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

Amrutha Manjunath[†]

November 21, 2024

Updated regularly; please click here for the latest version.

Abstract

This paper studies how language barriers impact internal migration, skill premium, and aggregate welfare using rich microdata from India applied to a quantitative spatial general equilibrium framework. I first document four empirical facts: (1) workers migrate less often to locations where they face high language barriers; (2) migrants with high language barriers are employed less often in speaking-intensive occupations; (3) migrants with high language barriers get a wage premium; and (4) these patterns are strongest for unskilled workers. To explain 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, augment skill premium, and reduce aggregate welfare. As economies shift towards services, language barriers increasingly impede aggregate gains due to the rising prevalence of speaking-intensive occupations. In the absence of language barriers—relative to observed changes—structural change would have increased aggregate welfare by 1.9 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, Sorting, Selection, Labor Markets, Internal Migration

JEL Codes: J24, J31, J61, R23, O15

[*] I am indebted to Fernando Parro, Kala Krishna, Jim Tybout, Jingting Fan, and Enghin Atalay for their advice and encouragement at every stage of research. I thank Salomé Baslandze, Jonathan Eaton, Simon Fuchs, Seula Kim, Veronika Penciakova, Joris Pinkse, Bradley Setzler, Stephen Yeaple, Kai-Jie Wu, and conference, seminar, and brownbag participants at Penn State, the Federal Reserve Bank of Atlanta, the Federal Reserve Bank of Philadelphia, and the 13th European Meeting of the Urban Economics Association in Copenhagen for their generous comments and suggestions. I thank the Data Dissemination Unit at the Office of Registrar General of India for helping me access confidential data and Hailey Gretz for excellent research assistance.

[†] Department of Economics, Pennsylvania State University, State College, PA 16801. Email: amrutha.manjunath@psu.edu.

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, 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 skill premium. This makes language a critical factor in understanding and mitigating economic disparities within multilingual countries.

This paper is the first to study the aggregate and distributional general equilibrium effects of language as both a spatial and a labor market friction. Specifically, it asks the following questions: How do language barriers impact internal migration, skill premium, and aggregate welfare?[3] How do language barriers impact gains from structural change in services, as characterized by a rise in speaking occupations? How could language policy mitigate these effects?

To answer these, 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 four empirical facts that relate language barriers to migration patterns and labor market outcomes of workers by skill. Second, to explain these facts, I develop and structurally estimate a quantitative spatial general equilibrium model of migration that embeds labor market effects of language in workers' migration choices. That is, workers sort across locations and occupations by skill and language, with wages accounting for worker selection and adjusting in general equilibrium. Third, I use the estimated model to perform counterfactual exercises that quantify the extent to which language barriers impact aggregate outcomes. I show that this impact becomes particularly significant as economies shift towards services and speaking-intensive occupations become more prevalent across space. I then calibrate costs of language program provision and learning and show that welfare benefits of implementing language programs would outweigh costs.

I begin by leveraging novel, highly disaggregated data on district-to-district migration from the Census of India (2001, 2011). This dataset consists of the universe of internal migrants and provides information on their origins, destinations, and key demographic characteristics. I complement this with other census data on languages spoken across districts and additional household and administrative datasets that provide information on workers' occupations and wages, among other demographic and labor market characteristics.

India's linguistic diversity, with over 130 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 distances 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 utility 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—widespread among college-educated workers but limited among those without college education. 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 four 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. The next three facts show how language barriers shape migrants' occupations and wages, and 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. I interpret this as language barriers diminishing workers' productivity in occupations that require communication. In the third fact, I show that among migrant workers in comparable occupations, those with high language barriers get a wage premium. I interpret this as language barriers increasing migration costs, such that only migrants that expect to get paid higher wages choose to move despite these barriers. In the fourth fact, I show that these patterns are less pronounced for skilled (college) than unskilled (non-college) workers. I understand this to reflect how college education in India provides proficiency in English, which may reduce 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 modeled both as a component of migration cost and as 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, through a nested CES structure of labor aggregation, I allow for complementarities by skill and occupation type in production. 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). I incorporate labor demand considerations by allowing firms to differentiate between workers based on skill and language (empirical fact 4). 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 removing the sorting effect of language barriers dominates the effects on wages.

I take the model to microdata from India and estimate novel elasticities of substitution between occupation and worker types. I find high degrees of complementarities between unskilled workers with and without language barriers in speaking occupations, but not so in non-speaking occupations or among skilled workers. I also estimate the average productivity of worker types in speaking occupations. I find that unskilled workers that face language barriers are nearly twice as productive in non-speaking than in speaking occupations, but this is not so for unskilled workers without language barriers or among skilled workers. Thus, the estimates corroborate my hypothesis that language barriers affect unskilled rather than skilled workers, and in speaking rather than non-speaking occupations.

Using the estimated model, I conduct three counterfactual exercises. In the first counterfactual, I quantify the impact of language barriers on internal migration, skill premium, and welfare. To do so, I remove language both as a driver of productivity and component of migration cost. This increases internal migration by 6.2 percentage points, decreases inequality by 1.9 percentage points, and increases welfare by 1.2 percent relative to the estimated model. To contextualize these magnitudes, I target the same aggregate welfare gains and find that removing language barriers is equivalent to (1) proportionally decreasing geographic barriers between every pair of locations by 56 percent or (2) increasing the share of college workers in each location by 34 percent.

In the second counterfactual, I consider the increasing importance of language barriers due to structural change, as characterized by the prevalence of speaking occupations. I quantify the impact of language barriers on internal migration, skill premium, and welfare when speaking occupations became more prevalent across space. Between 1987 (before India's trade liberalization) and 2011 (the timeline of my analysis), there was rapid growth in particularly consumer services across space (Fan et al., 2023). In other words, there was an increase in the prevalence of speaking occupations across space. This heightened the importance of language skills in the labor market.

To quantify the impact of language on aggregate outcomes when this transformation occurred, I use the estimated model to simulate internal migration, skill premium, and welfare. I compare the effects of the occupational shift over time in scenarios with and without language barriers. I find that relative to observed changes, and in the absence of language barriers, the increased prevalence of speaking occupations would have caused internal migration to be higher by 7.2 percentage points, skill premium to be lower by 3.4 percentage points, and welfare to be higher by 1.9 percent.

In the third counterfactual, 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 charging a uniform tax to all workers in the economy.

To weigh the benefits and costs of this language policy, I compare welfare under the policy to that under the benchmark estimated model. Welfare under the policy is the aggregate real income in the economy with total program cost subtracted, which has two components I calibrate separately. First, I measure the opportunity cost of learning using information on time taken to overcome language barriers and wages foregone, 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. The calibrated model shows that language programs generate welfare gains when costs stay below a certain threshold. I show that the cost-per-student calculated from CIIL reports falls well below this threshold. Thus, I argue that welfare benefits of language programs outweigh costs and language programs should indeed be implemented.

Broadly, this paper leverages India's linguistic diversity and the rise of speaking occupations to investigate the economic impact of language barriers on internal migration, skill premium, and welfare. It emphasizes the mechanisms through which language barriers shape labor market outcomes and drive aggregate outcomes in diverse, developing economies. The findings in this paper have important implications for understanding sources of inequality and designing policies to promote more inclusive economic growth in multilingual countries.

Related Literature: This paper is related to several strands of literature. First, it contributes to a broad line of work in economic geography and migration that studies aggregate implications of spatial frictions. Similar to Tombe and Zhu (2019), Fan (2019), and Allen et al. (2018), the model features multiple regions and sectors (occupations) with costly labor mobility in a static framework. Previous research, including dynamic migration models in Kennan and Walker (2011), Monras (2018), and Caliendo et al. (2019), have emphasized geographic barriers as an impediment to migration but ignored language. In contrast, this paper microfounds and studies the aggregate implications of language, which in part acts as a component of migration cost.

In quantifying aggregate gains from workers following comparative advantage in general equilibrium, this paper is related to Lagakos and Waugh (2013), Bryan and Morten (2019),

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

Hsieh et al. (2019), Burstein et al. (2019), Burstein et al. (2020), and Bratsberg et al. (2023). Relative to these, I study the impact of language barriers through how they shape internal migration. In this sense, this paper is closest to Wang (2024), who finds that migration increases with ethnolinguistic diversity due to the amenity value of cultural diversity outweighing communication barriers in Indonesia. In contrast, I model language as both a component of migration cost and a technological friction, with explicit and flexible complementarities between skill and language in occupations of varied speaking-intensity. This framework reveals labor market and distributional consequences of language in general equilibrium and finds an overall negative relationship with internal migration in India. Thus, this paper is the first in any context to study language as both a spatial and labor market friction in a quantitative framework.

Second, this paper is related to empirical work that study the role of language in trade, migration, and the labor market. Melitz (2008) introduced linguistic distance to empirical gravity models of international trade. Adserà and Pytliková (2015) document the empirical significance of language in shaping international migration patterns. Dustmann and Fabbri (2003) and Chiswick and Miller (2015) study the importance of language proficiency of migrants to the UK in shaping their labor market outcomes. Peri and Sparber (2009) explain the occupation choices of comparably skilled migrants away from communication-language tasks as the imperfect substitutability between migrants and natives in partial equilibrium. In the context of internal migration, Kone et al. (2018) and Imbert and Papp (2020) provide evidence on how Indian migration flows empirically relate to both distance and linguistic differences. This paper borrows insights from this literature to incorporate and study language barriers in general equilibrium.

Third, this paper contributes to the growing literature in macroeconomic development that studies the distributional effects of spatial and labor market forces. It is related to Fan et al. (2023), who study the unequal effects of service-led growth in India. Rather than factors that contribute to structural transformation, my focus is on understanding how language barriers shape distributional outcomes against the backdrop of growth in services. It is also related to Ghose (2024), who studies the role of migration cost for education and work in shaping distributional effects of globalization in India. In contrast, I study the role of language barriers in driving heterogeneous outcomes for workers of different skills and origins.

Road Map: The rest of this paper is structured as follows: Section 2 describes various sources of data on migration, language, and labor market outcomes; Section 3 lays out three empirical facts that relate language, locations, and the labor market; Section 4 develops a static spatial quantitative general equilibrium model of migration that features language barriers; Section 5 takes the model to the data; Section 6 uses the estimated model to perform counterfactual exercises; 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 confidential internal migration data from the Census of India and combine this with data on language and labor market outcomes from various sources. To obtain an occupational ranking of speaking-intensity, I create a concordance between occupations in the Indian household surveys and O*NET.

2.1 Data on Migration

The primary source of data on internal migration is the Census of India (2001, 2011). The publicly available data contains information, at each district, on migrants' education level, age, reasons for migration, the rural/urban nature of their previous residence, and duration of stay in the current residence since migration.

In order to analyze the effects of language barriers on migration decisions, it is crucial to identify the district of origin jointly with these variables. This information is available in the more disaggregated district-to-district migration data. The 2011 data records migration flows jointly by education level, age, reasons for migration, the rural/urban nature of their previous residence, and duration of stay. This allows for a rich analysis through cross-tabulations that are not possible in the 2001 data. I combine the 2001 and 2011 disaggregated data into a novel district-level panel dataset.[5]

2.2 Data on Language

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 130 languages (each spoken by over 10,000 people each) at home i.e., as a first language, in every district. Using this data and linguistic trees from Eberhard et al. (2024), I follow Spolaore and Wacziarg (2009) and construct a measure of language barriers for any region-pair o, d,

$$\text{Linguistic Distance}_{o,d} = \sum_{m}\sum_{n}(s_{o,m} \times s_{d,n} \times \text{dist. b/w languages}_{m,n}) \tag{1}$$

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 able 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.

