CAGE

Working Paper

774/2025
September 2025

A Century of Language Barriers to Migration in India

Latika Chaudhary,
Yannick Dupraz,
and James Fenske

ISSN: 2978-0276
Grant number: ES/7504701/1

UNIVERSITY OF WARWICK

UKRI Economic and Social Research Council

A CENTURY OF LANGUAGE BARRIERS TO MIGRATION IN INDIA

LATIKA CHAUDHARY*, YANNICK DUPRAZ†, AND JAMES FENSKE‡

Abstract. Combining detailed data on language and migration across colonial Indian districts in 1901 with a gravity model, we find origin and destination districts separated by more dissimilar languages saw less migration. We control for the physical distance between origin-destination pairs, several measures of dissimilarity in geographic characteristics, as well as origin and destination fixed effects. The results are robust to a regression discontinuity design that exploits spatial boundaries across language groups. We also find linguistic differences predict lower migration in 2001. Cultural channels are a small part of the link from linguistic diversity to lower migration. Rather, the evidence suggests communication and information channels are more important.

Keywords: Migration, Linguistic Diversity, India.
JEL Codes: N35, O15, Z13

*Associate Professor, Department of Defense Management, Naval Postgraduate School
†Permanent Researcher, Paris Dauphine University, PSL University, LEDA, CNRS, IRD
‡Professor, Department of Economics, University of Warwick
E-mail addresses: lhartman@nps.edu, yannick.dupraz@cnrs.fr, J.Fenske@warwick.ac.uk.
Date: September 29, 2025.
We are grateful to audiences at the AMSE Eco-lunch, the DIE informal meeting, the University of Warwick, the ASREC Europe 2024 Conference, the University of Paris Dauphine, the Economic History Society Conference 2025, the 2025 NICEP Conference, the Paris School of Economics, and the Saint-Etienne GATE Development Economics Workshop for comments. We also thank the Centre for Competitive Advantage in the Global Economy and the Leverhulme Trust for funding. We thank Anna Zhukova for formulating the problem of constructing a panel of consistent geographical units over time as a graph theory problem.

1. Introduction

Although migration confers large economic benefits on migrants and their families, migration rates vary significantly across space and time (Abramitzky et al., 2021; Chetty et al., 2016; Lagakos et al., 2023). Low rates of migration are especially striking on account of the large wage gaps between rich and poor countries, and between rich and poor parts of the same country. While linguistic divisions have been proposed as one explanation for this puzzle (Belot and Ederveen, 2012), most of the evidence evaluates international migration (Adsera and Pytlikova, 2015; Chiswick and Miller, 2015), or migration between regions of developed countries (Falck et al., 2012). We know less about the effect of language on migration within developing countries. Moreover, we know even less about the evolution of cultural barriers to migration over time as mass education and economic development have increased in these countries.

Our paper fills this gap by studying the relationship between language and migration in colonial and contemporary India in 1901 and 2001. Understanding the language spoken in a place likely increases the economic returns to migration, while reducing the cultural costs of migration. Yet the literature on language has made slow progress because of insufficient data and low linguistic diversity within countries. India offers an ideal research context on both fronts. Linguistic diversity is high, with more than 100 major languages. Moreover, the colonial and contemporary censuses enumerate different language speakers and migrants at the district level.

Apart from linguistic diversity, India is also characterized by low internal migration. Colonial administrators often commented on this low migration, with more than 90% of individuals recorded in the 1881 Indian census living in their birth districts (Collins, 1999). While internal migration has increased since then, especially in the past decade, the rate is still lower than other countries. Using a consistent definition of internal migration, Bell et al. (2015) find a migration intensity of 5% in India during the 2000-2010 decade. By this measure, India has the lowest migration intensity in their sample of 61 countries, well below the median of 18%.[1] Can linguistic diversity explain the low rate of internal migration in colonial India? If yes, does this relationship persist to the contemporary period in the presence of higher incomes, more schooling and the emergence of Hindi and English as pan-Indian languages?

We answer these questions using a new dataset of district-pair dyads of origin (o) and destination (d) districts in both 1901 and 2001. We measure migration as the stock of migrants born in o and residing in d. Following Esteban et al. (2012), we use the language

[1]There is some debate on the extent of internal migration within India in recent years with sources using railway passenger flows estimating higher migration flows than those using census data (Government of India, 2017). But even accounting for the higher estimates in Government of India (2017) would not change India's position as a low internal migration country.

trees recorded in Ethnologue to compute linguistic distance between languages. Our measure of linguistic distance captures the linguistic distance between the majority languages spoken in o and d. Using the majority language, rather than the full distribution of languages spoken in a district, avoids the problem of reverse causality whereby migration from o would directly increase the number of people in d speaking the languages of o. As migration is low and majority language speakers typically account for a large share of the district population, migration from o is unlikely to change the majority language spoken in d.

Armed with this dataset, we estimate a gravity model of migration between o and d as a function of the linguistic distance between the majority languages of o and d, the physical distance between o and d, an indicator for whether o and d are neighboring districts, indicators for whether o and d are in the same province, and a rich set of measures of their geographic dissimilarity. We also control for origin and destination fixed effects that capture unobservable characteristics of o and d that make them more or less attractive to migrants, and that may be correlated with their majority language. We find that a unit increase in linguistic distance — equivalent to the distance between unrelated languages such as Bengali and Tamil that are part of different language families — predicts a 38% decline in migration compared to migration between districts with the same majority language.

Since these gravity estimates may be confounded by omitted variables, we implement a new regression discontinuity (RD) design exploiting discontinuous spatial changes in language families within India. In the RD specification, we compare outmigration from a single origin o to two destinations d_1 and d_2 that are adjacent to each other and at similar distance from o, but where d_1 is in the same language region as o while d_2 is not. This exercise involves, for example, comparing outmigration between Dharwar (a Kannada-speaking district) and Bijapur (another Kannada-speaking district) to outmigration between Dharwar and Sholapur. All three districts are within the same colonial province – the Bombay Presidency. In Sholapur the majority language is Marathi, an Indo-European language. Kannada in Dharwar is, by contrast, a Dravidian language. Considering the linguistic border between Indo-European and Dravidian languages, we find migration drops discontinuously at the Indo-European and Dravidian border.[2]

We find a similar negative relationship between linguistic distance and migration using the 2001 census. A unit increase in linguistic distance predicts a 39% decline in migration compared to migration between districts with the same majority language. Since the borders of post-colonial India are different from the colonial period, we construct a bilateral dataset based on consistent geographical units where we observe migration in both 1901 and 2001. This exercise enables us to compare the 1901 and 2001 estimates. We find they are statistically similar, suggesting the importance of language as a barrier to internal migration

[2]In a balance test, we show this border does not overlap with any geographical boundary. It also does not overlap with colonial administrative boundaries separating British provinces and Princely States.

in India has not declined over the twentieth century. We also find comparable RD results using the 2001 data to those in 1901. But we hesitate to draw strong conclusions from the 2001 RD exercise as Indian state borders were redrawn in 1956 along linguistic lines.

Our results are stable across multiple robustness checks. We find similar results in alternative samples of colonial *od* pairs and in the 1921 census. The results are robust to different measures of linguistic distance, different measures of migration, and alternate controls for physical distance. The relationship between linguistic distance and migration in 2001 is also robust. Moreover, we find comparable results using the less detailed 2011 census. Taken together, our results highlight the persistence of language barriers to internal migration.

Why does more linguistic distance between o and d predict lower migration? We first consider the role of cultural channels such as caste and religion that may be related to language differences. Controlling for caste distance does not reduce the coefficient on linguistic distance. The coefficient on linguistic distance is also robust to controlling for a measure of religious distance. Including an ethnographic measure of cultural distance does not affect the results for 1901. Using individual data from the 1980s, we find the relationship between linguistic distance and migration is robust to controlling for the existing stock of migrants. We consider split samples by sex; especially in the colonial period, men mostly migrated for work, while women migrated for marriage or as part of a family. Yet we find no differential coefficients by sex, either in colonial or contemporary India.

Rather, the evidence suggests that language barriers constrain migration by impeding communication and information flows, especially those related to labor market opportunities. For example, potential migrants are likely to consider information on destination wages in their decision on where to migrate and they are more likely to have this information when they speak the destination language. After moving, migrants that speak the destination language may also find it easier to find and retain jobs. We find the predictive power of linguistic distance is greatest for *od* pairs where origin wages are lower than destination wages, suggesting language differences are more salient where economic returns to migration are higher. We also find migration between *od* pairs with mutually intelligible languages from the same dialectal chain is the same as for *od* pars with the same majority language. Exploiting rich data on subsidiary languages in 2001, we find distance between languages spoken, including second and third languages, is a better predictor of migration than the distance between mother tongues. The importance of second and third languages highlights communication is more important than mother tongue differences, which are likely a function of family and culture.

If language barriers constrain migration on account of communication, do education and official language policies mitigate the effects of linguistic distance? We do not find that to be the case. There are no differential effects of linguistic distance for high literacy origins

in 1901 or for more educated individuals in the 1980s. We also find no evidence that official languages mitigate language barriers: controlling for whether two districts share an official state language does not decrease the coefficient on linguistic distance.

1.1. **Contributions.** Our paper contributes to three literatures. First, we contribute to the literature on cultural barriers to migration. Cultural differences can both lower the returns to migration and reduce awareness of these returns. For example, differences in food cultures have been linked to lower caloric intake among migrants (Atkin, 2013). Culturally distant migrants are often not welcomed by native populations (Alesina and Tabellini, 2024). Regarding language, Adsera and Pytlikova (2015) and Chiswick and Miller (2015) find language is a barrier to international migration, Falck et al. (2012) find negative effects of German dialects on internal migration, and Wang (2025) finds an inverted U-shaped relationship between ethnolinguistic distance and internal migration in Indonesia. Our paper accords with Manjunath (2024), who finds contemporary Indian workers in 2001 and 2011 are less likely to migrate to places where they face high language barriers.

Unlike these studies that focus on recent years, we study the evolution of language barriers to migration between 1901 and 2001. Over this century, colonial India was partitioned and became independent, real GDP per capita tripled, education increased, and pan-Indian languages like Hindi and English became more important.[3] And yet language barriers to migration have endured. We also offer an important methodological contribution by exploiting a spatial RD along the boundaries of linguistic regions. Such a design offers a new credible strategy to identify the effect of language differences.

Second, we contribute to the literature on linguistic diversity and economic development, which finds that linguistic divisions, such as those that separate ethnic groups, predict worse development outcomes. Using cross-country variation, early work in this field finds countries with greater ethnolinguistic fractionalization experience lower rates of economic growth (Easterly and Levine, 1997). Such effects are especially severe when there are cleavages separating entire language families (Desmet et al., 2016). Subsequent work has — using cross-country and disaggregated data — identified mediating factors such as public goods provision, including schooling and infrastructure, financial systems, foreign exchange markets, political instability, conflict, efficiency of task allocation, effort provision, and government deficits (Alesina and La Ferrara, 2005; Bazzi and Gudgeon, 2021; Desmet et al., 2020; Dickens, 2018a; Lyons, 2017; Marx et al., 2021; Miguel and Gugerty, 2005).[4] Our paper

[3]These comparisons use GDP per capita from Maddison Project Database (Bolt and Van Zanden, 2024).
[4]Several moderating factors may mitigate these effects, including inter-group contact (Bazzi et al., 2019), in particular collaborative forms of contact (Lowe, 2021), training in perspective taking (Alan et al., 2021), and nation-building policies such as a national language (Blanc and Kubo, 2024; Carlitz et al., 2024; Miguel, 2004).

suggests linguistic diversity may lower growth by reducing internal migration, which leads to less efficient labor allocation within countries.

One branch of this literature has focused specifically on Indian linguistic diversity. Linguistic mismatch between the majority language of a district and the official language of the state reduces literacy and college graduation rates (Jain, 2017), while the growth in industrial employment between 1931 and 1961 increased bilingual speakers (Clingingsmith, 2014). Fenske and Kala (2021) find a negative correlation between linguistic distance and price integration across market pairs in colonial India. Laitin and Ramachandran (2016) show adverse economic and human capital outcomes for individuals more distant from the official state language, while Gupta et al. (2024) find that language barriers hamper the adoption of agricultural technologies. We document a robust relationship between language divisions and migration — an important development outcome that captures the integration of labor markets across space.

Finally, we contribute to the large literature on internal migration in South Asia (Bryan et al., 2014; Bryan and Morten, 2019; Tumbe, 2025).[5] While several studies offer different explanations for India's low rate of internal migration, no single explanation dominates in the literature. In the past, British officials argued that Indian "preferences" for village life constrained migration. While Hindu scriptures discourage travel across the foreign seas and transgressors may lose their caste privileges (Bates and Carter, 2021), such strictures generally apply to foreign travel, not movement within India.[6] Using individual migrant data, Imbert and Papp (2020) point to the high non-monetary costs of living and working in cities, Munshi and Rosenzweig (2016) highlight rural insurance through caste networks, Lagakos et al. (2023) find a role for incomplete information coupled with high risk, and Kone et al. (2018) emphasize state entitlement and employment programs. Much of this research exploits data on individual migrants in contemporary South Asia and identifies the effect of different push and pull factors on migration. It also tends to study rural to urban migration, albeit with exceptions (Kone et al., 2018). Unlike these studies on individual migrants, we use census data across Indian districts that are more suitable for identifying spatial barriers to inter-district migration. As far as we know, we are among the first to estimate the relationship between language differences and internal migration across two time periods separated by a century.[7]

[5]Tumbe (2012) is a rich narrative history of Indian migration including indentured labor that were sent on contracts to different parts of the British Empire in the colonial period.

[6]This relates to the Hindu taboo of crossing the ocean, i.e., *kala pani*.

[7]A recent Government of India (2017) report evaluates the importance of Hindi for internal migration in the 2010s. Government of India (2017) regresses railway passenger flows between 2011 and 2016 on an indicator for districts in Hindi-speaking states (an indicator for ten states in central and northern India where Hindi is commonly spoken) and other variables, including origin and destination fixed effects. The Hindi indicator conflates the prevalence of Hindi with other differences between these states and other Indian states. Our paper differs by studying detailed language differences between districts, by looking at migration

The rest of the paper is organised as follows. Section 2 provides historical background on migration and language in India. Section 3 describes our data, and is followed by an outline of our estimation strategies in Section 4. We report the results and show their robustness in Section 5, followed by a discussion of potential channels underlying the results in Section 6. We conclude in Section 7.

2. Historical Background

2.1. **Migration.** In this section, we describe features of colonial migration that inform our analysis. In the 1881 Census, 95% of individuals were enumerated in their birth districts. This number decreased to 91% in the 1901 census and remained around 90% up to the 1931 census. Migration over large distances within and outside India was uncommon (Visaria and Visaria, 1983). This is not to say Indians never moved; rather, the share of migrants was small relative to the total population, with most Indians living in their birth districts.

Though the migration rate was low, migration took multiple forms. Some migration was casual, with people moving to a neighboring village or town, or a temporary move to attend pilgrimages and marriages. Temporary migrants were also involved in the construction of canals and railways (Kerr, 2006). Famines and plagues pushed people to move, but such moves were not always permanent (Gait, 1913). Other forms of migration included seasonal labor, in which individuals moved to meet demand during the harvesting and sowing seasons, and permanent migration, in which individuals left their ancestral villages for marriage or better economic opportunities (Tumbe, 2018). Many permanent migrants would return home for marriages and funerals. Few migrants completely cut ties from their ancestral village. This practice of maintaining connections continues today (Tumbe, 2012).

While migration from one village to a neighboring village or town in the same district would not usually be counted as migration in the census, it would be considered migration if the destination was in a different district. Indeed, two-third of individuals enumerated as born outside their birth district in 1901 were born in a neighboring district. This pattern underscores the need to control for neighboring districts in the analysis.

Out-migration was more common from rural districts with high population density and often low wages, while in-migration was concentrated in more sparsely populated areas and industrial centers (Tumbe, 2012). Migrants worked in mines (Simmons, 1976), in Bengal jute mills (Sen, 1999), in Bombay cotton mills (Gupta, 2011; Morris, 1965), in Assam tea plantations (Gupta and Swamy, 2017) and in the newly irrigated tracts in Punjab (Agnihotri, 1996). To the extent certain locations were more attractive to all migrants or prone to out-migration, we control for origin and destination fixed effects.

---

more broadly rather than railway traffic, and by exploiting variation in migration flows between districts even within states.

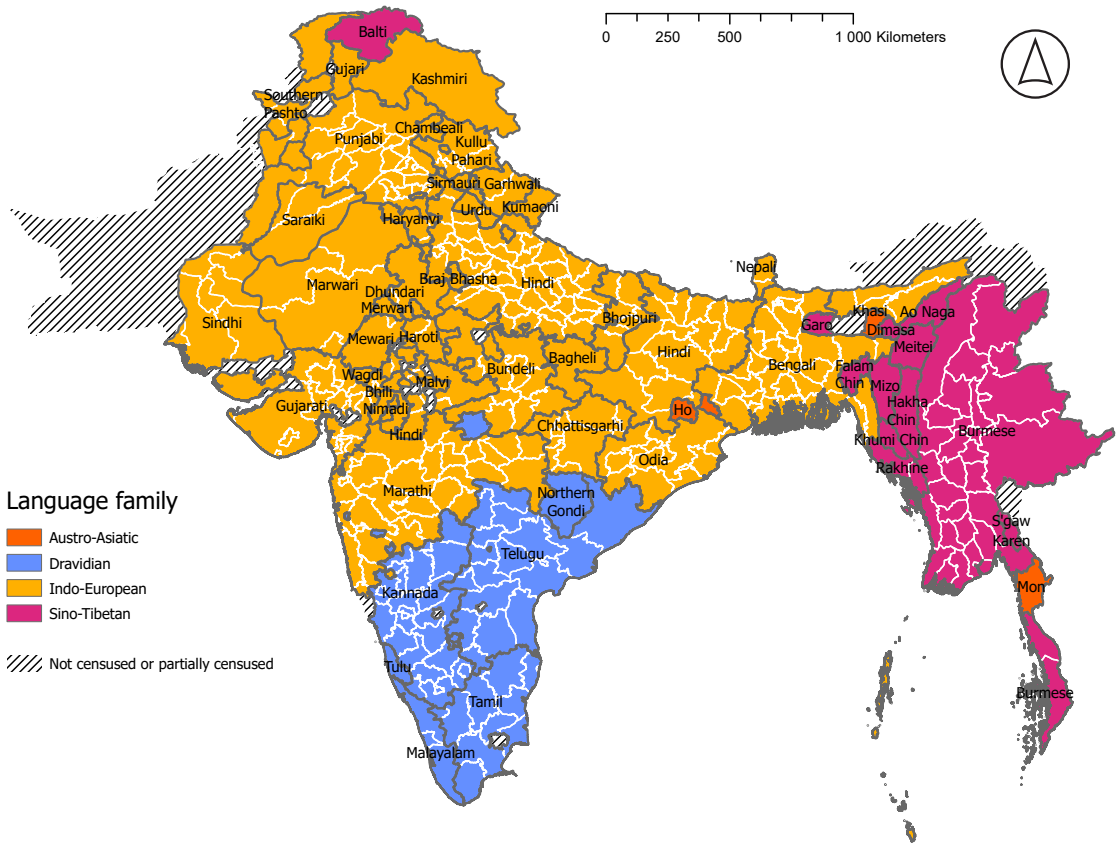
Apart from economic migration, women often moved for marriage, as village exogamy was common, especially in the North (Clark, 2000; Dyson and Moore, 1983; Miller, 1981). In contrast, men were more likely to move in response to economic conditions. Some male migrants moved with their families, but male migrants also moved to urban centers without them (Chaudhary and Fenske, 2023a). For these reasons, we estimate split samples by sex.

