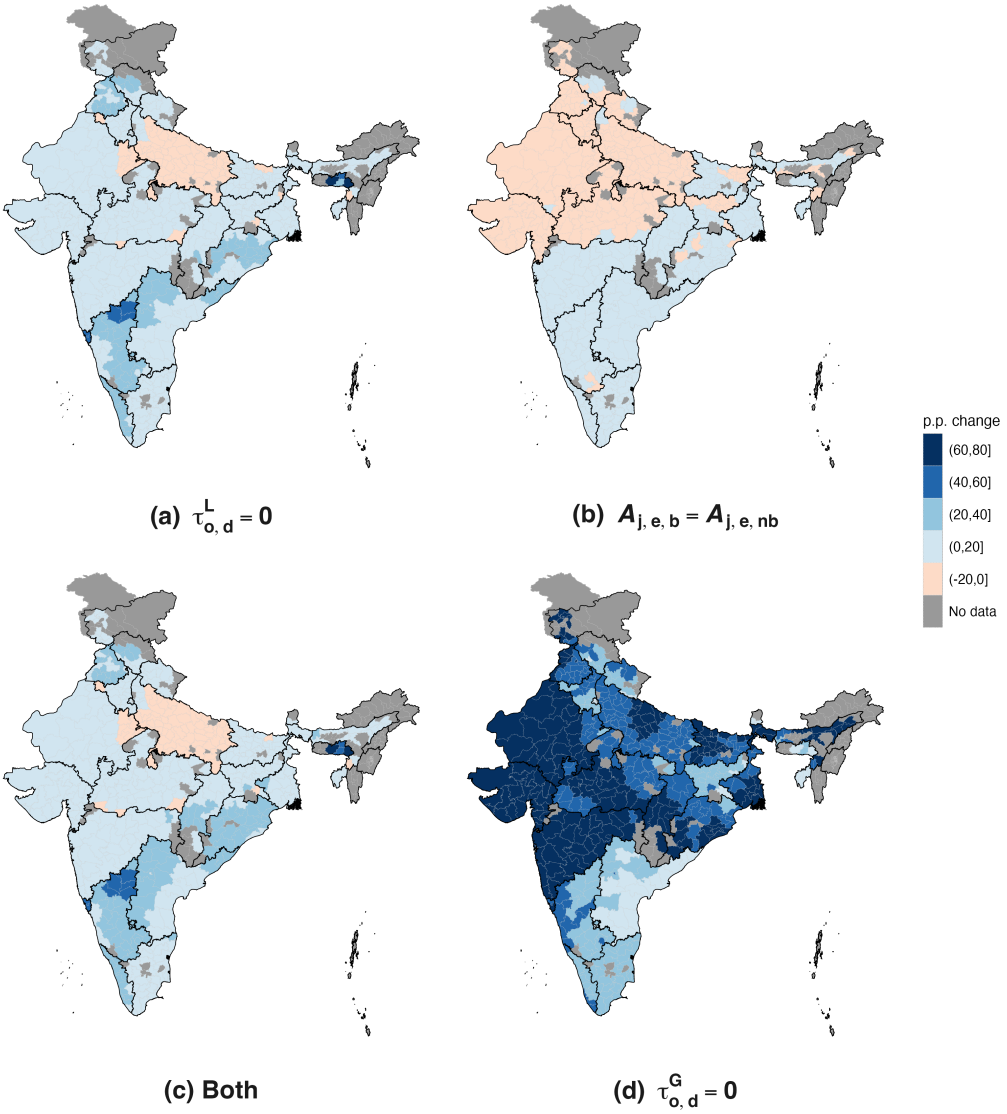
In-Migration: The Sorting Channel. Figure 6 shows the percentage-point change in the in-migration rate at each destination. Under counterfactuals 1 and 3, the largest inflows land on southern and northeastern destinations, with Karnataka and Kerala among the top receiving states; Goa also stands out. Under counterfactual 2 the map is nearly flat: the bilateral migration wedge is unchanged, and the improved productivity draws mostly induce local occupational switching rather than inter-state movement. The full removal of geographic costs in panel (d) distributes inflows far more evenly across central and northern districts (Chhattisgarh, Madhya Pradesh, Uttarakhand), the spatial signature of a friction that scales with distance rather than with language lines.

The migration-cost channel reallocates workers across space, and because $\eta^L_{nc} > \eta^L_c$, the inflow tilts toward non-college workers. General equilibrium caps it: as workers crowd in, $w_{d,j,e,b}$ falls through the inner CES nest and p_d falls through the regional aggregator, bringing the marginal migrant to indifference. The same logic applies symmetrically at origins, since language barriers are bilateral — southern states that are net destinations also ship workers out at a higher rate once the corridor opens in both directions.

The contrast between panels (a) and (d) is the spatial signature of language versus geographic frictions at the in-migration margin. Counterfactual 1 concentrates the response at high-barrier southern corridors, while the removal of geographic costs distributes it diffusely across the country. The same aggregate migration response can have very different spatial incidence depending on which friction is removed.

Efficiency per Worker: The Selection Channel. Figure 7 shows the percentage change in efficiency units per worker at each destination. Under counterfactual 2, destinations that host barrier-corridor non-college workers gain in per-worker efficiency, because those workers mechanically become more productive with no flow change to offset the gain through selection. Under counterfactual 1, two forces move this map in opposite directions. Within a (d, j, e, b) cell, the Fréchet selection effect lowers average quality as more workers crowd in, since average productivity $\tilde{\Gamma}(\pi_{o,d,j,e}/A_{j,e,b})^{-1/\theta}$ falls as π rises. Across cells, the same reallocation shifts workers toward high-productivity destinations and occupations, and this composition effect dominates: most of the largest receiving destinations gain in per-worker efficiency, with the selection dilution surviving only as a partial offset in a handful of receivers. Under counterfactual 3 the two forces stack:

Figure 6: Counterfactual Changes in In-Migration Rate by Destination

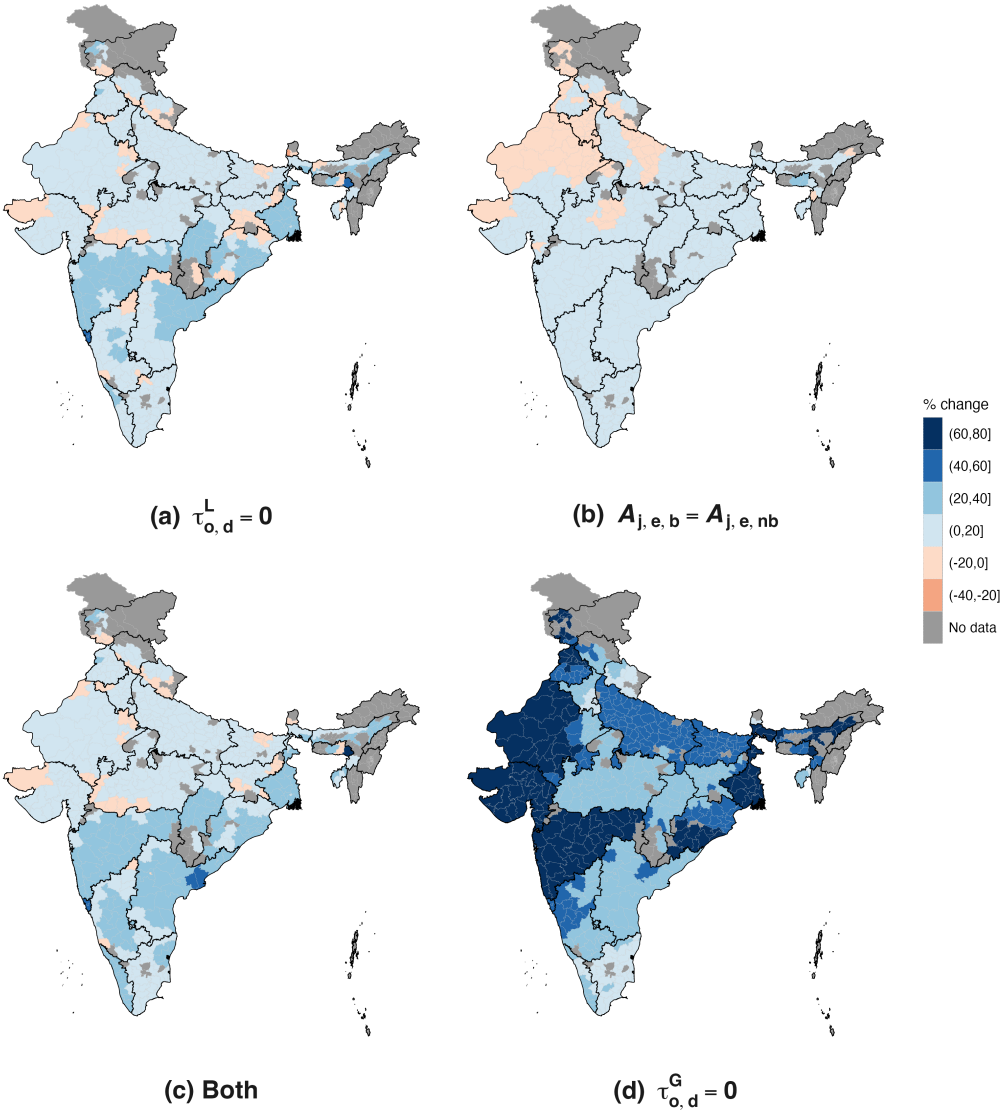


Notes: Each panel maps the percentage-point change in the gross in-migration rate at the district level under one of four counterfactuals. The gross in-migration rate at destination d is the number of cross-state migrants who arrive at d divided by the workforce at d. Panels (a)–(d) are as in Figure 5. Blue indicates positive changes; red indicates negative changes. Districts shaded grey have no data.

the productivity gain dominates the selection dilution at most corridors because the proportional A-jump is large relative to the dispersion implied by θ.

The contrast between Figure 6 and Figure 7 is between sorting and selection: sorting moves labor, while selection moves *effective* labor. The regional welfare map is approximately their sum, since regional welfare $p_d y_d$ scales with the workforce at d times the average efficiency of those workers. Under counterfactual 2 the welfare gain is almost entirely selection (efficiency rises at barrier corridors with no offsetting inflow); under counterfactual 1 the two channels move together:

Figure 7: Counterfactual Changes in Efficiency Units per Worker by Destination

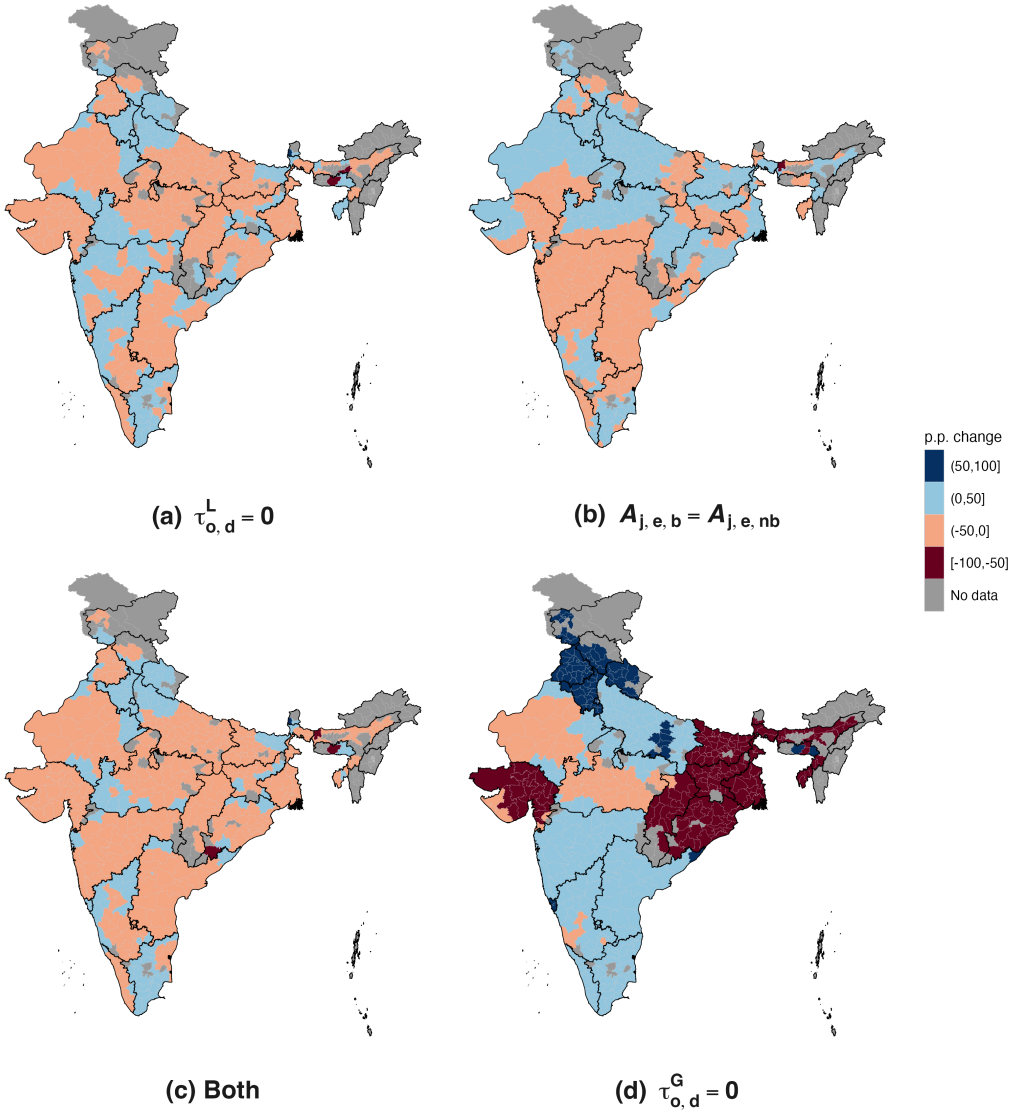


Notes: Each panel maps the percentage change in efficiency units per worker at the district level. Efficiency units per worker at destination d are total efficiency-unit labor supplied at d divided by the headcount of workers at d; the object isolates the selection (Fréchet quality) channel. Panels (a)–(d) are as in Figure 5. Blue indicates positive changes; red indicates negative changes. Districts shaded grey have no data.

workers crowd into southern corridors, and because the reallocation moves them toward better matches, per-worker efficiency rises at most of those destinations as well; under counterfactual 3 the two stack at southern barrier corridors, where the aggregate welfare gain concentrates.