[5] Research 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), Ghose (2024).

2.3 Data on Labor Market Outcomes

I use monthly data from January 2014 to December 2019 from the Consumer Pyramids Household Surveys (CPHS) to obtain information on individuals' state of origin, occupation choices, 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 match occupations to O*NET based on their descriptions to obtain a score of the importance of speaking in each occupation. I define an occupation as speaking-intensive if their rank is above the median.

Table 1: Occupations Categorized by O*NET Speaking Importance Levels

	Occupation	O*NET Speaking Importance Level
Speaking	Lawyers	91
	Teachers (School, University)	78
	Nurses	78
	Human Resources Managers	78
	Call Center Operators	78
	Sales Representatives	75
	Customer Service Representatives	72
	Receptionists	72
	Child Care Maids	69
	Journalists	63
Non-Speaking	Electricians	60
	Software Developers	56
	Carpenters	53
	Agricultural Laborers	50
	Masons, Brick Layers	50
	Truck, Bus Drivers	50
	Textile Workers (e.g., Weavers)	47
	Data Entry Operators	47
	Domestic Helpers, Cleaners	44
	Plumbers	38

Notes: Data sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. The occupation descriptions are matched to those in the O*NET database and ranked in ordinal fashion using a measure of the importance of speaking. The median rank is used to create two groups of occupations: speaking and non-speaking. The sample of 10 occupations in each group includes examples close to the median. The total number of occupations in the full sample is over 200.

Table 1 presents a sample of occupations categorized by their speaking intensity based on O*NET importance scores. Professional occupations like lawyers, teachers, and nurses have high speaking importance scores between 70-90, while manual jobs like plumbers and textile workers have scores below 50. Notably, the correlation between skill and speaking intensity is not perfect. That is, some high-skill occupations like software developers have relatively low speaking scores (56), while some traditionally lower-skill service jobs like call center operators and child care workers have high speaking scores (78 and 69 respectively). The classification uses the median score to divide occupations into speaking and non-speaking categories.

Table 2 shows wage premia and employment shares across speaking and non-speaking occupations by education level. Speaking occupations offer higher wages for both groups. That is, college workers earn 30 percent more and non-college workers earn 16 percent more in speaking compared to non-speaking occupations. However, employment is heavily concentrated in non-speaking occupations, which employ 79 percent of all workers. The majority of these are non-college workers, corresponding to 74 percent of total employment. On the other hand, college workers are more evenly distributed between speaking (7 percent) and non-speaking occupations (5 percent).

Table 2: Wage and Employment in Speaking and Non-Speaking Occupations

	Wage Ratio		Employment Share	
College	Non-College	College	Non-College	
Speaking	1.30	1.16	0.07	0.14
Non-Speaking			0.05	0.74

Notes: Data sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Wages are deflated using monthly state-level CPI from the Reserve Bank of India. Wage ratio is defined as the population-weighted mean wages in speaking vs. non-speaking occupations. College workers are paid 30 percent higher in speaking occupations than in non-speaking occupations. Non-college workers are paid 16 percent higher wages in speaking occupations than in non-speaking occupations. Employment share is defined as the number of college and non-college in each occupation type as a fraction of the total number of workers in the economy.

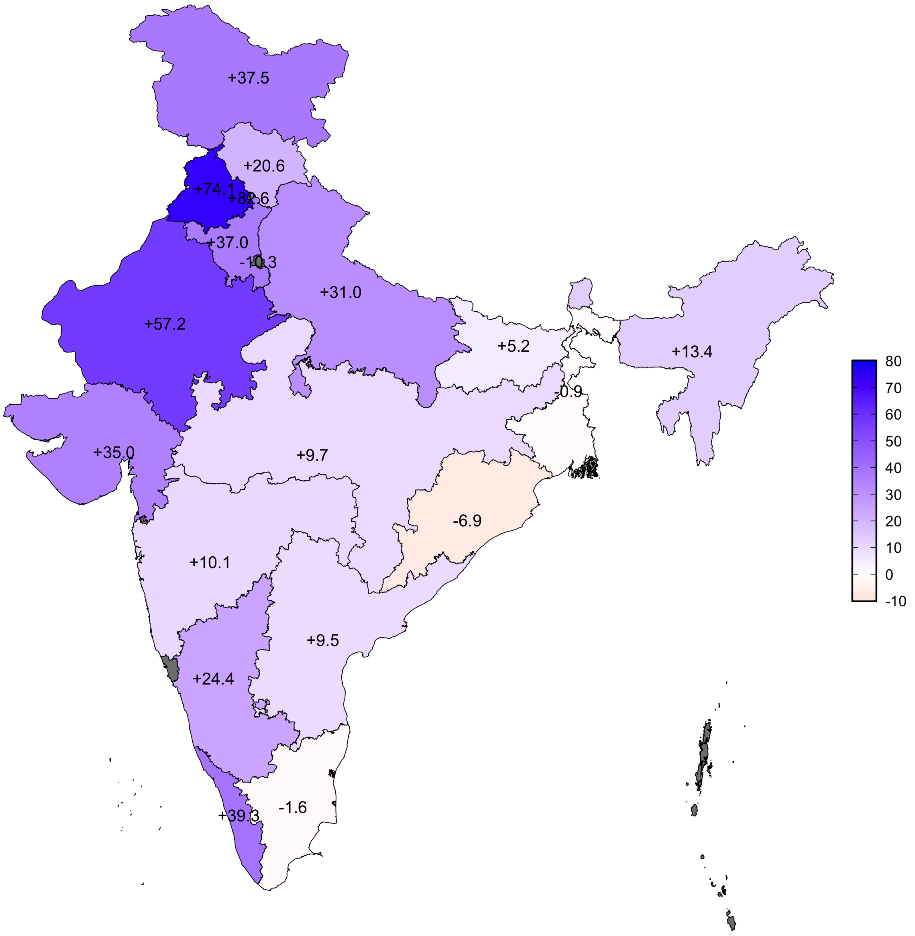
2.4 Other Data

I supplement the above data with two others. First, I use the Indian Human Development Survey-II, 2011-12 (Desai et al., 2018) to obtain information on English proficiency among college and non-college workers. I use the share of English speakers by education across states of India to build a skill-specific linguistic index for structural estimation.

Second, I use microdata from the Employment and Unemployment Rounds of the National Sample Survey (NSS) in 1987 and 2011, before and after India's trade liberalization. Once again, I match occupations to O*NET and construct employment shares in speaking and non-speaking occupations across states of India.

Figure 1 shows the heterogeneous rise of speaking employment across states of India. Most states exhibit positive changes, particularly in northern India, with Rajasthan (57.2 percent) and Punjab (74.1 percent) showing substantial growth. The southern states show more moderate changes, with Tamil Nadu showing a slight decrease (-1.6 percent). West Bengal and Orissa (now Odisha) are among the few states showing negative trends, at -0.9 percent and -6.9 percent respectively. The northeastern states, grouped together as "North East" in this visualization, show a moderate increase of 13.4 percent. However, these percentage changes mask baseline differences in levels. That is, large percent changes reflect small absolute changes in the northern states, while small percent changes reflect large absolute changes in southern states.

Figure 1: Changes in Employment Shares of Speaking Occupations between 1987 and 2011



Notes: Data is from the Employment-Unemployment Rounds of the National Sample Survey 1987 and 2011. The NCO codes are matched by description to O*NET to obtain a ranking of speaking-intensity. Change is computed as the percentage point difference in employment in each state. State boundaries are from 1987 except that northeastern states and Sikkim are pooled together.

3 Empirical Facts

In this section, I present four empirical facts about the relationship between language barriers, location choices, and labor market outcomes of internal migrant workers in India. I compare three patterns—on migration, occupations, and wages—for college and non-college migrant workers and find that each is attenuated for college relative to non-college workers. This is written up summarily as a fourth fact.

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

To understand how language barriers are related to the migration of college and non-college workers, I use district-to-district migration data from the 2011 Census jointly with language data from the 2001 Census. This is done to avoid issues of simultaneity, since measures of language barriers are constructed with the distribution of speakers of different languages in each district, which includes contemporaneous migrants. Further, since the 2011 Census contains information on reason for migration jointly with the education level of migrants, I am able to focus my analysis on economic migrants.

I measure bilateral language barriers, $\tau^L_{o,d}$, using the linguistic distance index defined in Section 2.2. For ease of comparison across two dimensions of heterogeneity (language, education), I demarcate migrant workers of each education type into two parts: those with a language barrier relative to their origin and those without, defined by $\mathbb{1}\{\text{Barrier}_{o,d}\} = 1$ if $\tau^L_{o,d} > \textit{median}$. I measure bilateral geographic barriers by the geodesic distance between geographic centers of districts o, d. I estimate the following regression specification,

$$\begin{aligned} \ln(\pi_{o,d,e}) = \beta_0 + \beta_1 \mathbb{1}\{\text{Barrier}_{o,d}\} + \beta_2 \mathbb{1}\{\text{College}\} + \beta_3 \mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}\} \\ + \beta_4 \ln(\text{Geo. Dist.}_{o,d}) + \beta_5 \ln(\text{Geo. Dist.}_{o,d}) \times \mathbb{1}\{\text{College}\} + \gamma_o + \gamma_d + \varepsilon_{o,d,e}, \end{aligned} \tag{2}$$

where $\ln(\pi_{o,d,e}) \equiv \ln(N_{o,d,e}/\sum_d N_{o,d,e})$ is the fraction of workers at origin o with education $e \in \{\text{College, Non-College}\}$ that migrate to destination d, which includes the origin. I add origin and destination fixed-effects, γ_o and γ_d, to control for time-invariant unobserved heterogeneity across districts. In the final specification, I also include an origin state × destination state fixed-effect, to control for any bilateral variation across states, exploiting linguistic and geographic variation across districts *within* states to capture the relationship between language barriers and migration patterns.

Table 3 presents the results of this regression. Columns 1 and 2 report specifications with only language barriers and geographic barriers, respectively. I include these to compare how much additional variation in migration flows is explained by each factor. In the data, linguistic and geographic barriers have a correlation coefficient of 0.35, indicating a moderate positive relationship. This correlation suggests that locations that are geographically distant tend to also have higher language barriers, and vice versa. However, the correlation is not so high as to preclude locations that are geographically close but linguistically distant, and vice versa.

Table 3: Language and Migration of College vs. Non-College

Dependent Variable: $\ln(\pi_{o,d,e})$	(1)	(2)	(3)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$	-2.21*** (.021)		-.617*** (.020)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}\}$	.074*** (.024)		.093*** (.022)
$\mathbb{1}\{\text{College}\}$	.892*** (.015)	.685*** (.030)	.639*** (.014)
$\ln(\text{Geo. Dist.}_{o,d})$		-1.97*** (.009)	-1.84*** (.008)
$\ln(\text{Geo. Dist.}_{o,d}) \times \mathbb{1}\{\text{College}\}$		.116*** (.013)	.104*** (.014)
Origin District FE	Yes	Yes	Yes
Dest. District FE	Yes	Yes	Yes
Origin State × Dest. State FE	Yes	Yes	Yes
R^2	0.180	0.314	0.316
Observations	277,379	277,379	277,379

Notes: Data on migration is sourced from the Census of India 2011. Data on languages is sourced from Census of India 2001. Standard errors are in parentheses and clustered at the destination district. *** means $p < 0.01$.