Internal migration remained low up to 1991. It began to increase in the early 1990s following reforms that liberalized the economy (Pandey, 2014). While contemporary migration has increased compared to the colonial period, India is not a high internal migration country even today (Bell et al., 2015; Kone et al., 2018). Rural to urban migration is more common now than it was in the early 20$^{\text{th}}$ century. In the past, temporary labor moved to support infrastructure projects like railways and canals. Today, migrant labor supports the construction of buildings, roads, and subways in many cities. Women continue to leave their ancestral villages at marriage, albeit at an older age due to an increase in age of marriage. Similar to the past, out-migrants often hail from states like Uttar Pradesh and Bihar with lower wages and high population densities (Pandey, 2014). Contemporary migration thus continues to share features with the past.

2.2. **Linguistic Diversity.** Linguistic diversity was high in colonial India. In 1901 there were 147 languages enumerated in the colonial census across four major language families (Risley and Gait, 1903, p. 248). The four language families are: Indo-European, which includes languages such as Hindi, Bengali and Punjabi that are spoken primarily in the North and North-West; Dravidian, encompassing languages such as Tamil, Telegu, and Malayalam that are spoken primarily in the South; Sino-Tibetan languages such as Balti that are spoken on the northern and eastern periphery, and; Austro-Asiatic languages such as Khasi and Munda. Figure 1 shows the distribution of the majority languages (by district) and their families in 1901. The probability of two randomly selected individuals speaking different languages was 90% in 1901. In 2001, the probability was 89% even though contemporary India is smaller than colonial India. Linguistic diversity has thus remained high over the 20$^{\text{th}}$ century.

Linguistic diversity is higher in the east, with fewer people speaking many different languages, compared to northern India where fewer languages are spoken by larger groups of people. While linguistic differences across regions are larger than within regions, there is remarkable diversity even within regions. Writing in 1901, the Census Commissioner described "tracts of country on the border-land between two languages, which are inhabited by both communities, living side by side and each speaking its own language" (Risley and Gait, 1903, p. 249). This observation points to linguistic distances even between neighboring areas. Such abrupt variation across space enables us to disentangle the separate influence

FIGURE 1. Majority Language by District in 1901



British district or Princely State borders are represented by a white line, while the grey lines represent the borders of language regions where a language region is a group of contiguous districts sharing the same majority language. We only label large language regions in the figure.

of linguistic distance on internal migration in contexts where there are no discontinuous changes in geography.

India's linguistic diversity has ancient roots shaped by large-scale migrations (Joseph, 2018). While Dravidian languages are concentrated in the South, new DNA and archaeological evidence suggests the ancient Harappan civilization spoke a form of proto-Dravidian. This civilization existed circa 2600 to 1900 BCE and was centered around the Indus river

valley in Pakistan and north western India.[8] As Indo-European migrants from the central Asian Steppes began arriving in northwestern India circa 2100 BCE, they pushed the proto-Dravidian speakers into peninsular India.[9] While the DNA evidence supporting these migrations is recent (Reich, 2018), Joseph (2018) argues that the distribution of Dravidian place names in Maharashtra, Gujarat and northwestern India supports the DNA evidence that Dravidian languages spread from northwestern parts of the subcontinent to south India (Joseph, 2018, p. 157). The geographic distribution of the four language families has remained similar for the past 1500 years or so (Joseph, 2018).

Austro-Asiatic languages arrived in India from South East Asia circa 2000 BCE, with Chinese farmers migrating to eastern India, while Sino-Tibetan language speakers migrated from East Asia around the same period (LaPolla, 2001; Lipson et al., 2018). Both these language families have been concentrated in Eastern and Northern India, in and near the Himalayas. Unlike the Dravidian languages, which are spoken almost entirely in South Asia or by the South Asian diaspora, the Indo-European languages along with the Austro-Asiatic and Sino-Tibetan languages are related to many European and East Asian languages.

Apart from this brief summary of the historical roots of India's linguistic diversity, recent work suggests that geography may explain the diversity of languages throughout the world (Dickens, 2022; Michalopoulos, 2012). To the extent that geography may correlate with linguistic divisions, we control for many observable differences in geography between origin and destination districts.

3. Data

In this section, we describe the construction of our variables and data sources. For the main analysis, we construct two datasets of district-pair dyads linking origins to destinations using the 1901 census and 2001 census. In 1901, we include both the British provinces and Princely States enumerated in the census.[10]

[8]The new DNA evidence suggests the most Indians trace their genetic roots to (1) First Indians represented by Out of Africa migrants, the first homo sapiens that arrived in India around 65,000 years ago, (2) Zagrosian migrants from Iran that arrived in India around 7000 BCE and intermixed with the First Indians, and (3) Steppe pastoralists from central Asia that arrived around 2100 BCE and that also mixed with the populations they encountered (Joseph, 2018; Reich, 2018). Such mixing across groups was more common between roughly 2200 BCE and 100 CE, after which endogamy within groups became the norm (Reich, 2018). Many scholars attribute this shift to the emergence and consolidation of the Hindu caste system. The linguistic evidence also corroborates the direction of these large scale ancient migrations, namely of the Zagrosian migrants to India, of proto-Dravidian speakers migrating from northwest to south India, and of Indo-European language speakers arriving in north India from the Central Asian steppes.

[9]There is debate on the precise timing of the Indo-European migrants to India, with recent genetic evidence dating the beginnings of admixture between Ancestral North Indians and Ancestral South Indians to the second millennium BCE (Narasimhan et al., 2019; Reich et al., 2009).

[10]We drop all dyads with either an origin or destination in the province of Baluchistan as the census only surveyed a small fraction of their population near military cantonments and railways.

3.1. **Linguistic Distance.** To measure linguistic distances, we follow Fenske and Kala (2021) using data from each volume of the 1901 census.[11] For each district, the census records the number of language speakers by mother tongue. To standardize languages, we match each census language with an ISO language code and aggregate district language speakers by ISO code. Then we compute the distance between languages using the language trees recorded in version 19 of the *Ethnologue* Global Dataset. This source classifies each language using a tree with up to 15 branches.

Our measure of linguistic distance builds on standard "cladistic" measures common in the literature (Desmet et al., 2012; Gomes, 2020a). Similar to Esteban et al. (2012), we compute the distance $lingdist_{mn}$ between any two languages m and n with $m \neq n$ as:

$$lingdist_{mn} = 1 - \left(\frac{SharedBranches}{15}\right)^{\delta} \tag{1}$$

The distance of a language with itself, $lingdist_{nn}$, is set equal to zero. We set $\delta = 0.5$ in order to have distances between languages evenly spread between 0 and 1.[12] We also show nonparametric estimates with indicators ranging from 0 shared branches (different language families) to 7 shared branches (the maximum similarity between two languages in our data).

In our baseline measure of linguistic distance between districts o and d, we identify the majority language spoken in each district and use the distances between these languages as given in equation (1). That is, $lingdist_{od} = lingdist_{m(o)n(d)}$, where $m(o)$ is the majority language in district o and $n(d)$ is the majority language in district d. By "majority," we mean the plurality language, i.e. the language with the largest number of speakers. For 90% of districts in 1901, the plurality language is spoken by at least half the population. This construction minimizes reverse causality, as migration of language's speakers from o to d would increase the number of that language's speakers in d. This migration would decrease linguistic distance between o and d were we to use the full distribution of languages spoken in both districts when computing linguistic distance. Using majority language speakers overcomes the problem because migration is low and majority language speakers account for a large share of the district population (80% on average in 1901).

[11]A volume corresponds to British provinces (e.g., Punjab), an agency (a group of Princely States like the Central India Agency), or a single Princely State (e.g., Cochin). We use the word "province" to refer to a geographical unit with its own census volume. Each province (or agency, or state) is divided into finer geographical units: these units can be districts (for British India), but also Princely States, or groups of Princely States. In the rest of the paper, we refer to these disaggregated geographical units as "districts".

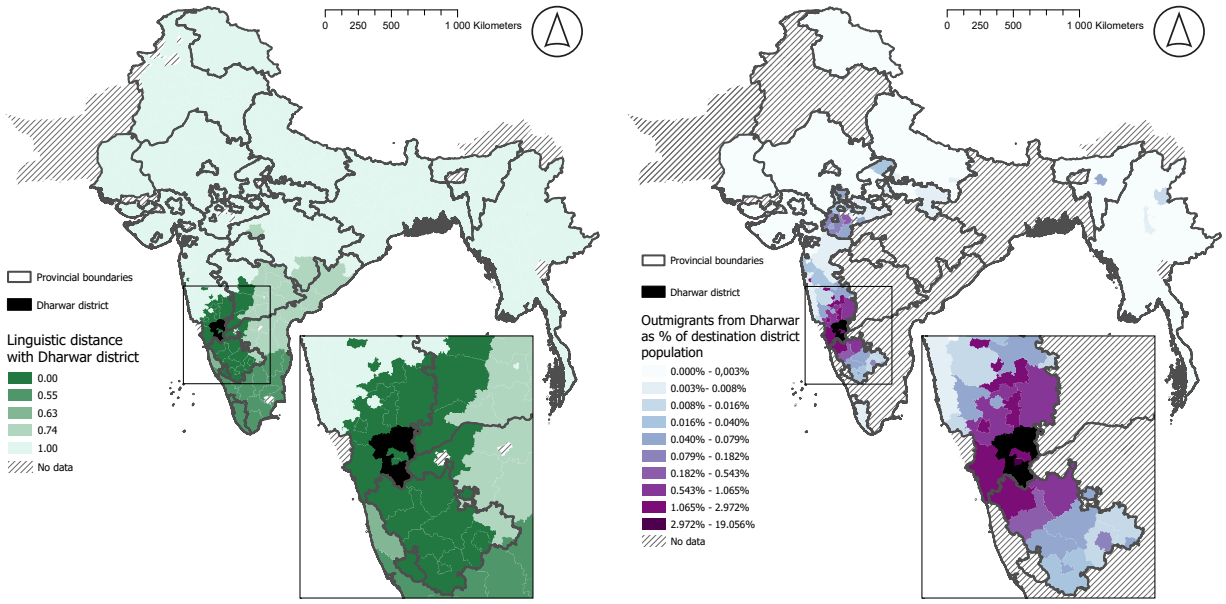
[12]Smaller values of δ emphasize differences between language families, while larger values of δ emphasize differences between close languages. Appendix Figure A.1 displays the distribution of linguistic distance between all pairs of Indian languages for $\delta = 0.05$, 0.5, and 5. When $\delta = 0.05$, the distance between languages part of the same family are very close to zero and the distance measure is akin to a binary indicator for languages from the same family. When $\delta = 5$, the distances between dissimilar languages are close to 1 and the distance measure is equivalent to a binary indicator for same languages.

For robustness, we compute an alternative measure of linguistic distance as the expected distance between the languages spoken by two individuals selected at random from each district. This alternative measure uses the full distribution of languages spoken in both districts and follows Spolaore and Wacziarg (2009). Equation (2) displays our alternative measure of linguistic distance between o and d:

$$lingdist^{A}_{od} = \sum_{m} \sum_{n} (s_{mo} \times s_{nd} \times lingdist_{mn}). \tag{2}$$

Here, s_{mo} is the share of the district o population that speaks language m, while s_{nd} is the share of the district d population that speaks language n. $lingdist_{mn}$ is the distance between the languages m and n, as given in equation (1).

FIGURE 2. Linguistic Distance and Outmigrants from Dharwar District in 1901



Left panel: linguistic distance is the cladistic distance between a district's majority language and the majority language of Dharwar (Kannada). Right panel: individuals born in Dharwar and residing in a district in 1901, as a percentage of that district's population. This map represents only the districts for which the census gives information on the number of residents born in Dharwar district. This is not the case for the districts of Punjab, Bengal, the Central Provinces, Berar, Hyderabad, and Madras, since in those provinces, the place of birth is given at the more aggregate level of the Bombay Presidency. Provincial boundaries are the boundaries of provinces, agencies, or princely states that have their own census volume. The highest concentration of migrants from Dharwar, where the majority language is Kannada, are observed in other districts where the majority language is Kannada. The share of migrants from Dharwar decreases discontinuously when we cross the border between Kannada and the Indo-European language region to the North.

To illustrate our distance measure, the left panel of Figure 2 maps the linguistic distance of all districts from Kannada speaking Dharwar district. While there is a correlation between linguistic and physical distance, several discontinuities are apparent. For example, north of Dharwar is the boundary of the majority Dravidian regions of south India. In districts

immediately north of this line, the majority language is Marathi, an Indo-European language. While Dharwar and the districts to its north are part of the same province, the majority language speakers across these districts belong to different language families. As linguistic distance correlates with physical distance, we control for physical distance between district pairs and other geographic measures of dissimilarity between districts. We also employ a regression discontinuity design exploiting the presence of linguistic boundaries.

For the 2001 analysis, we compute linguistic distance using Table C16 of the 2001 census titled "Population by Mother Tongue." Similar to the colonial exercise, we first match the census languages to ISO codes, compute the district population shares by ISO code, compute distances across languages using language trees, and finally compute the language distance between two districts as the distance between their majority languages.

3.2. **Migration.** We measure migration using the 1901 census, which enumerates both men and women by place of birth. In cases where individuals are born in the same province as the reported district, the census enumerates the exact district of birth. In cases where individuals are born outside the province, the census reports specific districts if many residents were born in that district or reports aggregate birth data, usually at the province level.[13]

We transform the census data as follows. When place of birth is reported at the district level, we compute language and geographic characteristics for each district (92% of bilateral pairs). When place of birth is aggregated to the province, we compute language and geographic variables for the province. In these cases, the bilateral pairs are between an origin province and a destination district. When the place of birth is reported for some but not all of the districts of a province, we allocate the residents with an unknown district of birth to the remaining districts in proportion to the known migrant population from each district.[14] We impute null migration from missing migration.[15]

We then construct two measures of migration. First, we measure the number of migrants born in origin o and living in destination d. Second, we construct the ratio of migrants from

[13]In the Punjab volume, aggregate birth data is given for "Bengal and Assam" and "Mysore and Coorg". We distribute to Bengal, Assam, Mysore, and Coorg (who each have their own census volume) in proportion to the known out-of-province population born in each province.

[14]For example, suppose district x has 1,000 residents born in province P, which has 4 districts (a, b, c, and d). We know that 300 were born in a, and 400 in b. The number of residents with unknown birth district is $1000 - 300 - 400 = 300$. To allocate these 300 people across districts c and d, we first compute m_c, the number of individuals born in c but currently living outside c, and m_d, the same for d. These are the known migrant populations from c and d. Importantly, they do not include individuals from c and d who are currently living in x, since these are the very numbers we aim to estimate. Because most migration is local, within-province migrants (whose population we always know) dominate these figures. We then allocate the residual births using the shares $m_c/(m_c + m_d)$ for c, and $m_d/(m_c + m_d)$ for d. Say $m_c = 100$ and $m_d = 200$, we will impute that $100/300 \times 300 = 100$ residents of x were born in c and $200/300 \times 300 = 200$ in d.

[15]For example, in the census volume for Hyderabad, Assam is not mentioned as a potential place of birth. We consider that no resident of Hyderabad was born in Assam.

origin o in destination d to the stayers born in o who remain in o; this latter approach is common in the migration literature (Adsera and Pytlikova, 2015; Chiswick and Miller, 2015).

In the right panel of Figure 2, we show outmigration from Dharwar to other districts in 1901.[16] While migration declines with distance, distance is not the only determinant of destination choices. In particular, comparing the maps of linguistic distance and migration in the adjacent panels of Figure 2, we see migration rates decline discontinuously when we cross the linguistic border between Dravidian and Indo-European languages.

We compute migration between districts in 2001 using the complete district-to-district dataset of Imbert and Papp (2020). For each district of residence, the data record the distribution of the migrant population by their district residence 10 years earlier.[17] Similar to the colonial data, we compute migration in 2001 as migrant counts and as the ratio of migrants to stayers.

3.3. **Geographic Controls.** We control for a wide set of control variables. To compute geographic variables for colonial districts, we begin with a shapefile of present-day subdistrict (e.g. tehsil) borders. We make a many-to-one correspondence between these subdistricts and the districts existing in the 1901 census and treat the union of subdistricts assigned to a colonial district as a polygon representation of that district.[18]

We compute the following controls directly. We compute distance between two districts using the great circle distance between the coordinates of their centroids. We use ArcGIS to compute indicators for neighboring districts. "Same Province" is an indicator for whether both districts are in the same 1901 province, while "Same State" is defined analogously for 2001.

For other variables, we use raster data to compute a value of the variable for each district.[19] Here, we largely follow Fenske and Kala (2021) and omit the details for brevity. For each district, we compute average elevation,[20] ruggedness,[21] temperature, precipitation,[22] malaria

[16]In many provinces, we do not have this information, because the census records place of birth at the provincial level. For example, in the census volume for the Madras Province, the number of residents born in Dharwar district is not given, only the number of residents born in the Bombay Presidency.
[17]We combine these data with district population to calculate the population of stayers.
[18]Our correspondence follows Fenske et al. (2025) and is based off several sources. Our primary source is Chandramouli et al. (2011). We also use the map from the 1931 census as digitized by Fenske and Kala (2021), maps from the Imperial Gazetteer of India taken from the Digital South Asia Library, and maps from the various census reports. We use the same method to obtain a map of 1921 census districts.
[19]When the origin is a province, we compute these variables at the level of the province.
[20]Source: STRM Digital Elevation Database; `https://srtm.csi.cgiar.org/srtmdata/`
[21]Source: Nunn and Puga (2012); `http://diegopuga.org/data/rugged/tri.zip`
[22]Source: WorldClim; `https://www.worldclim.org/data/index.html`

TABLE 1. Summary Statistics, 1901

	Bilateral dataset 1901 census				
	Obs.	Mean	St. Dev.	Min	Max
Migration variables					
Number migrants (born in *o*, residing in *d*)	51,058	517	4,319	0	524,817
Female migrants	50,834	263	2,208	0	252,422
Male migrants	50,834	256	2,198	0	268,698
Any migration indicator	51,058	0.38	0.48	0	1
(Migrants from *o* / Stayers in *o*)×100	51,058	0.09	1.16	0.00	100.00
ln((Migrants from *o* +1)/(Stayers in *o* +1))	51,058	-11.08	2.64	-18.17	0.00
Language variables					
Linguistic distance (majority languages)	51,058	0.64	0.33	0	1
Linguistic distance (all languages)	51,058	0.65	0.28	0	1
Indicator same majority language in *o* & *d*	51,058	0.12	0.33	0	1
Location variables					
Indicator *od* same province	51,058	0.26	0.44	0	1
Indicator *od* within post-1947 Indian borders	51,058	0.59	0.49	0	1
Indicator *o* is a province	51,058	0.08	0.28	0	1
Distance variables					
od neighbors	51,058	0.03	0.18	0	1
Geodesic distance between *od* (km)	51,058	1,049	712	6	3,769
ln distance	51,058	6.67	0.84	1.72	8.23
Travel time walking and sailing between *od* (hours)	51,058	249	148	2	715
Travel time with rail and river between *od* (hours)	51,058	124	72	1	363

The unit of observation is an *od* pair where *o* refers to origin and *d* refers to destination. A destination is a district (or Princely State) while an origin is either a district, Princely State, an aggregate group of districts and Princely States.

prevalence,[23] land quality,[24] and suitability for 15 crops.[25] We construct bilateral controls by taking the absolute difference across a district pair. We also compute the correlation of monthly rainfall and temperature between the two districts over the 20th century.[26] While the use of GIS sources forces us to use contemporary raster data for these variables, we anticipate that these sources will only induce measurement error in our covariates and not create any spurious correlation between linguistic distance and migration flows.

