Common tongue: The impact of language on economic performance

Jain, Tarun

Indian School of Business

1 November 2011

Common tongue: The impact of language on economic performance

Tarun Jain[*]

Indian School of Business

November 1, 2011

Abstract

This paper investigates the impact of language on economic performance. I use the 1956 reorganization of Indian states on linguistic lines as a natural experiment to estimate the impact of speaking the majority language on educational and occupational outcomes. I find that districts that spoke the majority language of the state during colonial times enjoy persistent economic benefits, as evidenced by higher educational achievement and employment in communication intensive sectors. After reorganization, historically minority language districts experience greater growth in educational achievement, indicating that reassignment could reverse the impact of history.

Keywords: Language, Communication costs, Education, Occupational choice, Reorganization of Indian states.

JEL Codes: I20, N95, O15, O43.

[*] Finance, Economics and Public Policy Area, Indian School of Business, Hyderabad, India. Email: `tj9d@virginia.edu`. This paper benefitted from detailed feedback from Tanika Chakraborty, Krishna Kumar, Pushkar Maitra, Subha Mani, Prachi Mishra, Kaivan Munshi and Ashima Sood. I also thank seminar and conference participants at the ISB CAF Winter Research Camp, BITS Pilani-Hyderabad and the Econometric Society European Meetings for helpful comments on this paper. Urvashi Jain, Tilak Mukhopadhyay and Lakshmi Sripada provided excellent research assistance. All errors are my own.

1 Introduction

Many differences in economic performance and human welfare can be traced to variations in institutional characteristics of different regions. A large literature in economics examines the impact of historically determined institutions on the trajectory of economic welfare. Pioneered by North (1990), economists such as Acemoglu, Johnson, and Robinson (2001) and Banerjee and Iyer (2005) have established the role of colonial institutions on economic performance. Along with formal institutions, social institutions such as caste, religion and kinship are perhaps equally important in determining individual welfare in the modern era (Pande and Udry 2006; Roy 2002). However, the role of colonial practices in shaping social institutions remains an open research question. This paper examines the impact of colonial and language-based post-colonial political organization on economic outcomes and welfare in India.

Language is a social characteristic with near-universal importance in determining economic outcomes. Workers who are fluent in the dominant language of a region receive higher returns in the labor market than workers who are not. As a result, for instance, immigrants expend substantial effort to learn the language of their new home countries. In the United States and United Kingdom, new immigrants enjoy higher incomes if they are fluent in English (Chiswick 1991; Dustmann and Fabbri 2003). In parallel, one consequence of globalization is increasing returns to and demand for learning English (Munshi and Rosenzweig 2006).

The mid-century reorganization of states in India on linguistic lines offers an opportunity to examine the impact of language on economic outcomes. Each state consists of a number of administrative units called districts. If residents of a district speak, read and write the same language as a majority of the residents of the state, they might experience better economic outcomes if they incur lower transaction costs, leading to greater inter-district economic activity (Lang 1986).

A rich literature supports this premise. Lazear (1999) presents evidence from the United States that individuals from small, minority communities are more likely to adopt the language of the majority than individuals from larger minority groups, presumably because the former have greater incentive to do so. Similarly, Carliner (1981) documents wage differentials among Francophone and Anglophone men in Quebec in 1971 when most of the economy was controlled by Anglophones. He reports substantial rewards to English-speaking Francophones but no significant wage premium for French-speaking Anglophones. In the rest of Canada, the wage difference between Anglophone and Francophone workers was smaller than in Quebec. As a result, Francophone workers in Quebec responded by learning English, with little response among Anglophones. As Francophones began to increasingly control Quebec's economy in the 1970s and 1980s, the wage gap between Francophone and Anglophone workers narrowed (Albouy 2008).

Similar evidence emerges from developing countries. Clingingsmith (2007) argues that economic growth in sectors where communication is relatively important increases the incentive to learn new languages. Consequently, growth of the manufacturing sector increased bilingualism in mid-century India, especially among minority language speakers. The next generation spoke only the economically dominant language, increasing linguistic homogeneity in the long run. Angrist and Lavy (1997) demonstrate the economic impact of increasing transaction costs associated with language. They report that when Morocco changed the language of instruction in schools from French to Arabic in 1983, the returns to post-secondary education declined by one-half, primarily because most organized economic activity was conducted in French.