Column 3 presents the most comprehensive specification, incorporating controls for both language barriers and geographic distance, along with their interactions with a college fixed-effect. The coefficient on $\mathbb{1}\{\text{Barrier}_{o,d}\}$ is -0.617 (p < 0.01), suggesting a negative association between the presence of a language barrier and migration. This correlation implies that, on average and holding other factors constant, the presence of a language barrier is associated with a decrease in the fraction of migrants by approximately $\exp(-0.617) - 1 = 46$ percent. Interestingly, this negative relationship appears to be weaker for college-educated individuals, as indicated by the positive coefficient on the interaction term, $\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}\}$, which is 0.093 (p < 0.01). When considering both coefficients together, the data suggest a $\exp(-0.617+0.093) - 1 = 41$ percent lower fraction of migrants in the presence of a language barrier for college-educated workers, a less pronounced difference compared to non-college workers.

It is worth noting two additional patterns that are consistent with the literature. First, college workers are more likely to migrate, as indicated by the positive coefficient on $\mathbb{1}\{\text{College}\}$, which is 0.639 ($p < 0.01$). This suggests that all else equal, college workers have an 89 percent higher fraction of migrants relative to non-college workers. Second, geographic distance is a significant impediment to internal migration. The coefficient on $\ln(\text{Geo. Dist.}_{o,d})$, which is -1.84 ($p < 0.01$), indicates that a 10 percent increase in geographic distance is associated with a 18.4 percent decrease in the fraction of migrants for non-college workers. However, college workers are less deterred by geographic distance, from whom the positive interaction term on $\ln(\text{Geo. Dist.}_{o,d}) \times \mathbb{1}\{\text{College}\}$, which is 0.104 ($p < 0.01$), suggests a 17.4 percent decrease in the fraction of migrants from the same 10 percent increase in distance. This indicates that while geographic distance is a significant deterrent for all workers, its effect is less pronounced for those with a college education.

Fact 1 is a statement of these findings, which show that lower fractions of migrant workers choose locations with high language barriers, but this correlation is less pronounced for college workers relative to non-college workers. This may be because college workers are more likely to speak a lingua franca like English and have access to occupations that are less dependent on local language proficiency.

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 understand how language barriers are related to the occupations of college and non-college migrant workers. I use 28 waves (Sep 2017 to Dec 2019) of the CPHS repeated cross-section data jointly with language data from the 2011 Census. The CPHS contains detailed information on labor market outcomes of individual workers, whose observed "state of origin" allows me to link them to language and geographic barriers. I also manually create a concordance between occupations in the CPHS panel and O*NET to obtain an exogenous ranking in intensity of "speaking," defined by O*NET as "talking to others to convey information effectively." I show that among migrant workers, those with high language barriers are less likely than workers with low language barriers to choose occupations that are intensive in "speaking."

In particular, I estimate the following linear probability specification 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, which account for the degree to which individuals' demographic characteristics and proxied ability may influence their occupation choice. I add origin-month and destination-month fixed-effects, $\gamma_{o,t}$ and $\gamma_{d,t}$, to control for time-varying factors in the states of origin and destination districts. I include these to minimize, to the extent possible, bias from omitted variables and isolate the relationship between language barriers and occupational choice for college and non-college workers.

The results from these regressions are in Table 4. Column 1 presents the relationship between college education and speaking-intensive occupations for native workers, serving as a point of comparison. Columns 2-4 focus on migrant workers, progressively adding controls aimed at isolating the effects of language barriers and education. The specification in Column 2 adds the indicator for language barrier and its interaction with college education, showing the primary relationships of interest without geographic controls. The specification in Column 3 adds the log of geographic distance to account for spatial factors that might influence occupational choices independent of language.

Table 4: Language and Occupation Choice of College vs. Non-College

	Natives	Migrants		
Dependent Variable: $\mathbb{1}\{\text{Speaking Occ.}_{i,o,d,j,t}\}$	(1)	(2)	(3)	(4)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		-.066*** (.007)		-.057*** (.007)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$		.078*** (.012)		.075*** (.012)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	.248*** (.0006)	.296*** (.004)	.121*** (.036)	.114*** (.040)
$\ln(\text{Geo. Dist.}_{o,d})$			-.005 (.005)	.002 (.005)
$\ln(\text{Wages}_{i,o,d,j,t-1})$	Yes	Yes	Yes	Yes
$\text{Controls}_{i,o,d,j,t}$	Yes	Yes	Yes	Yes
Origin State × Month FE	No	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes	Yes
R^2	0.246	0.274	0.273	0.292
Observations	4,457,316	109,591	109,591	109,591

Notes: Data sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Wages are deflated using monthly state-level CPI from the Reserve Bank of India. Data on languages sourced from the Census of India 2011. Standard errors are in parentheses and clustered at the destination district × month. *** means $p < 0.01$.

Column 4 includes all controls and fixed-effects and provides the most comprehensive specification. The coefficient on the language barrier indicator, which is -0.06 (p < 0.01) suggests that migrant workers facing a language barrier are 6 percentage points less likely to work in speaking-intensive occupations compared to those without a language barrier. The coefficient on the college indicator, which is 0.11 (p < 0.01), suggests that college-educated migrant workers are 11 percentage points more likely to work in speaking-intensive occupations than non-college educated workers. Finally, the interaction term between indicators for language barriers and college education, which is 0.075 (p < 0.01), shows that the negative correlation of language barriers on choosing speaking-intensive occupations is reduced by 7.5 percentage points for college-educated workers.

Fact 2 is a statement of these findings, which show that high language barriers are associated with a lower likelihood of choosing speaking-intensive occupations while college education increases the likelihood of choosing speaking-intensive occupations. Importantly, the interaction supports the second part of Fact 2, showing that the negative correlation of language barriers is indeed muted for college relative to non-college workers.

Fact 3 (Wages): *Among migrant workers, those with high language barriers receive higher wages, but this is attenuateed for college relative to non-college.*

To understand how language barriers are related to the wages of college and non-college migrant workers, once again, I use 28 waves (Sep 2017 to Dec 2019) of the CPHS repeated cross-section data jointly with language data from the 2011 Census. I deflate the nominal wages using the monthly CPI time-series released by the Ministry of Statistics and Programme Implementation, Government of India.

Then, I estimate the following specification,

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

where $\text{Controls}_{i,o,d,j,t}$ are individual i's gender, age, and caste, which account for the degree to which individuals' demographic characteristics may influence their wages. I add origin-month and destination-month fixed-effects, $\gamma_{o,t}$ and $\gamma_{d,t}$, to control for time-varying factors in the origin states and destination districts. Importantly, I include an occupation-month fixed-effect to control for underlying characteristics across occupations that may be contributing to observed differences in wages. These controls account for systematic differences in location, occupation, and demographic characteristics that might bias our estimates. I include these controls to minimize, to the extent possible, bias from omitted variables and isolate the relationship between language barriers and wages for college and non-college workers.

The results from these regressions are in Table 5. Once again, Column 1 showing the baseline relationship for native workers and Columns 2-4 focus on migrant workers. As before, the specification in Column 2 adds the indicator for language barrier and its interaction with college education, showing the primary relationships of interest without geographic controls. The specification in Column 3 adds the log of geographic distance to account for spatial factors that might influence occupational choices independent of language.

Table 5: Language and Wages of College vs. Non-College

	Natives	*Migrants*		
Dependent Variable: $\ln(\text{Wages}_{i,o,d,j,t})$	(1)	(2)	(3)	(4)
$\mathbb{1}\{\text{Barrier}_{o,d}\}$		.063*** (.013)		.055*** (.013)
$\mathbb{1}\{\text{Barrier}_{o,d}\} \times \mathbb{1}\{\text{College}_{i,o,d,j,t}\}$		-.063*** (.014)		-.062*** (.014)
$\mathbb{1}\{\text{College}_{i,o,d,j,t}\}$	.196*** (.001)	.279*** (.006)	.128*** (.048)	.134*** (.048)
$\ln(\text{Geo. Dist.}_{o,d})$			.036*** (.007)	.034*** (.007)
$\text{Controls}_{i,o,d,j,t}$	Yes	Yes	Yes	Yes
Occupation × Month FE	Yes	Yes	Yes	Yes
Origin State × Month FE	No	Yes	Yes	Yes
Dest. District × Month FE	Yes	Yes	Yes	Yes
R^2	0.607	0.676	0.673	0.676
Observations	4,105,305	103,442	103,442	103,442

Notes: Data sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Wages are deflated using monthly state-level CPI from the Reserve Bank of India. Data on languages sourced from the Census of India 2011. Standard errors are in parentheses and clustered at the destination district × month. *** means $p < 0.01$.

Column 4, which includes all controls and fixed-effects, provides the most comprehensive specification. The coefficient on the language barrier indicator, which is 0.055 (p < 0.01), suggests that migrant workers facing a language barrier earn approximately 5.5 percent higher wages compared to those without a language barrier. The coefficient on the college indicator, which is 0.134 (p < 0.01), suggests that migrant workers with a college education earn about 13.4 percent higher wages relative to migrant workers without a college education. The coefficient on the interaction term between language barriers and college education, which is -0.062 (p < 0.01), suggests that the positive correlation of language barriers on wages is reduced by approximately 6.2 percentage points for college-educated workers. Finally, it is worth noting that the coefficient on geographic distance between origin and destination, which is 0.034 (p < 0.01), suggests that migrants who travel farther earn a wage premium. This is consistent with the migration literature that has suggested that geographic barriers have a selection effect

(Bryan and Morten, 2019).

Fact 3 is a statement of these findings, which show that high language barriers are associated with higher wages among migrant workers. Importantly, the wage premium associated with language barriers is muted for college relative to non-college workers. This may be due to a selection effect of language barriers, which is attenuated for college workers. That is, workers who face language barriers may need higher idiosyncratic productivity to be able to overcome the barriers.

Fact 4 (College vs. Non-College): *Each of the patterns in Facts 1, 2, and 3 is attenuated for college relative to non-college.*

As shown in Tables 3, 4, and 5, the relationship between language and migration, occupations, and wages is attenuated for college relative to non-college workers. Fact 4 is a reiterated statement of this difference in patterns.

Summary: This section presents four empirical facts about the relationship between language barriers, location choices, and labor market outcomes for internal migrant workers in India, comparing patterns for college and non-college-educated individuals. First, workers are less likely to migrate to locations with high language barriers. Second, among migrants, those facing high language barriers are less likely to choose occupations intensive in speaking skills. Third, migrant workers facing high language barriers tend to receive higher wages. Fourth, each of these patterns is attenuated for college relative to non-college workers.

These patterns persist even when controlling for geographic distance and other relevant factors. College-educated workers are generally more mobile, more likely to choose speaking-intensive occupations, and earn higher wages than their non-college counterparts. The observed patterns suggest complex interactions between language barriers, education levels, and labor market outcomes. In the following section, I show how these facts can be explained in a spatial equilibrium model with multiple locations, multiple occupations, and heterogeneous workers that differ by skill, education, and language barriers.

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 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. The Appendix contains detailed derivations of the main equations of the model.

Environment: Let there be R regions in the economy, which represent the states of India. I denote the origin region by o and the destination region by d. 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 and without language barriers. This form allows the degree of substitution between labor inputs to be flexible. But we may expect, for instance, that skilled workers with and without language barriers are more substitutable than unskilled workers with and without language barriers. This is because skilled workers often have access to a common language (e.g., English in India) used across different labor markets, while unskilled workers typically require local language proficiency.

I assume that workers, who 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,

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

where the location parameter, $A_{j,e,\mathscr{L}(o,d)}$, governs absolute advantage and 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, $A_{spk,nc,b} < A_{spk,nc,nb}$.