Skill Premium. The two components of language have opposite implications for the spatial distribution of the skill premium, shown in Figure 8. Under counterfactual 2, the premium compresses across the destinations that host barrier-corridor non-college workers. While the resulting expansion in non-college effective labor triggers a general-equilibrium headwind in the middle nest, this

Figure 8: Counterfactual Changes in Skill Premium by Destination



Notes: Each panel maps the change in the skill premium at the district level. The skill premium at destination d is the percentage by which the population-weighted mean wage of college workers exceeds that of non-college workers earning income at d, and changes are in percentage points of this premium. Panels (a)–(d) are as in Figure 5. Blue indicates a wider premium at the destination; red indicates a narrower premium. Districts shaded grey have no data.

penalty is entirely overwhelmed by the direct productivity (A) gain that boosts non-college observed wages. Under counterfactual 1, the spatial premium response is mixed, and its direction tracks the skill composition of the flows a district gains or loses. Destinations absorbing a non-college-tilted inflow see the premium widen through relative-supply repricing; where the inflow is college-tilted, or where origins ship non-college workers out, the premium compresses. Districts in the eastern belt — Assam, Odisha, Chhattisgarh — show the strongest compression, while a few northern and northeastern districts widen.

Counterfactual 3 superimposes these opposing forces, and the resulting map is mixed in sign even though the underlying redistribution of relative supply is large. The premium response is itself a general equilibrium phenomenon: without the ρ-driven repricing of relative supply in the middle CES nest, this map would be flat. In the aggregate the premium compresses by 3.77 percentage points, so the friction-side compression dominates on average, even as sorting widens the premium locally at destinations whose inflow tilts toward non-college workers.

Thus, the effects of removing language barriers vary systematically with regions' baseline linguistic structure and with which channel is activated. The migration-cost channel reallocates labor across space, tilts the spatial composition of receiving destinations toward non-college workers, and generates most of the aggregate migration response, most of the aggregate welfare gain, and most of the spatial heterogeneity in the regional welfare map. The technological-friction channel mechanically lifts effective labor at barrier corridors and generates the largest direct contribution to the compression in the skill premium. Together, the two channels and their interaction account for a 10.9 percent gain in aggregate welfare and a 3.77-percentage-point compression in the skill premium — a welfare gain comparable to making India's geography as frictionless as the United States', and with a qualitatively different, regressive incidence across the skill distribution.

6.2 Language Policy

In this section, I weigh the costs and benefits of a language policy. The policy is motivated by cultural integration courses for foreign nationals in Germany, of which a major component is a language program. The program aims to bring migrants to working proficiency in German so that they can participate in the labor market without language barriers. This focus on language acquisition reflects the fundamental role that communication plays in economic integration and workplace productivity. A substantial body of evidence shows that language policy has first-order consequences for human development: distance from the official language of schooling and administration lowers educational, occupational, and wealth outcomes across countries, including in India (Laitin and Ramachandran, 2016, 2022).

In a similar vein, I introduce a simple extension to the model whereby non-college migrant workers, for whom language barriers bind in the estimated model, may enroll in language programs to overcome language barriers. To evaluate the potential benefits from this policy, I introduce two kinds of costs to the model. The first cost is faced by migrant workers: an opportunity cost of learning languages, $c_{o,d}$, which they incur if they choose to enroll in the program. If they choose to enroll in the program and incur the opportunity cost of learning, they may participate in the labor market without language barriers. The second cost is faced by the government: the expense of implementing language programs, which is ultimately borne by all workers in the economy in the form of a uniform income tax.

The language programs and associated costs introduce several changes to the model, which

are focused on region-pairs (o,d) that have a language barrier between them. When making the choice to migrate between any region-pair (o,d) that has a language barrier, non-college workers now make an additional choice of whether to enroll in the program. In other words, when language programs are made available to non-college workers, the language barrier becomes a function of whether or not they choose to enroll, i.e., $\mathscr{L}(o,d,\mathbb{1}\{\text{enroll}\})$. In particular, $\mathscr{L}(o,d,\mathbb{1}\{\text{enroll}\}) = b$ if $\mathbb{1}\{\text{enroll}\} = 0$ and $\mathscr{L}(o,d,\mathbb{1}\{\text{enroll}\}) = nb$ if $\mathbb{1}\{\text{enroll}\} = 1$. That is, language barriers remain for the worker if they choose not to enroll, but they are removed if they do.

For some migrants, it is optimal to migrate to a location where they face language barriers and incur the opportunity cost of learning. For them, the benefit of participating in the labor market without language barriers outweighs the opportunity cost of overcoming them. The fraction of non-college workers that choose to migrate from origin o to destination d to work in occupation j and learn the language is given by,

$$\pi_{o,d,j,nc}(\mathbb{1}\{\text{enroll}\} = 1) = \frac{A_{j,nc,nb}\left[(1-\tau^G_{o,d})^{\eta^G}\ \alpha_d\ (w_{d,j,nc,nb})\right]^{\theta}\ (c_{o,d})^{-\theta}}{(\Phi_{o,nc})^{\theta}}, \tag{15}$$

where $(\Phi_{o,nc})^{\theta} = \sum_{(d',j',\mathbb{1}\{\text{enroll}\})} A_{j',nc,\mathscr{L}(o,d',\mathbb{1}\{\text{enroll}\})}\ (\Lambda_{o,d',j',nc,\mathbb{1}\{\text{enroll}\}})^{\theta}$. The numerator captures that workers that would have previously had barrier productivity, $A_{j,nc,b}$, and incurred a language migration cost ($\tau^L_{o,d} > 0$) now have non-barrier productivity, $A_{j,nc,nb}$, and do not incur a language migration cost ($\tau^L_{o,d} = 0$). They do, however, incur the opportunity cost of learning, $c_{o,d}$, which enters as a multiplicative drag on the net wage: net of this cost, the effective wage of an enrolled worker scales by $1/c_{o,d}$.

However, for other migrants, it is optimal to migrate to a location where they face language barriers but choose *not* to participate in the language program. For them, the opportunity cost of overcoming language barriers outweighs the benefit of participating in the labor market without them. The fraction of non-college workers that choose to migrate from origin o to destination d to work in occupation j and do *not* learn the language is given by

$$\pi_{o,d,j,nc}(\mathbb{1}\{\text{enroll}\} = 0) = \frac{A_{j,nc,b}\left[(1-\tau^G_{o,d})^{\eta^G}(1-\tau^L_{o,d})^{\eta^L_{nc}}\ \alpha_d\ (w_{d,j,nc,b})\right]^{\theta}}{(\Phi_{o,nc})^{\theta}}, \tag{16}$$

where $(\Phi_{o,nc})^{\theta} = \sum_{(d',j',\mathbb{1}\{\text{enroll}\})} A_{j',nc,\mathscr{L}(o,d',\mathbb{1}\{\text{enroll}\})}\ (\Lambda_{o,d',j',nc,\mathbb{1}\{\text{enroll}\}})^{\theta}$. Notice that workers that would have previously had barrier productivity, $A_{j,nc,b}$, and incurred a language migration cost ($\tau^L_{o,d} > 0$) continue to do so. Since they opt not to participate in the language program, they do not incur the opportunity cost of learning, $c_{o,d}$. Workers sort into this case when $c_{o,d}$ is large relative to the wage gain from removing the language barrier. Since $c_{o,d}$ scales with Linguistic Distance$_{o,d}$, enrollment concentrates among migrants from origins with low linguistic distance to the destination.

To incorporate the cost of implementing language programs, I assume that the government finances the program by charging a proportional income tax at a uniform rate to all workers in the economy. A program outlay is a one-time cost, while income in the model is a per-period flow. I therefore first amortize each region's program outlay to a per-period equivalent by dividing it by the working-life annuity,

$$\overline{\text{Cost}}_r = \frac{\text{Cost of Language Programs}_r}{a_T}, \qquad a_T = \frac{1-\beta_m^{12T}}{1-\beta_m}, \qquad \beta_m = (1+r)^{-1/12}, \tag{17}$$

where β_m is the monthly discount factor implied by the annual interest rate r, T is the remaining working life, and a_T is the present value of one unit of income per period over the working life, so that the present value of the per-period flow $\overline{\text{Cost}}_r$ equals the one-time outlay. The tax rate is then set so that per-period revenue exactly covers the per-period program cost,

$$t = \frac{\sum_{r\in R} \overline{\text{Cost}}_r}{\sum_{d,j,e,\mathscr{L}} w_{d,j,e,\mathscr{L}}\, \ell_{d,j,e,\mathscr{L}}}, \tag{18}$$

where the numerator is the amortized per-period cost of implementing language programs across states of India and the denominator is aggregate labor income. Since the tax is levied at a common rate, it scales every alternative's net wage by the same factor $(1-t)$ and therefore cancels from the sorting probabilities in equations 15–16, leaving bilateral migration shares unchanged. It does, however, discount aggregate welfare.

I measure welfare under the policy as aggregate real income, consistent with the estimated model and the counterfactuals above. The opportunity cost of learning, $c_{o,d}$, enters this measure not as a separate deduction but through the supply of enrolled workers. An enrolled migrant spends part of the working life learning rather than producing, and therefore supplies only the fraction $1/c_{o,d}$ of a full non-barrier worker's efficiency units. This reduced supply is already booked in $\sum w\ell$ through market clearing, so the labor income forgone while learning requires no separate welfare term. The only additional deduction is the amortized program cost financed by the tax. Welfare under the policy is therefore

$$Y^{Policy} = \frac{1}{P}\left[\sum_{d,j,e,\mathscr{L}} w_{d,j,e,\mathscr{L}}\, \ell_{d,j,e,\mathscr{L}} - \sum_{r\in R} \overline{\text{Cost}}_r\right], \tag{19}$$

where the first term is aggregate nominal income, evaluated at the reduced efficiency units that enrolled workers supply while learning, and the second is the amortized cost of the language programs. Equivalently, $Y^{Policy} = (1-t)\sum w\ell/P$, since $\sum_{r\in R}\overline{\text{Cost}}_r = t\sum w\ell$ by equation 18. Because welfare remains aggregate real income, Y^{Policy} is directly comparable to the baseline Y and to the counterfactuals above.

Calibration of Costs: First, I calibrate $c_{o,d}$, the cost of learning the destination language. Learning takes a certain number of hours, during which the migrant forgoes labor income; once the language is learned, the migrant is no longer a barrier worker and earns the barrier-free wage thereafter. I express this cost in present-value terms, as the lifetime labor income a migrant forgoes while learning. To do so, I compute bilateral time costs using data from the United States State Department on the number of hours required for native English speakers to master 66 foreign languages up to "General Working Proficiency." Next, I compute the genealogical distance between English and each foreign language from Ethnologue language trees. Then, I estimate the following specification by least squares,

$$\text{Hours}_{\text{English, Foreign}} = \alpha \cdot \text{Distance}_{\text{English, Foreign}} + \varepsilon_{\text{English, Foreign}}, \tag{20}$$

where $\text{Distance}_{\text{English, Foreign}}$ is the genealogical distance between English and each foreign language from Ethnologue language trees. I fit the regression without a constant to be consistent with the fact that a native speaker of any language would need 0 hours to learn that language. The slope coefficient, $\hat{\alpha} = 1148$ hours, is the average time taken to learn languages. Then, I predict bilateral time costs by multiplying the coefficient from this regression, $\hat{\alpha}$, with the linguistic distance index between region-pairs,

$$\text{Time Cost}_{o,d} = \hat{\alpha} \cdot \text{Linguistic Distance}_{o,d}, \tag{21}$$

which measures the average number of hours that a person at origin o needs to overcome language barriers to destination d. Because migration is a once-and-for-all decision while the gains from learning the language persist over the worker's remaining career, I evaluate the cost of learning against the present value of lifetime income. A migrant who enrolls forgoes labor income while learning and earns the higher, barrier-free wage thereafter, whereas a migrant who does not enroll earns the barrier wage throughout. Let H denote annual hours worked,[11] so that learning takes $m_{o,d} = \text{Time Cost}_{o,d}/H$ years, let T denote the remaining working life, and let $\beta = 1/(1+r)$ be the annual discount factor. Discounting the income stream of each case and taking their ratio isolates the cost of enrolling as the dimensionless wedge

$$c_{o,d} = \frac{1-\beta^{T}}{\beta^{\,m_{o,d}}\left(1-\beta^{\,T-m_{o,d}}\right)} \;\geq\; 1, \tag{22}$$

which is the term that enters the enrollment probability in equation 15. The wedge captures both the income forgone during the $m_{o,d}$ years of learning and the discounting of the delayed wage gain: it equals one when learning is instantaneous ($m_{o,d} = 0$) and rises with linguistic distance through $\text{Time Cost}_{o,d}$. I set $T = 35$ years and $r = 0.05$. Because a one-time learning investment is small

[11] I compute annual hours worked from the National Sample Survey. Hours worked may differ between migrants and natives; since enrollees are migrants, I use the value for non-college migrants.

relative to a full career of higher wages, $c_{o,d}$ stays close to one, so that the enrollment decision is governed primarily by the wage gain from removing the language barrier rather than by the cost of learning itself.

Second, I calibrate the cost of language programs using budget and expenditure records of the CIIL, a subsidiary of the Ministry of Education, Government of India. The CIIL was set up in 1969 to promote linguistic harmony by teaching Indian languages to non-native learners. The CIIL, along with seven Regional Language Centers (RLCs), currently implements various language training programs in 22 languages.

In calibrating the cost of implementing language programs, I account for recurring costs of salaries for teachers, admin, and support staff, teacher training and curriculum design, and maintenance, as well as non-recurring infrastructure costs. I measure recurring costs as the average annual expenditure of the CIIL and its RLCs over 2014–2024, from budget and expenditure statements obtained under the Right to Information Act, expressed in 2018 rupees. I measure non-recurring costs from the 1969 Expenditure Finance Committee memorandum that sanctioned the language centers, converted to 2018 rupees using the consumer price index. Assuming that capital depreciates in 10 years, I distribute the non-recurring infrastructure costs across 10 years' worth of students. Each language center trains approximately 66 students per year and, on average, students need 1148 hours to learn a language, the average time cost implied by equation 20. Thus, I compute the cost per student to be \$5.2 per hour. As a sanity check, I compare this to the average cost of private English lessons in India, which is approximately \$5.6 per hour. The program cost sits just below this benchmark, which is reasonable: private tutors seek to make a profit, whereas language programs instituted by the government seek only to cover costs.