3.4. **Summary statistics.** We present summary statistics for the 1901 data in Table 1. Our observations are 51,058 origin-destination, or *od* pairs, with $o \neq d$. While each destination d is a district, some origins o are provinces aggregating several districts. Across pairs in 1901, the mean ratio of migrants to stayers is 0.09%, with a standard deviation of 1.16%, pointing to significant differences in migration between pairs. Linguistic distance ranges from 0 when

[23]Source: Kiszewski et al. (2004); we are grateful to Marcella Alsan for sharing these data.
[24]Source: Ramankutty et al. (2002); `https://nelson.wisc.edu/sage/data-and-models/atlas/maps.php?datasetid=19&includerelatedlinks=1&dataset=19`
[25]The crops are banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, and white potato. Source: FAO-GAEZ.
[26]Source: Matsuura and Willmott (2007); `http://climate.geog.udel.edu/climate`

o and d share the same majority language to 1 for od pairs where the majority languages are from different families. Across pairs, the linguistic distance between the majority languages averages 0.64, reflecting colonial India's high level of linguistic diversity. 26% of pairs are in the same province and fewer than 8% of dyads include an origin that is a province rather than a district. Since we consider 379 districts, if we had complete bilateral data, it would amount to $379 \times 378 = 143,262$ observations. Our sample is smaller because origins outside the destination province are often aggregated to the provincial level in the census data. About 3% of od pairs are neighbors.

In Appendix Table B.1, we also show summary statistics for 2001. Though colonial India included contemporary Bangladesh, Burma, and Pakistan, we have more observations in 2001, because we have almost complete district-to-district bilateral data, and because of the proliferation of new districts.[27] The ratio of migrants to stayers across origin-destination pairs in 2001 averages 0.01%.[28] Unlike in 1901, more females migrate than males in 2001. The linguistic distance between majority languages averages 0.69. Because of the creation of several new states in India, only 5% of the district pairs are now in the same state.

4. Estimation Strategy

4.1. **Gravity Model.** We first estimate a gravity model as shown in equation (3), where each observation is an origin (o) and destination (d) district pair, such that $d \neq o$. The expected value of the outcome, the number of migrants from o to d, m_{od}, depends exponentially on the linguistic distance between o and d ($lingdist_{od}$), the natural log of the physical distance between o and d ($distance_{od}$), an indicator for neighboring od pairs ($neighbors_{od}$), a vector x_{od} of the geographic differences between o and d, and separate fixed effects for o (α_o) and d (α_d):

$$\mathbb{E}(m_{od}) = \exp\Big\{\beta lingdist_{od} + \gamma \ln(distance_{od}) + \theta neighbors_{od} + \alpha_o + \alpha_d + x'_{od}\delta\Big\} \tag{3}$$

As described earlier, we use the majority languages of o and d in $lingdist_{od}$ to mitigate reverse causality problems. The vector, x_{od}, includes the absolute difference in geographic

[27]There were 593 districts in 2001, which should lead to $593 \times 592 = 351,056$ pairs with $o \neq d$. We have fewer observations because we aggregate the nine Delhi districts into a single origin/destination, we drop Ambedkar Nagar (missing data), Lakshadweep, Andaman and Nicobar islands, and origins with missing data on stayers.

[28]This figure is not directly comparable with the 1901 figure, because the set of origin-destination pairs is different. In particular, in 1901, origins sending a small number of migrants are more likely to be aggregated by province, and therefore contribute less to the simple average. In addition, migration is defined differently in 1901 and 2001: in 1901, migrants are individuals born in o and living in d. In 2001, migrants are individuals living in d who were living in o 10 years before.

characteristics between o and d, the correlation of monthly rainfall and temperature between o and d over the 20$^{\text{th}}$ century, and an indicator if o and d are in the same province in 1901.[29]

We estimate equation (3) using a pseudo-poisson maximum likelihood estimator (Correia et al., 2019, 2020) because m_{od} is a count with many zeroes and some large values. We cluster the standard errors two ways, by origin and destination (Cameron et al., 2011).

Our specification accounts for many confounding factors that may be correlated with linguistic distance and migration. In particular, we control for the physical distance between o and d and whether they are neighboring districts. We also include fixed effects, which control for unobservable features of o and d, for example a high population density o or an industrial center d, that may affect migration and linguistic distance between o and d. Finally, we control for rich geographical differences between o and d to identify the coefficient on linguistic distance, separate from geography.

Like in all gravity models, the coefficients of equation (3) must be interpreted in relative percentage changes. For example, the coefficient on linguistic distance will be interpreted in terms of the change in migration relative to district pairs for which the linguistic distance is zero, and so have the same majority language.

4.2. **Regression Discontinuity.** We complement the gravity model with an RD that exploits sharp spatial changes in languages spoken across the linguistic border between Indo-European and Dravidian languages, which lies within British Indian provinces and does not overlap with geographical boundaries.[30] Appendix Figure A.2 presents balance tests in geographic characteristics across the Indo-European and Dravidian border supporting the validity of the RD design. Across 25 characteristics, the estimated differences in geography between Indo-European and Dravidian language regions are small and largely insignificant at conventional levels.[31]

Figure 3 illustrates the intuition behind the RD strategy by displaying the major linguistic regions of British India. The RD involves comparing migration between a district like

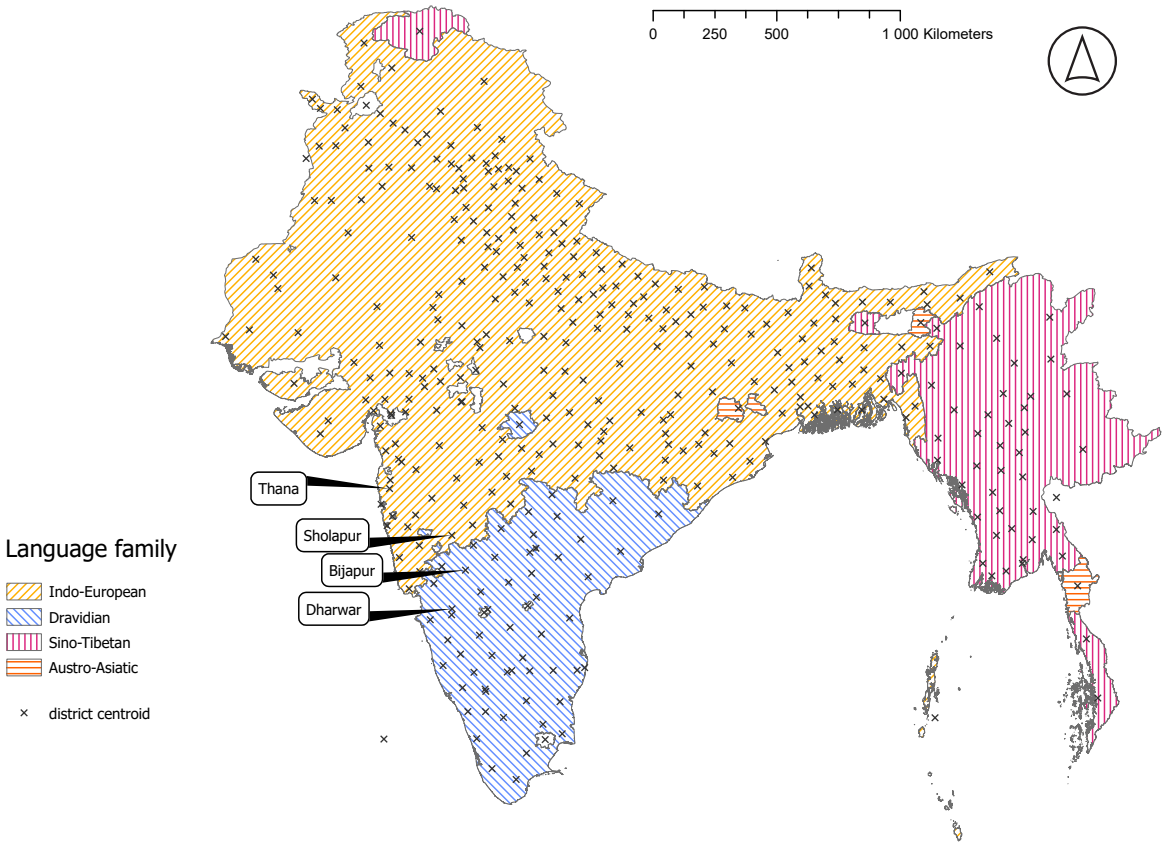
[29]The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato).

[30]We do not use the linguistic border between Indo-European and Sino-Tibetan regions because it corresponds to the geographic boundary between the Indo-Gangetic plains and the Barail Mountain Range.

[31]As shown in the Appendix Figure A.2, we estimate balance tests on two samples. First, we use data on 321 districts that are part of the Indo-European or Dravidian regions. In this case, we regress a geographic characteristic, normalized to unit variance, on an indicator for Dravidian regions and on the distance from the centroid of the district to the linguistic border. Second, we estimate balance tests using the bilateral dataset where normalized bilateral geographical controls are regressed again on an indicator if any part of the *od* pair lies in a Dravidian region along with the physical distance between o and d to the linguistic border. In this second sample, we find 3 out of 28 characteristics are significant at the 5% level, albeit with small coefficient sizes. These are the absolute *od* difference in longitude, distance to the coast, and malaria stability index.

Dharwar (o) in the Dravidian linguistic region and nearby Sholapur (d), on the other side of the linguistic border in the Indo-European region to migration between Dharwar (o) and Bijapur (d) on the same side of the linguistic border in the Dravidian region.

FIGURE 3. Major Linguistic Regions and Linguistic Borders, 1901



A linguistic region is a collection of contiguous districts whose majority languages belong to the same family. Linguistic borders are the borders between two linguistic regions. Crosses represent the centroids of districts (used to compute distance to linguistic borders).

The Indo-European and Dravidian regions define two *border experiments*. The first experiment involves origins in the Indo-European region and destinations in both the Indo-European and the Dravidian region (*treatment*). In this border RD, the running variable is the distance from d to the Indo-European and Dravidian border. The second experiment involves origins in the Dravidian region and destinations in both the Indo-European (*treatment*) and Dravidian region. We show the RD results aggregated across the two border experiments in the next section and separate results by experiment in the Appendix.[32]

Specially, we estimate the following on a sample of od district pairs assigned to a border experiment e:

[32]When the two experiments are considered together, some observations (od pairs) appear twice. This duplication however does not affect the precision of our estimates because we cluster by destination district.

$$\ln\left(\frac{m_{ode}}{m_{ooe}}\right) = \beta oth_{ode} + g(d_{ode}) + \alpha_e + P_{od} + \varepsilon_{ode} \tag{4}$$

In equation (3), m_{od} is the number of migrants born in o living in d, plus one. m_{oo} is the number of stayers born in o living in o, plus one.[33] Following the migration literature, the natural log of this ratio is our outcome. This outcome $\ln\left(\frac{m_{ode}}{m_{ooe}}\right)$ differs from equation (3) because standard RD estimators are not adapted for count variables. oth_{ode} is an indicator for o and d belonging to different linguistic regions. d_{ode} is the great circle distance of d from the Indo-European and Dravidian border, which is negative if d is in the same region as o, and positive if d is in a different region. Following Cattaneo and Titiunik (2022) and Calonico et al. (2020), $g(.)$ is a local linear function that captures the smooth relationship between migration and distance to the border d_{ode}. We use a mean square error (MSE) optimal bandwidth and a triangular kernel. α_e is a vector of border experiment fixed effects (indicators for Dravidian or European region o districts). P_{od} is an indicator if o and d are in the same 1901 province. We cluster the standard errors by destination.

If all the other factors affecting migration from o vary continuously at the border, we do not require additional controls. We however estimate the RD using the same controls as in the baseline gravity model. Since these RD estimates are not directly comparable to the gravity model, we also estimate a fuzzy RD using the border discontinuity as an instrument for linguistic distance.

Unlike the 1901 RD exercise where the Indo-European and Dravidian linguistic border cuts across colonial provinces and Princely States, Indian state borders were redrawn along linguistic lines in 1956. As a result, the linguistic border between the Indo-European and Dravidian regions matches the administrative border between the states of Andhra Pradesh and Karnataka in the south, and Maharashtra, Chhattisgarh and Orissa in the north. This redrawing makes it hard to disentangle the role of a linguistic border from an administrative (state) border. Nonetheless we estimate equation (4) using the 2001 data and show the results in the Appendix.

5. Results

In this section, we discuss the 1901 gravity model and RD estimates, followed by the 2001 gravity model, before outlining our robustness checks.

5.1. **1901 Results.**

[33]Adding 1 to the number of migrants enables us to include null migration flows (Adsera and Pytlikova, 2015).

5.1.1. *Gravity Model.* Table 2 presents the results of the gravity model. We sequentially add more controls from column (1) to column (5). In column (1) we begin with linguistic distance and an indicator for same province or Princely State and find a large, negative, and statistically significant coefficient on linguistic distance. Because physical distance is an important correlate of migration, the coefficient decreases in size to -0.834 and the pseudo R^2 almost doubles from 22% to 41% when we control for physical distance between o and d in column (2). In column (3) we add an indicator for whether o and d are neighbors; while this coefficient is positive, adding it does not diminish the coefficient on linguistic distance.

TABLE 2. Gravity Model, 1901

	(1)	(2)	(3)	(4)	(5)
	Number of Migrants				
Linguistic distance	-2.694***	-0.834***	-0.890***	-0.456**	-0.474***
	(0.211)	(0.211)	(0.214)	(0.179)	(0.132)
ln distance		-1.448***	-0.967***	-0.449***	-1.432***
		(0.091)	(0.087)	(0.111)	(0.150)
Indicator, neighboring *od*			1.523***	1.615***	1.269***
			(0.166)	(0.155)	(0.129)
Observations	51,058	51,058	51,058	51,058	51,058
Pseudo R-squared	.22	.41	.46	.57	.86
Same province dummy	✓	✓	✓	✓	✓
Origin geog. controls				✓	
Destination geog. controls				✓	
Bilateral geog. controls				✓	✓
Origin F.E.					✓
Destination F.E.					✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

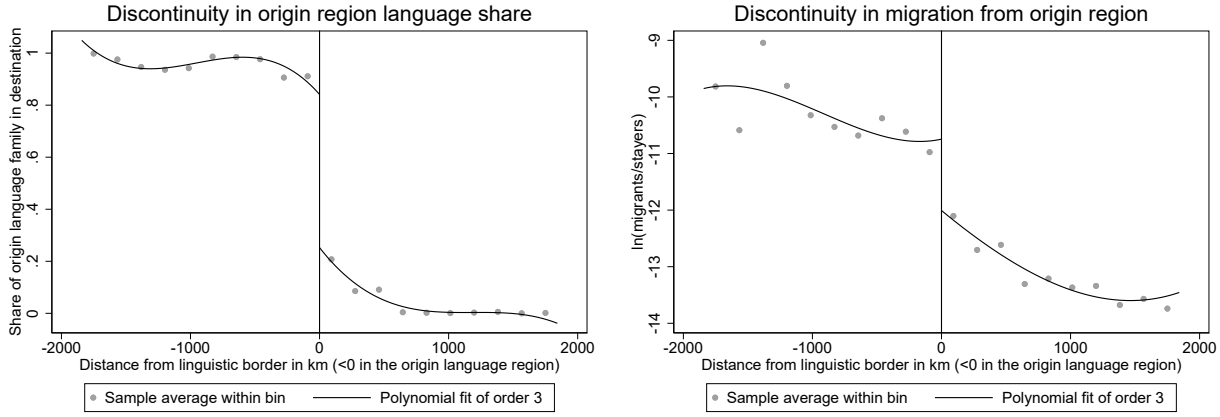
In column (4), we add controls for the geographic characteristics of o and d along with bilateral geographic differences between o and d, which decreases the coefficient to -0.456. Finally, in column (5) including o and d fixed effects shows that moving from *od* pairs with the same majority language to *od* pairs with unrelated languages – a one unit change in linguistic distance – predicts a 38% decline in migration.[34] While the coefficient is big, the

[34] $e^{-0.474} - 1 \approx -0.38$.

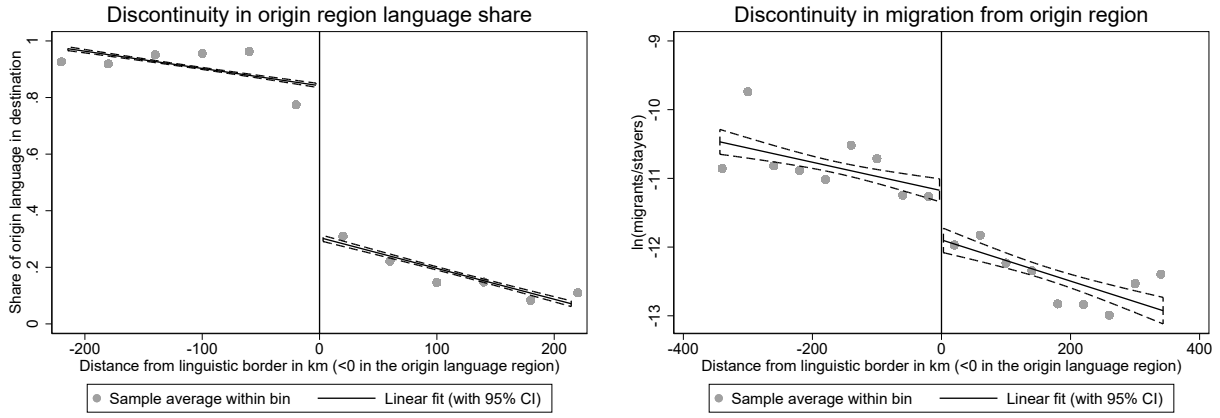
economic importance of linguistic distance is smaller than for physical distance, as we would expect. A standard deviation increase in physical distance between o and d predicts a 70% decline in migration, while a standard deviation increase in language differences predicts a 14% decline.[35]

FIGURE 4. RD along Indo-European and Dravidian Border, 1901

Panel A: Full Sample



Panel B: Border District Pairs



This figure displays the discontinuity in (left) the share of d resident speakers that speak the same language as o, and (right) the natural log of the ratio of the number of migrants (born in o and living in d) to the number of stayers (born and living in o). We use a vertical line at 0 to show the Indo-European and Dravidian linguistic region border. To the left of the vertical line are district pairs with o and d in the same language region averaged in bins of distance to the border. To the right of the vertical line are district pairs with an o and d in different language regions. averaged in bins of distance to the border. In Panel A, the black line is a visual polynomial of order 3. In panel B, we restrict the RD to observations within a data driven bandwidth of the border. We estimate a linear fit with 95% conference interval on each side of the border.

5.1.2. *Regression Discontinuity.* Figure 4 shows the RD results. In both panels, the horizontal axis shows the distance d from the Indo-European and Dravidian border. Negative values are within the same linguistic region as o. Positive values are in a different linguistic

[35]Using the SD of linguistic distance and of $\ln distance$ in Table 1, the predicted change of an SD change in linguistic distance is $e^{-0.474\times0.33} - 1 \approx -0.14$, and in $\ln distance$ is $e^{-1.432\times0.84} \approx -0.70$.

region. In Panel A, we show the full sample, while in Panel B we restrict the sample to border districts within the Calonico et al. (2020) mean square error (MSE) optimal bandwidth. There is a sharp drop at the cutoff in the share of d district residents that speak a language belonging to the same family as the o district's majority language (left side). We observe a similar discontinuous drop in the log migration rate (right side) when crossing the border to the treated, i.e., linguistically different region. We show the results for the separate border experiments in Appendix Figure A.3, which confirm the results in Figure 4 are not driven by a particular border experiment. The drop in migration from Indo-European o districts into Dravidian region destinations is of the same magnitude as the drop in migration from Dravidian o districts to Indo-European destinations.