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, $w_{d,j,e,\mathscr{L}(o,d)}$, 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 at occupation $j \in J$ at destination $d \in R$ is given by

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

where $\tau^G_{o,d}$ is the geographic barrier, $\tau^L_{o,d,e}$ is the language barrier (measured separately for college and non-college workers), and α_d are region-specific amenities. They choose (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

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

where $\Phi_{o,e} \equiv \sum_{(d',j')} A_{j',e,\mathscr{L}(o,d')}\ (\Lambda_{o,d',j',e})^{\theta}$. 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 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 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. This equation implies that as more workers from origin o with education e migrate to work at occupation j at destination d, their average productivity decreases because the marginal migrant is drawn from further down the left tail of the productivity distribution. The expression for the average wage of worker from origin o with education e in occupation j at destination d,

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

embeds this mechanism. This is the key selection equation and captures the idea that when few workers from origin o with education e choose occupation j at destination d, their average productivity is higher, and so their average wages are higher. [This is consistent with Fact 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 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, $\phi_{d,j}$,

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

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

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

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 $\nu_{j,e}$,

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

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. Language barriers enter the innermost nest because they represent a friction that modifies worker productivity within skill groups, rather than a fundamental technological distinction.

This ordering allows for occupation-specific elasticities between workers with and without language barriers ($\nu_{j,e}$). This captures that the substitutability of workers with language barriers likely varies by both occupation type and education level. This key feature would be harder to capture with alternative nesting structures. Importantly, if the innermost elasticities ($\nu_{j,e}$)

turn out to be large, this nesting structure naturally collapses to more conventional production functions that distinguish only between skilled and unskilled labor, nesting my model within the broader literature on education-based labor market sorting.

The regional firm takes wages and prices as given and maximizes profit. This is given by $p_d y_d - \sum_{j\in J}\sum_{e\in\{c,nc\}}\sum_{\mathscr{L}(o,d)\in\{b,nb\}} w_{d,j,e,\mathscr{L}(o,d)}\ell_{d,j,e,\mathscr{L}(o,d)}$. The first-order conditions determine labor demand,

$$\ell_{d,j,e,\mathscr{L}(o,d)} = \Big(MP_{d,j} \,\cdot\, MP_{d,j,e} \,\cdot\, w_{d,j,e,\mathscr{L}(o,d)}\Big)^{\frac{\nu_{j,e}-1}{\nu_{j,e}}} \tag{10}$$

where $MP_{d,j} = p_d(y_d)^{\frac{1}{\kappa-1}}(\phi_{d,j})^{\frac{1}{\kappa}}(\ell_{d,j})^{\frac{\rho}{\rho-1}\frac{\kappa-1}{\kappa}-1}$ is the value marginal product of labor for occupation j at region d, and $MP_{d,j,e} = (\ell_{d,j,e})^{\frac{\nu_{j,e}}{\nu_{j,e}-1}\frac{\rho-1}{\rho}-1}$ is the value marginal product of labor for workers with education e in occupation j at region d. I assume that markets are perfectly competitive, so prices are equal to marginal cost functions,

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

where

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

and

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

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 (so, less sensitivity) of the aggregate price index to regional price variations. The firm takes prices as given and maximizes profit, which is given by $PY - \sum_{r\in R} p_d y_d$. 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 $\left[\sum_{d\in R}(p_d)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}$. 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 numeraire i.e., $P = 1$.

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

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

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 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,

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

The equilibrium in this model is defined as follows. Given fundamentals, $\Theta \equiv \{A_{j,e,\mathscr{L}(o,d)}, \theta, \sigma, \phi_{d,j}, \kappa, \rho, \nu_{j,e}, N_{d,e}, \tau^G_{o,d}, \tau^L_{o,d,e}, \alpha_d\}_{o,d\in R, j\in J, e\in\{c,nc\}, \mathscr{L}(o,d)\in\{b,nb\}}$, an equilibrium is a set of wages and prices, $\{w_{d,j,e,\mathscr{L}(o,d)}, p_d, P\}_{o,d\in R, j\in J, e\in\{c,nc\}, \mathscr{L}(o,d)\in\{b,nb\}}$, 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).

5 Taking the Model to the Data

In this section, I outline the procedure by which I discipline the model parameters.[6] I proceed in three steps. First, I estimate the Fréchet dispersion parameter, θ, from a regression of migration flows against migration costs. Second, I jointly estimate the Fréchet location parameter, $A_{j,e,\mathscr{L}(o,d)}$, amenities, α_d, and CES elasticities from the outer, middle, and inner nests, κ, ρ, and $\nu_{j,e}$ in a GMM procedure. For this, I use moment conditions derived from the sorting

[6] See the Appendix for a summary of model parameters and their identification.

equation and labor demand equations. Third, I do various validation exercises and sensitivity checks on the estimated parameters.

I begin by using the CPHS data to construct migration flows between states of India, disaggregated by occupations and skill.[7] Using the origin and destination states, I link this data to geodesic distance and linguistic index. Since the data is a repeated cross-section, I use time periods as an additional source of variation. These are defined at a frequency of four months on monthly waves from January 2014 to December 2019. This definition results in $T+1 = 18$ time periods, of which the first is defined as the initial period ($t = 0$). I denote each subsequent period by $t = 1, ..., 17$.

In the first step, to estimate the Fréchet dispersion parameter, θ, I compute the Head-Reis index to derive a relationship between migration flows and migration costs (see the Appendix for the derivation). I parameterize the geographic component of migration cost, $\tau^G_{o,d}$, using geodesic distance between states of India. Similarly, I parameterize the linguistic component of migration cost, $\tau^L_{o,d,e}$, using linguistic distance between states of India. For this mapping, I build a skill-specific linguistic index, which assumes that different populations of college and non-college workers across states may speak English as a second language. The share of English speakers by education across states is computed from the IHDS-II (2011-12).

Using parameterizations of migration costs, I fit the relationship between migration flows and migration costs by least squares. For identification of θ, the specification allows for a constant and an error term that is assumed to be uncorrelated with migration costs in expectation. That is, I estimate θ from the following,

$$\ln\left(\sqrt{\frac{N_{o,d,j,e,t}}{N_{d,d,j,e,t}}\frac{N_{d,o,j,e,t}}{N_{o,o,j,e,t}}}\right) = \text{Constant} + \theta \ln\left[(1-\tau^G_{o,d})(1-\tau^L_{o,d,e})\right] + \varepsilon_{o,d,j,e,t},$$

where the source of identifying variation is over state-pairs, occupations, and time periods. Table 6 contains the regression results and my preferred specification estimates $\hat{\theta} = 5.4$.

Table 6: Estimation of θ

	(1)	(2)
$\ln\left[(1-\tau^G_{o,d})(1-\tau^L_{o,d,e})\right]$	5.35^{***}	5.47^{***}
	(.200)	(.170)
Destination × Occupation FE	N	Y
Observations	1,433	1,433

Notes: Data on migration flows sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Data on language sourced from Census 2011 and IHDS-II.

[7] The CPHS dataset contains information on destination districts and origin states, so the estimation could be done at this level of aggregation. However, since this would increase the dimensionality of the problem, I plan to do this in future iterations of this paper.

In the second step, I jointly estimate the Fréchet location parameter, $A_{j,e,\mathscr{L}(o,d)}$, amenities, α_d, and CES elasticities from the outer, middle, and inner nests, κ, ρ, and $\nu_{j,e}$ in a two-step GMM procedure with instrumental variables. For the identification of these parameters, I need to take a stand on the migration costs, $\tau^G_{o,d}$ and $\tau^L_{o,d,e}$, the CES elasticity from the aggregate nest, σ, and the CES share parameters from the outer nest, $\phi_{r,j}$.

As described above, I parameterize migration costs by geodesic distance and linguistic index between states. Additionally, I take the CES elasticity parameter of the aggregate nest, σ, from Bryan and Morten (2019) and follow them in presenting results for a value of $\sigma = 8$. I proxy for CES share parameters from the outer nest, $\phi_{r,j}$, from employment shares in speaking and non-speaking occupations.

The GMM objective function uses four sets of moment conditions to jointly estimate the Fréchet location parameter, $A_{j,e,\mathscr{L}(o,d)}$, amenities, α_d, and CES elasticities from the outer, middle, and inner nests, κ, ρ, and $\nu_{j,e}$. The first is the log-linearized sorting equation (corresponding to equation 7) to identify $A_{j,e,\mathscr{L}(o,d)}$ and α_d,

$$\left(\frac{1}{T}\right)\left(\frac{1}{|o:(o,d)\in\mathscr{L}(o,d)|}\right)\sum_{t=1}^{T}\sum_{(o,d)\in\mathscr{L}(o,d)}\Bigg(\ln(\pi_{o,d,j,e,t})$$

$$-\Big(\ln(A_{j,e,\mathscr{L}(o,d)})+\theta\Big[\ln(1-\tau^G_{od})+\ln(1-\tau^L_{o,d,e})+\ln(w_{d,j,e,\mathscr{L}(o,d),t})+\ln(\alpha_d)+\ln(\Phi_{o,e,t})\Big]\Big)\Bigg),$$

where these parameters are over-identified, with $8+R$ unknowns and $8R$ moment conditions. The sources of identifying variation are origin states with or without a language barrier to destination state d. Time periods are an additional source of variation.

The GMM objective function nests the model solution. That is, in each iteration, the model solution is computed using data, known parameters, and guess values for the parameters being estimated. The model solution for the particular iteration gives equilibrium values of $w_{d,j,e,\mathscr{L}(o,d)}$, which ensures that all components of the sorting equation are known, except for $A_{j,e,\mathscr{L}(o,d)}$ and α_d, which are being estimated.

The next three sets of moment conditions correspond to the outer, middle, and inner CES nests and identify the CES elasticities, κ, ρ, and $\nu_{j,e}$. Derived from taking ratios of labor demand equations, the moment conditions can be expressed as linear relationships between relative labor quantities and relative wages at three levels of aggregation.