In this paper, I examine the impact of speaking the dominant language on economic performance in India using district level data on educational achievement and occupational choice. India is a particularly appropriate setting for this study since over 3,000 languages are used in the country, with 18 languages claiming both wide speakership as well as constitutional recognition. Modern Indian states correspond to the areas where these languages are in common use. Regulators, the judiciary and other arms of the state government mandate use of official languages, in addition to English, in their correspondence with citizens. Additionally, public schools offer instruction using the medium of the state language, with some states such as Gujarat and West Bengal doing so exclusively.[1]

Bivariate comparisons of communities that either speak or do not speak the dominant language might not yield unbiased estimates because of potential endogeneity in community formation. Migrants may live in ethnic enclaves where they do not need to learn the new language (Chiswick and Miller 2005a). Communities that recognize the link between speaking a minority language and poorer economic outcomes may try to form their own political units. For example, minorities in developed countries such as Belgium, Spain and Canada and in developing countries such as Sri Lanka and Cameroon launched separatist movements based on language.[2] Thus, unbiased estimates require exogenous assignment of some com-

[1] Chakraborty and Kapur (2009) estimate the impact of introducing local language instruction on labor market outcomes in the state of West Bengal.

[2] The economic impact of such movements in Spain and Sri Lanka are discussed in Abadie and Gardeazabal (2003) and Abeyratne (2004), respectively.

munities to minority and majority languages.

To generate such estimates, I use the reorganization of Indian states in 1956 along strict linguistic lines as a natural experiment. Provincial boundaries in British India were determined either by the sequence of British military conquest, with provinces cobbled together as imperial rule extended from the coasts into the hinterland, or when the British decided to leave native rulers in place. Banerjee and Iyer (2005) discuss the sequence of colonial conquest and argue that the process was independent of linguistic concerns. In addition, the "Doctrine of Lapse" specified that princely states where the ruler died without a natural-born male heir would cede to direct colonial rule. Therefore, this policy determined the areas which would be ruled directly by the British (Iyer 2010). Since both these factors were exogenous to a district's dominant language, so was the provincial organization of British India.[3]

In 1955, the States Reorganization Commission recommended the reorganization of Indian states on strict linguistic lines, a principle which the central government followed while redrawing state boundaries in 1956, 1960, 1966 and 1971 (Govt. of India 1955). Table 1 shows the outcome of this principle. Despite half a century of migration, communication and integration, major languages in modern India are essentially spoken only in states formed by the districts that speak those languages.[4] Following from this, I classify each district as either a "majority" district, if before reorganization, it belonged to a state where the district's language was the majority language of the state, or a "minority" district, if it belonged to a state where the district's language was the minority language. After reorganization, new state boundaries encompassed those districts that shared a common language.

[3] Ban and Rao (2007) also use reorganization of Indian states as a source of exogenous variation.

[4] The major exception is Hindi/Urdu which is spoken all over Northern India.

Thus, a natural experiment is set up in the opposite sequence of a typical randomized control trial. Before the reorganization, each district was classified as either majority or minority by historical accident, and reassigned as majority after the reorganization.[5] The district-wise organization of the states in South India before and after 1956 is shown in Figures 1 and 2, respectively.

If native language (or "mother tongue") instruction facilitates schooling, I expect to find that majority districts should achieve persistently better educational outcomes compared to minority districts. If a common language decreases communication costs, I expect relatively lower shares of employment in the agricultural sector (where communication between workers is less important) compared to the manufacturing and services sectors in the majority districts. Growth rates for each of these measures should also be higher in minority districts after reorganization, as these districts "catch-up" after integrating into co-linguistic states.