Table 3 shows the RD results using a local linear nonparametric function of distance to the border with an MSE optimal bandwidth.[36] Column (1) includes the function for distance to the border, an indicator for same province, and border experiment fixed effect. We include additional controls for distance and geography in columns (2) and (3), and finally o and d fixed effects in column (4). While the RD estimate decreases in magnitude from column (1) to column (4), we cannot reject the equality of the four RD estimates. Using the column (4) estimate suggests that crossing the Indo-European or Dravidian language border reduces migration by 34%.[37]

For ease of interpretation, we estimate a fuzzy RD using the border discontinuity as an instrument for linguistic distance between o and d.[38] It shows that a unit increase in linguistic distance reduces migration by 50% (Appendix Table B.2).[39]. We also consider borders between finer linguistic regions than language families (Appendix C), which show that migration falls by more when we cross borders between more dissimilar linguistic regions.

5.2. **2001 Results.** Table 4 presents the results of the gravity model in 2001. As noted earlier, Pakistan, Myanmar (Burma), and Bangladesh were separated from colonial India and are not included in the 2001 data. Unlike the 1901 analysis, in which a small share of observations (8%) are between an origin province (aggregating several districts) and a destination district, the 2001 analysis uses the complete bilateral district-to-district data. We find remarkably similar results to those in 1901. Migration is lower between od pairs that are physically distant from each other, while migration is higher between neighboring district

[36]To facilitate the comparison of coefficients across columns, we compute the optimal bandwidth using equation (4) corresponding to column (1).

[37]$e^{-0.413} - 1 \approx -0.34$.

[38]Crossing the border increases linguistic distance by between 0.26 and 1. When the destination is in the origin region, o and d belong to the same family, and linguistic distance between o and d ranges from 0 (o and d share the same language) to 0.74 (the languages of o and d have only one branch in common on the tree, like Iranian and Indo-Aryan languages). When the destination is in the treated region, the languages of o and d belong to different families, so linguistic distance is always 1.

[39]$e^{-0.700} - 1 \approx -0.50$. The coefficient of -0.700 is remarkably similar in magnitude to the most comparable gravity estimate of -0.713 in Appendix Table B.7, column (6).

TABLE 3. Border RD Results along the Indo-European and Dravidian Border, 1901

	(1)	(2)	(3)	(4)
	ln(migrants/stayers)			
RD Estimate	-0.912***	-0.676**	-0.435**	-0.413***
	(0.281)	(0.283)	(0.194)	(0.056)
	[-1.525,-0.244]	[-1.378,-0.086]	[-0.907,-0.002]	[-0.603,-0.286]
Observations	36,985	36,985	36,985	36,985
Bandwidth (km)	348	348	348	348
Same province dummy	✓	✓	✓	✓
Border experiment F.E.	✓	✓	✓	✓
Distance controls		✓	✓	✓
Origin geog. controls			✓	
Destination geog. controls			✓	
Bilateral geog. controls			✓	✓
Origin F.E.				✓
Destination F.E.				✓

Standard errors clustered by destination district in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Robust Calonico et al. (2014) 95% confidence interval in [ ]. We combine the data of two border experiments. In the first one, we consider all origins in the Indo-European region (the origin region) and all destinations in both the Indo-European region and the Dravidian region (the treated region). In the second experiment, Dravidian is the origin region and Indo-European is the treated region.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

pairs compared to non-neighboring pairs. Most importantly, the coefficient on linguistic distance is large, negative and statistically significant: increasing linguistic distance from identical languages to unrelated languages predicts a 36% decline in migration.[40]

We cannot compare these 2001 estimates directly to the 1901 estimates because of changes to state and district borders. For example, the 2001 Indian states do not correspond to the 1901 British Indian provinces and Princely States. Nonetheless, we can construct alternative samples in 1901 that are closer to the 2001 data. Appendix Table B.3 shows the results of this exercise. We find comparable coefficients between 1901 and 2001 when we exclude districts in 1901 that became part of Bangladesh, Pakistan and Myanmar.[41] In columns (3) and (4), focusing on districts least affected by the Partition of India in 1947, the 2001 coefficient is smaller but we cannot reject the equality between the 1901 and 2001 coefficients.[42] And, we

[40] $e^{-0.442} - 1 \approx -0.36$.

[41] We assess the statistical difference between the coefficients by appending the 1901 and 2001 datasets. We then estimate the specification interacting all dependent variables, including fixed effects, with an indicator for the time period.

[42] In 1901 we restrict the sample to Ajmer-Merwara, Berar, Central India Agency States, Central Provinces, Cochin, Coorg, Gwalior, Hyderabad, Madras, Mysore, United Provinces, and Travancore. In 2001 we restrict

TABLE 4. Gravity Model Results, 2001

	(1)	(2)	(3)	(4)	(5)
	Number of Migrants				
Linguistic distance	-2.336***	-1.225***	-1.357***	-0.815***	-0.442***
	(0.283)	(0.212)	(0.226)	(0.263)	(0.098)
ln distance		-1.439***	-0.973***	-0.981***	-1.282***
		(0.071)	(0.066)	(0.067)	(0.141)
Indicator, neighboring *od*			1.336***	1.267***	1.132***
			(0.156)	(0.110)	(0.094)
Observations	335,820	335,820	335,820	335,820	335,820
Pseudo R-squared	.39	.56	.59	.71	.88
Same state dummy	✓	✓	✓	✓	✓
Origin geog. controls				✓	
Destination geog. controls				✓	
Bilateral geog. controls				✓	✓
Origin F.E.					✓
Destination F.E.					✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

find comparable results for a bilateral dataset of *od* pairs based on consistent geographical units in 1901 and 2001.[43]

the sample to Andhra Pradesh, Chhattisgarh, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Orissa, Tamil Nadu, Uttar Pradesh, and Uttaranchal.

[43]We use a graph theory-based algorithm to aggregate districts and create a set of consistent geographical units for which we have migration data in both 1901 and 2001. We begin by overlaying the district maps from 1901 and 2001. Whenever a district from 1901 overlaps with a district from 2001, they must belong to the same aggregated unit. Simple case: if district A (1901) was split into districts 1 and 2 (2001), we aggregate 1 and 2 back to A. More complex case: consider districts A and B in 1901. In 2001, district 1 comes from A, district 2 comes from B, and district 3 comes partly from A and partly from B. In this case, the aggregated unit must include both A and B (in 1901) and district 1, 2, and 3 (in 2001). The optimal aggregation has the minimal number of aggregated units while respecting these overlap rules. The problem can be represented as a graph, where districts in 1901 and 2001 are nodes. Two nodes are connected if the districts overlap. Finding the optimal aggregation of districts is equivalent to finding connected components in this graph (i.e., groups of nodes where each node can be reached from any other node in the group). To do this, we use the `connected_components()` function from the python NetworkX package. This function applies a Breadth First Search (BFS) algorithm to find all connected components. This process produces two bilateral origin-destination datasets (one for 1901, one for 2001) where origins and destinations are consistently defined. If migration data is missing for an origin-destination pair in one year, we drop that pair for the other year to maintain comparability. Two maximize the number of observations, we use the

Since a gravity model estimates relative migration costs, our results broadly suggest that migration between linguistically distant districts has not fallen over the 20$^{\text{th}}$ century relative to districts with the same majority language. The model, however, cannot capture generalized cultural integration that may increase migration proportionally within and across linguistic regions. This echoes the "distance puzzle" in international trade (Disdier and Head, 2008): gravity models cannot capture a generalized fall in distance costs that increases trade proportionally across all pairs of countries (Buch et al., 2004; Yotov, 2012). Note, however, that while there are no country pairs with a physical distance of zero, there are many Indian districts with a linguistic distance of zero (belonging to the same linguistic region), making the relative comparison meaningful in our case.

As Indian state borders were redrawn along linguistic lines in 1956, the Indo-European and Dravidian border corresponds with administrative borders in 2001.[44] Nonetheless we report the 2001 RD results for completeness in Appendix Table B.4, which also finds a negative link between linguistic distance and migration.

5.3. **Robustness to samples, alternative measures and estimators.** We find our results are robust to different ways of organizing the raw census data in 1901 (Appendix Table B.5). In particular we find the same results when we systematically distribute migration to districts whose origin is reported at the province level, when we use the census data as reported, including aggregated residuals (for example, the origin is reported as "Rest of Madras Province") and do not impute zeros from missing migration, when we restrict the analysis to observed district-to-district migration, and finally when we use only district-to-district migration within provinces. This is also true for the RD results (Appendix Table B.6).

While we use a Poisson pseudo-maximum likelihood estimator for the gravity model, the results are robust to using ordinary least squares (OLS) with the dependent variable expressed as the ratio of migrants to stayers in log form. They are also similar when we use extensive form measures of migration such as an indicator for any migration between o and d (Appendix Table B.7).

Our results are robust to different measures of linguistic distance (Appendix Table B.7) including the complete language distribution of speakers in o and d, and a lexicostatistical measure of linguistic distance as in Dickens (2018a).[45] They are also robust to different measures of constructing physical distance between od pairs such as a time-based measure of

full bilateral district-to-district version of the 1901 data, distributing provincial totals across districts when needed.

[44]The border between the states of Karnataka and Andhra Pradesh to the south, and Maharashtra, Chhattisgarh and Orissa to the north.

[45]We use the data of the Automatic Similarity Judgment Program or ASJP (Wichmann et al., 2022), containing 40 word lists for most languages in the world, transcribed into a standardized phonetic alphabet. We use these data and the software of Holman (2011) to compute the normalized lexicostatistical distance for each language pair, as defined in Bakker et al. (2009).

the distance, or a measure that allows for rail and river transport between o and d (Appendix Table B.7).[46]

The 2001 results are also robust to using different measures of linguistic and physical distance, and to alternative estimators (Appendix Table B.8). Finally, we replicate the results using 1921 (colonial) and 2011 (post-independence) data to ensure the results hold broadly and are not specific to certain years (Appendix Table B.9).[47]

6. Mechanisms from linguistic distance to migration

In this section, we consider the deeper mechanisms from higher linguistic distance to lower migration. While we find little evidence that culture explains our results, there is stronger evidence that communication plays a role. Education and regional language policy, however, do not mitigate the negative effect of linguistic distance.

6.1. **Cultural and network channels.**

Apart from the higher costs of communicating in a different language, people may hesitate to move to a different linguistic region because the culture there may be different. They may find it harder to interact with neighbors and forge closer friendships (Bleakley and Chin, 2010; Guven and Islam, 2015). Indeed, colonial administrators often attributed low migration to Indian culture, blaming the supposed:

> home-loving character of the Indian people, which is the result of economic and social causes, and of the immobility of an agricultural population rooted to the ground, fenced in by caste, language and social customs (Risley and Gait, 1903, p 83).

Yet we find weak evidence of cultural channels underlying the results on linguistic distance. We first consider the role of caste networks that offer informal insurance to their members. Migrants are likely to lose such informal insurance as other caste members may find it harder to observe their actions, which may increase the cost of migrating (Munshi, 2019; Munshi and Rosenzweig, 2006). Caste networks might drive the results on language differences if migration occurred within caste networks and castes were linguistically homogeneous. However, there are examples of multilingual castes such as Brahmans, Banias, Chamars, and Jats among others. Nonetheless, to address this concern, we use a novel measure of

[46]We use the Özak (2010, 2018) Human Mobility Index based on walking speeds by terrain and traditional methods of seafaring. Using this index, we compute the travel time using the least cost path — in terms of time — between the centroids of o and d. In the case of rail and river travel, we use railways data from Fenske et al. (2023), rivers from Natural Earth Data and rail plus river speeds from Donaldson (2018).

[47]We do not have full bilateral district-to-district migration data for 2011 and used Table D11, titled "Persons born and enumerated in districts of the state/UT." As this source only gives population by district of birth within a state, and by state of birth for individuals born in other states, we have fewer observations in 2011 compared to 2001.

caste distance between districts drawing on data reported in the 1901 census.[48] We include this measure of caste distance in the gravity model in Appendix Table B.10. As seen in columns (1) and (2), the coefficient on linguistic distance decreases by 28%, but remains large, negative, and significant.[49] While the coefficient on caste distance is negative and significant, it cannot fully account for language diversity.

In an analogous manner, we construct a measure of religious distance between o and d.[50] We find that controlling for religious distance reduces the coefficient on linguistic distance by only 15%. Finally, we construct two ethnographic measures of cultural distance between

[48]We begin by digitizing the Index to Caste Classification in the 1901 census. This index crosswalks caste names (roughly, *jatis*) from the province-specific Provincial Tables to the names recorded in the all-India Imperial Tables. For example, while the provincial caste name "Baniya" is listed as a "Main Caste," alternative names for the group – including Gelora, Kharwal, and Vani Dikshawanth, – are indicated as "Baniya" in the index. We use this source to harmonize caste names in the Provincial Tables. Our measure of caste distance, $castedist_{od}$, equals 1 minus the probability that a randomly drawn individual from district o and a randomly drawn individual from district d belong to the same caste. As there are many castes, the average caste distance is close to 1 (0.97) and 90% of observations fall between 0.94 and 0.99.

[49]In this specification, we use the full distribution of language speakers in o and d as we construct caste distance using the full distribution of castes in o and d. We do not have caste data for a few Princely States and ensure that our main results on language distance hold for the od pairs for which we have both caste and language data.

[50]Our measure of religious distance equals 1 minus the probability that a random individual from district o and a random individual from district d belong to the same religion accounting for seven potential religious groups, namely Hindus, Muslims, Christians, Sikhs, Jains, Buddhists and Tribals. In this specification, we also use the full distribution of language speakers to construct language distance as we use the full distribution of religious groups for religious distance. Since these measures may suffer from reverse causality, we also used a binary indicator for od pairs with different majority religions and linguistic diversity among majority languages. Our findings are unchanged: accounting for religious difference hardly affects the coefficient on linguistic distance (Appendix Table B.12).

linguistic groups using the extended *Ethnographic Atlas* of Walter (2025).[51] The *Ethnographic Atlas* is commonly used in the literature, and measures of cultural practices in this source correlate well with self-reported practices in survey data (Bahrami-Rad et al., 2021). Including these measures as a control in the gravity model does not diminish the coefficient on linguistic distance (columns (5) to (7) in Appendix Table B.10).

We also control for these measures in 2001, except for caste distance as the Indian government stopped reporting detailed caste data in the post-independence censuses. Controlling for religious distance does not reduce the coefficient on linguistic distance (columns (1) and (2) in Appendix Table B.11). Controlling for ethnographic measures of cultural distance reduces the coefficient on linguistic distance by about a quarter and makes it less precisely estimated (columns (3) to (5) in Appendix Table B.11). But we hesitate to draw strong conclusions on ethnographic distance because we do not observe similar results in 1901 and because of the incomplete coverage of the data.

Related to cultural networks, linguistic distance may reduce migration if it hurts the development of migration networks that facilitate chain migration. To assess such channels, we need data on both the stock and flow of migrants. Since the colonial and post-colonial censuses do not report them, we use the 1983 employment survey conducted by the National Sample Survey Office (NSSO). This survey identifies the region of origin and residence of

[51]This source extends the original Atlas from Murdock (1967) to include more ethnic groups. We begin by matching language groups from the censuses to ethnic groups from the Atlas. We are able to match 86% of majority language groups in 1901 to a group in Walter (2025), corresponding to 91% of districts. Similarly, we match 72% of groups in 2001, corresponding to 91% of districts. The matching is usually straightforward — for example, Walter (2025) has a group called "Hindi," a group called "Tamil," and a group called "Bengali." In a few cases, we match several languages to the same group. For example, we match all Bhil languages (Bhili, Pauri Bareli, Dhanki, Wagdi) to the group called "Bhil." For district pairs speaking different languages but belonging to the same ethnographic group, linguistic distance is positive and ethnographic distance is zero, but ethnographic distance might be measured with more error than linguistic distance. We exclude these pairs from the analysis (less than 1% of the sample), but including them does not modify our conclusions. For each society, the Atlas records 94 variables that describe cultural practices. For example, one variable records the type of transaction at marriage, another the degree and type of caste differentiation, and another the type of games played. We build two measures of ethnographic difference between two groups A and B. The first measure is the percentage of non-missing Atlas items that are different for group A and group B. To build the second measure, we undertake principal component analysis on a vector of variables that measure the absolute difference between A and B for the variable in the *Ethnographic Atlas*. For this measure, we use only the 29 variables that are missing for less than 10% of South Asian groups. We deal with the remaining missing values by replacing them with the average of all non-missing pairs. For example, if we do not know whether A and B agree on an item and the rate of agreement among all other pairs is 50%, we replace the missing variable with 0.5. We run the principal component analysis on only the 122 South Asian groups, giving about 15,000 pairs. Our two measures of ethnographic distance measured across these pairs of groups have a correlation of 0.82.

migrants along with their date of migration.[52] Using this survey, we build a district-to-district dataset aggregating the individual data by district of residence and district of origin. One disadvantage of the survey data is the small number of migration stocks and flows: 92% of district pairs out of 80,014 have zero migrants.

TABLE 5. Gravity Model, Flows and Stocks, 1983 Survey

	(1)	(2)	(3)	(4)
	Number of Migrants			
	last 5 years		last 10 years	
Linguistic distance	-0.341*	-0.311*	-0.376**	-0.358**
	(0.178)	(0.175)	(0.155)	(0.152)
Indicator, neighboring *od*	0.894***	0.827***	0.911***	0.807***
	(0.090)	(0.086)	(0.078)	(0.072)
ln distance	-0.631***	-0.564***	-0.660***	-0.555***
	(0.123)	(0.120)	(0.109)	(0.107)
ln stock of migrants		0.041***		0.055***
as share of *o* pop		(0.014)		(0.010)
Observations	73,111	73,111	76,291	76,291
Pseudo R-squared	.76	.76	.8	.8
Same state dummy	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

The dependent variable is the number of migrants residing in d at the date of the survey who came from o within the previous 5 years in columns (1) & (2), and within the previous 10 years in columns (3) & (4). The stock of migrants is the number of migrants living in d at the date of the survey who came from o more than 5 years before the date of the survey in columns (1) & (2), and more than 10 years in columns (3) & (4).

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

We separate the flow of recently arrived migrants from the stock that arrived in the past. Because 98% of districts have zero migrants in the last year, we define the flow of migrants

[52]While recent censuses collect more data on migration, this is not part of the publicly shared data (Kone et al., 2018). The NSSO regions are smaller than states but larger than districts. We are able to identify the district of residence of 449,264 out of 623,494 individuals in the survey. The survey asks migrants for their district of origin, including for migrants who migrated from a place in the same district. It is also possible to identify groups of households residing in the same village using the fact that household numbers reset to 1 in each village. So long as there is at least one person in a given village who migrated from a different village in the same district, the district of residence can be identified for the whole village.

as the number that arrived in the last 5 years and the stock as those that arrived more than 5 years ago (95% of district pairs have zero migrants in the last 5 years). As an alternative, we also define the flow as migrants who arrived in the last 10 years, and the stock as those that arrived more than 10 years ago (94% of district pairs have zero migrants in the last 10 years).[53]

Using the gravity model, we estimate the flow of migrants as a function of linguistic distance, the same controls as in Table 2, and the natural log of (one plus) the stock of migrants from o in d, normalized by the population in o (Adsera and Pytlikova, 2015). In odd columns of Table 5, we replicate the results in this new dataset without including the migrant stock, while in even columns, we include the natural log of the migrant stock. The coefficient on linguistic distance is negative and significant at the 10% level.[54] Including the stock of migrants does not affect the coefficient on linguistic distance (columns (1) and (2)). As we would expect, the coefficient on the stock of migrants is significant. But it does not drive the results on linguistic distance. The results are robust to changing the window to 10 years. When we change the window to 1 year (Appendix Table B.13), the results are quantitatively similar, but imprecise as expected, because of the small number of district pairs with migrants in the last year.