For the outer nest, I take the ratio of labor demand for two different occupations j and j' at the same destination d. Taking logs and time differences eliminates destination-specific terms that appear multiplicatively in the labor demand equation. This yields equation 15, which relates relative wages to relative quantities with the elasticity parameter κ,

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R}\right)\sum_{t=1}^{T-1}\sum_{d=1}^{R}\left(\Delta_t \ln\left(\frac{w_{d,j,t}}{w_{d,j',t}}\right)-\left(\frac{1-\kappa}{\kappa}\right)\cdot\Delta_t \ln\left(\frac{\ell_{d,j,t}}{\ell_{d,j',t}}\right)\right). \tag{15}$$

For the middle nest, I take the ratio of labor demand for college and non-college workers within the same occupation j at destination d. Taking logs and time differences again eliminates occupation-destination specific terms. This yields equation 16, which relates relative wages to relative quantities with the elasticity parameter ρ,

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R}\right)\sum_{t=1}^{T-1}\sum_{d=1}^{R}\left(\Delta_t \ln\left(\frac{w_{d,j,c,t}}{w_{d,j,nc,t}}\right)-\left(\frac{1-\rho}{\rho}\right)\Delta_t \ln\left(\frac{\ell_{d,j,c,t}}{\ell_{d,j,nc,t}}\right)\right)\cdot\Delta_t \ln\left(\frac{\ell_{d,j,c,t}}{\ell_{d,j,nc,t}}\right)\Bigg). \tag{16}$$

Similarly, for the inner nest, I take the ratio of labor demand for workers with and without language barriers, within the same education level e and occupation j at destination d. Taking logs and time differences eliminates education-occupation-destination specific terms. This yields equation 16, which relates relative wages to relative quantities with the elasticity parameter $\nu_{j,e}$,

$$\left(\frac{1}{T-1}\right)\left(\frac{1}{R}\right)\sum_{t=1}^{T-1}\sum_{d=1}^{R}\left(\Delta_t \ln\left(\frac{w_{d,j,e,b,t}}{w_{d,j,e,nb,t}}\right)-\left(\frac{1-\nu_{j,e}}{\nu_{j,e}}\right)\Delta_t \ln\left(\frac{\ell_{d,j,e,b,t}}{\ell_{d,j,e,nb,t}}\right)\right)\cdot\Delta_t \ln\left(\frac{\ell_{d,j,e,b,t}}{\ell^{d,j,e,nb,t}}\right)\Bigg). \tag{17}$$

Further, in equations 15, 16, and 16, I substitute for $w_{d,j,e,\mathscr{L}(o,d),t}$ from the selection equation (corresponding to equation 9),

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

and $\ell_{d,j,e,\mathscr{L}(o,d),t}$ from labor market clearing condition (corresponding to equation 12),

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

so that $w_{d,j,e,\mathscr{L}(o,d),t}$ and $\ell_{d,j,e,\mathscr{L}(o,d),t}$ are written in terms of data, known parameters, and the unknown parameters being estimated, $\nu_{j,e}$, ρ, κ, and $A_{j,e,\mathscr{L}(o,d)}$. The CES elasticities are exactly identified, with 6 unknown parameters and 6 moment conditions. The source of identifying variation are regions for the inner and outer nests and regions and occupations for the middle nest. Time periods, as before, are an additional source of variation.

For the identification of the CES elasticity parameters, there is reason to be concerned about simultaneity between wages and the dependent variable in each of the moment conditions. That is, higher relative wages may increase the influx of migrants at a location, but an increase in the relative supply of workers may depress wages. Further, wages of workers might be influenced by unobserved skills or by existing migrant networks at the destination, that also affect their propensity to migrate, resulting in omitted variable bias. Amenities at the destination may be endogenous, as they might improve due to the influx of migrants.

To allay these concerns, and based on arguments in the literature on migrant networks, e.g., Munshi (2003), I use data on past migration shares and relative labor quantities to build shift-share instrumental variables following Altonji and Card (1991) and Card (2001). The instrumental variables for the CES moment conditions are, respectively,

$$\frac{z_{d,j,t}}{z_{d,j',t}} = \frac{\sum_{o\in R}\left(\frac{N_{o,d,j,t=0}}{N_{d,j,t=0}}\ \Delta_t N_{o,-d,-j}\right)}{\sum_{o\in R}\left(\frac{N_{o,d,j',t=0}}{N_{d,j',t=0}}\ \Delta_t N_{o,-d,-j'}\right)},$$

$$\frac{z_{d,j,c,t}}{z_{d,j,nc,t}} = \frac{\sum_{o\in R}\left(\frac{N_{o,d,j,c,t=0}}{N_{d,j,c,t=0}}\ \Delta_t N_{o,-d,-j,c}\right)}{\sum_{o\in R}\left(\frac{N_{o,d,j,nc,t=0}}{N_{d,j,nc,t=0}}\ \Delta_t N_{o,-d,-j,nc}\right)},$$

$$\frac{z_{d,j,e,b,t}}{z_{d,j,e,nb,t}} = \frac{\sum_{o\in R}\left(\frac{N_{o,d,j,e,b,t=0}}{N_{d,j,e,b,t=0}}\ \Delta_t N_{o,-d,-j,e,b}\right)}{\sum_{o\in R}\left(\frac{N_{o,d,j,e,nb,t=0}}{N_{d,j,e,nb,t=0}}\ \Delta_t N_{o,-d,-j,e,nb}\right)}.$$

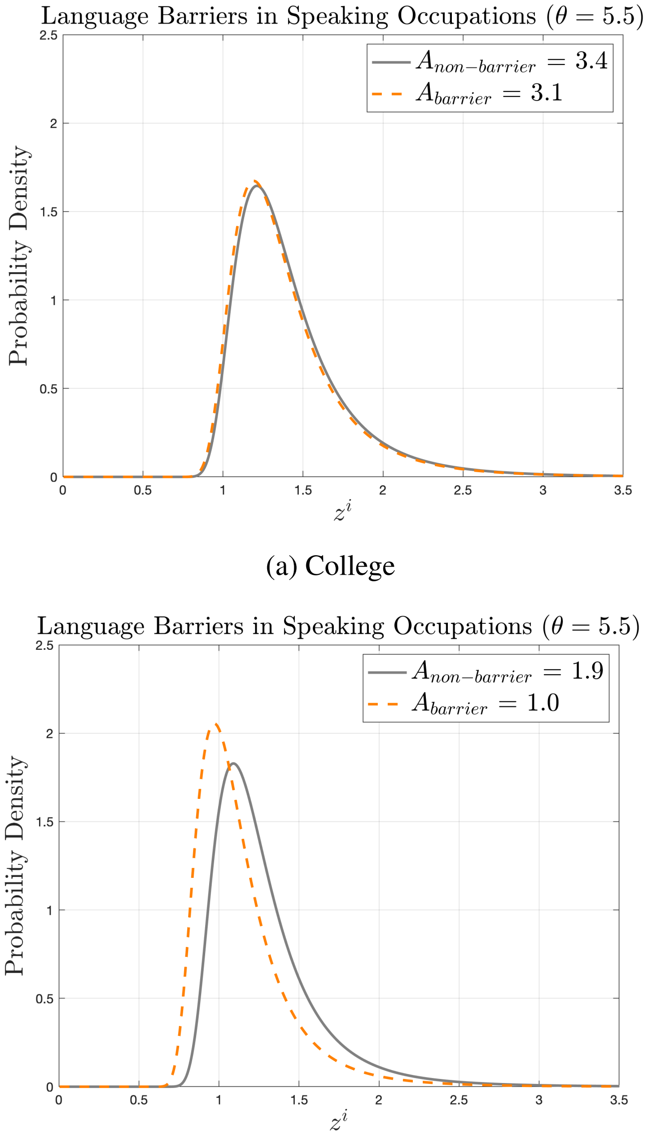
In the first stage, I regress these instruments on potentially endogenous regressors (relative labor quantities) and compute residuals. In the second stage, I implement the GMM procedure described above with residuals entering multiplicatively in the moment conditions. These instruments meet the exclusion restriction if past migration flows are uncorrelated with current labor demand conditions. The identification assumption is that historical settlement patterns of migrants only affect current wages through their influence on current migration, not through other channels.

This is plausible because historical migration networks were established for reasons unrelated to current local labor demand shocks. The shift component, which uses changes in labor supply from origin regions to other destinations, further helps isolate variation that is likely exogenous to local demand conditions at the destination. Further, residuals from the first-stage regressions mitigate bias from omitted variables that might simultaneously affect wages and migration decisions, such as unobserved worker quality or destination amenities.

The estimates of Fréchet location parameters and CES elasticities from the GMM procedure validate my hypothesis that language barriers affect unskilled rather than skilled workers, and in speaking rather than non-speaking occupations. These estimates are robust to alternative specifications and different initial values in the GMM procedure.

I normalize the Fréchet location parameter estimate of non-college workers with language barriers in speaking occupations, $\hat{A}_{spk,b,nc}$ to be 1. I find that $\hat{A}_{spk,b,nc} = 1.9$, so non-college workers without language barriers are nearly twice as productive on average in speaking occupations than those with language barriers. The size of these productivity differences reveals the substantial economic costs that arise when language barriers shape workers' sorting patterns across occupations. The precision of these estimates allows us to make strong inferences about the role of language barriers in determining worker productivity. I plot these estimates in Figure 2.

Figure 2: Estimates of $A_{j,e,\mathscr{L}(o,d)}$ for College and Non-College in Speaking Occupations



(a) College

(b) Non-College

Notes: Using estimates of $A_{j,e,\mathscr{L}(o,d)}$ and θ, I simulate the probability density functions by taking draws of z^i for college and non-college workers with and without language barriers in speaking occupations. The figure shows that $A_{spk,nb,c} = 1.1 \cdot A_{spk,b,c}$ whereas $A_{spk,nb,nc} = 1.9 \cdot A_{spk,b,nc}$.

For college workers in speaking occupations, the ratio between the average productivity of workers without language barriers and those with language barriers is more even, at 1.1. Similarly, this ratio is close to 1 for both college and non-college workers in non-speaking occupations. This pattern aligns with the observation that college education in India often provides English language skills, reducing the impact of local language barriers. Moreover, these findings suggest that education and language ability act as complementary factors in determining worker productivity in speaking-intensive jobs. Thus, productivity differences between workers with and without language barriers are more pronounced for non-college workers in speaking occupations.

The estimates of CES elasticities, which are summarized in Table 7 corroborate this story. For the outer nest, I find that the elasticity of substitution between labor in speaking and non-speaking occupations is $\hat{\kappa} = 2.65$. This is a novel estimate with no benchmark in the literature. It has an intuitive interpretation. In this context, both front-office occupations that involve interaction with customers and back-office occupations that involve less interaction with the customers complement each other in production. For example, both a waiter, who interacts with the customer, and a chef, who provides the food, complement each other in producing the restaurant service.

Table 7: Estimates of CES Elasticities

Description		Sub. Elasticity (x)	Point Estimate ($\frac{1-x}{x}$)	Standard Error
Occupation	$\hat{\kappa}$	2.65	-0.623^{**}	(0.308)
Skill	$\hat{\rho}$	1.68	-0.405^{**}	(0.206)
Language Barrier	$\hat{\nu}_{non\text{-}spk,c}$	18.4	-0.946^{**}	(0.473)
	$\hat{\nu}_{spk,c}$	15.6	-0.936^{**}	(0.468)
	$\hat{\nu}_{non\text{-}spk,nc}$	19.8	-0.949^{**}	(0.478)
	$\hat{\nu}_{spk,nc}$	4.6	-0.783^{**}	(0.389)

Notes: Data on migration flows sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Data on language sourced from Census 2011 and IHDS-II. ** means p < 0.05.

For the middle nest, I find that the elasticity of substitution between college and non-college workers is $\hat{\rho} = 1.68$ across occupations and regions. Reassuringly, it is close to benchmark values in the literature. For example, Katz and Murphy (1992) find an estimate of 1.4 in the context of the US and Khanna and Morales (2023) find an estimate of 1.7 in the context of the India. The similarity of my estimate to these established benchmarks lends credibility to the estimation strategy. This relatively low elasticity suggests limited substitutability between college and non-college workers in production. The consistency of this finding across developed and developing country contexts points to fundamental complementarities between skill types in modern production processes.

For the inner nest, I find varying elasticities of substitution between workers with and without language barriers ($\nu_{j,c}$, $\nu_{j,nc}$) across different occupations and skills. These are also novel elasticities with no benchmark for comparison in the literature. In non-speaking occupations, the elasticity of substitution is relatively high for both college ($\nu_{\text{non-spk},c} = 18.4$) and non-college workers ($\nu_{\text{non-spk},nc} = 19.8$), indicating that workers with and without language barriers are quite substitutable in these roles. For speaking occupations, college-educated workers still show a high degree of substitutability ($\nu_{\text{spk},c} = 15.6$). However, there is a notable difference for non-college workers in speaking occupations, where the elasticity of substitution is substantially lower ($\nu_{\text{spk},nc} = 4.6$). This suggests that in speaking occupations, non-college workers with and without language barriers are much less substitutable, highlighting the importance of language skills in these roles for workers without college education.