Using a district level panel dataset based on the Census of India, I find support for the hypothesis that shared language within a state potentially lowers communication costs, increasing schooling and facilitating the switch to non-agricultural professions. The analysis shows that colonial mis-assignment for minority districts is associated with lower rates of educational achievement. The impact is greater on primary and secondary schooling, which in most states is conducted in the vernacular, than on college education, where the medium of instruction is typically English. In addition, the results show that workers in these districts are employed disproportionately in agriculture rather than communication intensive industries such as

[5] In a classic randomized control trial, subjects are drawn from a mixed sample and assigned to either a treatment or control group. In this case, districts were assigned to minority or majority status by the British, and then placed in a pure sample of majority-only states by reorganization.

commerce and transportation. The results are robust to a number of alternate specifications that use the linguistic distance and the degree of language polarization in the population to indicate minority status. I also find evidence that reassigning a minority district to the state where it is part of the linguistic majority can reverse the impact of history. Minority districts experienced greater growth in educational achievement after reorganization as they "caught up" with the previously majority districts. Workers in these districts shifted away from agriculture towards communication intensive industries such as manufacturing and transportation, although this result is not significant. Finally, I calculate the impact of historical language status on modern income differences between districts.

2 Historical Background

2.1 British Conquest of India, 1757-1857

In this section, I present the history of British conquest that determined the pre-1956 boundaries of Indian states. Direct and indirect British colonial rule in India lasted 190 years. British trading companies arrived in India in 1600, but historians date the political conquest of the Indian subcontinent from 1757, when the East India Company gained control over the province of Bengal in the Battle of Plassey. The areas around Bengal, i.e., Jalpaiguri, Darjeeling and Orissa, were integrated to form the Bengal Presidency with Calcutta as the capital.

In the next 90 years, the Company extended its control over the rest of India. It obtained feudal control over the Southern coast from the Nawab of Carnatic where it established the trading post of Fort St. George (later called Madras and now Chennai) in 1640. The geographically contiguous areas around Madras were acquired and integrated into the Madras Presidency.[6]

The Company acquired seven islands from Portugal as part of Catherine de Braganza's dowry in her marriage to Charles II of England in 1661. The seven islands became the East India Company headquarters in western India in 1668 and renamed Bombay (now Mumbai). This and subsequent territorial acquisitions in Western India, notably the Maratha territories obtained in 1817-18, were integrated to form the Bombay Presidency.[7] Similarly, the United Provinces were formed around the districts that were acquired from the Nawab of Oudh in 1775, and the province of Punjab was organized around the territories acquired after defeating Ranjit Singh in 1848. The city of Delhi, which became the capital of the British Empire in India in 1911, came under *de facto* British control after the Battle of Delhi in 1803.

The rest of India, approximately 38% of the area, was administered indirectly by the British through the agency of native kings and princes (Roy 1994). Major examples were Kashmir in the North, Hyderabad in Central India and Mysore and Travancore in South India. Iyer (2010) shows that determination of regions annexed for direct rule, and regions administered by native rulers was not a function of the economic or social characteristics of those places. Instead, a policy called "Doctrine of Lapse" specified that a territory was annexed if the ruler died without a natural male heir. Since the gender of a child is random, endogenous assignment into direct or indirect colonial rule is not a significant concern.

[6] Northern Circars, 1765; districts ceded by Mysore, 1792 and 1800; Tanjore, 1799; districts ceded by the Nawab of Carnatic, 1801.

[7] Surat, 1800; Gujarat, 1803 and 1818; Maratha territories, 1818; Satara, 1848.

The East India Company's administration ended after the mutiny of 1857 and India was ruled directly as part of the British Empire. British India consisted of nine major and six minor provinces, as well as a multitude of princely states. The major provinces were the presidencies of Bengal, Bombay and Madras, the lieutenant-governorships of Bihar, Burma, Orissa, Punjab and United Provinces, and the Chief Commissionerships of Assam and Central Provinces. Chelmsford (1918) describes the process of both conquest and organization of the administrative structure of colonial India.

> [T]he present map of British India was shaped by the military, political or administrative exigencies or conveniences of the moment, and with small regard to the natural affinities or wishes of the people.

This sentiment was echoed 12 years later by the Simon Commission (1930) which was established to review the constitutional structure of British India

> [there were in India] only a number of administrative areas [which had] grown up almost haphazard as the result of conquest, supersession of former rules or administrative convenience

The commission recommended reorganization of states to reflect a more coherent administrative picture.

> Although we are well aware of the difficulties encountered in all attempts to alter boundaries and of the administrative and financial complications that arise, we are making a definite recommendation for reviewing, and if possible resettling, the provincial boundaries of India at as early a date as possible.