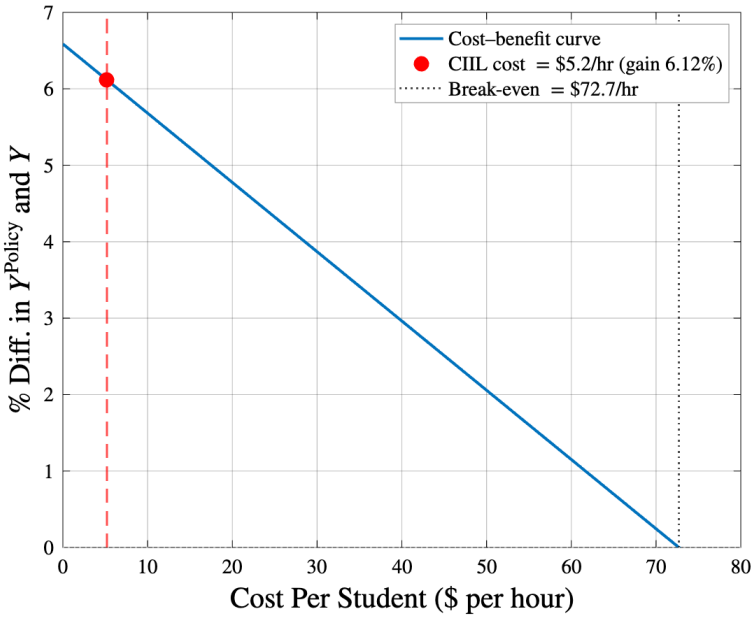
Cost vs. Benefit of Language Programs: To evaluate how the cost of language programs weighs against the benefit, I solve the model under a range of program costs. Tracing welfare across this range identifies the break-even cost above which the policy reduces aggregate welfare. At each cost level, I use the model to predict how many non-college migrant workers would incur $c_{o,d}$ and enroll in each region. I then take a weighted sum of cost per student across origins, with weights given by the number of hours that workers from each origin need to overcome language barriers at the destination (Time Cost$_{o,d}$ from equation 21).

For each value of the program cost, I compute aggregate welfare under the policy, Y^{Policy}, from equation 19, netting out the amortized program cost summed across regions. The program cost feeds welfare only through the income tax in equation 18, since the enrollment decision depends on the worker opportunity cost $c_{o,d}$ and not on the program cost. Because the financing tax cancels from the sorting probabilities, the equilibrium allocation, and hence the total instruction hours that enrolled workers require, does not depend on the program cost. The productivity gains from removing language barriers for the enrolled population are therefore pinned by $c_{o,d}$, and welfare is

exactly linear in the cost per student-hour. A single equilibrium solve thus traces the entire cost-benefit curve and pins the break-even cost analytically. I define the benefit of the program as the percent gap between Y^{Policy} and welfare in the baseline estimated model without the policy, Y.

The negative relationship in Figure 9 illustrates how welfare gains from language programs decline as implementation costs rise. The break-even cost, at which the welfare gap crosses zero, is \$73 per hour of instruction. Above this threshold, costs outweigh benefits, and it is not worth implementing such a policy. My calculation of \$5.2 per hour, derived from budget and expenditure records of the CIIL, falls substantially below this break-even threshold. At this cost of \$5.2 per person per hour, the policy generates a welfare gain of 6.1 percent.

Figure 9: Cost vs. Benefit of Language Policy



Notes: Cost is defined as the expense per student (\$ per hour) of instituting language programs. The expense per student is aggregated using information on how many workers that face language barriers will enroll in the program in each state and the number of hours it would take them to overcome the language barrier they face. Benefit is defined as the percent difference between Y^{Policy}, which is welfare under the policy (computed as aggregate real income net of the amortized cost of implementing language programs), and Y, which is welfare under the estimated model. The cost-benefit relationship is exactly linear because the financing tax cancels from the sorting probabilities. The red marker denotes the CIIL benchmark cost of \$5.2 per hour, at which the policy raises welfare by 6.1 percent; the dotted vertical line denotes the break-even cost of \$73 per hour.

To contextualize the magnitude of this 6.1 percent gain, I compare it to the 10.9 percent welfare gain under counterfactual 3, which removes language barriers entirely. The feasible policy at calibrated costs captures approximately 56 percent of this upper bound. The remaining gap reflects the narrower scope of the policy. For an enrolled non-college migrant, the program removes both channels of the language barrier, the technological friction ($A_{j,nc,b} \to A_{j,nc,nb}$) and the migration cost ($\tau^L_{o,d} \to 0$), exactly as counterfactual 3 does. The difference is the population reached: the

program operates only on non-college migrants who choose to enroll, whereas counterfactual 3 removes both channels for all workers, including college workers and non-migrants at barrier corridors. The actual cost also leaves a buffer of approximately $67 per hour below break-even, which is a margin of approximately 93 percent.

This buffer guards against implementation challenges or cost overruns. The analysis also likely understates true benefits, since it does not account for potential positive externalities from linguistic integration, such as the human capital gains that language-of-instruction reforms have generated in other multilingual developing countries (Ramachandran, 2017; Laitin et al., 2019), or long-term gains from enhanced labor market mobility. This counterfactual highlights that policies which directly address language barriers, even when restricted to a subset of the workforce, can generate substantial aggregate welfare gains at costs well below break-even. This finding supports the policy intervention of scaling the existing CIIL infrastructure to non-college migrants moving along high-barrier corridors, where the worker opportunity cost $c_{o,d}$ is largest and the productivity gain from enrollment is highest.

7 Conclusion

In this paper, I ask how language barriers shape internal migration, the skill premium, and aggregate welfare in a multilingual economy. I answer this question with Indian microdata, including confidential district-to-district migration records from the Census of India used here for the first time, and a linguistic distance index constructed from language trees. Three empirical facts anchor the analysis. Workers migrate less often to destinations where they face high language barriers. Migrants who face high barriers work less often in speaking-intensive occupations. Within occupations, migrants who face high barriers earn higher wages, consistent with positive selection at linguistic boundaries. These patterns are concentrated among unskilled workers; for college-educated workers, they are attenuated overall and eliminated entirely on the occupational margin.

Motivated by these facts, I develop a static quantitative spatial general equilibrium model in which language enters twice: as a component of bilateral migration costs and as a technological friction whose impact varies with the speaking intensity of the destination occupation. I estimate the model jointly by GMM with shift-share instruments. The estimates confirm the paper's central hypothesis. Unskilled workers who face language barriers lose roughly 40 percent of their efficiency in speaking-intensive occupations, but remain essentially unaffected in language-light work. Skilled workers are insulated altogether. Conditional on occupation and skill, workers with and without barriers are close substitutes in production, so the labor market cost of language operates through productivity rather than through imperfect substitutability. On the mobility side, the utility cost of crossing a language line is larger for unskilled than for skilled workers.

The counterfactuals quantify the aggregate stakes. Removing language barriers raises aggre-

gate welfare by 10.9 percent, increases internal migration by 14.6 percentage points, and compresses the skill premium by 3.8 percentage points. These two channels play distinct roles. The migration-cost channel drives the welfare gain and essentially all of the migration response, because it reallocates workers across space toward their best matches. The technological-friction channel contributes little to aggregate output on its own, but is the larger direct force compressing the skill premium. Their joint effect exceeds the sum of the parts: the workers who newly cross barriers once the migration cost is removed are the same workers whose productivity penalty the friction channel removes. The welfare gain from removing language barriers is slightly larger than the 8.4 percent gain achieved by reducing India's geographic frictions to United States levels. Moreover, removing language barriers has the opposite incidence, compressing rather than widening the skill premium. The gains concentrate in the linguistically distinct south and northeast, where barriers to the rest of the country are highest.

These results motivate language policy as a spatial policy. I extend the model with government-provided language programs for unskilled migrants and calibrate both the opportunity cost of learning and the cost of program provision. The break-even cost is approximately $73 per hour of instruction, while the cost calibrated from the budget records of the CIIL is $5.2 per hour. At this cost, the policy raises welfare by 6.1 percent, roughly 56 percent of the gain from removing language barriers entirely. Language training for unskilled migrants along high-barrier corridors is therefore a cost-effective complement to conventional place-based and transport policies.

This analysis points to several unexplored directions for future research. While the current model is static, a dynamic framework could explore the transition path of migration and the gradual acquisition of language skills over a worker's life. Future work might also incorporate unemployment or account for the intrinsic value of linguistic diversity. More broadly, the implications of this framework extend beyond India. Many multilingual countries in South Asia and Sub-Saharan Africa are undergoing structural transformation toward communication-intensive services, precisely where language frictions bind. Additionally, this lens is highly relevant to the European Union, where language programs are increasingly central to refugee integration. Measuring these barriers and evaluating language policy across such diverse settings is a promising agenda for future work.

References

Adda, J., C. Dustmann, and J.-S. Görlach (2022). The Dynamics of Return Migration, Human Capital Accumulation, and Wage Assimilation. *The Review of Economic Studies 89*(6), 2841–2871.

Adserà, A. and M. Pytliková (2015). The Role of Language in Shaping International Migration. *The Economic Journal 125*(586), F49–F81.

Allen, T., C. d. C. Dobbin, and M. Morten (2018). Border Walls. Working Paper 25267, National Bureau of Economic Research.

Altonji, J. G. and D. Card (1991). The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives. In J. M. Abowd and R. B. Freeman (Eds.), *Immigration, Trade, and the Labor Market*, pp. 201–234. University of Chicago Press.

Azam, M., A. Chin, and N. Prakash (2013). The Returns to English-Language Skills in India. *Economic Development and Cultural Change 61*(2), 335–367.

Bleakley, H. and A. Chin (2004). Language Skills and Earnings: Evidence from Childhood Immigrants. *Review of Economics and Statistics 86*(2), 481–496.

Bratsberg, B., A. Moxnes, O. Raaum, and K. H. Ulltveit-Moe (2023). Opening the Floodgates: Partial and General Equilibrium Adjustments to Labor Immigration. *International Economic Review 64*(1), 3–21.

Bryan, G. and M. Morten (2019). The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia. *Journal of Political Economy 127*(5), 2229–2268.

Burstein, A., G. Hanson, L. Tian, and J. Vogel (2020). Tradability and the Labor-Market Impact of Immigration: Theory and Evidence from the United States. *Econometrica 88*(3), 1071–1112.

Burstein, A., E. Morales, and J. Vogel (2019). Changes in Between-Group Inequality. *American Economic Journal: Macroeconomics 11*(2), 348–400.

Caliendo, L., M. Dvorkin, and F. Parro (2019). Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock. *Econometrica 87*(3), 741–835.

Card, D. (2001). Immigrant Inflows, Native Outflows, and the Local Market Impacts of Higher Immigration. *Journal of Labor Economics 19*(1), 22–64.

Census of India (2001). Migration Tables. Government of India, Office of the Registrar General & Census Commissioner.

Census of India (2011). Migration Tables. Government of India, Office of the Registrar General & Census Commissioner.

Centre for Monitoring Indian Economy (2019). Consumer Pyramids Household Survey. Mumbai: Centre for Monitoring Indian Economy Pvt. Ltd. Monthly income tables, waves September 2017–December 2019, accessed under license.

Chaudhary, L., Y. Dupraz, and J. Fenske (2025). A Century of Language Barriers to Migration in India. Working Paper 774, CAGE Online Working Paper Series.

Chiswick, B. R. and P. W. Miller (2015). International Migration and the Economics of Language. In *Handbook of the Economics of International Migration*, pp. 211–269. Elsevier.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Desai, S., R. Vanneman, and National Council of Applied Economic Research (2018). India Human Development Survey-II (IHDS-II), 2011–12. Inter-university Consortium for Political and Social Research [distributor]. Dataset, https://doi.org/10.3886/ICPSR36151.v6.

Desmet, K., E. Ghani, S. O'Connell, and E. Rossi-Hansberg (2015). The Spatial Development of India. *Journal of Regional Science 55*(1), 10–30.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2012). The Political Economy of Linguistic Cleavages. *Journal of Development Economics 97*(2), 322–338.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2017). Culture, Ethnicity, and Diversity. *American Economic Review 107*(9), 2479–2513.

Desmet, K. and F. Parro (2025). Spatial Dynamics. In *Handbook of Regional and Urban Economics*, Volume 6. Elsevier.

Dustmann, C. and F. Fabbri (2003). Language Proficiency and Labour Market Performance of Immigrants in the UK. *The Economic Journal 113*(489), 695–717.

Eberhard, D. M., G. F. Simons, and C. D. Fennig (2024). *Ethnologue: Languages of the World* (Twenty-seventh ed.). Dallas, Texas: SIL International. Online version.

Fan, J. (2019). Internal Geography, Labor Mobility, and the Distributional Impacts of Trade. *American Economic Journal: Macroeconomics 11*(3), 252–288.

Fan, T., M. Peters, and F. Zilibotti (2023). Growing Like India—The Unequal Effects of Service-Led Growth. *Econometrica 91*(4), 1457–1494.

Fearon, J. D. (2003). Ethnic and Cultural Diversity by Country. *Journal of Economic Growth 8*(2), 195–222.

Fenske, J. and N. Kala (2021). Linguistic Distance and Market Integration in India. *The Journal of Economic History 81*(1), 1–39.

Ghose, D. (2026). Trade, Internal Migration, and Human Capital: Who Gains from India's IT Boom? *The Review of Economics and Statistics*. Forthcoming.

Government of India (2017). Eighth Schedule to the Constitution of India. Ministry of Home Affairs, Government of India.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The Allocation of Talent and U.S. Economic Growth. *Econometrica 87*(5), 1439–1474.

Imbert, C. and J. Papp (2020). Costs and Benefits of Rural-Urban Migration: Evidence from India. *Journal of Development Economics 146*, 102473.

Katz, L. F. and K. M. Murphy (1992). Changes in Relative Wages, 1963–1987: Supply and Demand Factors. *The Quarterly Journal of Economics 107*(1), 35–78.

Kennan, J. and J. R. Walker (2011). The Effect of Expected Income on Individual Migration Decisions. *Econometrica 79*(1), 211–251.

Khanna, G. and N. Morales (2023). The IT Boom and Other Unintended Consequences of Chasing the American Dream. Working paper, Federal Reserve Bank of Richmond.

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*(4), 729–759.

Lagakos, D. and M. E. Waugh (2013). Selection, Agriculture, and Cross-Country Productivity Differences. *American Economic Review 103*(2), 948–980.

Laitin, D. D. and R. Ramachandran (2016). Language Policy and Human Development. *American Political Science Review 110*(3), 457–480.

Laitin, D. D. and R. Ramachandran (2022). Linguistic Diversity, Official Language Choice and Human Capital. *Journal of Development Economics 156*, 102811.

Laitin, D. D., R. Ramachandran, and S. L. Walter (2019). The Legacy of Colonial Language Policies and Their Impact on Student Learning: Evidence from an Experimental Program in Cameroon. *Economic Development and Cultural Change 68*(1), 239–272.

Melitz, J. (2008). Language and Foreign Trade. *European Economic Review 52*(4), 667–699.

Melitz, J. and F. Toubal (2014). Native Language, Spoken Language, Translation and Trade. *Journal of International Economics 93*(2), 351–363.

Monras, J. (2018). Economic Shocks and Internal Migration. Discussion Paper DP12977, Centre for Economic Policy Research.

Peri, G. and C. Sparber (2009). Task Specialization, Immigration, and Wages. *American Economic Journal: Applied Economics 1*(3), 135–169.

Rai, A. (2023). Temporary vs. Permanent Migration in a Spatial Economy: Evidence from India. SSRN Working Paper No. 4404189.

Ramachandran, R. (2017). Language Use in Education and Human Capital Formation: Evidence from the Ethiopian Educational Reform. *World Development 98*, 195–213.