Finally, we estimate split samples by sex in Table 6. We find no evidence of heterogeneity in the colonial or contemporary period. The coefficient on an indicator for *od* pairs being neighbors is larger for women, especially in 2001. Women may be more likely to migrate after marriage, which supports historical accounts of female marriage migration (Tumbe, 2012). If cultural channels related to patrilocal marriage were underlying the results on linguistic distance, we would expect a differential magnitude for women. We do not find evidence of this difference.

6.2. **Communication channels.** Understanding the language of a destination region increases the benefits and decreases the costs of migration along multiple dimensions. Linguistic homogeneity may enable better information flows about labor market conditions before and after migration (Dickens, 2018b). Speaking the local language may also reduce the costs of finding and retaining employment (Cohen-Goldner and Eckstein, 2008; Dustmann and Fabbri, 2003; Lochmann et al., 2019; McManus, 1985), navigating local markets, and finding cheaper goods (Chiswick and Miller, 2015). Apart from economic benefits, linguistic similarity between o and d may also improve infant health (Auer and Kunz, 2025;

[53]Because of the large number of district pairs with zero migration, we add 1 to the stock of migrants and to the population of the origin. As we do not have language data for 1983, we use the 2001 census data on linguistic distance. Majority languages rarely change within districts, which suggests any measurement error is small.

[54]This loss of precision is expected; we are using a survey, which increases measurement error in migration, and we observe fewer district pairs.

TABLE 6. Gravity Model, Split Samples by Sex, 1901 & 2001

	(1)	(2)	(3)	(4)
	1901		2001	
	Number of Migrants			
	Male	Female	Male	Female
Linguistic distance	-0.509***	-0.453***	-0.392***	-0.437***
	(0.137)	(0.136)	(0.110)	(0.095)
Indicator, neighboring *od*	1.140***	1.404***	0.781***	1.295***
	(0.125)	(0.135)	(0.092)	(0.087)
ln distance	-1.320***	-1.497***	-1.087***	-1.394***
	(0.149)	(0.153)	(0.156)	(0.127)
Observations	50,834	50,834	335,820	335,820
Same state/province dummy	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato).

The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20$^{\text{th}}$ century.

Gomes, 2020b) and schooling outcomes (Bleakley and Chin, 2004). For contemporary India, Manjunath (2024) finds migrants with high language barriers are employed less often in speaking-intensive occupations, suggesting language barriers reduce worker productivity when communication is a key part of the job.

We show three pieces of evidence that suggest language distance is more about communication and information bottlenecks. First, we estimate split samples for *od* pairs where the agricultural wage in d is higher than in o and where the agricultural wage in o is higher than in d.[55] We conjecture that communication and information barriers reduce migration between *od* pairs when they reinforce economic gaps and constrain information flows about jobs and wages, i.e., for district pairs where d wage $> o$. Indeed, we find a larger coefficient on linguistic distance for *od* pairs when d wages exceed those in o (column (2) Table 7). In column (1), we find no significant coefficient on linguistic distance on *od* pairs when d wages do not exceed those in o. In columns (3) to (6), we split the sample by quartile differences in the *od* wage gap. While the coefficient on linguistic distance is statistically indistinguishable

[55]The wage data, from Fenske et al. (2022), include agricultural and rural non-agricultural wages for 80 districts. This reduces our sample size of *od* pairs and prevents us from evaluating rural-urban wage gaps.

TABLE 7. Gravity Model: Split Samples of 1901 Agricultural Wages

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of Migrants					
			od wage difference (quartile)			
	d wage < o wage	d wage > o wage	1st	2nd	3rd	4th
Linguistic Distance	0.899	-1.107**	0.242	1.408	1.387	-1.977***
	(1.108)	(0.479)	(1.085)	(1.146)	(1.696)	(0.506)
Indicator, neighboring *od*	1.978***	1.817***	1.188*	1.504***	2.200**	1.801**
	(0.312)	(0.322)	(0.609)	(0.193)	(0.952)	(0.770)
Observations	694	1,038	442	439	446	440
Pseudo R-squared	0.97	0.95	0.96	0.98	0.96	0.98
Same province dummy	✓	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

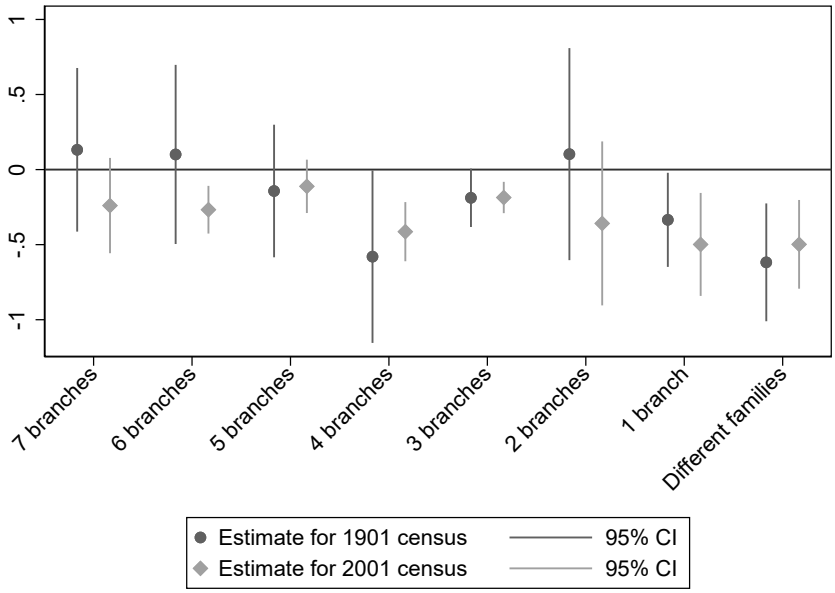
from zero in the first three quartiles, it is large and significant for the last quartile where the average wage in d is much higher than in o (column (6)).

Second, we find linguistic differences are most salient at higher levels of dissimilarity, i.e. district pairs with languages from different families have lower migration compared to district pairs with more closely related, albeit different, languages. As discussed in Section 3.1, we measure cladistic distance based on the number of branches two languages have in common in the language tree. Another way to capture linguistic distance is to use indicators for the number of branches shared by the majority languages in o and d. This enables us to run the following regression where the reference group is district pairs with the same majority language, regardless of number of branches:

$$\mathbb{E}(m_{od}) = \exp\Big\{\sum_{i=0}^{7} \beta_i \mathbb{1}(\neq \text{ languages }, i \text{ common branches})_{od} + \gamma \ln(distance_{od}) + \theta neighbors_{od} + \alpha_o + \alpha_d + x'_{od}\delta\Big\} \tag{5}$$

Figure 5 shows the estimated β coefficients of this non-parametric approach. As the number of common branches increases, the coefficient on the number of shared branches becomes smaller. For languages with 7 branches in common, the coefficient is statistically indistinguishable from zero. Languages with 7 common branches are often part of the same dialectal chain and mutually intelligible, such as the various Rajasthani languages. For languages belonging to different families, the 1901 coefficient is large and significant at -0.62.

FIGURE 5. Non-Parametric Linguistic Distance – Gravity Model, 1901 & 2001



The controls are the same as in Tables 2 and 4, column (5).
We replace cladistic linguistic distance with a vector of 8 indicators for the number of branches in common along the linguistic phylogenetic tree. The reference category is having the same majority language.
The figure shows the coefficients and 95% confidence intervals for the 8 indicators. Standard errors are clustered two ways by origin and destination districts.

Finally, we find differences in languages spoken are more salient than differences in mother tongue. We use the 2001 census for this exercise because it enumerates the distribution of second and third subsidiary languages for each state and mother tongue. This information was not collected in 1901. Using these data, we construct the distance between the most spoken language in o and the most spoken language in d, inclusive of subsidiary languages.[56] Considering subsidiary languages changes the majority language in 5% of districts and linguistic distance in 5% of district pairs.[57] We also compute the share of the o district's

[56]The data on subsidiary languages is given at the state × mother tongue level, not district × mother tongue level. To estimate data at the district level, we combine the mother tongue distribution at the district level and the subsidiary language distribution at the state × mother tongue level.
[57]For most districts, the majority language becomes Hindi.

TABLE 8. Mechanisms: Subsidiary Languages, 2001

	(1)	(2)	(3)	(4)
	Number of Migrants			
Majority mother tongues *od* different				-0.102
				(0.146)
LD majority mother tongues	-0.442***		0.217	-0.430*
	(0.098)		(0.299)	(0.259)
LD most spoken languages		-0.579***	-0.750**	
		(0.117)	(0.293)	
Share *o*'s majority mother tongue speakers				0.808***
also speaking *d*'s majority mother tongue				(0.186)
Share *d*'s majority mother tongue speakers				0.999***
also speaking *o*'s majority mother tongue				(0.203)
Indicator, neighboring *od*	1.132***	1.123***	1.121***	1.121***
	(0.094)	(0.094)	(0.093)	(0.093)
ln distance	-1.282***	-1.295***	-1.299***	-1.286***
	(0.141)	(0.138)	(0.134)	(0.142)
Observations	335,820	335,820	335,820	335,820
Pseudo R-squared	.88	.88	.88	.88
Same state dummy	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.
The 2001 census data on subsidiary languages is given at the state × mother tongue level, not the district × mother tongue level. To estimate data on subsidiary languages spoken at the district level, we combine the mother tongue distribution at the district level and the subsidiary language distribution at the state × mother tongue level.

majority mother tongue speakers that speak the majority mother tongue of d as a second or third language, and vice-versa.

Column (1) in Table 8 shows the results of Table 4 column (5) for comparison. Column (2) replaces linguistic distance between majority mother tongues with linguistic distance between the most spoken languages. Column (3) includes both measures of linguistic distance. The coefficient on linguistic distance between languages spoken is larger than the linguistic distance between majority mother tongues (column (2)). Moreover, the coefficient on linguistic distance between majority mother tongues is statistically insignificant compared to the significant coefficient on linguistic distance between languages spoken. In column (4), we

also include the share of $o's$ majority mother tongue speakers that speak $d's$ majority mother tongue as a second or third language and vice versa.[58] The coefficients on both variables are positive, large, and statistically significant. While migration is lower between o and d if their majority mother tongues differ, subsidiary languages play an offsetting role.

6.3. **Education and regional languages.** If, as we argue, language barriers constrained migration by impeding communication and information transmission, we may expect education and official language policy to mitigate the negative effect of linguistic distance. However, it is unclear theoretically whether education would attenuate linguistic barriers to migration. While schooling may enable people to learn more languages, literacy may also tie labor market skills to a specific language and script. An agricultural migrant may not need to master the d language as much as a teacher, for example.[59]

To assess heterogeneity by education, we estimate split samples of the gravity model for o districts with above and below median literacy in 1901 (Table B.14).[60] We find no significant differences in the coefficient on linguistic distance between high and low literacy districts. We also use individual data from the 1983 NSSO employment survey to compute bilateral od migration flows for illiterate migrants, migrants with up to primary schooling, and migrants with secondary schooling or higher.[61] Similar to the 1901 results, we find no evidence of heterogeneous coefficients on linguistic distance by education.

As education has increased over the 20$^{\text{th}}$ century in India, English and Hindi have also emerged as pan-Indian languages. Both languages are used by the Central Government, while the Indian constitution recognizes 22 scheduled languages i.e., regional languages tied to different states.[62] We test for whether the existence of the official and scheduled languages attenuates the coefficient on linguistic distance using data for 2001.[63] Column (1) of appendix Table B.15 shows our baseline specification from Table 4, column (5). We replace the distance between majority languages with the distance between official languages in column

[58]Note that when both districts share the same majority mother tongue, these variables are not defined and arbitrarily set to zero. It is inconsequential, as we include a binary indicator if the majority mother tongues of o and d are different.

[59]The empirical research is also ambiguous on this question. Berman et al. (2003) and Lochmann et al. (2019) find higher returns to learning a language among high skilled migrants, while Hall and Farkas (2008) finds high returns to learning English for example among low skilled immigrants.

[60]We use literacy data from Chaudhary and Fenske (2023b) for this exercise.

[61]We compute migration flows using a 10 year time window and obtain similar results when using the shorter 5 year time window.

[62]There were originally 14 languages, and new languages were added over time. Hindi is included in the list of scheduled languages. English is not.

[63]These data are from the 42$^{\text{nd}}$ Report of the Commissioner for Linguistic Minorities (2003-2004) and the 52$^{\text{nd}}$ Report of the Commissioner for Linguistic Minorities (2014-2015).

(2).[64] Column (3) includes distance between majority languages and official languages as two separate variables. While the coefficient on the distance between majority languages increases in magnitude, we find no significant coefficient on distance between official languages. In column (4), we interact linguistic distance with indicators for o and d sharing an official or scheduled language. Although linguistic distance is associated with a larger decrease in migration when o and d do not share an official or scheduled language, the two coefficients are not statistically different from one another (p-value of 0.55). Official languages may not attenuate linguistic difference perhaps because they are not widely spoken by minority language speakers. For example, in the 2001 census, only 42% of minority language speakers speak one of their state's official languages as a second or third language.[65]

7. Conclusion

Using new data on language and migration across the districts of colonial India, we find district pairs with more dissimilar languages experience lower migration than pairs with similar languages. This pattern endures up to 2001. Moreover, the results are robust to including rich geographic controls, fixed effects, an RD design, and other tests. We find limited evidence that cultural channels drive the results on linguistic distance. Rather, we find more support for an information and communication channel whereby language differences may reduce the economic returns to migration.

Barriers to migration are barriers to the integration of labor markets, and we have shown that linguistic diversity does have a role in explaining the low rates of internal migration in both colonial and postcolonial India – a role that persists in the present. The correlation between language differences and migration survives controlling for distance, geographic and fixed effects, which suggests that this correlation is not spurious. Our regression discontinuity design adds credibility to our findings and offers a novel strategy that can be used to evaluate the importance of linguistic barriers in other contexts.

We have shown that cultural barriers to migration have not diminished over time, despite mass education, economic development, state language policies, and the emergence of pan-Indian languages. The failure of these factors to reduce language as a barrier to migration poses a challenge for nation-building efforts, which have elsewhere succeeded in unifying countries (Rohner and Zhuravskaya, 2024). Why communication frictions remain, despite mass education and the spread of both Hindi and English, remains as an avenue for future research.

[64]Each state has one or more official languages, and two districts in different states can share an official or scheduled language without sharing a majority language. For states that have several official languages, we compute the distance between all pairwise combinations of official languages and compute a simple average.
[65]Minority language speakers are defined as individuals who do not speak as a mother tongue one of the states' official languages.

References

Abramitzky, R., Boustan, L., Jácome, E., and Pérez, S. (2021). Intergenerational mobility of immigrants in the United States over two centuries. *American Economic Review*, 111(2):580–608.

Adsera, A. and Pytlikova, M. (2015). The role of language in shaping international migration. *The Economic Journal*, 125(586):F49–F81.

Agnihotri, I. (1996). Ecology, land use and colonisation: the canal colonies of Punjab. *The Indian Economic & Social History Review*, 33(1):37–58.

Alan, S., Baysan, C., Gumren, M., and Kubilay, E. (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics*, 136(4):2147–2194.

Alesina, A. and La Ferrara, E. (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A. and Tabellini, M. (2024). The political effects of immigration: Culture or economics? *Journal of Economic Literature*, 62(1):5–46.

Atkin, D. (2013). Trade, tastes, and nutrition in India. *The American Economic Review*, 103(5):1629–1663.

Auer, D. and Kunz, J. S. (2025). Communication barriers and infant health: the intergenerational effect of randomly allocating refugees across language regions. *American Economic Journal: Economic Policy*, 17(3):71–106.

Bahrami-Rad, D., Becker, A., and Henrich, J. (2021). Tabulated nonsense? Testing the validity of the Ethnographic Atlas. *Economics Letters*, 204:109880.

Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C., Brown, P., Egorov, D., Mailhammer, R., Grant, A., and Holman, E. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181.

Bazzi, S., Gaduh, A., Rothenberg, A. D., and Wong, M. (2019). Unity in diversity? How intergroup contact can foster nation building. *American Economic Review*, 109(11):3978–4025.

Bazzi, S. and Gudgeon, M. (2021). The political boundaries of ethnic divisions. *American Economic Journal: Applied Economics*, 13(1):235–266.

Bell, M., Charles-Edwards, E., Ueffing, P., Stillwell, J., Kupiszewski, M., and Kupiszewska, D. (2015). Internal migration and development: Comparing migration intensities around the world. *Population and Development Review*, 41(1):33–58.

Belot, M. and Ederveen, S. (2012). Cultural barriers in migration between OECD countries. *Journal of Population Economics*, 25:1077–1105.

Berman, E., Lang, K., and Siniver, E. (2003). Language-skill complementarity: returns to immigrant language acquisition. *Labour Economics*, 10(3):265–290.

Blanc, G. and Kubo, M. (2024). French. *Working Paper, The University of Manchester*.

Bleakley, H. and Chin, A. (2004). Language skills and earnings: Evidence from childhood immigrants. *Review of Economics and Statistics*, 86(2):481–496.

Bleakley, H. and Chin, A. (2010). Age at arrival, English proficiency, and social assimilation among US immigrants. *American Economic Journal: Applied Economics*, 2(1):165–192.

Bolt, J. and Van Zanden, J. L. (2024). Maddison-style estimates of the evolution of the world economy: A new 2023 update. *Journal of Economic Surveys*.

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. *Econometrica*, 82(5):1671–1748.

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

Buch, C. M., Kleinert, J., and Toubal, F. (2004). The distance puzzle: on the interpretation of the distance coefficient in gravity equations. *Economics Letters*, 83(3):293–298.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Carlitz, R., Morjaria, A., Mueller, J., and Osafo-Kwaako, P. (2024). State building in a diverse society. *Forthcoming, Review of Economic Studies*.

Cattaneo, M. D. and Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, 14(1):821–851.

Chandramouli, C., Singh, A., and Sethi, R. (2011). *Census of India 2011: Administrative Atlas of India*. Office of the Registrar General & Census Commissioner, India, Ministry of Home Affairs, Government of India.

Chaudhary, L. and Fenske, J. (2023a). Colonial cities and urbanization. In *Forthcoming: The Cambridge Economic History of Modern South Asia: 1750-1947*. Cambridge University Press.

Chaudhary, L. and Fenske, J. (2023b). Railways, development, and literacy in India. *The Journal of Economic History*, 83(4):1139–1174.

Chetty, R., Hendren, N., and Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902.

Chiswick, B. R. and Miller, P. W. (2015). International migration and the economics of language. In *Handbook of the economics of international migration*, volume 1, pages 211–269. Elsevier.

Clark, S. (2000). Son Preference and Sex Composition of Children: Evidence from India. *Demography*, 37(1):95–108.

Clingingsmith, D. (2014). Industrialization and bilingualism in India. *Journal of Human Resources*, 49(1):73–109.

Cohen-Goldner, S. and Eckstein, Z. (2008). Labor mobility of immigrants: Training, experience, language, and opportunities. *International Economic Review*, 49(3):837–872.

Collins, W. J. (1999). Labor mobility, market integration, and wage convergence in late 19th century India. *Explorations in Economic History*, 36(3):246–277.