Despite the Simon Commission's recommendations, the colonial government undertook no systematic reorganization of administrative units in India. British rule in India ended in 1947, concurrent with the Partition of the Indian Empire into India and Pakistan, which was formed from the Muslim-majority areas of Punjab, Bengal and Bombay. The provincial boundaries of independent India in 1947 reflected geographical continuity in the pattern of British military conquest in the eighteenth and nineteenth centuries, with little consideration towards the cultural or social characteristics that united or divided the provinces.

2.2 Education Provision in Colonial India

Education in pre-British India followed indigenous systems without standardization or significant state patronage, and was restricted to the social and economic elites (Acharya 1978). As British officials focused on administration of conquered territories in the eighteenth century, they introduced formal education both to train potential employees for clerical positions as well as to create acceptance of Western traditions and colonial rule (Evans 2002). A rich debate emerged on the language of instruction in government-aided schools between the Orientalists, who favored instruction in English, and the Vernaculars, who advocated instruction in local languages. Inspired by Macaulay's (1835) famous *Minute on Indian Education*, Governor-General Lord William Bentick decided initially to use English as the medium of instruction in mass education. However, instruction exclusively in English proved expensive. Consequently, his successor, Lord Auckland, accepted in 1839 Wood's recommendations (outlined in his *Dispatch*, which Radhakrishnan (1948) called the "Magna Carta of English Education in India") that the government adopt vernacular languages for instruction in primary and secondary schools and English for higher education (Windhausen 1964; Evans 2002).

Colonial officials at the province level, rather than central or local administrators, had significant influence on the development of schools, and educational outcomes. Chaudhary (2010) reports that public financing was the backbone of the school system, accounting for nearly half the expenditures during colonial rule. This increased to 60% by 1947, with the rest from annual school fees levied on students.[8] With these funds, the Bombay Presidency constructed a large network of public schools whereas Bengal, Bihar and Orissa relied on private schools that were incorporated into the state system.

As a result of colonial educational investments, 93% of schools in 1931 were either public schools or privately managed schools receiving public funds. Total enrollment, which was 10% in 1891, increased to 30% in 1931 (Chaudhary 2010). Crude literacy rates in 1931 ranged from 17.4% in Madras Presidency to 8.9% in Bihar and Orissa. Educational achievement primarily reflected facility with vernacular languages, with only 14.3% of all literates also able to read and write in English. Within this, the dominant language was the majority vernacular of the province. For example, Mohanty (2002) reports that the entire Orissa division of Bengal had only seven Oriya schoolteachers. The majority of teachers were Bengalis, and Bengali language textbooks were used for instruction.

At the post-secondary level, the British established universities in Calcutta, Bombay and Madras starting from 1857 (Radhakrishnan 1948). Annamalai (2004) reports that unlike primary and secondary schools, these universities employed a

[8] In regions that experienced indirect colonial rule (native states), education policy was determined by local rulers (Chaudhary 2010).

number of British faculty members, and the medium of instruction was English. Growing demand for higher and professional education led to the establishment of the University of Allahabad in 1887, as well as 21 other universities in the twentieth century. With the exception of Osmania University in Hyderabad where undergraduate classes were taught in Urdu, the medium of instruction remained English (Radhakrishnan 1948).

2.3 Reorganization of Indian States, 1956-1971

The nationalist leadership in India before Independence recognized the value of reorganization of states. The Indian National Congress, the main nationalist party, organized its regional committees on the basis of language in 1921 and endorsed the principle of the linguistic provinces (Arora 1956). However, India's Independence was accompanied by Partition on religious lines, which dampened enthusiasm for further division on an ethnic or cultural basis and the national leadership showed no interest in pursuing this reorganization (Guha 2008). The first Home Minister, Vallabhbhai Patel wrote (Dar 1948)

> [T]he first and last need of India at the present moment is that it should be made a nation . . . Everything which helps the growth of nationalism has to go forward and everything which throws obstacles in its way has to be rejected or should stand over. We have applied this test to linguistic provinces also, and judged by this test, on our opinion [they] cannot be supported.