Santos Silva, J. M. C. and S. Tenreyro (2006). The Log of Gravity. *The Review of Economics and Statistics 88*(4), 641–658.

Shastry, G. K. (2012). Human Capital Response to Globalization: Education and Information Technology in India. *Journal of Human Resources 47*(2), 287–330.

Spolaore, E. and R. Wacziarg (2009). The Diffusion of Development. *The Quarterly Journal of Economics 124*(2), 469–529.

Tombe, T. and X. Zhu (2019). Trade, Migration, and Productivity: A Quantitative Analysis of China. *American Economic Review 109*(5), 1843–1872.

Wang, Y. (2026). Cultural Distance and Internal Migration: Evidence from Indonesia. *Economic Development and Cultural Change 74*(2), 565–606.

Appendix Table of Contents

A Data **56**

A.1 Census Migration Data . . . 56

A.2 O*NET Speaking Intensity Index . . . 57

B Empirical Facts: Additional Results **59**

B.1 Fact 1: Language and Migration of College vs. Non-College . . . 59

B.1.1 Continuous Linguistic Distance . . . 59

B.1.2 Historical Migration Control . . . 59

B.2 Fact 2: Language and Occupations of College vs. Non-College . . . 61

B.2.1 Population-Weighted Threshold . . . 61

B.2.2 Continuous Speaking Score . . . 62

B.2.3 State-State vs. State-District Language Barriers . . . 63

B.3 Fact 3: Language and Wages of College vs. Non-College . . . 64

B.3.1 Continuous Linguistic Distance . . . 64

B.3.2 State-State vs. State-District Language Barriers . . . 65

C Model **67**

C.1 Model Derivations . . . 67

C.2 Model Solution . . . 72

D Estimation Appendix **74**

D.1 Summary of Parameters and Identifying Variation . . . 74

D.2 GMM Moment Conditions and Objective . . . 74

D.3 CES Moment Condition: Outer Nest (κ) . . . 76

D.4 CES Moment Condition: Middle Nest (ρ) . . . 77

D.5 CES Moment Condition: Inner Nest (ν) . . . 77

D.6 Substitution Formulas . . . 78

D.7 Shift-Share Instrumental Variables . . . 79

D.8 Imputation of Speaking-Occupation Status for the Pre-Sample Period . . . 80

D.9 Model Fit . . . 83

A Data

A.1 Census Migration Data

I obtain district-to-district internal migration tables from the 2001 and 2011 Censuses of India. The public release reports migrants only by destination district. To identify origin and destination jointly, I use a confidential extract from the Office of the Registrar General. The 2011 tables are cross-classified jointly by reason for migration, education, age, gender, and duration of residence, whereas the 2001 tables are marginal, with each dimension tabulated separately. For this reason I report the education breakdown below for 2011 only. Throughout the paper I restrict attention to work migrants, defined as individuals reporting work, employment, or business as their reason for migration.

Table 8 situates work migration within total internal migration in 2011. Work and business together account for only about 11 percent of migrants, while marriage is the single largest reason at 44 percent, followed by moves with family. Work migrants nonetheless stand out on every margin central to this paper: they are the most likely to move across state lines (27 percent, against 8 percent for marriage migrants), the most male (86 percent), and the most educated. These contrasts motivate the paper's focus on migration for work or business, the flows most exposed to language barriers.

Table 8: Composition of Internal Migration by Reason, 2011

Reason for Migration	Share of Migrants	% Inter-State	% College	% Male
Work/Employment	9.7	27.1	10.8	86.0
Business	0.9	21.0	9.3	77.0
Marriage	43.9	7.5	2.8	2.7
Moved with Family	22.7	13.6	5.3	51.3
Education	1.5	11.4	8.8	60.5
Other	21.4	7.1	4.4	55.2

Notes: Author's calculations from the confidential 2011 Census district-to-district migration tables. Each row is a reason for migration. "Share of Migrants" is the reason's percentage of all internal migrants. "% Inter-State" is the share of that reason's migrants who moved to another state. "% College" and "% Male" are, respectively, the share with a college degree and the share who are male.

Table 9 decomposes work migrants by the geography of their move. Most work migration is local: 65.9 percent of work migrants stayed within their origin state in 2001, rising to 73.4 percent in 2011, split between moves within the origin district and moves to another district in the same state. Inter-state migration, the flows most likely to cross a linguistic boundary, accounts for roughly a quarter of work migration in 2011 and a third in 2001, and the predominance of intra-state moves holds for both non-college and college workers.

Table 9: Geographic Distribution of Work Migrants

	2001	2011		
	All	All	Non-College	College
Within own district	31.8	37.1	37.6	33.4
Within state, other district	34.1	36.3	35.6	42.2
Intra-state (subtotal)	65.9	73.4	73.1	75.7
Inter-state	34.1	26.6	26.9	24.3
Migrants (millions)	30.8	74.1	66.3	7.9

Notes: Author's calculations from the confidential 2001 and 2011 Census district-to-district migration tables. The sample is work migrants, defined by a reason for migration of work, employment, or business. Each column reports the percentage distribution of work migrants across move types: within the origin district, to another district in the same state, and to another state. Intra-state is the sum of the first two rows. The 2011 columns split migrants by education; the 2001 tables are not cross-classified by education, so only the total is shown. The final row reports the number of work migrants in millions.

A.2 O*NET Speaking Intensity Index

To classify occupations by their speaking intensity, I construct a composite index from four O*NET Importance measures, each capturing a distinct dimension of oral communication at work.

Components. The index is the equal-weighted average of:

1. **Speaking Importance** (O*NET Skills): Defined as "talking to others to convey information effectively." Scores from `speaking.csv` are pre-normalized to a 0–100 scale.

2. **Active Listening Importance** (O*NET Skills): "Giving full attention to what other people are saying, taking time to understand the points being made." Raw scores on the 1–5 importance scale.

3. **Oral Comprehension Importance** (O*NET Abilities, Element ID 1.A.1.a.1): The ability to listen to and understand information and ideas presented through spoken words and sentences. Raw scores on the 1–5 importance scale.

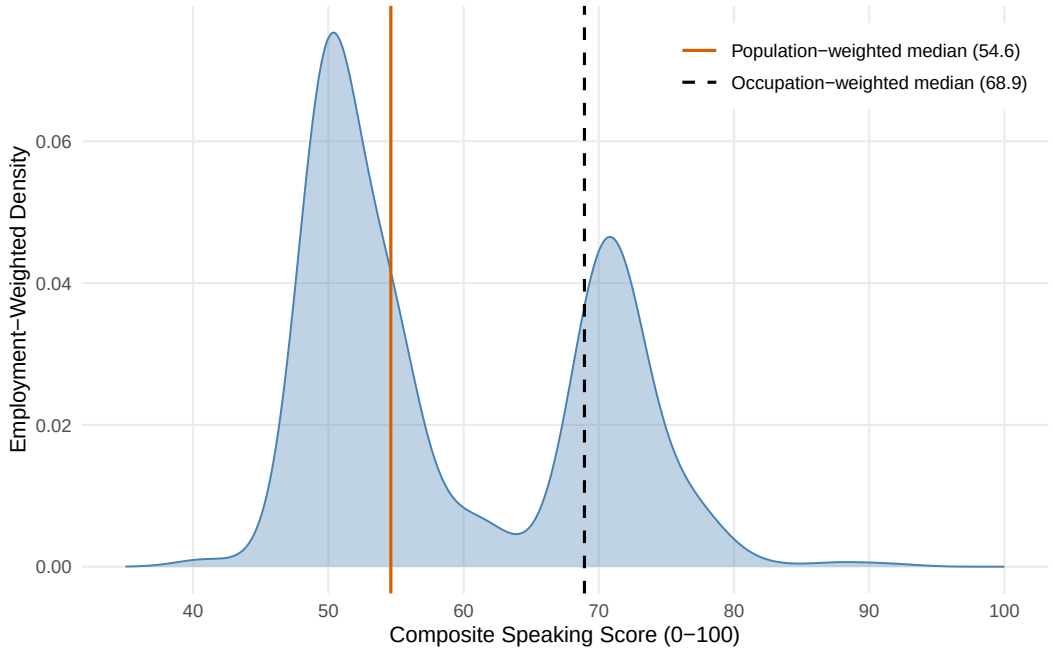
4. **Oral Expression Importance** (O*NET Abilities, Element ID 1.A.1.a.3): The ability to communicate information and ideas in speaking so others will understand. Raw scores on the 1–5 importance scale.

Components 2–4 are normalized from the 1–5 importance scale to 0–100 via $(x-1)/4 \times 100$ before averaging. The composite is the simple mean of all available components; 87 percent of matched

occupations have all four. The index is drawn from O*NET database version 23.1, covering 873 occupations in the Skills data and 967 in the Abilities data.

CPHS–O*NET Concordance. Of 218 CPHS occupation categories, 213 are matched to specific O*NET occupation codes through a hand-constructed concordance. Five non-worker and unmatched categories (Home Makers, Retired/Aged, Students, Unoccupied, and Not Applicable) are excluded. Residual "All Other" O*NET categories (e.g., 51-9199.00 Production Workers, All Other), which are not individually surveyed by O*NET, are replaced with the most specific applicable code.

Figure 10: Employment-Weighted Distribution of Speaking Scores



Notes: Author's calculations from CPHS and O*NET. The curve is a kernel density of the composite speaking score across the 213 O*NET-matched occupations, each weighted by its share of national employment (the sum of CPHS sampling weights). The composite score is the equal-weighted average of the four O*NET oral-communication importance measures described above. Vertical lines mark the occupation-weighted median (68.9), which defines the speaking-intensive indicator in Fact 2, and the population-weighted median (54.6), used in robustness checks.

Threshold. The binary speaking-intensive indicator used in Fact 2 is defined as an occupation's composite score exceeding the occupation-weighted median (68.9), computed with each of the 213 matched occupations counting once. This is the stricter of the two candidate cutoffs: the population-weighted median, which instead weights each worker by the CPHS sampling weight, is lower at 54.6. Robustness results using this looser threshold and the continuous score are in Appendix B.2.

Figure 10 plots the employment-weighted distribution of the composite score, with each occupation weighted by its share of national employment. The distribution is bimodal: a large mass of

manual and agricultural labourers sits near 50, and a second mass of service, trade, and white-collar occupations sits near 70, separated by a sparse middle. Since the low-scoring occupations employ a disproportionate share of workers, the population-weighted median (54.6) falls in the lower mass, whereas the occupation-weighted median (68.9) falls in the upper one. By adopting the higher, stricter occupation-weighted cutoff, Fact 2 classifies only the upper mass as speaking-intensive.

B Empirical Facts: Additional Results

B.1 Fact 1: Language and Migration of College vs. Non-College

B.1.1 Continuous Linguistic Distance

Table 10 replicates the main result replacing the binary barrier indicator $\mathbb{1}\{\text{Barrier}_{o,d}\}$ with the raw continuous linguistic distance index $\tau^L_{o,d} \in [0,1]$. All three columns include a control for the log 2001 outmigration share, $\ln(2001\ \text{Outmigration Share}_{o,d})$, which absorbs bilateral network effects and historical migration patterns that would otherwise confound the continuous slope estimate. The three columns build up controls progressively: Column (1) includes only linguistic distance and the 2001 network control; Column (2) adds geographic distance; Column (3) adds the education interaction.

The results confirm the main finding. Column (3) shows that a one-unit increase in the linguistic distance index reduces migration by 1.460 log points for non-college workers ($p < 0.01$). The interaction term $\tau^L_{o,d} \times \mathbb{1}\{\text{College}\}$ is $+0.170$ ($p < 0.05$), indicating that the negative slope is attenuated for college workers: the implied effect for college workers is $-1.460 + 0.170 = -1.290$, roughly 12 percent smaller in absolute value. The attenuation is consistent in sign and significance with the binary threshold specification in the main text.

B.1.2 Historical Migration Control

Table 11 replicates Table 3 while controlling for historical migration patterns using 2001 census data. This is a demanding robustness check that addresses several potential concerns. The 2001 migration data includes *all* migrants (not just work migrants), capturing the full stock of historical migration across all reasons. This broad measure absorbs time-invariant bilateral factors including cultural ties, linguistic networks, family connections, chain migration patterns, and historical settlement patterns that may correlate with contemporary migration and education.

Critically, because these networks accumulated over decades before 2011, they are predetermined with respect to the 2011 work migration outcomes, addressing concerns about reverse causality. By including this control, I account for the possibility that certain origin-destination pairs simply have higher migration propensities for reasons unrelated to language barriers, or that

college workers systematically select into corridors with established networks.

Table 10: Language and Migration of College vs. Non-College (Fact 1)
Robustness: Continuous Linguistic Distance

	Dep. Var.: $N_{o,d,e,g}$		
	(1)	(2)	(3)
$\tau^L_{o,d}$	-1.891***	-1.440***	-1.460***
	(0.110)	(0.095)	(0.097)
$\tau^L_{o,d} \times \mathbb{1}\{\text{College}\}$			0.170**
			(0.080)
$\ln(\text{Geo. Dist.}_{o,d})$		-0.346***	-0.361***
		(0.052)	(0.052)
$\ln(\text{2001 Outmigration Share}_{o,d})$	1.048***	0.934***	0.925***
	(0.013)	(0.016)	(0.017)
$\mathbb{1}\{\text{College}\}$			0.087**
			(0.034)
Origin District FE	Yes	Yes	Yes
Destination District FE	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes
Observations	1,352,568	1,352,568	1,352,568
Pseudo R-squared	0.956	0.958	0.959

Notes: The gravity model is estimated using the PPML estimator on the universe of bilateral economic migration flows, with fixed-effects for origin district, destination district, gender, and offset by 2011 origin district population. The continuous linguistic distance index is used to measure bilateral language barriers. Standard errors are clustered two-way at origin and destination districts. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The results provide compelling evidence for Fact 1. The language barrier interaction is highly significant (Column 3: $p < 0.001$), with a magnitude economically similar to the main text. This indicates that the differential effect by education is not driven by omitted variables, historical networks, or path dependence. The coefficient on log 2001 migration share is 0.944 ($p < 0.001$), confirming strong persistence in migration patterns over time and validating that the control captures meaningful bilateral variation.

Strikingly, the attenuation of the language barrier for college workers more than doubles relative to Table 3. In Column (3), college workers face a language barrier effect that is 53 percent smaller than non-college workers, compared to 25 percent in the main table. This amplification reveals that the 2001 control absorbs general bilateral mobility differences between education groups, leaving a larger and sharper language-specific differential. Separately, when I add distance $\times$ college as an additional interaction (not shown), it is economically and statistically negligible ($p = 0.264$), while the language interaction remains highly significant. Thus, language represents

a specific friction, not a proxy for general mobility costs. College workers are better equipped to overcome language barriers precisely because of channels such as English proficiency and access to occupations less dependent on local language fluency, supporting the mechanisms explored in Facts 2 and 3.