Correia, S., Guimarães, P., and Zylkin, T. (2019). Verifying the existence of maximum likelihood estimates for generalized linear models. *arXiv preprint arXiv:1903.01633*.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast Poisson estimation with high-dimensional fixed effects. *The Stata Journal*, 20(1):95–115.

Desmet, K., Gomes, J. F., and Ortuño-Ortín, I. (2020). The geography of linguistic diversity and the provision of public goods. *Journal of Development Economics*, 143:102384.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2012). The political economy of linguistic cleavages. *Journal of Development Economics*, 97(2):322–338.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2016). Linguistic Cleavages and Economic Development. In *The Palgrave Handbook of Economics and Language*, pages 425–446. Springer.

Dickens, A. (2018a). Ethnolinguistic Favouritism in African Politics. *American Economic Journal: Applied Economics*, 10(3):370–402.

Dickens, A. (2018b). Population relatedness and cross-country idea flows: evidence from book translations. *Journal of Economic Growth*, 23(4):367–386.

Dickens, A. (2022). Understanding ethnolinguistic differences: The roles of geography and trade. *The Economic Journal*, 132(643):953–980.

Disdier, A.-C. and Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Dustmann, C. and Fabbri, F. (2003). Language proficiency and labour market performance of immigrants in the UK. *The Economic Journal*, 113(489):695–717.

Dyson, T. and Moore, M. (1983). On Kinship Structure, Female Autonomy, and Demographic Behavior in India. *Population and Development Review*, 9(1):35–60.

Easterly, W. and Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, 112(4):1203–1250.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and conflict: An empirical study. *The American Economic Review*, 102(4):1310–1342.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2):225–239.

Fenske, J., Gupta, B., and Neumann, C. (2025). Missing Women in Colonial India. *Forthcoming in the Economic History Review*.

Fenske, J., Gupta, B., and Yuan, S. (2022). Demographic shocks and women's labor market participation: Evidence from the 1918 influenza pandemic in India. *The Journal of Economic History*, 82(3):875–912.

Fenske, J. and Kala, N. (2021). Linguistic distance and market integration in India. *The Journal of Economic History*, 81(1):1–39.

Fenske, J., Kala, N., and Wei, J. (2023). Railways and cities in India. *Journal of Development Economics*, 161:103038.

Gait, E. A. (1913). *Census of India 1911. Vol. 1, India. Pt. 1, Report*. Office of the Superintedent of Government Printing, Calcutta.

Gomes, J. F. (2020a). The health costs of ethnic distance: Evidence from Sub-Saharan Africa. *Journal of Economic Growth*, 25:195–226.

Gomes, J. F. (2020b). The health costs of ethnic distance: evidence from Sub-Saharan Africa. *Journal of Economic Growth*, 25(2):195–226.

Government of India (2017). India on the move and churning: New evidence. In *Economic Survey 2016-17*, pages 264–284. Government of India.

Gupta, A., Ponticelli, J., and Tesei, A. (2024). Language barriers, technology adoption and productivity: Evidence from agriculture in India. *Forthcoming, Review of Economics and Statistics*.

Gupta, B. (2011). Wages, unions, and labour productivity: evidence from Indian cotton mills. *The Economic History Review*, 64:76–98.

Gupta, B. and Swamy, A. V. (2017). Reputational Consequences of Labor Coercion: Evidence from Assam's Tea Plantations. *Journal of Development Economics*, 127:431–439.

Guven, C. and Islam, A. (2015). Age at migration, language proficiency, and socioeconomic outcomes: evidence from Australia. *Demography*, 52(2):513–542.

Hall, M. and Farkas, G. (2008). Does human capital raise earnings for immigrants in the low-skill labor market? *Demography*, 45(3):619–639.

Holman, E. W. (2011). Programs for calculating asjp distance matrices (version 2.1).

Imbert, C. and Papp, J. (2020). Costs and benefits of rural-urban migration: Evidence from India. *Journal of Development Economics*, 146:102473.

Jain, T. (2017). Common tongue: The impact of language on educational outcomes. *Journal of Economic History*, 77(2):473–510.

Joseph, T. (2018). *Early Indians: the story of our ancestors and where we came from*. Juggernaut, New Delhi.

Kerr, I. J. (2006). On the move: circulating labor in pre-colonial, colonial, and post-colonial India. *International Review of Social History*, 51(S14):85–109.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene*, 70(5):486–498.

Kone, Z. L., Liu, M. Y., Mattoo, A., Ozden, C., and Sharma, S. (2018). Internal borders and migration in India. *Journal of Economic Geography*, 18(4):729–759.

Lagakos, D., , Mobarak, A. M., and Waugh, M. E. (2023). The Welfare Effects of Encouraging Rural-Urban Migration. *Econometrica*, 91(3):803–837.

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

LaPolla, R. J. (2001). The role of migration and language contact in the development of the Sino-Tibetan language family. *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, pages 225–254.

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T. O., Willis, A., Matsumura, H., Buckley, H., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, 361(6397):92–95.

Lochmann, A., Rapoport, H., and Speciale, B. (2019). The effect of language training on immigrants' economic integration: Empirical evidence from France. *European Economic Review*, 113:265–296.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, 111(6):1807–1844.

Lyons, E. (2017). Team production in international labor markets: Experimental evidence from the field. *American Economic Journal: Applied Economics*, 9(3):70–104.

Manjunath, A. (2024). Language Barriers, Internal Migration, and Labor Markets in General Equilibrium. *Working Paper*.

Marx, B., Pons, V., and Suri, T. (2021). Diversity and team performance in a Kenyan organization. *Journal of Public Economics*, 197:104332.

Matsuura, K. and Willmott, C. (2007). Terrestrial Air Temperature and Precipitation: 1900-2006 Gridded Monthly Time Series, Version 1.01. *University of Delaware*.

McManus, W. S. (1985). Labor market costs of language disparity: An interpretation of Hispanic earnings differences. *The American Economic Review*, 75(4):818–827.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *The American Economic Review*, 102(4):1508–1539.

Miguel, E. (2004). Tribe or nation? Nation building and public goods in Kenya versus Tanzania. *World Politics*, 56(3):327–362.

Miguel, E. and Gugerty, M. K. (2005). Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics*, 89(11-12):2325–2368.

Miller, B. (1981). *The Endangered Sex*. Cornell.

Morris, D. M. (1965). *The emergence of an industrial labor force in India: A study of the Bombay cotton mills, 1854-1947*. University of California Press.

Munshi, K. (2019). Caste and the Indian economy. *Journal of Economic Literature*, 57(4):781–834.

Munshi, K. and Rosenzweig, M. (2006). Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy. *American Economic Review*, 96(4):1225–1252.

Munshi, K. and Rosenzweig, M. (2016). Networks and misallocation: Insurance, migration, and the rural-urban wage gap. *American Economic Review*, 106(1):46–98.

Murdock, G. P. (1967). Ethnographic atlas: a summary. *Ethnology*, 6(2):109–236.

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. *Science*, 365(6457):eaat7487.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*, 94(1):20–36.

Özak, Ö. (2010). The Voyage of Homo-Economicus: Some Economic Measures of Distance. *Working Paper: Department of Economics, Southern Methodist University*.

Özak, Ö. (2018). Distance to the technological frontier and economic development. *Journal of Economic Growth*, 23(2):175–221.

Pandey, A. K. (2014). Spatio-temporal changes in internal migration in India during post reform period. *Journal of Economic & Social Development*, 10(1):107–116.

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and Biogeography*, 11(5):377–392.

Reich, D. (2018). *Who we are and how we got here: ancient DNA and the new science of the human past*. Oxford University Press, Oxford, United Kingdom, first edition edition.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature*, 461(7263):489–494.

Risley, H. H. and Gait, E. A. (1903). *Census of India 1901. Vol. 1, India. Pt. 1, Report*. Office of the Superintedent of Government Printing, Calcutta.

Rohner, D. and Zhuravskaya, E. (2024). The economics of nation-building: Methodological tool kit and policy lessons. *Annual Review of Economics*, 17:453–478.

Sen, S. (1999). *Women and labour in late colonial India: The Bengal jute industry*. Cambridge University Press.

Simmons, C. (1976). Recruiting and Organizing an Industrial Labour Force in Colonial India: The Case of the Coal Mining Industry, c. 1880-1939. *The Indian Economic & Social History Review*, 13(4):455–485.

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

Tumbe, C. (2012). Migration persistence across twentieth century India. *Migration and Development*, 1(1):87–112.

Tumbe, C. (2018). *India moving: A history of migration*. Penguin Random House India Private Limited.

Tumbe, C. (2025). The Internal and International Diasporas of India. *Forthcoming, Sociological Bulletin*.

Visaria, L. and Visaria, P. (1983). V - population (1757-1947). In Kumar, D. and Desai, M., editors, *The Cambridge Economic History of India: Volume 2 c. 1757-c. 1970*, pages 463–532. Cambridge University Press.

Walter, A. (2025). The Extended Ethnographic Atlas. *Working Paper*.

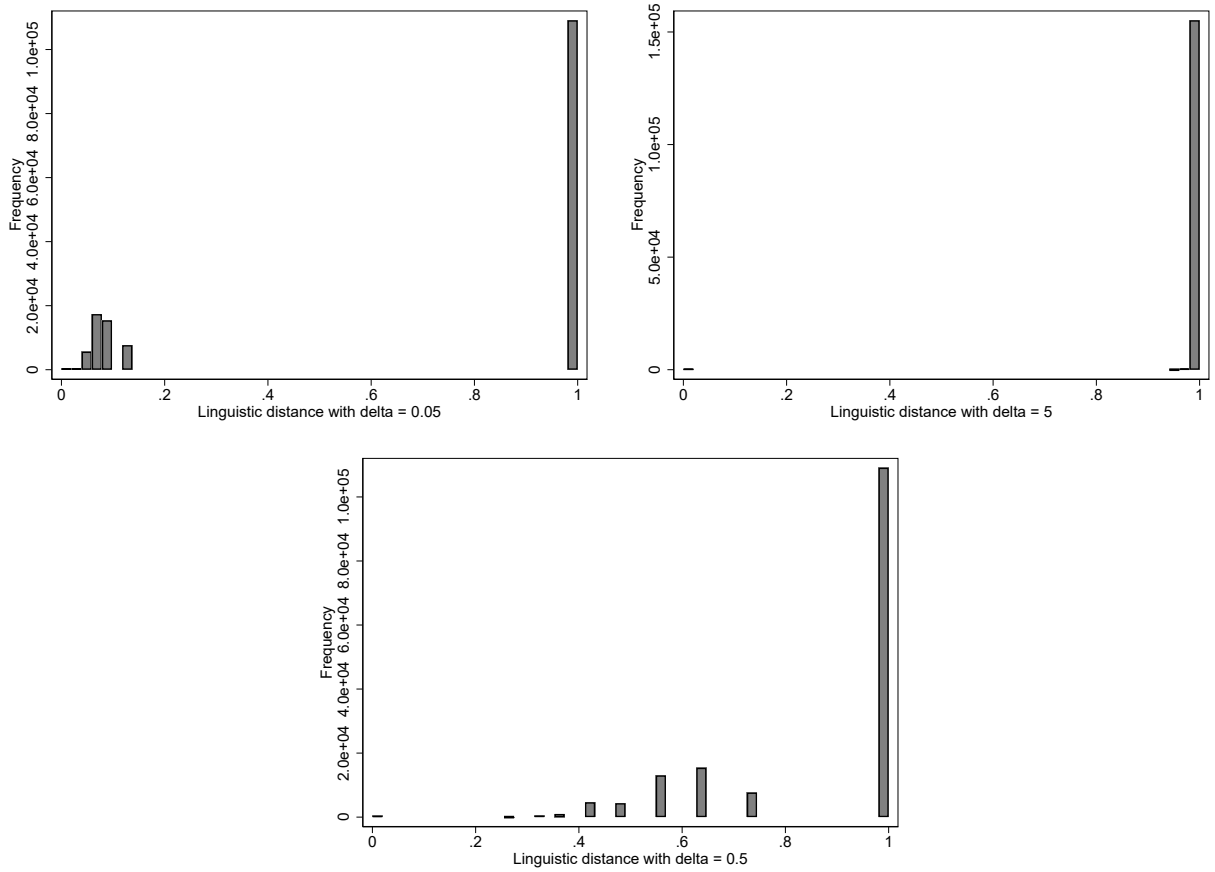
Wang, Y. (2025). Cultural Distance and Internal Migration: Evidence from Indonesia. *Forthcoming, Economic Development and Cultural Change*.

Wichmann, S., Holman, E. W., and Brown, C. H. (2022). The asjp database (version 20).

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters*, 117(3):794–798.

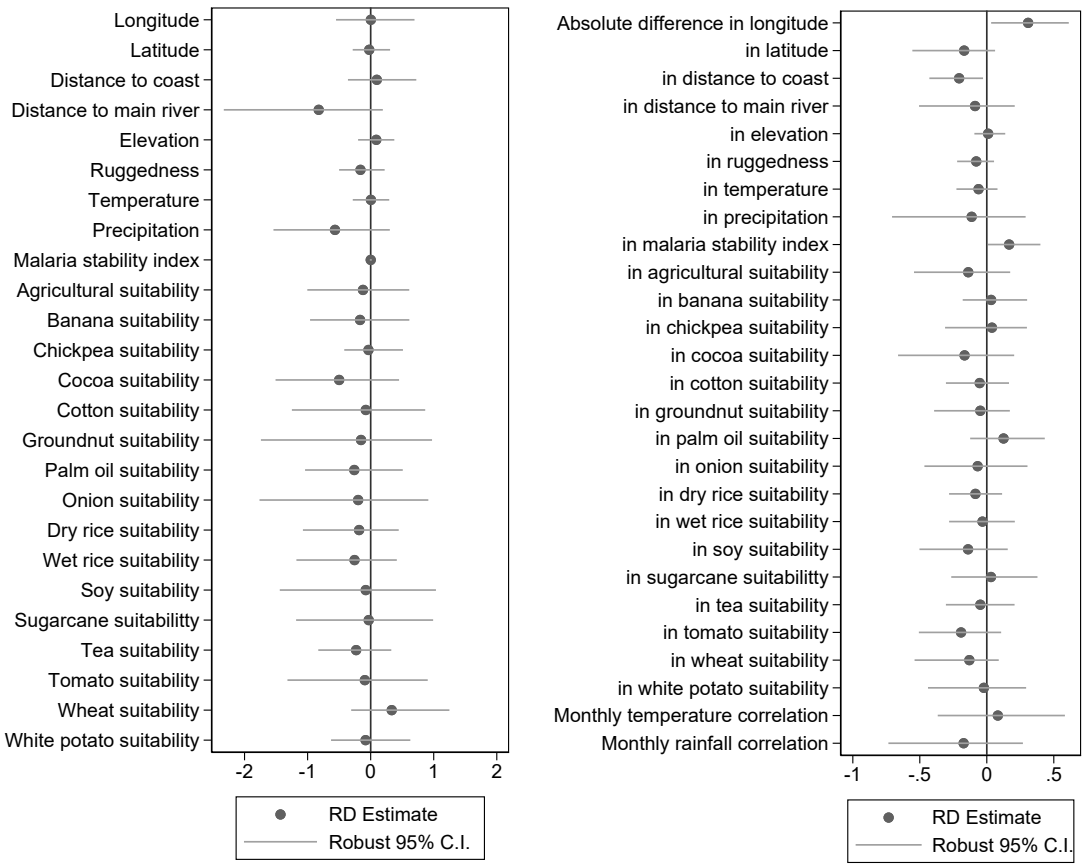
Appendix A. Appendix Figures

Figure A.1. Distribution of linguistic distance between all pairs of languages in India using different values of δ



We consider all 395 languages present in the 1901, 1921, 1931, 2001 and 2011 censuses, harmonized and matched with the *Ethnologue* dataset. Each observation is a pair of languages, hence there are $395^2 = 156,025$ observations. The linguistic distance between two languages m and n is $d_{mn} = 1 - \left(\frac{SharedBranches_{mn}}{15}\right)^{\delta}$. In the first panel, $\delta = 0.05$. In the second panel, $\delta = 5$. In the second panel, $\delta = 0.5$

FIGURE A.2. RD Balance Tests, 1901: Discontinuities in Geographic Characteristics

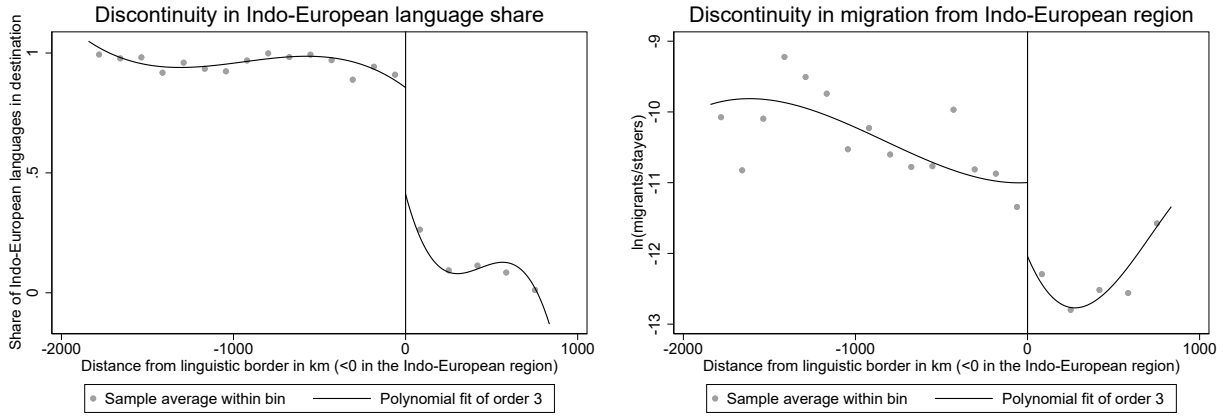


We estimate robust Calonico et al. (2014) 95% confidence interval.

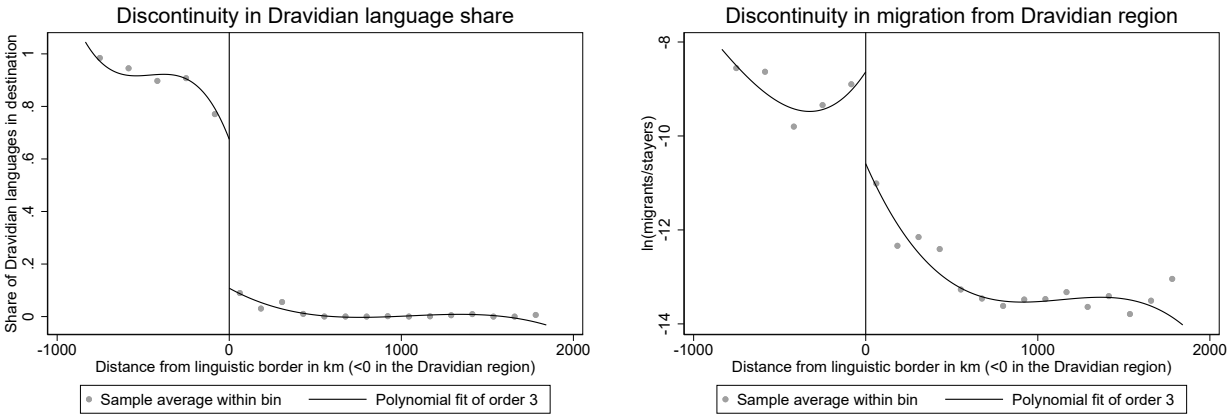
In Panel A, we collapse our bilateral data to the district level and estimate a separate discontinuity for each of the 25 geographic district characteristics, normalized to unit standard deviation. The forcing variable is the distance from the district to the Indo-European/Dravidian border, negative in the Indo-European region, and positive in the Dravidian region. The treatment variable is an indicator for being in the Dravidian region. In Panel B, we use the full bilateral dataset, combining the two border experiments, and estimate a separate discontinuity for each of 27 bilateral controls (normalized to have unit standard deviation): the absolute difference between *od* in all geographic characteristics of Panel A, as well as the correlation in monthly rainfall and temperature in the origin and destination over the 20th century. The forcing variable is the distance from the destination district to the Indo-European/Dravidian border, negative if origin and destination are in the same linguistic region, positive if the destination is in a different linguistic region. The treatment variable is a binary for origin and destination belonging to different linguistic regions. In both panels, discontinuities are estimated using a local linear non-parametric function of distance to the border, with a triangular kernel and MSE optimal bandwidth. In Panel B, we also control for border experiment fixed effects and an indicator if *od* are in the same province.