Nonetheless, many regional movements advocated the idea of linguistic autonomy, particularly in the southern states. In 1952, following the death of an activist protesting for a separate state for Telugu speakers, the central government announced the formation of Andhra Pradesh from the Telugu speaking districts of Madras Province as well as a States Reorganization Commission (Govt. of India 1955). This commission recommended redrawing state boundaries entirely on linguistic principles, explicitly recognizing the role of shared language in reducing transaction costs ("Indian states, if linguistically constituted, will be able to achieve internal cohesiveness because language is a vehicle for communion of thoughts"), especially through education in vernacular schools ("educational activity can be stimulated by giving regional languages their due place"), leading to increasing administrative links within the state ("linguistic homogeneity as an important factor conducive to administrative convenience and efficiency"). South India was reorganized in 1956, West India in 1960 and the rest of India in 1965 and 1970.

Table 2 lists the states that were formed as a result of the Commission's recommendation from 1956 through 1971. Table 3 shows the impact of the reorganization, comparing the Herfindahl-Hirschman Index (HHI), a measure of concentration, for the share of different languages for the 1911 provinces and princely states with the "successor" states in 2001.[9] The table reveals that despite 90 years of population mixing, reorganization yielded states that were systematically more concentrated in spoken languages than the predecessor states.[10]

[9] $HHI = \sum_{i=1}^{N} s_i^2$ where s_i is the percent share of language i in the state. $HHI = 10,000$ indicates that a single language is used in the state and therefore a high degree of concentration.

[10] The major exception is Bengal, which was partitioned into West Bengal and the more populous Bangladesh. Since Hindi speakers were concentrated in the West before Independence, the successor

3 Data Description

Estimating the impact of language requires district level data on demographic and economic characteristics before and after 1956. The data should contain variables that represent the outcomes of interest, as well as a rich set of covariates representing factors that might impact performance. Also critical is that each district in the data should be classified as part of the linguistic minority or majority in its state before the mid-century reorganization.

The primary source of data that meets the above requirements is the decadal Census of India conducted by the Ministry of Home Affairs of the Government of India. I use the 1951, 1961, 1971, 1981 and 1991 waves of the Census. Data at the district level from the last four waves is compiled into a panel by Barnes and Vanneman (2000). This version of the Census contains data on population characteristics such as literacy, educational achievement and source of livelihood. Each variable is reported separately for all persons, men, rural residents and rural male residents in the district. In addition, the 1981 and 1991 Census contain data on the number of schools and colleges at the district level.

The Barnes and Vanneman (2000) dataset is augmented with data on mother tongue, education, religious and caste composition, and occupational choice from the 1951 Census of India. This allows us to measure the baseline rates of education and occupational choice before 1956, and estimate the difference in outcomes as a result of the change. The sources of this data are the economic tables, and the district census handbooks. While the economic tables report population size variables for all 321 districts in 1951, the district census handbooks report a more detailed set

state has a more diverse set of languages.

of variables, including data on educational achievement and occupational choice, for 140 districts. I am not concerned about selective data reporting since the remaining handbooks were destroyed by humidity and pests, and these factors are unlikely to be correlated with economic outcomes in 1951.

I add a number of district-level geographic controls that might impact economic performance since India is primarily an agricultural economy. The Indian Meteorological Department provides monthly rainfall and temperature readings at the sub-division level.[11] I calculate the mean and variance of the January and July measures for each census decade and include these in the dataset. In addition, the World Bank's India Agriculture and Climate Data Set reports the latitude, longitude, elevation and soil type for each district.

I restrict the study to states of South India for three reasons. First, the first wave of reorganization in 1956 took place in South India only. A potential concern with subsequent waves of reorganization (listed in Table 2) is that states bargained over districts and therefore the natural experiment is not as clean. For example, Guha (2008) recounts considerable political bargaining between the states of Maharashtra and Gujarat over the city of Bombay. Second, Kumar and Somanathan (2009) document changes in district boundaries across many of the Census waves in my dataset. Most of these changes were in North India, whereas the district boundaries in South India have remained stable over time. Although they offer a method to correct for population totals, there is no way to correct for other variables such as educational achievement or occupational choice in districts outside South India. Third, the modern states of Andhra Pradesh, Karnataka, Kerala and Tamil Nadu exhibit significant

[11] Each district is matched to a sub-division.

diversity in languages unlike North India where Hindi is widespread, allowing for cleaner identification of the effects of language on economic outcomes.