Table 11: Language and Migration of College vs. Non-College (Fact 1)
Robustness: Add 2001 Migration Control

	Dep. Var.: $N_{o,d,e,g}$		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$	-0.782***	-0.604***	-0.645***
	(0.099)	(0.083)	(0.077)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}\}$			0.339***
			(0.054)
$\ln(\text{Geo. Dist.}_{o,d})$		-0.431***	-0.445***
		(0.049)	(0.050)
$\ln(\text{2001 Outmigration Share}_{o,d})$	1.118***	0.953***	0.944***
	(0.013)	(0.014)	(0.015)
$\mathbb{1}\{\text{College}\}$			0.110***
			(0.028)
Origin District FE	Yes	Yes	Yes
Destination District FE	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes
Observations	1,352,568	1,352,568	1,352,568
Pseudo R-squared	0.953	0.957	0.957

Notes: The gravity model is estimated using the PPML estimator on the universe of bilateral economic migration flows, with fixed-effects for origin district, destination district, gender, and offset by 2011 origin district population. The 2001 district-to-district migration flows are used as an additional control to capture bilateral network effects. Standard errors are clustered two-way at origin and destination districts. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Fact 2: Language and Occupations of College vs. Non-College

B.2.1 Population-Weighted Threshold

Table 12 reclassifies speaking-intensive occupations using the *population*-weighted median (54.6) of the composite score, weighting each worker observation by the CPHS sampling weight, in place of the occupation-weighted median ($\approx$68.9) used in the main text. This looser cutoff classifies a larger set of moderate-communication occupations as speaking-intensive. The sorting pattern is preserved and, if anything, sharper: the barrier effect is -0.088 ($p < 0.01$) and the interaction is $+0.054$ ($p < 0.01$), with a college coefficient of 0.273. That the result holds, indeed strengthens,

at the alternative threshold confirms that the occupational-sorting mechanism is not a knife-edge artifact of the stricter cutoff.

Table 12: Language and Occupations of College vs. Non-College (Fact 2)
Robustness: Population-Weighted Threshold (54.6)

	Dep. Var.: $\mathbb{1}\{\text{Speaking Occ.}_{i,o,d,j,t}\}$		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		-0.075***	-0.088***
		(0.019)	(0.020)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$			0.054***
			(0.011)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.281***	0.281***	0.273***
	(0.007)	(0.007)	(0.008)
$\ln(\text{Geo. Dist.}_{o,d})$	0.001	0.011**	0.009*
	(0.005)	(0.005)	(0.005)
$\ln(\text{Wages}_{i,o,d,j,t-1})$	0.008	0.008	0.008
	(0.007)	(0.007)	(0.007)
Origin State × Month FE	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes
Observations	96,014	96,014	96,014
Adj. R-squared	0.205	0.205	0.205

Notes: The linear probability model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, lagged log real wage, and fixed-effects for origin state × month, destination district × month. The dependent variable equals one for occupations whose composite O*NET speaking score exceeds the population-weighted median (54.6). Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2.2 Continuous Speaking Score

Table 13 replicates the full three-column sequence from Table 4 using the continuous composite speaking score (0–100) as the dependent variable in place of the binary indicator, avoiding any threshold choice. In the full specification, the barrier coefficient is -1.50 ($p < 0.01$) and the interaction is $+1.20$ ($p < 0.01$), implying a net barrier effect for college migrants of approximately -0.30 ($-1.50 + 1.20$). The college coefficient is $+6.73$ ($p < 0.01$), indicating that college workers score about 7 points higher on the 0–100 composite scale. The pattern is consistent with the binary results: language barriers reduce speaking intensity for non-college migrants, college education strongly increases it, and the interaction substantially attenuates (though does not fully offset) the barrier for college workers.

Table 13: Language and Occupations of College vs. Non-College (Fact 2)
Robustness: Continuous Speaking Score

	Dep. Var.: Speaking Occ. Index$_{i,o,d,j,t}$ (0–100)		
	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		-1.209***	-1.498***
		(0.393)	(0.407)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$			1.203***
			(0.373)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	6.899***	6.896***	6.728***
	(0.187)	(0.187)	(0.188)
$\ln(\text{Geo. Dist.}_{o,d})$	0.050	0.212	0.154
	(0.130)	(0.132)	(0.133)
$\ln(\text{Wages}_{i,o,d,j,t-1})$	2.641***	2.645***	2.648***
	(0.197)	(0.197)	(0.196)
Origin State × Month FE	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes
Observations	96,014	96,014	96,014
Adj. R-squared	0.512	0.512	0.512

Notes: The linear regression is estimated using ordinary least squares on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, lagged log real wage, and fixed-effects for origin state × month, destination district × month. The dependent variable is the continuous composite O*NET speaking index (0–100 scale). Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2.3 State-State vs. State-District Language Barriers

Table 14 replicates the full specification using state-to-state linguistic and geographic distances in place of the state-to-district measures used in the main table. State-to-state measures are a coarser aggregation that may be appropriate if migrants' destination choices operate at the state level, or if district-level language composition is measured with error. The state-to-state barrier effect is -0.035 ($p < 0.05$) and the interaction is $+0.061$ ($p < 0.01$), close to the state-district estimates of -0.049 and $+0.057$, confirming that the findings are robust to the level of geographic aggregation at which linguistic distance is measured. The college main effect (0.235) and the lagged-wage coefficient (0.090) are identical across the two columns. The implied net barrier for college migrants stays near zero under the coarser measure ($-0.035 + 0.061 = +0.026$). This is larger than the state-district net effect of $+0.008$, but both indicate that college education roughly neutralizes the occupational-sorting penalty. The one material change is geographic distance, whose coefficient falls from 0.014 ($p < 0.05$) to an insignificant 0.008, consistent with the state-state measure

carrying less spatial information.

Table 14: Language and Occupations of College vs. Non-College (Fact 2)
Robustness: State-State vs. State-District Language Barriers

	(1) State-District	(2) State-State
$\mathbb{1}\{\text{Barrier}_{o,d}\}$	-0.049***	-0.035**
	(0.016)	(0.015)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.057***	0.061***
	(0.014)	(0.014)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.235***	0.235***
	(0.010)	(0.010)
$\ln(\text{Geo. Dist.}_{o,d})$	0.014**	0.008
	(0.007)	(0.008)
$\ln(\text{Wages}_{i,o,d,j,t-1})$	0.090***	0.090***
	(0.007)	(0.007)
Origin State × Month FE	Yes	Yes
Dest. District × Month FE	Yes	Yes
Observations	96,014	96,014
Adj. R-squared	0.184	0.184

Notes: The linear probability model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, lagged log real wage, and fixed-effects for origin state × month, destination district × month. Linguistic index and geographic distance are constructed at the state-district aggregation in Column (1) and state-state aggregation in Column (2). The dependent variable equals one for occupations whose composite O*NET speaking score exceeds the occupation-weighted median (≈68.9). Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.3 Fact 3: Language and Wages of College vs. Non-College

B.3.1 Continuous Linguistic Distance

Table 15 replaces the binary barrier indicator $\mathbb{1}\{\text{Barrier}_{o,d}\}$ with the raw continuous linguistic distance index $\tau^L_{o,d} \in [0,1]$, building up over three columns to a full specification that corresponds to Column 3 of Table 5. This avoids discretizing the distribution at an arbitrary threshold and tests whether the attenuation result is an artifact of the binary coding.

The results are consistent with the main specification. A one-unit increase in linguistic distance, moving from no distance to the maximum, is associated with an approximately 35 percent wage premium for non-college migrant workers (0.303 log points, $p < 0.01$). The interaction term $\tau^L_{o,d} \times \mathbb{1}\{\text{College}\}$ is -0.196 ($p < 0.01$), indicating that the effect is substantially attenuated for

college workers, whose implied wage effect is approximately 11 percent (0.107 log points). The attenuation ratio is similar in magnitude to the binary specification. The coefficient on geographic distance is small and insignificant in this specification, consistent with the continuous linguistic distance measure absorbing most of the cross-corridor variation in barriers.[12]

Table 15: Language and Wages of College vs. Non-College (Fact 3)
Robustness: Continuous Linguistic Distance

	Dep. Var.: $\ln(\text{Wage}_{i,o,d,j,t})$		
	(1)	(2)	(3)
$\tau^L_{o,d}$	0.244***	0.251***	0.303***
	(0.033)	(0.042)	(0.043)
$\tau^L_{o,d} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$			-0.196***
			(0.023)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.262***	0.262***	0.354***
	(0.011)	(0.011)	(0.015)
$\ln(\text{Geo. Dist.}_{o,d})$		-0.002	0.006
		(0.007)	(0.007)
Occupation × Month FE	Yes	Yes	Yes
Origin State × Month FE	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes
Observations	101,207	101,207	101,207
Adj. R-squared	0.671	0.671	0.672

Notes: The linear regression model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, and fixed-effects for occupation × month, origin state × month, destination district × month. The continuous linguistic distance index is used to measure bilateral language barriers. Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.3.2 State-State vs. State-District Language Barriers

Table 16 presents robustness checks for the wage regression (Fact 3) using alternative measures of linguistic and geographic distance. The main specification measures linguistic distance at the state-to-district level, capturing the expected language barrier a migrant from origin state o would face in destination district d. An alternative approach measures distance at the state-to-state level,

[12] The college coefficient in the continuous specification (0.354) differs from the binary specification (0.269) because the base groups differ. In the continuous specification, the college effect is identified at $\tau^L_{o,d} = 0$ (zero linguistic distance). In the binary specification, it is identified at $\mathbb{1}\{\text{Barrier}_{o,d}\} = 0$, which corresponds to corridors below the median linguistic distance. These are not the same base group, so the magnitudes are not directly comparable.

which may be appropriate if migrants' location choices operate at a coarser geographic level or if district-level language composition is measured with error.

Column 1 reproduces the main specification using state-district measures for both linguistic distance and geographic distance. Column 2 uses state-state measures for both. The results are highly consistent: the barrier coefficient is positive and highly significant in both cases (0.082 for state-district; 0.086 for state-state, both $p < 0.01$), and the interaction is negative and significant in both (-0.054 and -0.053, respectively, both $p < 0.01$). The attenuation of the language barrier for college workers is robust to the level of geographic aggregation.

Table 16: Language and Wages of College vs. Non-College (Fact 3)
Robustness: State-State vs. State-District Language Barriers

	Dep. Var.: $\ln(\text{Wage}_{i,o,d,j,t})$	
	(1) State-District	(2) State-State
$\mathbb{1}\{\text{Barrier}_{o,d}\}$	0.082***	0.086***
	(0.018)	(0.017)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	-0.054***	-0.053***
	(0.015)	(0.015)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	0.269***	0.269***
	(0.011)	(0.011)
$\ln(\text{Geo. Dist.}_{o,d})$	0.019***	0.010
	(0.006)	(0.006)
Occupation × Month FE	Yes	Yes
Origin State × Month FE	Yes	Yes
Dest. District × Month FE	Yes	Yes
Observations	101,207	101,207
Adj. R-squared	0.671	0.671

Notes: The linear regression model is estimated using the OLS estimator on the sample of migrants (origin state $\neq$ destination state) from monthly individual income tables in CPHS waves Sep 2017 to Dec 2019, with controls for gender, age, caste, and fixed-effects for occupation × month, origin state × month, destination district × month. Linguistic index and geographic distance are constructed at the state-district aggregation in Column (1) and state-state aggregation in Column (2). Standard errors are clustered at destination district × month. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Model

C.1 Model Derivations

This appendix presents the derivation of the main equations of the model. First, I explain the notation used in the paper and list the model parameters for reference. Second, I derive the sorting and selection equations from the workers' utility maximization problem. Third, I derive the price equations and demand conditions from the optimization problems of the regional and aggregate firms.

Notation: In the table below, I describe the notation used in the paper to denote regions, occupations, and workers.

Regions ($r \in R$)	Origin	$o \in R$
	Destination	$d \in R$
Occupations ($j \in J$)	$j \in \{$Speaking, Non-Speaking$\}$	$j \in \{$spk, non-spk$\}$
Workers (i)	Education $\in \{$College, Non-College$\}$	$e \in \{c, nc\}$
	Language Barrier $\in \{$Barrier, Non-Barrier$\}$	$\mathscr{L}(o,d) \in \{b, nb\}$

Model Parameters: In the table below, I list the model parameters and describe what they mean.

Parameter	Description	Dimensions
$N_{o,e}$	worker endowments	$R \times 2$
$\tau^G_{o,d}$	geographic barriers	$R \times R$
$\tau^L_{o,d}$	language barriers	$R \times R$
η^G	sensitivity to geographic barriers	1×1
η^L_e	sensitivity to language barriers	2×1
α_d	amenities	$R \times 1$
$A_{j,e,\mathscr{L}(o,d)}$	Fréchet, location ($\mathscr{L}$-independent for $e = c$)	6
θ	Fréchet, dispersion	1×1
σ	elasticity parameter, aggregate nest	1×1
κ	elasticity parameter, outer nest	1×1
$\phi_{d,j}$	share parameter, outer nest	$R \times J$
ρ	elasticity parameter, middle nest	1×1
ν	elasticity parameter, inner nest	1×1

Utility and Sorting: The key sorting equation is derived from the probability that worker i from origin o with education e chooses occupation j at location d. Let $z^i_{j,e,\mathscr{L}(o,d)} \equiv z$ and $z^i_{j',e,\mathscr{L}(o,d')} \equiv z'$. From the properties of the Fréchet distribution,

$$
\begin{aligned}
\pi_{o,d,j,e} &= \Pr\Big(\Lambda_{o,d,j,e}\, z \geq \Lambda_{o,d',j',e}\, z' \;\; \forall (d',j') \neq (d,j)\Big) \\
&= \Pr\Big(z' \leq \frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z \;\; \forall (d',j') \neq (d,j)\Big) \\
&= \int_0^\infty \prod_{(d',j')\neq(d,j)} G_{j',e,\mathscr{L}(o,d')}\Big(\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z\Big)\, dG_{j,e,\mathscr{L}(o,d)}(z) \\
&= \int_0^\infty \prod_{(d',j')\neq(d,j)} \exp\Big[-A_{j',e,\mathscr{L}(o,d')}\Big(\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z\Big)^{-\theta}\Big]\, dG_{j,e,\mathscr{L}(o,d)}(z) \\
&= \int_0^\infty \exp\Big[-(\Lambda_{o,d,j,e} z)^{-\theta} \sum_{(d',j')\neq(d,j)} A_{j',e,\mathscr{L}(o,d')} (\Lambda_{o,d',j',e})^{\theta}\Big] A_{j,e,\mathscr{L}(o,d)} \theta z^{-\theta-1} \\
&\quad \cdot \exp\Big(-A_{j,e,\mathscr{L}(o,d)} z^{-\theta} (\Lambda_{o,d,j,e})^{-\theta} (\Lambda_{o,d,j,e})^{\theta}\Big) dz \\
&= \int_0^\infty \exp\Big[-(\Lambda_{o,d,j,e} z)^{-\theta} \sum_{(d',j')\neq(d,j)} A_{j',e,\mathscr{L}(o,d')} (\Lambda_{o,d',j',e})^{\theta} - \\
&\quad (\Lambda_{o,d,j,e} z)^{-\theta} A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta}\Big] A_{j,e,\mathscr{L}(o,d)} \theta z^{-\theta-1} dz \\
&= \int_0^\infty \exp\Big[-(\Lambda_{o,d,j,e} z)^{-\theta} \underbrace{\sum_{(d',j')} A_{j',e,\mathscr{L}(o,d')} (\Lambda_{o,d',j',e})^{\theta}}_{\equiv (\Phi_{o,e})^{\theta}}\Big] A_{j,e,\mathscr{L}(o,d)} \theta z^{-\theta-1} dz \\
&= \frac{A_{j,e,\mathscr{L}(o,d)}}{(\Lambda_{o,d,j,e})^{-\theta} (\Phi_{o,e})^{\theta}} \\
&\quad \times \int_0^\infty (\Lambda_{o,d,j,e})^{-\theta} (\Phi_{o,e})^{\theta} \theta z^{-\theta-1} \exp\Big(-(\Lambda_{o,d,j,e})^{-\theta} (\Phi_{o,e})^{\theta} z^{-\theta}\Big) dz \\
&= \frac{A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta}}{(\Phi_{o,e})^{\theta}}
\end{aligned}
$$