FIGURE A.3. Discontinuities Indo-European/Dravidian Border, 1901 – Separate Border Experiments

Panel A: origin district in the Indo-European region



Panel B: origin district in the Dravidian region



This figure displays graphical representations of the discontinuity in (left) the share of residents of the destination district speaking a language belonging to the same family as the origin region, and (right) the natural logarithm of the ratio of the number of migrants (born in the origin and living in the destination in 1901) to the number of stayers (born in the origin and living in the origin in 1901). We add 1 to the numerator and denominator, following Adsera and Pytlikova (2015). The border between the Indo-European and Dravidian linguistic regions is represented by the vertical line at 0. On the left of the vertical line are district pairs with an origin in the origin region and a destination in the origin region, averaged in bins of distance to the border. On the right of the vertical line are district pairs with an origin in the origin region and a destination in the treated region (belonging to a different language family), averaged in bins of distance to the border. In Panel A, the whole sample is represented, and the black line is a polynomial of order 3, estimated separately on each side of the border. The polynomial is purely indicative, as in our estimation we use a local linear nonparametric function, rather than a polynomial. In panel B, we zoom in on the border, considering only the observations within a data driven bandwidth on the right and left of the border. We estimate a linear fit with 95% confidence interval on each side of the border. This is more similar to our main specification, though the figure excludes the triangular kernel that gives more weight to observations closer to the border.

Appendix B. Appendix Tables

Table B.1. Summary Statistics, 2001

	Bilateral dataset 2001 census				
	Obs.	Mean	St. Dev.	Min	Max
Migration variables					
Number migrants (residing in *d*, living in *o* 10 years before)	335,820	119	1,570	0	609,956
Female migrants	335,820	69	866	0	297,649
Male migrants	335,820	49	748	0	312,307
Any migration indicator	335,820	0.50	0.50	0	1
(Migrants from *o* / Stayers in *o*)×100	335,820	0.01	0.10	0.00	27.87
ln((Migrants from *o* +1)/(Stayers in *o* +1))	335,820	-12.57	1.90	-16.05	-1.28
Language variables					
Linguistic distance (majority languages)	335,820	0.69	0.30	0	1
Linguistic distance (all languages)	335,820	0.68	0.25	0	1
Indicator same majority language in *o* & *d*	335,820	0.08	0.27	0	1
Location variables					
Indicator if *od* are in same state	335,820	0.05	0.22	0	1
Indicator *od* are within post-1947 Indian borders	335,820	1.00	0.00	1	1
Indicator if *o* is a state	335,820	0.00	0.00	0	0
Distance variables					
od neighbors	335,820	0.01	0.10	0	1
Geodesic distance between *od* (km)	335,820	1,112	591	9	2,979
ln distance	335,820	6.82	0.70	2.15	8.00
Travel time walking and sailing between *od* (hours)	335,820	265	124	3	616
Travel time with rail and river between *od* (hours)	335,820	128	62	1	364

The unit of observation is an *od* pair where *o* refers to origin district and *d* refers to destination district.

TABLE B.2. Fuzzy RD, 1901

	(1)	(2)	(3)	(4)
	Panel A. First Stage			
	Linguistic distance	Linguistic distance	Linguistic distance	Linguistic distance
RD Estimate	0.573***	0.562***	0.585***	0.590***
	(0.016)	(0.017)	(0.011)	(0.010)
	[0.545,0.616]	[0.535,0.611]	[0.561,0.612]	[0.580,0.626]
	Panel B. Second Stage			
	ln(migrants /stayers)	ln(migrants /stayers)	ln(migrants /stayers)	ln(migrants /stayers)
Linguistic distance	-1.591***	-1.204**	-0.743**	-0.700***
	(0.491)	(0.501)	(0.328)	(0.093)
	[-2.640,-0.405]	[-2.423,-0.135]	[-1.541,-0.008]	[-1.002,-0.474]
Observations	36,985	36,985	36,985	36,985
Bandwidth (km)	348	348	348	348
Border experiment F.E.	✓	✓	✓	✓
Same province dummy	✓	✓	✓	✓
Distance controls		✓	✓	✓
Origin geog. controls			✓	
Destination geog. controls			✓	
Bilateral geog. controls			✓	✓
Origin F.E.				✓
Destination F.E.				✓

The linguistic border discontinuity is used as an instrument to estimate the effect of linguistic distance on migration. Panel A displays the results of the first stage: border discontinuities in the linguistic distance between origin and destination. Linguistic distance is the cladistic distance between the majority languages of the origin and destination. A linguistic distance of zero means the majority languages are the same. A linguistic distance of one means the majority languages belong to different language families. Panel B displays the second stage: estimates of the effect of a unit increase in linguistic distance on migration. The dependent variable is the natural logarithm of the ratio of the number of migrants born in the origin and living in the destination in 1901 to the number of stayers born in the origin and living in the origin in 1901. We add 1 to the numerator and the denominator, following Adsera and Pytlikova (2015). An observation is a pair composed of a destination district (or Princely State) and an origin. The origin can be either a district or a province that aggregates several districts. We combine the data of two border experiments. In the first one, we consider all origins in the Indo-European region (the origin region) and all destinations in both the Indo-European region and the Dravidian region (the treated region). In the second experiment, Dravidian is the origin region and Indo-European is the treated region.The forcing variable is the distance from the destination district to the Indo-European/Dravidian border, which is negative in the origin region and positive in the treated region. Discontinuities are estimated using a local linear non-parametric function of distance to the linguistic border, with a triangular kernel and MSE optimal bandwidth. To facilitate the comparison of coefficients across columns, we compute the optimal bandwidth with only the same province dummy and border experiment fixed effects as covariates. We then keep the same bandwidth in the specifications of column (2) to (4). Distance controls are the natural logarithm of geodesic distance between origin and destination and a binary variable equal to 1 if the origin and destination districts are neighbors. The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.Standard errors clustered by destination district in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Robust Calonico et al. (2014) 95% confidence interval in [ ].

TABLE B.3. Alternate Samples: Gravity Model Results 1901 and 2001

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of Migrants					
	2001 Indian Borders		Excluding regions affected by Partition		Same geographical units in 1901 & 2011	
	1901	2001	1901	2001	1901	2001
Linguistic distance	-0.402***	-0.443***	-0.593***	-0.269	-0.495***	-0.608***
	(0.127)	(0.099)	(0.181)	(0.186)	(0.184)	(0.175)
Indicator, neighboring *od*	1.301***	1.129***	1.198***	0.995***	1.422***	1.390***
	(0.150)	(0.092)	(0.157)	(0.124)	(0.134)	(0.134)
ln distance	-1.531***	-1.288***	-1.405***	-1.400***	-1.322***	-1.078***
	(0.157)	(0.140)	(0.181)	(0.171)	(0.193)	(0.182)
Observations	30,019	325,470	8,652	90,902	27,082	27,082
Same state/province dummy	✓	✓	✓	✓		
Bilateral geog. controls	✓	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

In columns (1) and (2), we focus on the districts of present-day India, excluding from the 1901 sample origins and destinations belonging to present-day Pakistan, Bangladesh, and Myanmar/Burma. We also exclude from the 2001 sample the former French and Portuguese possessions that were not part of British India in 1901 (the states of Dadra and Nagar Haveli, Daman and Diu, Goa, and Pondicherry).

In Column (3) we restrict the sample to to Ajmer-Merwara, Berar, Central India, Central Provinces, Cochin, Coorg, Gwalior, Hyderabad, Madras, Mysore, North-Western Provinces and Oudh, and Travancore. In column (4), we restrict the sample to Andhra Pradesh, Chhattisgarh, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Orissa, Tamil Nadu, Uttar Pradesh, and Uttaranchal.

In Columns (5) and (6), we build a bilateral dataset of origin-destination pairs composed of the exact same geographical units in 1901 and 2001. We use a graph theory-based algorithm to aggregate districts and create a set of consistent geographical units for which we have migration data in both 1901 and 2001.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

TABLE B.4. Border RD along the Indo-European and Dravidian Border, 2001

	(1)	(2)	(3)	(4)	(5)
			ln(migrants/stayers)		
RD Estimate	-0.452**	-0.498**	-0.461***	-0.524***	-0.219***
	(0.212)	(0.238)	(0.139)	(0.046)	(0.046)
	[-0.893,0.099]	[-1.053,0.049]	[-0.771,-0.116]	[-0.652,-0.389]	[-0.333,-0.071]
Observations	266,966	266,966	266,966	266,966	266,966
Bandwidth (km)	296	296	296	296	296
Border experiment F.E.	✓	✓	✓	✓	✓
Distance controls		✓	✓	✓	✓
Origin geog. controls			✓		
Destination geog. controls			✓		
Bilateral geog. controls			✓	✓	✓
Origin F.E.				✓	✓
Destination F.E.				✓	✓
Same state dummy					✓

Standard errors clustered by destination district in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Robust Calonico et al. (2014) 95% confidence interval in [ ]. We combine the data of two border experiments. In the first one, we consider all origins in the Indo-European region (the origin region) and all destinations in both the Indo-European region and the Dravidian region (the treated region). In the second experiment, Dravidian is the origin region and Indo-European is the treated region.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20$^{\text{th}}$ century.

TABLE B.5. Gravity Model, 1901 – Data Organization Check

	(1)	(2)	(3)	(4)	(5)
			Number of Migrants		
	Baseline	All aggregates distributed	Data reported	Observed district to district migration	Within province migration
Linguistic distance	-0.474***	-0.473***	-0.455***	-0.418***	-0.340***
	(0.132)	(0.115)	(0.134)	(0.118)	(0.116)
ln distance	-1.432***	-1.260***	-1.375***	-1.283***	-1.201***
	(0.150)	(0.125)	(0.153)	(0.150)	(0.151)
Indicator if *od* are neighbors	1.269***	1.352***	1.349***	1.639***	1.533***
	(0.129)	(0.102)	(0.116)	(0.094)	(0.079)
Observations	51,058	139,561	25,771	20,581	13,366
Pseudo R-squared	.86	.84	.84	.87	.88
Same province dummy	✓	✓	✓	✓	
Origin F.E.	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20$^{\text{th}}$ century.

Column (1) displays results from Table 2 column (4). In column (2), we distribute all aggregated place of birth to districts to obtain complete bilateral district-to-district migration. In column (3), we use the data as reported: some of the origins are residuals corresponding to not always contiguous collections of districts for which geographic and linguistic variables are computed by averaging. In column (4), the sample is composed only of observed district-to-district migration and we exclude observations where the origin is a province and where the origin is a residual aggregate of several districts. In column (5), the sample is composed only of within-province district-to district-migration.

Table B.6. RD 1901 – Data Organization Check

	(1)	(2)	(3)	(4)
	ln(migrants/stayers)			
	Panel A. Baseline sample			
RD Estimate	-0.912***	-0.676**	-0.435**	-0.413***
	(0.281)	(0.283)	(0.194)	(0.056)
	[-1.525,-0.244]	[-1.378,-0.086]	[-0.907,-0.002]	[-0.603,-0.286]
Observations	36,985	36,985	36,985	36,985
Bandwidth (km)	348	348	348	348
	Panel B. All aggregates distributed			
RD Estimate	-0.364	-0.264	-0.334**	-0.226***
	(0.252)	(0.204)	(0.150)	(0.071)
	[-0.862,0.272]	[-0.704,0.224]	[-0.690,0.042]	[-0.318,0.032]
Observations	101,790	101,790	101,790	101,790
Bandwidth (km)	391	391	391	391
	Panel C. Data as reported			
RD Estimate	-0.858***	-0.291	-0.532**	-0.488***
	(0.302)	(0.255)	(0.210)	(0.063)
	[-1.490,-0.082]	[-0.918,0.264]	[-1.045,-0.088]	[-0.657,-0.333]
Observations	12,168	12,168	12,168	12,168
Bandwidth (km)	594	594	594	594
	Panel D. Only observed district-to-district migration			
RD Estimate	-0.794**	-0.031	-0.262	-0.175***
	(0.369)	(0.270)	(0.203)	(0.041)
	[-1.632,0.093]	[-0.744,0.505]	[-0.788,0.136]	[-0.341,-0.107]
Observations	8,475	8,475	8,475	8,475
Bandwidth (km)	555	555	555	555
Same province dummy	✓	✓	✓	✓
Border experiment F.E.	✓	✓	✓	✓
Distance controls		✓	✓	✓
Origin geog. controls			✓	
Destination geog. controls			✓	
Bilateral geog. controls			✓	✓
Origin F.E.				✓
Destination F.E.				✓

Standard errors clustered by destination district in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Robust Calonico et al. (2014) 95% confidence interval in [ ]. We combine the data of two border experiments. In the first one, we consider all origins in the Indo-European region (the origin region) and all destinations in both the Indo-European region and the Dravidian region (the treated region). In the second experiment, Dravidian is the origin region and Indo-European is the treated region.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

The sample of origin-destination pairs varies in each panel. Panel A displays results from the baseline sample in Table 3. In Panel B, we distribute all aggregated place of birth to districts to obtain complete bilateral district-to-district migration. In Panel C, we use the data as is: some of the origins are residuals corresponding to not always contiguous collections of districts for which geographic and linguistic variables are computed by averaging. In Panel D, the sample is composed only of observed district-to-district migration; we discard observations where the origin is a province and where the origin is a residual aggregate of several districts.

Table B.7. Alternate Measures of Distance and Estimators, 1901

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of Migrants					ln((migrants+1) /(stayers+1))	Indicator, Any Migration
Linguistic distance (majority languages)	-0.474*** (0.132)			-0.482*** (0.133)	-0.485*** (0.132)	-0.713*** (0.127)	-0.073*** (0.024)
Linguistic distance (all languages)		-1.588*** (0.263)					
Lexicostatistical distance (majority languages)			-0.340*** (0.129)				
Indicator, neighboring *od*	1.269*** (0.129)	1.219*** (0.126)	1.433*** (0.114)	1.293*** (0.123)	1.595*** (0.115)	2.056*** (0.106)	-0.109*** (0.018)
ln distance	-1.432*** (0.150)	-1.418*** (0.145)	-1.334*** (0.166)			-1.881*** (0.099)	-0.221*** (0.021)
ln travel time walking & sailing				-1.421*** (0.135)			
ln travel time rail & river					-1.014*** (0.120)		
Observations	51,058	51,058	31,850	51,058	51,058	51,058	51,058
(Pseudo) R-squared	.86	.86	.88	.86	.85	.78	.54
Same province dummy	✓	✓	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

In columns (1) to (5), we estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

Column (1) displays the baseline model of Table 2 column (5). In column (2), we use a measure of linguistic distance accounting for the entire language distribution in *o* and *d*: the expected cladistic distance between the mother tongues of a random individual from *o* and a random individual from *d*. In column (3), we use lexicostatistical linguistic distance between the majority languages of *o* and *d*, computed using word list data from the Automatic Similarity Judgment Program. Lexicostatistical distance takes values between 0 and 1.

In column (4), we replace geodesic distance between *o* and *d* by travel time walking or sailing, in hours, computed using the Özak (2010, 2018) Human Mobility Index (HMI). We use the CostDistance tool in ArcGIS to compute the least cost path between the centroids of *o* and *d*, using the HMI raster as the cost raster. In column (5), we control for travel time allowing for railway and river travel. We add to the HMI raster pixels along the 1901 railways and along navigable rivers. We use a speed of 50km per hour for railroads and of 3.3 km per hours for navigable rivers (Donaldson, 2018).

In columns (6) and (7), the model is estimated by OLS. In column (6), the dependent variable is the logarithm of the ratio of the number of migrants (born in the origin and living in the destination in 1901) on the number of stayers (born in the origin and living in the origin in 1901). This logarithm is not defined for origin-pairs with zero migration (38% of observations). Following Adsera and Pytlikova (2015), we add 1 to the number of migrants and to the number of stayers. In columns (7), we focus on the intensive margin of migration: the dependent variable is a binary variable equal to 1 if there is any migration between the origin and destination.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20$^{\text{th}}$ century.

TABLE B.8. Alternate Measures of Distance and Estimators, 2001

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of Migrants					ln((migrants+1) /(stayers+1))	Indicator, Any Migration
Linguistic distance (majority languages)	-0.442*** (0.098)			-0.457*** (0.096)	-0.361*** (0.101)	-0.377*** (0.055)	-0.092*** (0.012)
Linguistic distance (all languages)		-1.557*** (0.174)					
Lexicostatistical distance (majority languages)			-0.411*** (0.135)				
Indicator, neghboring *od*	1.132*** (0.094)	1.104*** (0.088)	1.132*** (0.110)	1.102*** (0.086)	1.516*** (0.075)	1.798*** (0.065)	-0.238*** (0.011)
ln distance	-1.282*** (0.141)	-1.310*** (0.133)	-1.142*** (0.159)			-1.637*** (0.051)	-0.220*** (0.011)
ln travel time walking & sailing				-1.359*** (0.117)			
ln travel time rail & river					-0.805*** (0.070)		
Observations	335,820	335,820	199,228	335,820	335,820	335,820	335,820
(Pseudo) R-squared	.88	.88	.89	.88	.87	.73	.47
Same state dummy	✓	✓	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

In columns (1) to (5), we estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

Column (1) displays the baseline model of Table 4 column (5). In column (2), we use a measure of linguistic distance accounting for the entire language distribution in *o* and *d*: the expected cladistic distance between the mother tongues of a random individual from *o* and a random individual from *d*. In column (3), we use lexicostatistical linguistic distance between the majority languages of *o* and *d*, computed using word list data from the Automatic Similarity Judgment Program. Lexicostatistical distance takes values between 0 and 1.

In column (4), we replace geodesic distance between *o* and *d* by travel time walking or sailing, in hours, computed using the Özak (2010, 2018) Human Mobility Index (HMI). We use the CostDistance tool in ArcGIS to compute the least cost path between the centroids of *o* and *d*, using the HMI raster as the cost raster. In column (5), we control for travel time allowing for railway and river travel. We add to the HMI raster pixels along the 2001 railways and along navigable rivers. We use a speed of 50km per hour for railroads and of 3.3 km per hours for navigable rivers (Donaldson, 2018).