In the third step, I conduct various validation exercises to ensure that the estimates are robust. First, I compare targeted and untargeted moments from the estimated model and data, as shown in Table 8. The targeted moments are migration shares by skill level, which show correlations above 0.94 between model and data. The untargeted moments—average wages and population counts by origin-destination-occupation-skill cells—also show strong correlations above 0.83, suggesting the model captures important patterns in the data beyond what it was explicitly fitted to match.

Table 8: Correlation of Targeted and Untargeted Moments from Data and Estimated Model

	Moment	Correlation of Model vs. Data
Targeted	$\pi_{o,d,j,c}$	0.94
	$\pi_{o,d,j,nc}$	0.97
Untargeted	$\bar{w}_{o,d,j,c}$	0.88
	$\bar{w}_{o,d,j,nc}$	0.83
	$N_{o,d,j,c}$	0.93
	$N_{o,d,j,nc}$	0.96

Notes: Data on migration flows sourced from the monthly individual income tables from the Consumer Pyramids Household Surveys. Wages are deflated using monthly state-level CPI from the Reserve Bank of India. Data on language sourced from Census 2011 and IHDS-II.

Second, I replicate the empirical facts using the estimated model to verify that it reproduces the key patterns documented in Section 3. Third, I compute CES shares using estimated CES elasticities and labor demand conditions and compare them to the corresponding shares computed from data. I find that their correlation is 0.7, indicating that the model reasonably captures the substitution patterns between different types of workers observed in the data. Together, these validation exercises suggest that the estimated model provides a good fit to both targeted and untargeted features of the data.

6 Counterfactuals

In this section, I use the estimated model to conduct three counterfactual exercises. First, I quantify the impact of language barriers on internal migration, skill premium, and welfare. Second, I quantify the impact of language barriers on the same aggregate outcomes when the spatial distribution of speaking occupations changed across space between 1987 and 2011. Third, I introduce language programs for unskilled migrants and weigh the welfare cost against the benefit of the policy.

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

To quantify the impact of language barriers on aggregate outcomes, I shut down language barriers in the estimated model, while holding other estimated parameters and model fundamentals constant. That is, I set language as a component of migration cost between all region-pairs to 0, i.e., $\tau^L_{o,d,e} = 0 \; \forall o, d \in R$. Simultaneously, I also set the average productivity of non-college workers in speaking occupations that face language barriers to the average productivity of those that do not face language barriers, i.e., $A_{j,b,e} = A_{j,nb,e}$ if $j =$ speaking-intensive. I then compare internal migration, skill premium, and welfare relative to the corresponding outcomes in the estimated model.

The results are in the first panel of Table 9. I find that removing language barriers increases internal migration by 6.2 percentage points, decreases inequality by 1.9 percentage points, and increases welfare by 1.3 percent. This is because when language barriers are removed, workers can relocate more easily to take advantage of economic opportunities and be more productive in speaking-intensive jobs once they arrive. This greater labor mobility enables higher productivity in speaking-intensive occupations, leading to higher aggregate output.

Ex ante, it is not clear that removing language barriers would decrease inequality. On the one hand, the sorting of workers across space and occupations would improve, which would increase their wages. This is particularly so for non-college workers who faced larger language barriers in speaking occupations. However, these workers would no longer face language migration costs and be less selected, which would decrease their wages. Further, in general equilibrium, higher supply of workers that previously faced language barriers, which would would decrease their wages. The quantitative results suggest that sorting gains dominate the general equilibrium and selection forces. I find that language barriers contribute to, approximately, 6.5 percent of skill premium, and so their removal leads to a reduction in skill premium.

To contextualize the magnitude of impact of language barriers, I target the same welfare gains by reducing geographic barriers and quantify the extent to which bilateral geographic barriers, $\tau^G_{o,d}$, need to decrease or the spatial distribution of college workers, $N_{r,c}$, need to increase to achieve this. These comparisons are particularly relevant as they correspond to concrete policy levers. Reducing geographic barriers maps to investments in highways and transportation networks that reduce travel time between regions. Increasing college education maps to expanding higher education through new institutions and financial aid.

Table 9: Effects of Removing Language Barriers

	Internal Migration	Skill Premium	Welfare
Remove Language Barriers	+6.2 p.p.	-1.9 p.p.	+1.3 percent
Decreasing Geographic Barriers (prop. $\downarrow \tau^G_{o,d}$ by $56 percent$)	+6.9 p.p.	-1.6 p.p.	+1.3 percent
Increasing College Share (prop. $\uparrow N_{r,c}$ by $34 percent$)	+6.7 p.p.	-2.1 p.p.	+1.3 percent

Notes: This table contains three panels with change in internal migration, skill premium, and welfare relative to the benchmark from the estimated model. Internal migration is defined as the share of the population that migrated outside their origin state. Skill premium is defined as the percent difference in the real income of the two groups. Welfare is defined as aggregate real income in the economy, which is equal to Y in the model. The first panel shows results from removing language barriers. The second and third panel show results from targeting the same welfare gains as in the first panel. However, this is achieved by decreasing $\tau^G_{o,d}$, geographic barriers between region-pairs o,d, and increasing $N_{r,c}$, share of college workers in each region r, respectively, with no change to language barriers.

The second panel shows that removing language barriers is equivalent, in terms of welfare gains, to proportionally decreasing geographic barriers between each pair of regions by 56 percent. By doing so, internal migration increases by 6.9 percentage points and inequality decreases by 1.6 percentage points. This is more than what is achieved by removing language barriers alone. The larger effect on migration suggests that geographic barriers may be more direct impediments to mobility than language barriers, even though both generate similar welfare losses. This comparison helps quantify the economic importance of language barriers—they are as costly to the economy as having regions be roughly 50 percent further apart from each other, which is significant.

The third panel shows that removing language barriers is equivalent, in terms of welfare gains, to proportionally increasing the share of college workers across regions by 34 percent. By doing so, internal migration increases by 6.7 percentage points and inequality decreases by 2.1 percentage points. This is also more than what is achieved by removing language barriers alone. The fact that removing language barriers generates welfare gains equivalent to such significant improvements in physical and educational infrastructure underscores their economic importance. Moreover, the similar effects on inequality across these counterfactuals suggest that language barriers act as significant constraints on economic opportunity, comparable to constraints from limited transportation access or educational attainment.

6.2 Quantifying the Effect of Language Barriers on Gains from Structural Change

In the second counterfactual, I use the estimated model to understand how language barriers may attenuate gains from structural change, as characterized by a rise in the prevalence of speaking occupations. In the decades following the 1991 liberalization, India experienced a marked expansion of its service sector. According to Fan et al. (2023), the share of employment in the service sector rose from 19 percent in 1987 to 28 percent in 2011, with a particular surge in consumer services in both rural and urban areas. Related to this sectoral shift, there has been a dramatic increase in the prevalence of speaking occupations, which involve a high degree of interpersonal communication, customer interaction, and information exchange. The share of employment in such occupations rose from 2.4 percent in 1987 to 23 percent in 2011.

This shift towards speaking occupations occurred against the backdrop of India's significant linguistic diversity. As the economy increasingly rewards communication skills, the ability to speak multiple languages fluently may have been a crucial determinant of labor market prospects. The interplay between the rising demand for communication skills and India's linguistic heterogeneity potentially affects various economic outcomes, including migration patterns, inequality between workers, spatial disparities in economic development, and overall welfare. Understanding these effects is crucial for assessing the full impact of India's economic transformation and for informing potential policy responses to address language-related barriers in the labor market.

In this exercise, I quantify the impact of language barriers internal migration, skill premium, and welfare when speaking occupations became more prevalent across space. To implement this, I use the estimated model to conduct a model-based difference-in-differences exercise. The first difference is between aggregate outcomes from the benchmark estimated model and from setting speaking occupations across space to their prevalence in 1987. To do so, I proxy for the prevalence of speaking occupations using employment shares and set the share parameters of the CES function, $\phi_{r,spk}$, to levels in 1987. The second difference is between this change occurring with language barriers present and one without.

Let $\mathbb{Y}$ denote the outcomes: internal migration, skill premium, and welfare. I compute these outcomes in four scenarios, which are summarized in Table 10. First, $\mathbb{Y}_{2011,\ barriers}$ comes directly from the estimated model that matches the 2011 data, where language barriers are present and speaking occupations are distributed across space as observed. Second, $\mathbb{Y}_{1987,\ barriers}$ is computed from a counterfactual where I keep all parameters from the estimated model but change the spatial distribution of speaking occupations ($\phi_{r,spk}$) to match their observed 1987 levels. This simulates how the economy would look with current language barriers but the pre-liberalization occupational structure. Third, $\mathbb{Y}_{2011,\ no\ barriers}$ is computed from a counterfactual where I remove language barriers (setting $\tau^L o, d = 0$ and $A_{spk,b} = A_{spk,nb}$) but maintain the current spatial distribution of speaking occupations. Fourth, $\mathbb{Y}_{1987,\ no\ barriers}$ combines both changes: I remove language barriers and set speaking occupations to their 1987 distribution.

Table 10: Model-Based Difference-in-Differences

	Language Barriers ✓	Language Barriers ×
Speaking Occupations in 2011 ($\phi_{r,spk,2011}$)	$Y_{2011,\ est.\ model}$	$Y_{2011,\ cf}$
Speaking Occupations in 1987 ($\phi_{r,spk,1987}$)	$Y_{1987,\ cf}$	$Y_{1987,\ cf}$

Notes: This table outlines the two margins of difference in this exercise. The first is between 1987 and 2011, which are periods before and after trade liberalization in India. The CES share parameter in the outer nest is proxied from NSS data by employment shares in speaking and non-speaking occupations across states of India. The second margin of difference simulates this difference after removing language barriers.

The impact of language barriers is given by,

$$\mathbb{Y}_{DID} = (\mathbb{Y}_{2011,\ barriers} - \mathbb{Y}_{1987,\ barriers}) - (\mathbb{Y}_{2011,\ no\ barriers} - \mathbb{Y}_{1987,\ no\ barriers}),$$

where the first difference captures how outcomes changed between 1987 and 2011 with language barriers present, and the second difference captures how outcomes would have changed over the same period if language barriers were absent. This difference-in-differences reveals how language barriers affected the economy's response to the rising prevalence of speaking occupations. I find that language barriers substantially impacted the economy's transition: they decreased internal migration by 7.6 percentage points, increased skill premium by 3.4 percentage points, and decreased welfare by 1.9 percent. This is summarized in Table 11.

Table 11: Effects of Language Barriers on Gains from Structural Change

	Internal Migration	Skill Premium	Welfare
Y_{DID}	-7.6 p.p.	+3.4 p.p.	-1.9 percent

Notes: Internal migration is defined as the share of the population that migrated outside their origin state. Skill premium is defined as the percent difference in the real income of skilled and unskilled workers. Welfare is defined as the real income in the economy.

In this exercise, the comparison is between the estimated model and a world without language barriers, but where speaking occupations increased in the same way across space between 1987 and 2011. The latter is not likely to have occurred as firms and workers probably made their location and human capital acquisition choices, respectively, to access opportunities from globalization. As Shastry (2012) shows, some Indian districts have a more elastic supply of English language human capital, which is particularly relevant for service exports. These districts attracted more export oriented skilled jobs, specifically in information technology, and experienced greater growth in schooling.