Expected Productivity and Selection: The key selection equation is derived from the expected productivity of worker i from origin o and education e conditional on having chosen destination d and occupation j. Let $z^i_{j,e,\mathscr{L}(o,d)} \equiv z$ and $z^i_{j',e,\mathscr{L}(o,d')} \equiv z'$. From the properties of the Fréchet distribution,

$$
E\Big[z \mid z' \leq \frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z \;\; \text{s.t. } (d',j') \neq (d,j)\Big]
$$

$$
\begin{aligned}
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z\, g_{j,e,\mathscr{L}(o,d)}(z) \prod_{(d',j')\neq(d,j)} G_{j',e,\mathscr{L}(o,d')}\Big(\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z\Big)\, dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z\, A_{j,e,\mathscr{L}(o,d)} \theta z^{-\theta-1} \exp\Big(-A_{j,e,\mathscr{L}(o,d)} z^{-\theta}\Big) \\
&\quad \times \exp\Big(-(\Lambda_{o,d,j,e})^{-\theta} z^{-\theta} \sum_{(d',j')\neq(d,j)} A_{j',e,\mathscr{L}(o,d')} (\Lambda_{o,d',j',e})^{\theta}\Big)\, dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z\theta A_{j,e,\mathscr{L}(o,d)} z^{-\theta-1} \exp\Big(-(\frac{\Phi_{o,e}}{\Lambda_{o,d,j,e}})^{\theta} z^{-\theta}\Big)\, dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z\theta A_{j,e,\mathscr{L}(o,d)} z^{-\theta-1} \exp\Big(-\frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta}\Big)\, dz \\
&= \int_{z=0}^{\infty} z\theta \frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta-1} \exp\Big(-\frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta}\Big)\, dz \\
&= \Gamma(1-1/\theta)(\pi_{o,d,j,e}/A_{j,e,\mathscr{L}(o,d)})^{-1/\theta}
\end{aligned}
$$

Expected Utility: The expected utility of worker i from origin o with education e is derived across all possible destinations d and occupations j. Let $z^i_{j,e,\mathscr{L}(o,d)} \equiv z$. Given the properties of the Fréchet distribution and the multiplicative nature of the utility function,

$$
\begin{aligned}
E\Big[U^i_{o,e}\Big] &= E\Big[\max_{(d,j)} \Lambda_{o,d,j,e}\, z^i_{j,e,\mathscr{L}(o,d)}\Big] \\
&= \int_0^{\infty} \Big[1 - \prod_{(d,j)} \Pr\Big(\Lambda_{o,d,j,e}\, z \leq t\Big)\Big] dt \\
&= \int_0^{\infty} \Big[1 - \prod_{(d,j)} \Pr\Big(z \leq \frac{t}{\Lambda_{o,d,j,e}}\Big)\Big] dt \\
&= \int_0^{\infty} \Big[1 - \prod_{(d,j)} \exp\Big(-A_{j,e,\mathscr{L}(o,d)} (\frac{t}{\Lambda_{o,d,j,e}})^{-\theta}\Big)\Big] dt \\
&= \int_0^{\infty} \Big[1 - \exp\Big(-\sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\frac{t}{\Lambda_{o,d,j,e}})^{-\theta}\Big)\Big] dt \\
&= \int_0^{\infty} \Big[1 - \exp\Big(-t^{-\theta} \sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta}\Big)\Big] dt \\
&= \int_0^{\infty} \Big[1 - \exp\Big(-\underbrace{\sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta}}_{\equiv (\Phi_{o,e})^{\theta}} t^{-\theta}\Big)\Big] dt
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^{\infty} \left[1 - \exp\left(-(\Phi_{o,e})^{\theta}\, t^{-\theta}\right)\right] dt \\
&= \left[(\Phi_{o,e})^{\theta}\right]^{1/\theta}\, \Gamma(1-1/\theta) \\
&= \Gamma(1-1/\theta)\, \Phi_{o,e}.
\end{aligned}
$$

The penultimate line uses the mean of a Fréchet maximum: for a random variable with c.d.f. $\exp(-c\, t^{-\theta})$, we have $\int_0^{\infty} \left[1-\exp(-c\, t^{-\theta})\right]\, dt = c^{1/\theta}\, \Gamma(1-1/\theta)$, here with $c = (\Phi_{o,e})^{\theta}$.

Wages: Wages are pinned down by the labor demand conditions, which are derived from the first-order conditions of the profit maximization problem of the firm in region d,

$$
\max_{\{\ell_{d,j,e,\mathscr{L}}\}} p_d y_d - \sum_{j\in J, e\in\{c,nc\}, \mathscr{L}\in\{b,nb\}} w_{d,j,e,\mathscr{L}} \ell_{d,j,e,\mathscr{L}},
$$

where the firm takes wages, $w_{d,j,e,\mathscr{L}}$, as given and chooses labor inputs, $\ell_{d,j,e,\mathscr{L}}$. From the first-order conditions of this problem,

$$
w_{d,j,e,\mathscr{L}} = MP_{d,j} \cdot MP_{d,j,e} \cdot (\ell_{d,j,e,\mathscr{L}})^{-\frac{1}{\nu}},
$$

where $MP_{d,j} = p_d (y_d)^{\frac{1}{\kappa}} (\phi_{d,j})^{\frac{1}{\kappa}} (\ell_{d,j})^{\frac{1}{\rho}-\frac{1}{\kappa}}$ is the marginal product of labor for occupation j at region d and $MP_{d,j,e} = (\ell_{d,j,e})^{\frac{1}{\nu}-\frac{1}{\rho}}$ is the marginal product of labor for workers with education e in occupation j at region d.

Regional Prices: Prices are pinned down by the marginal cost function, which is derived from the first-order conditions of the cost minimization problem of the firm in region d,

$$
\begin{aligned}
&\min_{\{\ell_{d,j,e,\mathscr{L}}\}} \sum_{j\in J, e\in\{c,nc\}, \mathscr{L}\in\{b,nb\}} w_{d,j,e,\mathscr{L}} \ell_{d,j,e,\mathscr{L}} \\
&\text{s.t. } y_d = \left[\sum_{j\in J} (\phi_{d,j})^{\frac{1}{\kappa}} (\ell_{d,j})^{\frac{\kappa-1}{\kappa}}\right]^{\frac{\kappa}{\kappa-1}}.
\end{aligned}
$$

First, the cost minimization problem of the inner nest is,

$$
\min_{\ell_{d,j,e,b}, \ell_{d,j,e,nb}} w_{d,j,e,b}\ell_{d,j,e,b} + w_{d,j,e,nb}\ell_{d,j,e,nb} \text{ s.t. } \ell_{d,j,e} = \left[(\ell_{d,j,e,b})^{\frac{\nu-1}{\nu}} + (\ell_{d,j,e,nb})^{\frac{\nu-1}{\nu}}\right]^{\frac{\nu}{\nu-1}},
$$

which allows us to solve for the cost of $\ell_{d,j,e}$,

$$w_{d,j,e} = \left[(w_{d,j,e,b})^{1-\nu} + (w_{d,j,e,nb})^{1-\nu} \right]^{\frac{1}{1-\nu}}.$$

Second, the cost minimization problem of the middle nest is,

$$\min_{\ell_{d,j,c},\ell_{d,j,nc}} w_{d,j,c}\ell_{d,j,c} + w_{d,j,nc}\ell_{d,j,nc} \text{ s.t. } \ell_{d,j} = \left[(\ell_{d,j,c})^{\frac{\rho-1}{\rho}} + (\ell_{d,j,nc})^{\frac{\rho-1}{\rho}} \right]^{\frac{\rho}{\rho-1}}$$

which allows us to solve for the cost of $\ell_{d,j}$,

$$w_{d,j} = \left[(w_{d,j,c})^{1-\rho} + (w_{d,j,nc})^{1-\rho} \right]^{\frac{1}{1-\rho}}.$$

Finally, the cost minimization problem of the outer nest is,

$$\min_{\{\ell_{d,j}\}} \sum_{j \in J} w_{d,j}\ell_{d,j} \text{ s.t. } y_d = \Big[\sum_{j \in J} (\phi_{d,j})^{\frac{1}{\kappa}} (\ell_{d,j})^{\frac{\kappa-1}{\kappa}} \Big]^{\frac{\kappa}{\kappa-1}},$$

which allows us to solve for the cost of y_d, the final cost function,

$$C(y_d) = y_d \Big[\sum_{j \in J} \phi_{d,j} (w_{d,j})^{1-\kappa} \Big]^{\frac{1}{1-\kappa}}.$$

Thus, price equals the marginal cost function,

$$p_d = \frac{\partial C(y_d)}{\partial y_d} = \Big[\sum_{j \in J} \phi_{d,j} (w_{d,j})^{1-\kappa} \Big]^{\frac{1}{1-\kappa}},$$

where

$$w_{d,j} = \left[(w_{d,j,c})^{1-\rho} + (w_{d,j,nc})^{1-\rho} \right]^{\frac{1}{1-\rho}},$$

and

$$w_{d,j,e} = \left[(w_{d,j,e,b})^{1-\nu} + (w_{d,j,e,nb})^{1-\nu} \right]^{\frac{1}{1-\nu}}, \text{ where } e \in \{c, nc\}.$$

Aggregate Price: The assumption that the aggregate good is traded costlessly across regions implies that each region faces the same price index. This follows from the profit maximization problem of the aggregate firm,

$$\max_{\{y_d\}} PY - \sum_{d \in R} p_d y_d.$$

The first-order condition yields the demand for regional goods,

$$y_d = (p_d/P)^{-\sigma} Y,$$

which can be substituted back into the production function to get that $P = \left[\sum_{d \in R} (p_d)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}$. This is a natural price index and represents the cost of producing one unit of the aggregate good Y given the prices of the regional goods p_d. Further, the expression for the price index is equivalent to the derivative of the cost function of the aggregate firm, which is also equal to P due to perfect competition. To see this, consider the aggregate firm's cost minimization problem,

$$\min_{\{y_d\}} \sum_{d \in R} p_d y_d$$

$$\text{s.t. } Y = \left[\sum_{d \in R} (y_d)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}},$$

which allows us to solve for the cost of Y,

$$C(Y) = Y \cdot \left[\sum_{d \in R} (p_d)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}.$$

Thus, the price equals the marginal cost function,

$$P = \left[\sum_{d \in R} (p_d)^{1-\sigma}\right]^{\frac{1}{1-\sigma}},$$

which is normalized to 1 by the assumption that the aggregate good is the numeraire.

C.2 Model Solution

The equilibrium is characterized by a system of nonlinear equations in wages, regional prices, and aggregate output that must be solved simultaneously. I employ an iterative fixed-point algorithm that iterates on wages until labor markets clear. The key insight is that, given wages, all other equilibrium objects can be computed sequentially: prices follow from the nested CES cost functions, output follows from income, and labor demand follows from Shephard's lemma. If labor markets clear at convergence, then by Walras' law, goods markets must also clear.

The algorithm proceeds as follows:

1. *Initialize wages.* Set wages $w_{d,j,e,\mathscr{L}}$ to an initial guess.

2. *Compute labor supply.* Given wages, compute sorting probabilities $\pi_{o,d,j,e}$ from equation 7, which depend on wages through $\Lambda_{o,d,j,e}$. Then compute average productivity for each origin-destination-occupation-education cell from equation 8, which exhibits positive selection: as

fewer workers choose a particular (d, j), their average productivity increases. Labor supply in efficiency units for each $(d, j, e, \mathscr{L})$ cell is obtained from the left-hand side of the labor market clearing condition (equation 12).

3. *Compute prices.* Compute regional prices p_d and the aggregate price P from the nested CES cost functions derived above. These depend only on wages and ensure zero profits under perfect competition. Normalize all wages and prices by P and set $P = 1$.

4. *Compute aggregate output.* Compute aggregate output as total labor income,

$$Y = \sum_{d,j,e,\mathscr{L}} w_{d,j,e,\mathscr{L}} \, \ell^{\text{supply}}_{d,j,e,\mathscr{L}},$$

   where $\ell^{\text{supply}}_{d,j,e,\mathscr{L}}$ is the labor supply in efficiency units from step 2. This step implicitly assumes that the aggregate goods market clears, so that total output equals total income.

5. *Compute regional goods demand.* Using the aggregate firm's first-order condition (equation 11), compute the demand for each regional good, $y_d = (p_d/P)^{-\sigma} Y$. This step implicitly assumes that regional goods markets clear, so that regional production equals regional demand.

6. *Compute labor demand using Shephard's lemma.* Given target output y_d from step 5 and factor prices (wages) from the current guess, compute the cost-minimizing labor allocation by applying Shephard's lemma to the nested CES cost function. This yields conditional factor demands that work top-down from y_d using only wages and prices, avoiding circular dependencies on labor supply aggregates. The conditional factor demands at each nest are:

$$\ell_{d,j} = y_d \cdot \phi_{d,j} \cdot \left(\frac{w_{d,j}}{p_d} \right)^{-\kappa},$$

$$\ell_{d,j,e} = \ell_{d,j} \cdot \left(\frac{w_{d,j,e}}{w_{d,j}} \right)^{-\rho},$$

$$\ell_{d,j,e,\mathscr{L}} = \ell_{d,j,e} \cdot \left(\frac{w_{d,j,e,\mathscr{L}}}{w_{d,j,e}} \right)^{-\nu}.$$

7. *Check labor market clearing.* Compare labor demand from step 6 with labor supply from step 2 for each $(d, j, e, \mathscr{L})$ cell. If excess demand is positive, increase wages; if excess supply, decrease wages. Apply damping for stability: $w^{\text{new}} = \alpha w^{\text{updated}} + (1 - \alpha) w^{\text{old}}$, where $\alpha \in (0, 1)$ is the damping parameter.