In columns (6) and (7), the model is estimated by OLS. In column (6), the dependent variable is the logarithm of the ratio of the number of migrants (born in the origin and living in the destination in 2001) on the number of stayers (born in the origin and living in the origin in 2001). This logarithm is not defined for origin-pairs with zero migration (50% of observations). Following Adsera and Pytlikova (2015), we add 1 to the number of migrants and to the number of stayers. In columns (7), we focus on the intensive margin of migration: the dependent variable is a binary variable equal to 1 if there is any migration between the origin and destination.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

TABLE B.9. Gravity Model, 1921 and 2011

	(1)	(2)	(3)	(4)
	Number of Migrants			
	1921		2011	
Linguistic distance	-0.339	-0.505***	-1.753***	-0.583***
	(0.251)	(0.122)	(0.344)	(0.113)
ln distance	-0.949***	-1.652***	-0.870***	-1.635***
	(0.104)	(0.173)	(0.095)	(0.137)
Indicator, neighboring *od*	1.693***	1.082***	1.510***	0.981***
	(0.206)	(0.108)	(0.122)	(0.084)
Observations	62,085	62,085	38,544	38,544
Pseudo R-squared	.47	.86	.4	.89
Same province/state dummy	✓	✓	✓	✓
Bilateral geog. controls		✓		✓
Origin F.E.		✓		✓
Destination F.E.		✓		✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

In columns (1) and (2), we use 1921 census data. In columns (2) and (3), we use 2011 census data, which only gives population by districts of birth within a state, and by state of birth for individuals born in other states.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

Table B.10. Cultural Channels, 1901

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of Migrants						
Linguistic distance, z-score (all languages)	-0.364*** (0.072)	-0.263*** (0.065)	-0.398*** (0.084)	-0.338*** (0.075)			
Caste distance, z-score		-0.325*** (0.081)					
Religious distance, z-score				-0.400*** (0.149)			
Linguistic distance, z-score (majority languages)					-0.119*** (0.040)	-0.258*** (0.097)	-0.224*** (0.079)
Ethnographic distance, z-score (% of EA items different)						0.156 (0.096)	
Ethnographic distance, z-score (PCA index)							0.146 (0.090)
Indicator, neighboring *od*	1.370*** (0.111)	1.356*** (0.109)	1.377*** (0.111)	1.371*** (0.107)	1.367*** (0.113)	1.366*** (0.113)	1.367*** (0.113)
ln distance	-1.350*** (0.148)	-1.334*** (0.144)	-1.319*** (0.182)	-1.337*** (0.173)	-1.354*** (0.156)	-1.373*** (0.157)	-1.364*** (0.157)
Observations	42,646	42,646	22,164	22,164	43,223	43,223	43,223
Pseudo R-squared	.86	.86	.88	.88	.87	.87	.87
Same province dummy	✓	✓	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

In column (1), we replicate the results of Table B.7, column (1) on the sample for which we know caste distance. In column (2), we control for caste distance: 1 minus the probability that a random individual from o and a random individual from d belong to the same caste (jati). We normalize linguistic and caste distances to have unit variance.

In column (3), we replicate the results of Table B.7, column (1) on the sample for which we know religious distance. In column (4), we control for religious distance: 1 minus the probability that a random individual from o and a random individual from d have the same religion. We normalize linguistic and religious distances to have unit variance.

In column (5), we replicate the results of Table 2, column (5) on the sample for which we know ethnographic distance. In columns (6) and (7), we control for two different measures of ethnographic distance between the majority groups of o and d, built using the extended *Ethnographic Atlas* of Walter (2025). In column (6), we control for the percentage of *Ethnographic Atlas* items that are different between the two groups (among items non-missing for both groups). In column (7), we use an index built by principal component analysis on a vector of variables indicating whether the groups agree or differ on each item.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20$^{\text{th}}$ century.

Table B.11. Cultural Channels, 2001

	(1)	(2)	(3)	(4)	(5)
	Number of Migrants				
Linguistic distance, z-score (all languages)	-0.396***	-0.415***			
	(0.044)	(0.045)			
Religious distance, z-score		0.224***			
		(0.083)			
Linguistic distance, z-score (majority languages)			-0.141***	-0.117*	-0.101
			(0.035)	(0.067)	(0.062)
Ethnographic distance, z-score (% of EA items different)				-0.023	
				(0.044)	
Ethnographic distance, z-score (PCA index)					-0.048
					(0.046)
Indicator, neighboring *od*	1.104***	1.106***	1.172***	1.172***	1.172***
	(0.088)	(0.088)	(0.102)	(0.103)	(0.102)
ln distance	-1.310***	-1.314***	-1.203***	-1.202***	-1.202***
	(0.133)	(0.132)	(0.150)	(0.150)	(0.150)
Observations	335,820	335,820	275,902	275,902	275,902
Pseudo R-squared	.88	.88	.88	.88	.88
Same province dummy					
Origin F.E.	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

In column (1), we replicate the results of Table B.8, column (1) on the sample for which we know religious distance. In column (2), we control for religious distance: 1 minus the probability that a random individual from o and a random individual from d have the same religion. We normalize linguistic and religious distances to have unit variance.

In column (3), we replicate the results of Table 4, column (5) on the sample for which we know ethnographic distance. In columns (4) and (5), we control for two different measures of ethnographic distance between the majority groups of o and d, built using the extended *Ethnographic Atlas* of Walter (2025). In column (4), we control for the percentage of *Ethnographic Atlas* items that are different between the two groups (among items non-missing for both groups). In column (5), we use an index built by principal component analysis on a vector of variables indicating whether the groups agree or differ on each item.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

TABLE B.12. Additional Results on Religion, 1901 & 2001

	(1)	(2)	(3)	(4)
	Number of Migrants			
	1901		2001	
Linguistic distance, z-score (majority languages)	-0.159***	-0.157***	-0.134***	-0.138***
	(0.047)	(0.047)	(0.030)	(0.030)
Indicator, *od* have different majority religions		-0.069		0.284**
		(0.121)		(0.117)
Indicator, neighboring *od*	1.429***	1.429***	1.132***	1.131***
	(0.116)	(0.116)	(0.094)	(0.094)
ln distance	-1.329***	-1.329***	-1.282***	-1.283***
	(0.194)	(0.194)	(0.141)	(0.141)
Observations	22,164	22,164	335,820	335,820
Pseudo R-squared	.87	.87	.88	.88
Same province/state dummy	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

TABLE B.13. Gravity Model, Flows and Stocks, 1983 Survey, One-Year Window

	(1)	(2)
	Number of Migrants last year	
Linguistic distance	-0.296	-0.270
	(0.240)	(0.238)
Indicator, neighboring *od*	1.099***	1.058***
	(0.136)	(0.127)
ln distance	-0.380**	-0.330**
	(0.162)	(0.165)
ln stock of migrants		0.031
as share of *o* pop		(0.031)
Observations	62,388	62,388
Pseudo R-squared	.67	.67
Same state dummy	✓	✓
Origin F.E.	✓	✓
Destination F.E.	✓	✓
Bilateral geog. controls	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

The dependent variable is the number of migrants residing in d at the date of the survey who came from o in the previous year. The stock of migrants is the number of migrants living in d at the date of the survey who came from o more than a year before the date of the survey.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

TABLE B.14. Gravity Model: Role of Education, 1901& 1983

	(1)	(2)	(3)	(4)	(5)
	Number of Migrants				
	1901		Migrants in last 10 years, 1983		
	o low literacy	*o* high literacy	illiterate	up to primary	secondary or higher
Linguistic distance	-0.426**	-0.766***	-0.338*	-0.430***	-0.420**
	(0.200)	(0.218)	(0.186)	(0.164)	(0.212)
Indicator, neighboring *od*	1.257***	1.643***	0.930***	0.705***	0.573***
	(0.130)	(0.166)	(0.101)	(0.088)	(0.118)
ln distance	-1.969***	-0.712***	-0.654***	-0.703***	-0.392***
	(0.175)	(0.206)	(0.147)	(0.126)	(0.136)
ln stock migrants as share of *o* pop.			0.093***	0.046***	-0.004
			(0.014)	(0.012)	(0.013)
Observations	17,881	17,531	69,428	64,633	62,066
Pseudo R-squared	0.89	0.89	0.77	0.75	0.72
Same province dummy	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.

In columns (1) and (2), the migration data comes from the 1901 census, and the dependent variable is the number of migrants (born in *o* and living in *d* in 1901). In column (1), we restrict the sample to origins with above median literacy (using the literacy data from Chaudhary and Fenske (2023b)). In column (2) we restrict the sample to origins with below median literacy.

In columns (3) to (5), the migration data comes from the 1983 employment survey conducted by the National Sample Survey Office (NSSO). The dependent variable is the number of migrants of a certain education level residing in *d* at the date of the survey who came from *o* within the previous 10 years. In column (3), we consider only migrants with no education, in column (4) migrants with primary education, and in column (5) migrants with secondary education or higher. The stock of migrants is the number of migrants living in *d* at the date of the survey who arrived from *o* more than 10 years before the date of the survey.

The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

Table B.15. Scheduled Languages, 2001

	(1)	(2)	(3)	(4)	(5)
	Number of Migrants				
Linguistic distance	-0.442***		-0.481***		
	(0.098)		(0.083)		
Distance scheduled languages		-0.213	0.091		
		(0.161)	(0.165)		
Linguistic distance × *od* no common scheduled language				-0.578**	-0.576**
				(0.225)	(0.225)
Linguistic distance × *od* share scheduled language				-0.435***	
				(0.087)	
Linguistic distance × *od* different states, common scheduled language					-0.390***
					(0.132)
Linguistic distance × *od* in the same state					-0.464***
					(0.106)
Indicator *od* share scheduled language				-0.154	-0.172
				(0.130)	(0.137)
Indicator, neighboring *od*	1.132***	1.135***	1.131***	1.129***	1.128***
	(0.094)	(0.095)	(0.093)	(0.092)	(0.092)
ln distance	-1.282***	-1.290***	-1.279***	-1.284***	-1.284***
	(0.141)	(0.145)	(0.144)	(0.142)	(0.143)
Observations	335,820	335,820	335,820	335,820	335,820
Pseudo R-squared	.88	.88	.88	.88	.88
Same state dummy	✓	✓	✓	✓	✓
Origin F.E.	✓	✓	✓	✓	✓
Destination F.E.	✓	✓	✓	✓	✓
Bilateral geog. controls	✓	✓	✓	✓	✓

Standard errors in parentheses clustered two ways by origin and destination districts.***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
We estimate the gravity model using Poisson pseudo-maximum likelihood estimator.
Distance between scheduled languages is the distance between the official or scheduled languages of o and d. For states that have several official languages, we compute the distance between all pairwise combinations of official languages and compute a simple average.
The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.

Appendix C. Regression Discontinuity with Finer Linguistic Regions

Our main RD specification exploits the linguistic border between Indo-European and Dravidian languages. In this appendix, we consider the borders between linguistic regions finer than the language family. Using the language composition of each district in 1901, we define language regions by aggregating languages in subfamilies 4 branches down from the root of the tree. A language region is a contiguous collection of districts speaking languages belonging to the same subfamily. The goal is to obtain linguistic regions large enough for a border discontinuity analysis to make sense, given that our unit of analysis is the district. Resulting linguistic regions are displayed in Figure C.1. We exclude regions composed of fewer than 4 districts (for example we do not consider the two regions speaking Gondi, because they each consist of a single district). We also exclude Sino-Tibetan regions, whose borders tend to coincide with geographical boundaries. We end up with 33 border experiments (ordered pairs of contiguous regions). For example in one border experiment, the sending region is the region speaking Kannada and the treated region is the region speaking Indo-Aryan outer languages to the north of the Kannada region (Marathi region). In the mirror border experiment, the sending region is the Marathi region and the treated region is the Kannada region.[66] To have enough statistical power, we aggregate the data of all experiments in 3 groups: the experiments comparing linguistic regions belonging to different families (Dravidian and Indo-European, 8 experiments), the ones comparing regions sharing 1 branch on the language tree (4 experiments), and the ones comparing regions sharing 4 branches on the language tree (21 experiments).[67] We report results from estimating equation (4) separately for each of these three groups.

Given the smaller sample sizes, and the aggregation of many different border experiments, results presented in Table C.1 are noisier and less stable across specifications that the main results of Table 3. However, we find that, overall, migration falls by more when we cross borders between more dissimilar linguistic regions. In Panel C, where we considers linguistic regions sharing 4 branches on the language tree, we estimate a negative border discontinuity, but it is only (faintly) statistically different from 0 in the last specification (with all the controls), and the magnitude is smaller than when considering linguistic borders between language families (-0.205: crossing a linguistic border decreases migration by $\exp(-0.205) - 1 = 19\%$).

[66] Only one border experiment has no mirror experiment because in this mirror experiment there is only one origin in the sending region, aggregated at the province level.

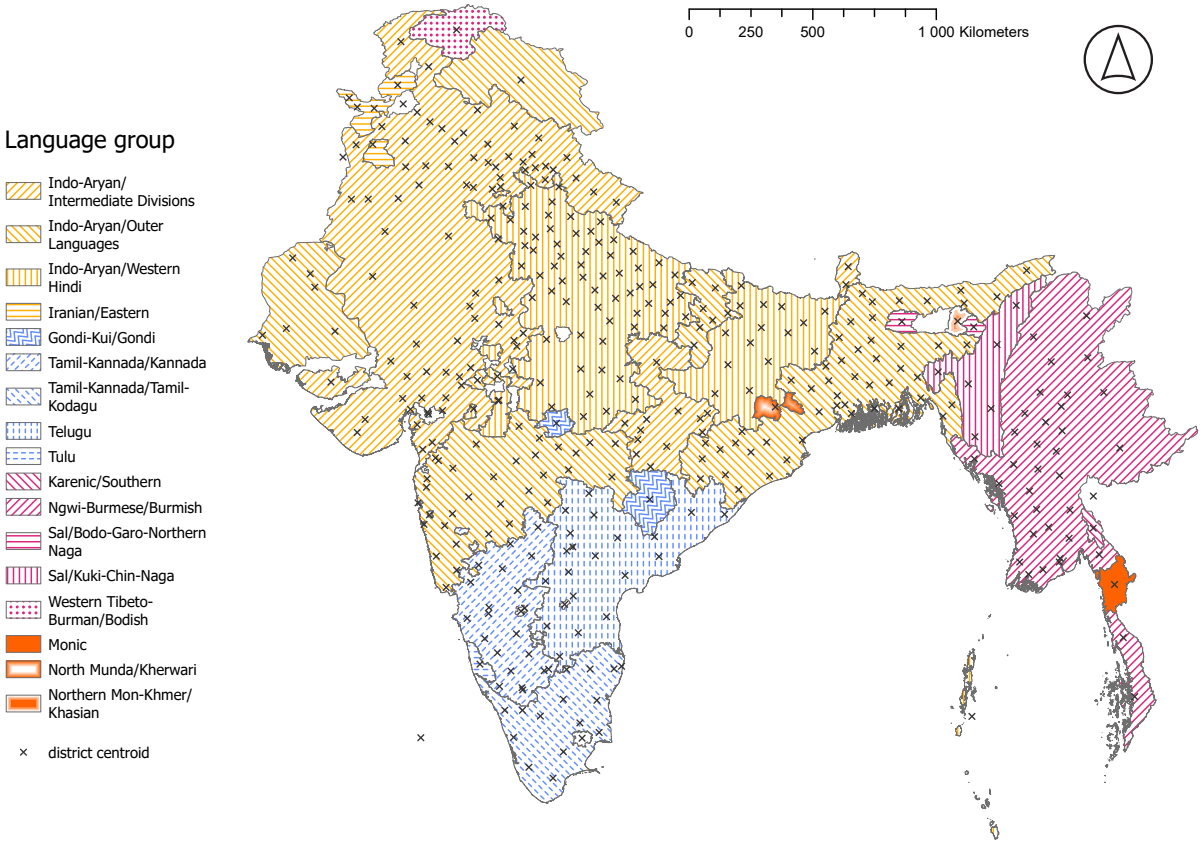
[67] Once we exclude very small linguistic regions and regions belonging to the Sino-Tibetan language family, there are no borders between linguistic regions sharing 2 or 3 branches on the language tree.

TABLE C.1. RD 1901 – Finer Linguistic Regions

	(1) ln(migrants /stayers)	(2) ln(migrants /stayers)	(3) ln(migrants /stayers)	(4) ln(migrants /stayers)
	Panel A. Linguistic regions belonging to different families (8 experiments)			
RD Estimate	-1.395***	-0.650**	-0.595***	-0.326***
	(0.394)	(0.282)	(0.179)	(0.091)
	[-2.189,-0.342]	[-1.282,0.015]	[-0.898,-0.019]	[-0.359,0.090]
Observations	3,685	3,685	3,685	3,685
Bandwidth (km)	391	391	391	391
	Panel B. Linguistic regions sharing 1 branch on the language tree (4 experiments)			
RD Estimate	-0.385	-0.120	-0.525***	-0.858***
	(0.581)	(0.535)	(0.196)	(0.173)
	[-1.785,1.071]	[-1.365,1.132]	[-2.600,1.309]	[-1.643,-0.376]
Observations	988	988	988	988
Bandwidth (km)	91	91	91	91
	Panel C. Linguistic regions sharing 4 branches on the language tree (21 experiments)			
RD Estimate	-0.049	-0.068	-0.134	-0.205*
	(0.322)	(0.304)	(0.215)	(0.114)
	[-0.662,0.828]	[-0.726,0.675]	[-0.638,0.351]	[-0.485,0.069]
Observations	31,151	31,151	31,151	31,151
Bandwidth (km)	177	177	177	177
Same province dummy	✓	✓	✓	✓
Border experiment F.E.	✓	✓	✓	✓
Distance controls		✓	✓	✓
Origin geog. controls			✓	✓
Destination geog. controls			✓	✓
Bilateral geog. controls			✓	✓
Origin F.E.				✓
Destination F.E.				✓

The dependent variable is the natural logarithm of the ratio of the number of migrants born in the origin and living in the destination in 1901 to the number of stayers born in the origin and living in the origin in 1901. We add 1 to the numerator and the denominator, following Adsera and Pytlikova (2015). An observation is a pair composed of a destination district (or Princely State) and an origin. The origin can be either a district or a province that aggregates several districts. A linguistic region is a contiguous collection of districts speaking languages belonging to the same subfamily 4 branches down from the root of the tree. A border experiment is an ordered couple (A, B) of contiguous linguistic regions. A is the origin region and B is the "treated" region. For each border experiment, we keep all district pairs having their origin in the origin region A and their destination either in the origin region A or the "treated" region B. Since the data of several experiments are aggregated, some observations (origin-destination pairs) appear twice. The forcing variable is the distance from the destination district to the linguistic border between regions A and B, which is negative in the origin region and positive in the treated region. Discontinuities are estimated using a local linear non-parametric function of distance to the linguistic border, with a triangular kernel and MSE optimal bandwidth. To facilitate the comparison of coefficients across columns, we compute the optimal bandwidth with only the same province dummy and border experiment fixed effects as covariates. We then keep the same bandwidth in the specifications of column (2) to (4). Distance controls are the natural logarithm of geodesic distance between origin and destination and a binary variable equal to 1 if the origin and destination districts are neighbors. The geographic controls are latitude and longitude, distance to the coast and to the nearest river, average elevation, ruggedness, temperature, precipitation, malaria stability index, agricultural suitability, suitability for 15 crops (banana, chickpeas, cocoa, cotton, groundnuts, palm oil, onion, dryland rice, wetland rice, soy, sugarcane, tea, tomato, wheat, white potato). The bilateral geographic controls are: (1) absolute differences between *od* in the geographic controls; (2) the correlation between *od* in monthly rainfall and temperature over the 20^{th} century.Standard errors clustered by destination district in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Robust Calonico et al. (2014) 95% confidence interval in [ ].

FIGURE C.1. Finer Linguistic Regions and Linguistic Borders in 1901



Linguistic regions are defined by aggregating languages in sub-families 4 branches down from the root of the tree. A linguistic region is a collection of contiguous districts whose majority languages belong to the same sub-family. Linguistic borders are the borders between two linguistic regions. Crosses represent the centroids of districts (used to compute distance to linguistic borders).