With this restriction, the dataset yields 370 district-year observations. Table 4 summarizes some of the variables of interest. Minority districts are 23 out of 67 districts. Scheduled Castes represent 14.3% of the population. The cumulative literacy rate over the decades was 37.2%, with the literacy rate for men approximately double than for women. As expected, completion rates decline for higher education levels, with only 1.2% of the population reporting a college degree. The census reported an occupation for 52% of the population, with the majority (30%) directly engaged in agriculture and 4.1% in communications intensive industries such as commerce and transportation.

4 Empirical Analysis

The objective of the empirical exercise is to estimate the impact of language on economic outcomes. A district's language identifies it as either a "majority" or a "minority" district. The language used by the majority of residents within a district is relatively stable over time. Conversely, state boundaries changed as a result of the reorganization exercise. Thus, a *minority* district was in the linguistic minority of the state before reorganization, and reassigned to a state formed on the basis of its language after reorganization. In a *majority* district, the primary language was the state's majority language both before and after reorganization. Since the classification of a district depends on the exogenous province-formation by the British, and subsequent allocation to states is on strict linguistic lines, the difference in outcomes between minority and majority districts identifies the impact of language ability on economic outcomes.

In the empirical tests presented in Section 4.1, I therefore use two types of outcomes that are strongly correlated with household and per capita income. The first type of outcomes is educational achievement, which includes literacy and primary school, high school and college completion rates as these are likely to be directly impacted by the language used for instruction. High rates of educational achievement are in turn linked to greater economic growth. For example, using cross-country panel data, Benhabib and Spiegel (1994) find that human capital, represented by the level of education, has a positive and significant impact on total factor productivity. In a country-specific study using school enrollment and factory employment data from nineteenth century Prussian districts, Becker, Hornung, and Woessmann (2011) found that higher initial levels of education were associated with significantly higher levels of technological adoption, leading to greater economic development during the Industrial Revolution. In an industry-specific study using state level data from India, Arora and Bagde (2010) find that greater investment in educational infrastructure, especially private engineering colleges, is strongly correlated with the volume of software exports.

The second type of outcomes are employment variables. Lower communication costs might induce a shift in occupational choice away from agriculture, where arguably communication between workers does not play a major role, and towards sectors such as manufacturing, commerce and transportation, where communication between workers is important for productivity. Murphy, Shleifer, and Vishny (1991) present a model, with complementary empirical evidence, where employment in a sector depends on the returns to ability and to scale in each sector. Thus, I expect greater employment in the secondary and tertiary sectors, which have experienced significantly greater growth in post-Independence India compared to employment in the agricultural sector.

Benhabib and Spiegel (2005) also develop a model where countries above a threshold level of human capital are able to "catch-up" with technologically advanced countries once they are able to access the technology as well. If language is viewed as a "social technology", then districts should experience catch-up economic growth once they are reassigned to states where they are part of the linguistic majority. In Section 4.2, I therefore test whether minority districts catch-up in educational achievement and shift to occupations in the secondary and tertiary sectors after reorganization by estimating differences-in-differences in outcomes before and after reorganization.

4.1 Do minority districts have persistently poorer outcomes?

In this section, I estimate whether districts with historical minority language status have persistently poorer economic outcomes. In addition to a test of differences between minority and majority districts, I conduct three robustness checks. First, I estimate the first differences model restricting my sample to the set of districts at the border of states, where I expect stronger results since I exclude districts that are far from the state borders. Second, I expect that increasing the exogenously determined linguistic distance between the district's language and state's official language will lead to relatively poorer economic outcomes. Third, I expect that a more polarized district, where the fraction of minority language speakers is large, will have relatively poorer outcomes compared to a district where the number of speakers of each language is evenly matched.

4.1.1 Test using minority status

The first specification tests for the difference in outcomes between minority and majority districts. The key identifying assumption in this model is that the initial assignment of districts as minority or majority districts is not correlated with outcomes. In section 2.1 I argue that this is indeed the case and therefore specify the following first differences model.

$$y_{it} = \beta_0 + \beta_1 minority_i + \beta_2 \mathbf{Z}_i + \beta_3 \mathbf{X}_{it} + D_t + \epsilon_{it} \tag{1}$$