8. *Iterate.* Repeat steps 2–7 until wages converge within a specified tolerance.

9. *Verify all markets clear.* After convergence, verify that the labor market clearing conditions (equation 12) are satisfied. Then verify regional goods market clearing by checking that production from the CES production function using equilibrium labor inputs equals demand y_d from step 5, and verify aggregate goods market clearing (equation 14) by checking that total output equals total income.

The structure of the algorithm ensures that if labor markets clear, goods markets must also clear. This follows from three properties: first, prices satisfy the zero-profit condition by construction (step 3); second, labor demand is computed as the cost-minimizing allocation to produce y_d (step 6), so if labor markets clear, the resulting production equals y_d; third, by Walras' law, if labor markets and regional goods markets clear, the aggregate goods market must also clear.

D Estimation Appendix

This appendix summarizes the sources of identifying variation for each parameter, collects the full system of GMM moment conditions, derives the CES moment conditions used in the GMM estimation procedure (Section 5), derives the substitution formulas that express unobserved model objects in terms of data and parameters, details the shift-share IVs, describes the imputation of pre-sample speaking-occupation status, and reports model-fit diagnostics.

D.1 Summary of Parameters and Identifying Variation

Table 17 lists every parameter of the model, distinguishing those estimated jointly by GMM from those calibrated or taken directly from the data, and records the dimension of the data that identifies each. Reading down the table makes clear that the parameters are separately identified: each draws on a distinct combination of variation across origins, destinations, occupations, education groups, language-barrier legs, and time periods.

D.2 GMM Moment Conditions and Objective

The estimator stacks two sets of population moment conditions: a set of PPML conditions on the sorting equation that identify the Fréchet location terms, the migration-cost elasticities, and the destination amenities; and three shift-share IV conditions on the nested labor-demand system that identify the CES elasticities. This subsection states the full system. The per-nest residuals are derived in Appendices D.3–D.5, the substitutions that make every term computable from data and parameters in Appendix D.6, and the instruments in Appendix D.7.

Table 17: Summary of Parameters and Identification

Description	Parameter	Identifying Variation / Source
Panel A: Migration Costs		
Geographic Elasticity	η^G	Origins, Destinations, Time
Linguistic Elasticity	η^L_e	Origins, Destinations, Time
Panel B: Technological Productivity		
Avg. Productivity	$A_{j,e,\mathscr{L}(o,d)}$	Occupations, Education, Origins, Destinations within $\mathscr{L}(o,d)$, Time
Panel C: Other Parameters		
Occupation Sub. Elasticity	κ	Districts, Time
Education Sub. Elasticity	ρ	Occupations, Destinations, Time
Barrier Sub. Elasticity	ν	Occupations, Education, Districts, Time
Amenities	α_d	Districts, Time
Panel D: Calibrated Parameters		
Migration Elasticity	$\theta = 3.5$	Fan (2019)
Sub. Elasticity b/w Districts	$\sigma = 8$	Bryan and Morten (2019)
Panel E: Data Inputs		
Migration Shares	$\pi_{o,d,j,e,t}$	CPHS Waves 2014–19
Average Wages	$\bar{w}_{o,d,j,e,t}$	CPHS Waves 2014–19
Labor Endowments	$N_{o,e,t}$	CPHS Waves 2014–19
Occupation Shifters	$\phi_{d,j}$	CPHS Waves 2014–19
Language Migration Cost	$\tau^L_{o,d}$	Linguistic Distance
Geographic Migration Cost	$\tau^G_{o,d}$	Geodesic Distance

Notes: The table lists all parameters and data inputs of the model. **Panels A–C** report parameters estimated jointly by GMM; for these the final column gives the primary dimension of identifying variation. **Panel D** reports parameters calibrated from external sources, with the source citation in the final column. **Panel E** reports the data inputs, with the data source in the final column. Origins are states ($R_o = 27$), destinations are districts ($R_d = 522$), and *Time* denotes the seven four-month periods of the estimation window (September 2017–December 2019), aggregated from the monthly CPHS waves. The pre-sample waves (January 2014–August 2017) enter only through the shift-share instrument shares (Appendix D.7), which is why the data span runs from 2014.

Sorting (PPML) moment. Let $\pi^{\text{data}}_{o,d,j,e,t}$ be the empirical migration share and $\pi_{o,d,j,e,t}$ its model counterpart from the sorting equation (equation 7). The PPML moment conditions set the level residual, weighted by each element of the regressor vector $\mathbf{x}_{o,d,j,e}$, to zero,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{o,d,j,e}\left(\pi^{\text{data}}_{o,d,j,e,t} - \pi_{o,d,j,e,t}\right)\cdot \mathbf{x}_{o,d,j,e} = 0, \tag{23}$$

where $\mathbf{x}_{o,d,j,e} \equiv \left[\mathbb{1}_{j,e,\mathscr{L}(o,d)},\ \ln(1-\tau^G_{o,d}),\ \ln(1-\tau^L_{o,d})\cdot\mathbb{1}_e,\ \mathbb{1}_d\right]$ collects the $(j,e,\mathscr{L})$ indicators (identifying $A_{j,e,\mathscr{L}(o,d)}$), log geographic accessibility (identifying η^G), education-specific log linguistic accessibility (identifying η^L_e), and destination indicators (identifying α_d). I use PPML following Santos Silva and Tenreyro (2006), which is consistent under heteroskedasticity and handles the share structure of π naturally.

CES (IV) moments. Taking ratios of the labor-demand first-order conditions across the two legs of each nest, then logging and time-differencing, yields a structural residual at each nest; interacting it with the corresponding shift-share instrument delivers the orthogonality condition. For the outer, middle, and inner nests, respectively,

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R_d}\right)\sum_{t=2}^{T}\sum_{d=1}^{R_d}\left(\Delta_t \ln\left(\frac{w_{d,j,t}}{w_{d,j',t}}\right)+\frac{1}{\kappa}\Delta_t \ln\left(\frac{\ell_{d,j,t}}{\ell_{d,j',t}}\right)\right)\cdot Z^{\kappa}_{d,t}=0, \tag{24}$$

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R_d}\right)\sum_{t=2}^{T}\sum_{d=1}^{R_d}\sum_{j}\left(\Delta_t \ln\left(\frac{w_{d,j,c,t}}{w_{d,j,nc,t}}\right)+\frac{1}{\rho}\Delta_t \ln\left(\frac{\ell_{d,j,c,t}}{\ell_{d,j,nc,t}}\right)\right)\cdot Z^{\rho}_{d,j,t}=0, \tag{25}$$

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R_d}\right)\sum_{t=2}^{T}\sum_{d=1}^{R_d}\sum_{j}\left(\Delta_t \ln\left(\frac{w_{b,d,j,t}}{w_{nb,d,j,t}}\right)+\frac{1}{\nu}\Delta_t \ln\left(\frac{\ell_{b,d,j,t}}{\ell_{nb,d,j,t}}\right)\right)\cdot Z^{\nu}_{d,j,t}=0, \tag{26}$$

where the unobserved efficiency wages $w_{d,j,e,\mathscr{L},t}$ and labor inputs $\ell_{d,j,e,\mathscr{L},t}$ are replaced by their data-and-parameter expressions from Appendix D.6, the instruments $Z^{\kappa}_{d,t}$, $Z^{\rho}_{d,j,t}$, and $Z^{\nu}_{d,j,t}$ are defined in Appendix D.7, and the single inner-nest condition pools the two education leaves of each barrier leg and sums over occupations (Appendix D.5).

Estimation and objective. These conditions cannot be estimated separately, because the moments depend on unobserved, structurally consistent efficiency wages. I therefore stack them and minimize a joint GMM criterion. A nested algorithm alternates between an inner loop that solves the model equilibrium at the current parameter guess to recover the implied efficiency wages and price indices, and an outer loop that evaluates the stacked objective: the PPML objective for the sorting block in equation 23, and a just-identified 2SLS projection for the three CES blocks in equations 24–26. Standard errors follow the multiplier (Bayesian) cluster bootstrap over the 522 destination districts described in Section 5, with the full joint model re-estimated in each replication.

D.3 CES Moment Condition: Outer Nest (κ)

The outer nest CES production function is

$$y_d = \Big[\sum_{j\in J} (\phi_{d,j})^{\frac{1}{\kappa}} (\ell_{d,j})^{\frac{\kappa-1}{\kappa}}\Big]^{\frac{\kappa}{\kappa-1}}.$$

From the first-order condition of the regional firm's profit maximization, the wage index for occupation j at destination d satisfies

$$w_{d,j} = \lambda_d \cdot y_d^{1/\kappa} \cdot \phi_{d,j}^{1/\kappa} \cdot \ell_{d,j}^{-1/\kappa},$$

where $\lambda_d = p_d$ under perfect competition. Taking the ratio for occupations j and j' at the same destination d eliminates p_d and y_d:

$$\frac{w_{d,j}}{w_{d,j'}} = \Big(\frac{\phi_{d,j}}{\phi_{d,j'}}\Big)^{1/\kappa} \Big(\frac{\ell_{d,j}}{\ell_{d,j'}}\Big)^{-1/\kappa}.$$

Taking logs and then time differences eliminates the share parameters $\phi_{d,j}$ (which are time-invariant):

$$\Delta_t \ln\Big(\frac{w_{d,j}}{w_{d,j'}}\Big) = -\frac{1}{\kappa} \cdot \Delta_t \ln\Big(\frac{\ell_{d,j}}{\ell_{d,j'}}\Big).$$

Rearranging yields the moment condition:

$$\Delta_t \ln\Big(\frac{w_{d,j}}{w_{d,j'}}\Big) + \frac{1}{\kappa} \cdot \Delta_t \ln\Big(\frac{\ell_{d,j}}{\ell_{d,j'}}\Big) = 0.$$

D.4 CES Moment Condition: Middle Nest (ρ)

The middle nest CES is

$$\ell_{d,j} = \Big[(\ell_{d,j,c})^{\frac{\rho-1}{\rho}} + (\ell_{d,j,nc})^{\frac{\rho-1}{\rho}}\Big]^{\frac{\rho}{\rho-1}}.$$

This nest has equal unit shares. Taking the ratio for college and non-college workers within the same (d, j) and following the same steps as above yields:

$$\frac{w_{d,j,c}}{w_{d,j,nc}} = \Big(\frac{\ell_{d,j,c}}{\ell_{d,j,nc}}\Big)^{-1/\rho}.$$

Taking logs and time differences yields the moment condition:

$$\Delta_t \ln\Big(\frac{w_{d,j,c}}{w_{d,j,nc}}\Big) + \frac{1}{\rho} \cdot \Delta_t \ln\Big(\frac{\ell_{d,j,c}}{\ell_{d,j,nc}}\Big) = 0.$$

D.5 CES Moment Condition: Inner Nest (ν)

The inner nest CES combines barrier and no-barrier workers within each (d, j, e) cell using a single elasticity ν, common across occupations and education,

$$\ell_{d,j,e} = \left[(\ell_{d,j,e,b})^{\frac{\nu-1}{\nu}} + (\ell_{d,j,e,nb})^{\frac{\nu-1}{\nu}} \right]^{\frac{\nu}{\nu-1}}.$$

Because ν is a single scalar, the barrier vs. no-barrier variation from every (j,e) cell identifies one moment, so I pool the two education leaves of each barrier leg $\mathscr{L} \in \{b, nb\}$ into an education-pooled wage and labor input at (d,j):

$$\ell_{\mathscr{L},d,j} = \sum_e \ell_{d,j,e,\mathscr{L}}, \qquad w_{\mathscr{L},d,j} = \frac{\sum_e w_{d,j,e,\mathscr{L}} \ell_{d,j,e,\mathscr{L}}}{\sum_e \ell_{d,j,e,\mathscr{L}}}.$$

The labor input sums efficiency units across education, and the wage is the wage-bill-preserving (efficiency-unit-weighted) mean of the leaf wages, so that $w_{\mathscr{L},d,j}\,\ell_{\mathscr{L},d,j} = \sum_e w_{d,j,e,\mathscr{L}}\,\ell_{d,j,e,\mathscr{L}}$. This aggregation keeps the inner moment independent of the middle-nest elasticity ρ. Taking the ratio of the pooled barrier to no-barrier wages, then logging and time-differencing, yields the residual

$$\Delta_t \ln\left(\frac{w_{b,d,j}}{w_{nb,d,j}}\right) + \frac{1}{\nu} \cdot \Delta_t \ln\left(\frac{\ell_{b,d,j}}{\ell_{nb,d,j}}\right) = 0.$$

Because ν is common across occupations and education, the orthogonality contributions are summed over both occupations j and all periods into a single moment for ν,

$$\sum_{t\geq 2}\sum_{d,j}\left[\Delta_t \ln\left(\frac{w_{b,d,j}}{w_{nb,d,j}}\right) + \frac{1}{\nu} \cdot \Delta_t \ln\left(\frac{\ell_{b,d,j}}{\ell_{nb,d,j}}\right)\right] \cdot Z^{\nu}_{d,j,t} = 0,$$

where the inner-nest instrument $Z^{\nu}_{d,j,t}$ is defined in Appendix D.7.

D.6 Substitution Formulas

The CES moment conditions require $w_{d,j,e,\mathscr{L}}$ (the wage per efficiency unit) and $\ell_{d,j,e,\mathscr{L}}$ (labor in efficiency units), which are not directly observed. These are expressed in terms of observables $(\bar{w}_{o,d,j,e}, \pi_{o,d,j,e}, N_{o,e})$ and parameters $(A_{j,e,\mathscr{L}(o,d)}, \theta)$ as follows.

Wage per efficiency unit. From the selection equation (equation 9),

$$\bar{w}_{o,d,j,e} = \tilde{\Gamma}\left(\frac{\pi_{o,d,j,e}}{A_{j,e,\mathscr{L}(o,d)}}\right)^{-1/\theta} \cdot w_{d,j,e,\mathscr{L}}.$$

Solving for $w_{d,j,e,\mathscr{L}}$:

$$w_{d,j,e,\mathscr{L}} = \frac{\bar{w}_{o,d,j,e}}{\tilde{\Gamma}\left(\pi_{o,d,j,e}/A_{j,e,\mathscr{L}(o,d)}\right)^{-1/\theta}}, \tag{27}$$

for any origin o such that $\mathscr{L}(o,d) = \mathscr{L}$. Since $w_{d,j,e,\mathscr{L}}$ is a destination-specific object, this expression yields the same value for any qualifying origin o. In practice with noisy data, one may

average across origins.