Though the parallel trends assumption in this difference-in-differences exercise is possibly violated, had language barriers not existed, firms are likely to have sorted across space according to their comparative advantage. In other words, the effects I find are likely lower bounds on what the impact of language barriers may be if we allow spatial sorting of firms and endogenous human capital acquisition for workers. In any case, this exercise shows that language barriers became more salient for labor market success when structural change and growth in services led to a higher prevalence of speaking occupations.

6.3 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.

In a similar vein, I introduce a simple extension to the model whereby non-college migrant workers 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 for migrant workers, $c_{o,d}$, 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 ad valorem 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, language barrier becomes a function of whether or not they choose to enroll i.e., $\mathscr{L}(o,d,\mathbb{1}\{\text{enrol}\})$. In particular, $\mathscr{L}(o,d,\mathbb{1}\{\text{enrol}\}) = b$ if $\mathbb{1}\{\text{enrol}\} = 0$ and $\mathscr{L}(o,d,\mathbb{1}\{\text{enrol}\}) = nb$ if $\mathbb{1}\{\text{enrol}\} = 1$. That is, language barriers remain for the worker if they choose not to enrol, 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,

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

where $\Phi_{o,nc} = \sum_{(d',j',\mathbb{1}\{\text{enrol}\})} A_{j',nc,\mathscr{L}(o,d',\mathbb{1}\{\text{enrol}\})}\ (\Lambda_{o,d',j',nc,\mathbb{1}\{\text{enrol}\}})^{\theta}$. The numerator captures that workers that would have previously had barrier productivity, $A_{j,nc,b}$ and incurred a language migration cost ($\tau^{L}_{o,d,nc} > 0$) now have non-barrier productivity, $A_{j,nc,nb}$, and do not incur a language migration cost ($\tau^{L}_{o,d,nc} = 0$). They do, however, incur the opportunity cost of learning, $c_{o,d}$.

However, for other migrants, it is optimal 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 overcomes 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

$$\pi_{o,d,j,nc}(\mathbb{1}\{\text{enrol}\} = 0) = \frac{A_{j,nc,b}\left[(1-\tau^{G}_{o,d})(1-\tau^{L}_{o,d,e})\ \alpha_d\ (w_{d,j,e,b})\right]^{\theta}}{\Phi_{o,nc}},$$

where $\Phi_{o,nc} = \sum_{(d',j',\mathbb{1}\{\text{enrol}\})} A_{j',nc,\mathscr{L}(o,d',\mathbb{1}\{\text{enrol}\})}\ (\Lambda_{o,d',j',nc,\mathbb{1}\{\text{enrol}\}})^{\theta}$. 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, $c_{o,d}$.

To incorporate the cost of implementing language programs, I assume that the government finances the program by charging a uniform ad valorem tax to all workers in the economy. The tax is charged to every worker i in the economy,

$$\text{Language Tax}^{i} = \frac{\sum_{r\in R} \text{Cost of Language Programs}_{r}}{\sum_{r\in R}\sum_{e\in\{c,nc\}} N_{r,e}},$$

where the the numerator is the total cost of implementing language programs across states of India and the denominator is the total number of workers in the economy. Since the tax is equally borne by each worker in the economy, it not affect bilateral migration shares. However, it does discount aggregate welfare. In particular, I subtract the total cost of language programs from the aggregate nominal income in the economy, which is divided by the aggregate price index, P, to obtain Y^{Policy}.

Calibration of Costs: In this section, I describe how I calibrate costs in the model. First, I calibrate $c_{o,d}$ using information on the time taken to learn languages and wages foregone as a consequence. 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,

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

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,

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

which measures the average number of hours that a person at origin o needs to overcome language barriers to destination d. Finally, I multiply this by the average wage per hour[8] of workers to obtain the opportunity cost of learning language i.e., $c_{o,d} = \text{Time Cost}_{o,d} \cdot \bar{w}_{o,d}$.

Second, I calibrate the cost of language programs using budget and expenditure reports from the Center Institute of Indian Languages (CIIL), a subsidiary of the Ministry of Education, Government of India. This was setup in 1970 to promote linguistic harmony by teaching 20 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 non-recurring infrastructure costs and recurring costs of teacher training and curriculum design, salaries for teachers, admin, and support staff, and maintenance. Each language center trains approximately 66 students per year and, on average, students need 1200 hours to learn a language. Assuming that capital depreciates in 10 years, I distribute the non-recurring infrastructure costs across 10 years' worth of students. Thus, I compute the cost per student to be \$3.6 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. This appears reasonable, as private lessons in English are likely to be more expensive than language programs. This is because English teachers not as widely available as teachers in local languages and private tutors seek to make profit, whereas language programs instituted by the government seeks only to cover costs.

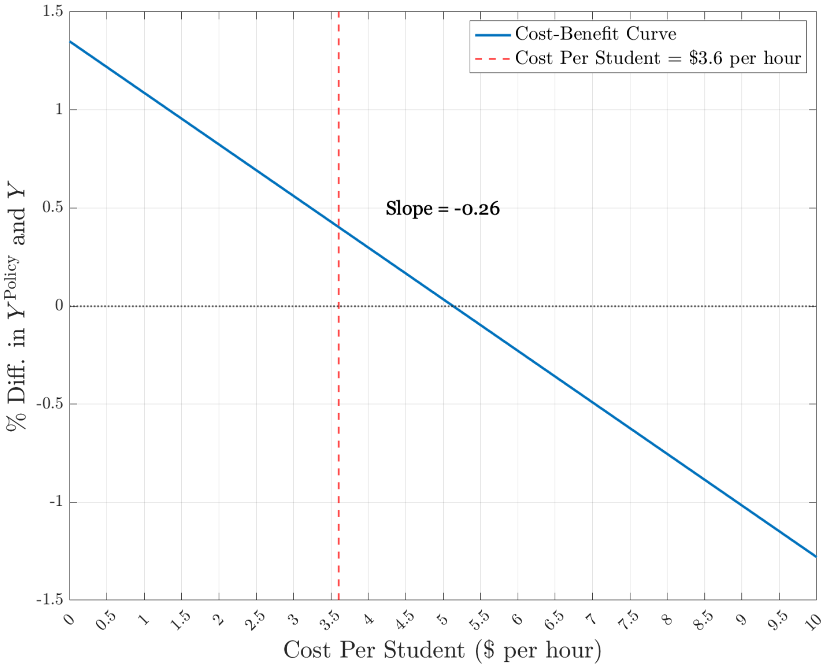
Cost vs. Benefit of Language Programs: To evaluate how the cost of language programs weigh against the benefit, I consider a range of values of the language program to meet the demand computed from the calibrated model. Using the model, I predict how many workers would incur $c_{o,d}$ and enroll in the program at each region. I take a weighted sum of cost per student to account for how many hours workers from different origins would need to learn language of a given destination ($\text{Time Cost}_{o,d}$).

For each value of the cost of the program, I compute aggregate welfare, Y^{Policy}. To do

[8] I compute this from data on monthly wages, which is available in the CPHS waves, and data on the average number of hours worked by occupation, which is available in the NSS.

so, I add up the costs in each region and subtract the total cost of language programs from the aggregate income in the economy. The cost of language programs is borne equally by all workers in the economy through a uniform ad valorem tax. I define benefit from the program as the percent gap between Y^{Policy} and the welfare in the baseline estimated model without language policy, Y.

Figure 3: 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 in Y^{Policy}, which is welfare under the policy (computed by subtracting the aggregate real income by the cost of implementing language programs), and Y, which is welfare under the estimated model.

The negative relationship in Figure 3 illustrates how welfare gains from language programs decline as implementation costs rise. The point at which the difference goes below 0 can be considered the threshold cost of the program, which is at $5.2 per hour of instruction. Above this threshold, costs outweigh the benefits, and it is not worth implementing such a policy. My calculation of $3.6 per hour, derived from budget reports of India's Central Institute of Indian Languages (CIIL), falls substantially below this break-even threshold. At this cost of $3.6 per person per hour, the policy generates a welfare gain of 0.4 percent.

The gap between actual and break-even costs provides a significant buffer against potential implementation challenges or cost overruns. Moreover, this analysis likely understates the true benefits as it does not account for potential positive externalities from increased linguistic integration or long-term gains from enhanced labor market mobility. So, I argue that language policy should indeed be implemented.

7 Conclusion

This paper provides the first structural analysis of language barriers as both a spatial and labor market friction, quantifying their aggregate and distributional effects in general equilibrium. By leveraging India's linguistic diversity and its economic transformation following the 1991 trade liberalization, this research uncovers how language shapes internal migration, skill premium, and aggregate welfare.

The empirical analysis reveals several key patterns about language, migration, and labor markets. Workers systematically avoid migrating to locations with high language barriers. When they do migrate, those facing language barriers tend to sort away from speaking-intensive occupations. These migrants receive wage premiums within their chosen occupations, suggesting only those expecting higher wages choose to move despite language barriers. Notably, these effects are less pronounced for skilled workers, likely due to their English proficiency through college education.

The quantitative spatial general equilibrium model developed in this paper captures the interaction between sorting and selection mechanisms in general equilibrium. By incorporating language as both a component of migration cost and a technological friction, the model reveals how language barriers affect workers' migration decisions, occupational choices, and productivity. The estimated model shows high degrees of complementarities between unskilled workers with and without language barriers in speaking occupations, while finding that unskilled workers facing language barriers are nearly twice as productive in non-speaking than speaking occupations.

The counterfactual analyses yield several important findings. First, removing language barriers would increase internal migration by 6.2 percentage points, enhance welfare by 1.2 percent, and reduce inequality by 1.9 percentage points—effects equivalent to reducing geographic barriers by 56 percent or increasing college worker share by 34 percent. Second, in the context of India's structural transformation toward services, language barriers significantly impeded gains from growing consumer service employment. Without these barriers, the increased prevalence of speaking occupations would have led to 7.2 percentage points higher internal migration, 1.9 percent higher welfare, and 3.4 percentage points lower inequality. Third, the analysis of language program implementation suggests that such interventions could be cost-effective policy tools for mitigating these barriers.

These findings have important implications for policy design in multilingual countries experiencing structural transformation. As economies shift toward service-oriented sectors with speaking-intensive occupations, the economic costs of language barriers become increasingly significant. The results suggest that targeted language programs for unskilled migrant workers could generate substantial welfare gains that outweigh implementation costs.

Several promising directions for future research emerge from this work. First, investigating the role of language in educational choices and outcomes could provide crucial insights into intergenerational mobility. Specifically, research could examine how the medium of instruction in schools affects students' educational trajectories and college choices across regions. This analysis would be particularly relevant in India, where language policies in education vary substantially across states.

Second, future work could explore the dynamic aspects of language acquisition and its long-term implications for human capital accumulation. This might include studying how expectations about future structural change influence current investments in language skills, particularly among young workers. Understanding these dynamic mechanisms could inform the optimal timing and targeting of language programs. Such research could also examine how parents' language choices affect their children's economic mobility across generations. This intergenerational perspective is particularly relevant in multilingual countries where language skills can serve as a gateway to better economic opportunities in growing sectors.

Third, research could investigate the interaction between language barriers and other forms of social and economic friction, such as caste networks or gender norms. This could help explain why some groups might be more affected by language barriers than others. For instance, access to social networks in destination regions might differentially affect how workers overcome language barriers, with some caste groups having stronger support systems that facilitate language learning. Gender norms could also intersect with language barriers in important ways, as women might face additional constraints in accessing language learning opportunities or participating in speaking-intensive occupations. Understanding these intersecting barriers would be particularly valuable for designing targeted interventions that account for the complex social fabric of developing economies.