Labor in efficiency units. From the labor market clearing condition (equation 12),

$$\ell_{d,j,e,\mathscr{L}} = \sum_{o:\mathscr{L}(o,d)=\mathscr{L}} \tilde{\Gamma}\Big(\frac{\pi_{o,d,j,e}}{A_{j,e,\mathscr{L}(o,d)}}\Big)^{-1/\theta} \cdot \pi_{o,d,j,e} \cdot N_{o,e}, \tag{28}$$

which sums over all origins o with the same language barrier status $\mathscr{L}$ at destination d.

Building the CES wage indices. Given $w_{d,j,e,\mathscr{L}}$ from equation 27, the CES wage indices are computed bottom-up:

$$w_{d,j,e} = \Big[w_{d,j,e,b}^{1-\nu} + w_{d,j,e,nb}^{1-\nu}\Big]^{1/(1-\nu)},$$
$$w_{d,j} = \Big[w_{d,j,c}^{1-\rho} + w_{d,j,nc}^{1-\rho}\Big]^{1/(1-\rho)}.$$

Similarly, the CES labor aggregates are computed bottom-up from $\ell_{d,j,e,\mathscr{L}}$ using the CES production functions.

D.7 Shift-Share Instrumental Variables

The CES moment conditions involve potentially endogenous relative labor quantities. I instrument them with time-differenced shift-share (Bartik) instruments following Altonji and Card (1991) and Card (2001), built from a predetermined pre-sample origin composition interacted with leave-out origin-supply growth.

Notation. Let $N_{o,d,j,e,t} = \pi_{o,d,j,e,t} N_{o,e,t}$ be the worker count in cell (o,d,j,e) at period t, and $N_{o,d,j,t} = \sum_e N_{o,d,j,e,t}$ its education-pooled counterpart. The pre-sample window is the first T_{pre} waves (January 2014–August 2017), which precede the estimation window (September 2017–December 2019); shares are formed over the pre-sample window and shifts over the estimation window.

Shares (predetermined). For each nest, the share is the pre-sample time-average of an origin's composition of the relevant destination cell,

$$\text{outer: } s^{\text{pre}}_{o,d,j} = \frac{1}{T_{\text{pre}}} \sum_{\tau \le T_{\text{pre}}} \frac{N_{o,d,j,\tau}}{\sum_{o'} N_{o',d,j,\tau}}, \qquad \text{middle: } s^{\text{pre}}_{o,d,j,e} = \frac{1}{T_{\text{pre}}} \sum_{\tau \le T_{\text{pre}}} \frac{N_{o,d,j,e,\tau}}{\sum_{o'} N_{o',d,j,e,\tau}},$$
$$\text{inner: } s^{\text{pre}}_{o,d,j,\mathscr{L}} = \frac{1}{T_{\text{pre}}} \sum_{\tau \le T_{\text{pre}}} \frac{N_{o,d,j,\tau}\,\mathbb{1}\{\mathscr{L}(o,d)=\mathscr{L}\}}{\sum_{o'} N_{o',d,j,\tau}\,\mathbb{1}\{\mathscr{L}(o',d)=\mathscr{L}\}}, \qquad \mathscr{L} \in \{b, nb\}.$$

Because $\mathscr{L}(o,d)$ does not depend on education, the barrier and no-barrier inner shares have dis-

joint origin support and each sums to one over its own support.

Shift (leave-out, time-differenced). The shift is the time-differenced log of leave-out origin labor supply, where the leave-out drops the own destination d only (not the occupation). For the outer and inner nests it is education-pooled,

$$g_{o,d,j,t} = \Delta_t \ln N_{o,-d,j,t}, \qquad N_{o,-d,j,t} = \sum_{d' \neq d} N_{o,d',j,t},$$

and for the middle nest it is education-specific, $g_{o,d,j,e,t} = \Delta_t \ln N_{o,-d,j,e,t}$ with $N_{o,-d,j,e,t} = \sum_{d' \neq d} N_{o,d',j,e,t}$. Time-differencing aligns the instrument with the time-differenced CES moments and removes the time-invariant origin-occupation scale.

Instruments. Each instrument sums origin-level share $\times$ shift products, contrasting the two legs of its nest:

$$\begin{aligned}
Z^{\kappa}_{d,t} &= \sum_o \left[s^{\text{pre}}_{o,d,1}\, g_{o,d,1,t} - s^{\text{pre}}_{o,d,2}\, g_{o,d,2,t} \right], \\
Z^{\rho}_{d,j,t} &= \sum_o \left[s^{\text{pre}}_{o,d,j,c}\, g_{o,d,j,c,t} - s^{\text{pre}}_{o,d,j,nc}\, g_{o,d,j,nc,t} \right], \\
Z^{\nu}_{d,j,t} &= \sum_o \left(s^{\text{pre}}_{o,d,j,b} - s^{\text{pre}}_{o,d,j,nb} \right) g_{o,d,j,t}.
\end{aligned}$$

The outer instrument contrasts the speaking ($j = 1$) and non-speaking ($j = 2$) occupation legs; the middle instrument contrasts college and non-college within occupation j; and the inner instrument weights the common occupation-j shift $g_{o,d,j,t}$ by the difference between the barrier and no-barrier origin shares. The outer and inner nests share the same shift $g_{o,d,j,t}$. The single ν moment sums the orthogonality contributions $Z^{\nu}_{d,j,t}$ over both occupations.

Identification assumptions. The instruments are valid under: (1) *Relevance*: the predetermined origin composition interacted with origin-supply growth predicts the time-differenced relative quantities in each nest, which is testable via first-stage regressions. (2) *Exclusion*: the pre-sample shares are predetermined relative to the estimation window, and the leave-out of destination d ensures the supply shift reflects origin-driven growth at *other* destinations, orthogonal to demand shocks at d.

D.8 Imputation of Speaking-Occupation Status for the Pre-Sample Period

The CPHS occupation module—which records a detailed occupation code that can be matched to O*NET—was introduced in September 2017 (Wave 12 of the survey). Waves 1–11 (January 2014–August 2017) contain demographic and employment-status variables but not occupation codes. Since the shift-share instruments require predetermined allocation shares, the period before

September 2017 is the natural base. I therefore impute the binary speaking-occupation indicator, speaking_occ$_{it}$, for the relevant pre-sample worker observations. This appendix describes the imputation method and its validation.

Worker definition. The CPHS member-income file records all resident household members, including non-workers. The speaking-occupation indicator is only meaningful for employed workers. I define a worker as any household member whose `nature_of_occupation` is not one of the following non-employment categories: *Home Maker*, *Retired/Aged*, *Student*, or *Unoccupied*. The validity of this definition is confirmed using the post-sample period (September 2017–December 2019), where the O*NET match outcome provides a ground truth. The four excluded categories have O*NET match rates of 0.06%, 0.04%, 0.07%, and 0.02%, respectively, while all other employment categories have match rates of 92–97%. The worker filter captures 99.9% of true workers (sensitivity) with 94.6% precision. Approximately 42% of person-wave observations in the pre-sample period are classified as workers.

Imputation strategy. For each pre-sample worker observation, speaking_occ$_{it}$ is imputed using the following priority rule.

Step 1: Panel carry-over. The CPHS is a rotating household panel, and individual members are tracked across waves by a unique household-member identifier. For each pre-sample worker observation (i,t), I search for the same individual i in the post-sample period (September 2017 onward). If individual i has at least one post-sample observation with a matched speaking-occupation value, I assign the speaking_occ value from the earliest such post-sample observation to all pre-sample worker waves for that individual. This carry-over is an exact assignment of the individual's actual speaking-occupation classification rather than a prediction. It is applied only when the individual was also classified as a worker in the pre-sample wave; a pre-sample homemaker or student wave of a person who subsequently became an employed worker does not inherit the later classification.

Of the 19.2 million pre-sample worker observations, 7.6 million belong to individuals who also appear in the post-sample as workers (39.6%). These are assigned the carry-over value. The remaining pre-sample worker observations (60.4%) are imputed by logit.

Step 2: Logit imputation. For pre-sample worker observations that cannot be matched to a post-sample counterpart, I estimate a logit model on post-sample workers with observed speaking_occ and predict for pre-sample workers. The model is:

$$\begin{aligned}\Pr(\text{speaking_occ}_{it}=1) = \Lambda\big(&\alpha + \delta^{\text{noc}}\mathbf{1}[\text{NOC}_{it}] + \delta^{\text{ind}}\mathbf{1}[\text{industry}_{it}] + \delta^{\text{edu}}\mathbf{1}[\text{education}_{it}] \\ &+ \beta_1 \text{male}_{it} + \beta_2 \text{urban}_{it}\big),\end{aligned} \tag{29}$$

where $\Lambda(\cdot)$ is the logistic function, $\mathbf{1}[\text{NOC}_{it}]$ are dummies for the 18 worker occupation-status categories (nature of occupation), $\mathbf{1}[\text{industry}_{it}]$ are dummies for industry of occupation, $\mathbf{1}[\text{education}_{it}]$

are dummies for education level, and male_{it}, urban_{it} are binary indicators. The training sample consists of 4,854,625 post-sample worker observations with matched O*NET speaking classifications. The model achieves a pseudo-R^2 of 0.69 and correctly classifies 91.7% of training observations. Non-worker observations in the pre-sample are excluded from the construction of shift-share allocation shares.

Validation. Three checks assess the quality of the imputation.

Check 1: Holdout accuracy. I re-estimate model (29) on 2018–2019 data only and predict speaking_occ for the September–December 2017 observations, where the true value is observed. Out-of-sample accuracy is 87.9% (sensitivity 85.4%, specificity 90.5%). The small gap relative to in-sample accuracy (91.7%) reflects modest overfitting to occupation–industry combinations that are rare in 2017 but more common by 2018–2019.

Check 2: Geographic stability. The Bartik identification strategy requires that the pre-sample worker allocation shares be orthogonal to subsequent demand shocks. A useful diagnostic is whether the cross-state variation in imputed shares resembles the cross-state variation in observed post-sample shares. The Pearson correlation between the imputed state-level speaking share (averaged over 2014) and the observed speaking share among post-sample O*NET-matched workers (averaged over 2018–2019) is 0.35. This moderate correlation partly reflects genuine demographic change over the five-year gap and partly reflects that the post-sample observed share is computed over formal-sector workers matched to O*NET—a subsample that differs in composition from the full pre-sample worker population. Bartik identification requires cross-state variation, not absolute level stability, so this comparison is informative but not a binding requirement.

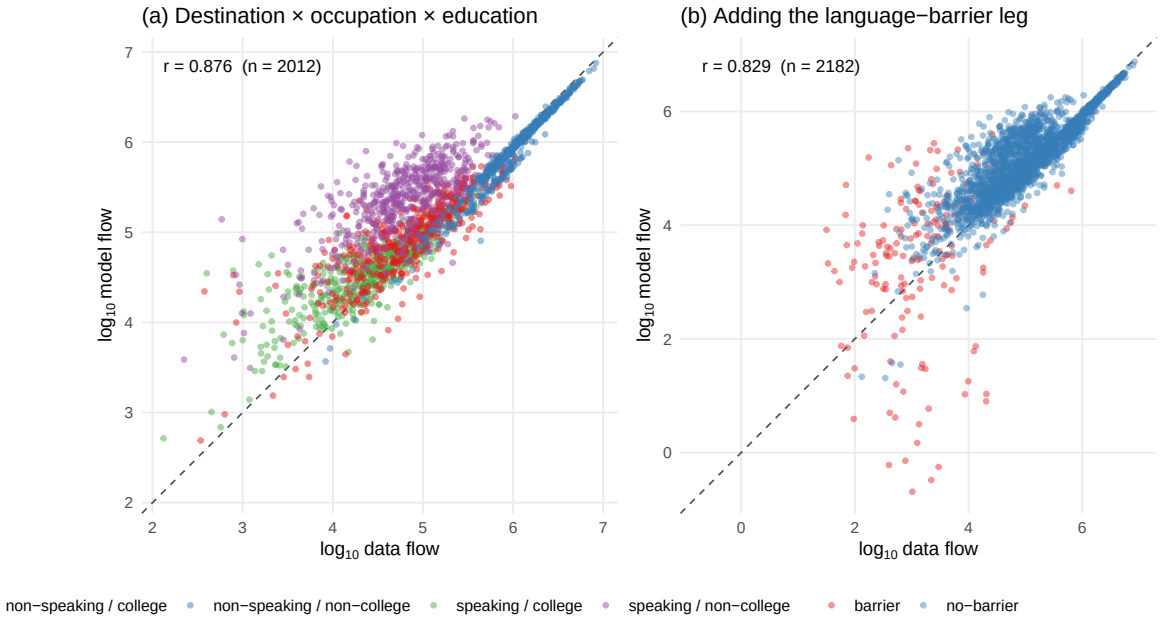
Check 3: Occupation consistency within individuals. Among the 279,310 workers who appear in two or more post-sample waves with a matched O*NET observation, 63.8% have the same speaking_occ value in every observed wave. Workers who switch primarily do so once between adjacent waves, consistent with measurement noise rather than genuine career changes. This supports the carry-over assumption: a worker's speaking-occupation status observed in 2018 is a reliable signal of the same individual's status in 2014–2017.

Output characteristics. Among pre-sample workers, the imputed speaking share is approximately 54% in 2014, declining to 50–51% in 2015–2017, consistent with the post-sample observed rate of 51%. The slightly higher 2014 share reflects positive selection in the carry-over group: individuals who remain continuously employed through 2017 have modestly higher rates of speaking-intensive work than the full pre-sample workforce.

D.9 Model Fit

Figure 11 compares model and data worker flows on the sorting margin targeted by the PPML moment conditions of Appendix D.2. Panel (a) aggregates the bilateral (o,d,j,e) flows to the destination-by-occupation-by-education level; panel (b) splits destinations by the language-barrier leg $\mathscr{L}(o,d)$, the margin on which the counterfactuals operate. In both panels the log flows track the forty-five-degree line, with Pearson correlations of 0.876 and 0.829 respectively. The model reproduces the no-barrier mass tightly; the thinner barrier leg is noisier, as expected given the smaller counts, but unbiased around the diagonal.

Figure 11: Model vs. Data: Sorting Flows (Targeted)



Notes: Each point compares $\log_{10}$ model and data worker flows. Panel (a) aggregates bilateral (o,d,j,e) flows to the destination $\times$ occupation $\times$ education level ($n = 2{,}012$); panel (b) splits destinations by the barrier leg $\mathscr{L}(o,d)$ ($n = 2{,}182$). The dashed line is the forty-five-degree line, and r is the Pearson correlation of the plotted logs. These are the flows targeted by the sorting (PPML) moment conditions in Appendix D.2. The full bilateral origin-to-destination matrix, which the moments do not directly target, fits less tightly (correlation about 0.59).