As developing economies continue to urbanize and shift toward service-oriented sectors, understanding and addressing language barriers becomes increasingly crucial for promoting inclusive growth. This paper's findings suggest that language policy should be considered an integral component of development strategy, particularly in diverse, multilingual societies undergoing structural transformation. The magnitude of welfare gains from removing language barriers—comparable to major investments in transportation infrastructure or higher education—underscores their centrality to economic development. As the global economy becomes increasingly integrated and communication-intensive, the economic costs of language barriers are likely to grow even more significant, making their resolution an urgent priority for development policy.

References

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

Allen, Treb, Cauê de Castro Dobbin, and Melanie Morten, "Border Walls," Working Paper 25267, National Bureau of Economic Research November 2018.

Altonji, Joseph G. and David Card, "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives," in John M. Abowd and Richard B. Freeman, eds., *Immigration, Trade, and the Labor Market*, University of Chicago Press, 1991, pp. 201–234.

Bratsberg, Bernt, Andreas Moxnes, Oddbjørn Raaum, and Karen Helene Ulltveit-Moe, "Opening the Floodgates: Partial and General Equilibrium Adjustments to Labor Immigration," *International Economic Review*, 2023, *64* (1), 3–21.

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

Burstein, Ariel, Eduardo Morales, and Jonathan Vogel, "Changes in Between-Group Inequality," *American Economic Journal: Macroeconomics*, 2019, *11* (2), 348–400.

_ , Gordon Hanson, Lin Tian, and Jonathan Vogel, "Tradability and the Labor-Market Impact of Immigration: Theory and Evidence From the United States," *Econometrica*, 2020, *88* (3), 1071–1112.

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

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

Census of India, "Migrants by Place of Last Residence, Duration of Residence and Reason for Migration, Census 2001 India and States," Government of India, Office of the Registrar General & Census Commissioner 2001.

_ , "Migration, Census 2011," Government of India, Office of the Registrar General & Census Commissioner 2011.

Chiswick, Barry R. and Paul W. Miller, "International Migration and the Economics of Language," in "Handbook of the Economics of International Migration," Elsevier, 2015, pp. 211–269.

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

Desai, Sonalde, Reeve Vanneman, and National Council of Applied Economic Research, "India Human Development Survey-II (IHDS-II), 2011-12," https://doi.org/10.3886/ICPSR36151.v6 Aug 2018. [distributor].

Dustmann, Christian and Francesca Fabbri, "Language Proficiency and Labour Market Performance of Immigrants in the UK," *The Economic Journal*, 2003.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, *Ethnologue: Languages of the World*, twenty-seventh ed., Dallas, Texas: SIL International, 2024. Online version.

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

Fan, Tianyu, Michael Peters, and Fabrizio Zilibotti, "Growing Like India—the Unequal Effects of Service-Led Growth," *Econometrica*, 2023, *91* (4), 1457–1494.

Fearon, James D, "Ethnic and cultural diversity by country," *Journal of economic growth*, 2003, *8*, 195–222.

Ghose, Devaki, "Trade, Internal Migration, and Human Capital: Who Gains from India's IT Boom?," 2024.

Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow, "The Allocation of Talent and U.S. Economic Growth," *Econometrica*, 2019, *87* (5), 1439–1474.

Imbert, Clément and John Papp, "Costs and benefits of rural-urban migration: Evidence from India," *Journal of Development Economics*, 2020, *146*, 102473.

Katz, Lawrence F. and Kevin M. Murphy, "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," 1992.

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

Khanna, Gaurav and Nicolas Morales, "The IT Boom and Other Unintended Consequences of Chasing the American Dream," 2023.

Kone, Zovanga L., Maggie Y. Liu, Aaditya Mattoo, Caglar Ozden, and Siddharth Sharma, "Internal borders and migration in India," *Journal of Economic Geography*, 2018, *18* (4), 729–759.

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

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

Munshi, Kaivan, "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market," *The Quarterly Journal of Economics*, 2003, *118* (2), 549–599.

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

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

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

Wang, Yao, "Linguistic Distance, Internal Migration and Welfare: Evidence from Indonesia," 2024. SSRN Working Paper.

Appendix: Model Derivations

This appendix presents the derivation of 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 Destination	$o \in R$ $d \in R$
Occupations ($j \in J$)	$j \in$ {Speaking, Non-Speaking}	$j \in$ {spk, non-spk}

Workers (i)	Education $\in$ {College, Non-College}
	Language Barrier $\in$ {Barrier, Non-Barrier}
		$e \in \{c, nc\}$
		$\mathscr{L}(o,d) \in b, nb$

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

Parameter	Description	Dimensions
$\tau^G_{o,d}$	geographic barriers	$R \times R$
$\tau^L_{o,d,e}$	language barriers	$R \times R \times 2$
α_r	amenities	$R \times 1$
$A_{j,e,\mathscr{L}(o,d)}$	Fréchet, location	$J \times 2 \times 2$
θ	Fréchet, dispersion	1×1
σ	elasticity parameter, aggregate nest	1×1
κ	elasticity parameter, outer nest	1×1
$\phi_{r,j}$	share parameter, outer nest	$R \times J$
ρ	elasticity parameter, middle nest	1×1
$\nu_{j,e}$	elasticity parameter, inner nest	$J \times 1$

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,

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

Expected Productivity and Selection: The key selection equation is derived from the expected productivity of worker i of 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,

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

$$
\begin{aligned}
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} \int_{z'=0}^{\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z} \; z \, g_{j,e,\mathscr{L}(o,d)}(z) \; g_{j',e,\mathscr{L}(o,d')}(z') \; dz \, dz' \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} \; z \, g_{j,e,\mathscr{L}(o,d)}(z) \; \Big(\int_{z'=0}^{\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z} \; g_{j',e,\mathscr{L}(o,d')}(z') \; dz' \Big) \; dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z \, g_{j,e,\mathscr{L}(o,d)}(z) \; g_{j',e,\mathscr{L}(o,d')} \Big(\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z \Big) \; dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z \theta A_{j,e,\mathscr{L}(o,d)} z^{-\theta-1} \exp \Big(A_{j,e,\mathscr{L}(o,d)} z^{-\theta} \Big) \; \exp \Big(A_{j',e,\mathscr{L}(o,d')} (\frac{\Lambda_{o,d,j,e}}{\Lambda_{o,d',j',e}} z)^{-\theta} \Big) \; dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z \theta A_{j,e,\mathscr{L}(o,d)} z^{-\theta-1} \exp \Big(- (\frac{\Phi^{oe}}{\Lambda_{o,d,j,e}})^{\theta} z^{-\theta} \Big) \; dz \\
&= \frac{1}{\pi_{o,d,j,e}} \int_{z=0}^{\infty} z \theta A_{j,e,\mathscr{L}(o,d)} z^{-\theta-1} \exp \Big(- \frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta} \Big) \; dz \\
&= \int_{z=0}^{\infty} z \theta \frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta-1} \exp \Big(- \frac{A_{j,e,\mathscr{L}(o,d)}}{\pi_{o,d,j,e}} z^{-\theta} \Big) \; dz \\
&= \Gamma(1 - 1/\theta) (\pi_{o,d,j,e} / A_{j,e,\mathscr{L}(o,d)})^{-1/\theta}
\end{aligned}
$$

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,

$$
\begin{aligned}
E\Big[U^{i}_{o,d,j,e}\Big] &= \; E\Big[\max_{(d,j)} \; \Lambda_{o,d,j,e} \; z\Big] \\
&= \int_{0}^{\infty} \Big[1 - \prod_{(d,j)} \Pr \Big(\Lambda_{o,d,j,e} \; z \leq t \Big) \Big] dt \\
&= \int_{0}^{\infty} \Big[1 - \prod_{(d,j)} \Pr \Big(z \leq \frac{t}{\Lambda_{o,d,j,e}} \Big) \Big] dt \\
&= \int_{0}^{\infty} \Big[1 - \prod_{(d,j)} \exp \Big(- A_{j,e,\mathscr{L}(o,d)} (\frac{t}{\Lambda_{o,d,j,e}})^{-\theta} \Big) \Big] dt \\
&= \int_{0}^{\infty} \Big[1 - \exp \Big(- \sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\frac{t}{\Lambda_{o,d,j,e}})^{-\theta} \Big) \Big] dt \\
&= \int_{0}^{\infty} \Big[1 - \exp \Big(- t^{-\theta} \sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta} \Big) \Big] dt \\
&= \int_{0}^{\infty} \Big[1 - \exp \Big(- \underbrace{\sum_{(d,j)} A_{j,e,\mathscr{L}(o,d)} (\Lambda_{o,d,j,e})^{\theta} t^{-\theta}}_{\equiv (\Phi_{o,e})^{\theta}} \Big) \Big] dt
\end{aligned}
$$

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

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,

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

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

$$w_{d,j,e,\mathscr{L}(o,d)} = MP_{d,j} \cdot MP_j^{d,e} \cdot (\ell_j^{d,e,\mathscr{L}(o,d)})^{\frac{\nu_{j,e}}{\nu_{j,e}-1}},$$

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

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

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

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

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

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

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

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,

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

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

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

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

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

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,

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

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

which allows us to solve for the cost of Y,

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

Thus, the price equals the marginal cost function,

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

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

Estimation of θ: I infer the migration elasticity, θ, from the relationship between cross-migration flows and migration costs. In particular, I derive the Head-Reis index,

$$\begin{aligned}
\frac{N_{o,d,j,e}}{N_{d,d,j,e}} \frac{N_{d,o,j,e}}{N_{o,o,j,e}} &= \frac{\pi_{o,d,j,e}}{\pi_{d,d,j,e}} \frac{N_{o,e}}{N_{d,e}} \frac{\pi_{d,o,j,e}}{\pi_{o,o,j,e}} \frac{N_{d,e}}{N_{o,e}} \\
&= \Big(\frac{\Lambda_{o,d,j,e}}{\Lambda_{d,d,j,e}} \frac{\Lambda_{d,o,j,e}}{\Lambda_{o,o,j,e}}\Big)^{\theta} \\
&= \Big[\frac{(1-\tau^G_{o,d})(1-\tau^L_{o,d,e})}{(1-\tau^L_{d,d})(1-\tau^L_{d,d,e})} \frac{(1-\tau^G_{d,o})(1-\tau^L_{d,o,e})}{(1-\tau^G_{o,o})(1-\tau^L_{o,o,e})}\Big]^{\theta} \\
&= \Big[(1-\tau^G_{o,d})(1-\tau^L_{o,d,e})\Big]^{\theta},
\end{aligned}$$

where the final step follows from the symmetry in geographic and language barriers, $\tau^G_{o,d} = \tau^G_{d,o}$, $\tau^L_{o,d,e} = \tau^L_{d,o,e}$, and from the absence of barriers for workers that do not migrate, $\tau^G_{d,d} = \tau^L_{d,d,e} = \tau^G_{o,o} = \tau_{o,o,e} = 0$. For identification of θ, the specification allows for a constant and an error term that is assumed to be uncorrelated with migration costs in expectation. That is, I estimate θ from the following,

$$\ln\left(\sqrt{\frac{N_{o,d,j,e,t}}{N_{d,d,j,e,t}}\frac{N_{d,o,j,e,t}}{N_{o,o,j,e,t}}}\right) = \text{Constant} + \theta \ln\left[(1-\tau^{G}_{o,d})(1-\tau^{L}_{o,d,e})\right] + \varepsilon_{o,d,j,e,t},$$

where the source of identifying variation is over state-pairs, occupations, and time periods. Table 6 contains the regression results.