The variable y_{it} represents outcomes where I expect systematic differences between minority and majority districts. The coefficient of interest is β_1, which represents the marginal impact of a district that was in the linguistic minority before the reorganization of states. $minority_i$ is an indicator variable that is 1 if district i was reassigned from one state to another in the reorganization of 1956 and 0 otherwise. I cannot introduce district fixed effects because $minority_i$ is constant over time. Hence, I introduce Z_i, which is a vector of observed district characteristics, mainly geographical variables, that are invariant over time and potentially impact y_{it}. I also include X_{it} as a vector of observed district-level time varying characteristics such as the average and variance in January and July rainfall over the decade and the fraction of residents who are from historically disadvantaged Scheduled Caste (SC) backgrounds. Finally, I include decade fixed effects (D_t) to account for all observed and unobserved decade characteristics that impact outcomes for all districts, as well as an i.i.d. normal error term (ϵ_{it}) to represent unobserved time and district varying characteristics.

First, I measure differences in educational achievement rates. I expect that majority language districts will have higher rates of literacy, middle school completion and matriculation rates, i.e., $\beta_1 < 0$ for these outcomes. While the qualitative impact on graduation rates predicted by the theory are the same, the coefficients for these outcomes might be smaller since English rather than the local language is commonly used as the medium of instruction at higher levels, mitigating the impact of historical differences in local language use. I expect that minority districts have a larger fraction of workers who are cultivators and agricultural laborers ($\beta_1 > 0$), and a smaller fraction who are employed in the non-farm sector ($\beta_1 < 0$). Within the non-farm sector, I expect stronger results for workers in the commerce and transportation sectors which might benefit disproportionately from ease of communication.[12]

Table 5 presents estimates for the OLS coefficient on minority district status (β_1) based on equation 1. In this table, the coefficients for literacy rates are positive and statistically significant. The literacy rate is 24.0% lower for minority districts and 27.4% lower in the rural areas of the same districts. Middle school completion and matriculation rates are similarly lower. Although college graduation rates are also lower in minority districts, the coefficient (-19.3%) is different from the school level results. One explanation for the negative coefficient is that the supply of students for colleges is lower in minority districts since fewer complete high school. However, the coefficient for college graduation is smaller than for school completion since

[12] Specifically, I take the log fraction of the population that is employed in each sector.

the medium of instruction in college is English, and therefore impacted less by a district's minority status.

Table 6 shows that a district's minority status has a negligent impact on employment in the agricultural sector. The fraction of cultivators is lower by 0.9%, but the fraction of agricultural laborers is higher by 10.6%. Both coefficients are statistically insignificant at the 10% level. As expected, Table 6 also reports that minority districts have a significantly lower rates of non-farm sector employment (-12.8% for all workers). The magnitude of the β_1 coefficient is -16.6% for commercial sector employment and -32.9% for transport sector employment, suggesting that the impact of minority status is stronger in sectors where the influence of language is potentially larger. The magnitude of the coefficients is smaller for rural areas perhaps because the commercial and transport sectors are not as large as in urban regions.

One concern with these results is that they are driven by variables omitted from the specification. To address this concern, I follow the strategy presented in Banerjee and Iyer (2005) and consider the subset of minority and majority districts that share a geographical border. Table 7 shows that minority districts report even worse education achievement rates (-27.4% for literacy, -43.4% for middle school completion, -32.3% for matriculation and -30.9% lower for graduation) when considering bordering districts only. The results for occupational choice show that minority districts have a greater share of the workforce in the agricultural sector. The fraction of cultivators is 21.1% higher, which is statistically significant at the 5% level. Simultaneously, the fraction of the non-farm workforce declines by 23.4%, with a large difference of 31.1% for workers in the transportation sector.

Thus, the test of first differences between minority and majority language districts presents evidence that historical mis-assignment has persistent impact on modern educational achievement and occupation outcomes. The additional test using a restricted sample of border districts only suggests that omitted variables do not drive this result, and that the language status of the district directly affects economic outcomes.

4.1.2 Test using linguistic distance measure

A potential concern with the interpretation of the results represented in the previous section is that they may represent systematic cultural differences between minority and majority districts, instead of linguistic differences by district. To address this concern, I propose a test using a measure of linguistic distance that is logically orthogonal to economic outcomes. This measure, developed by Lewis (2009), is constructed by counting the number of nodes between each pair of languages on the family tree of Indo-European and Dravidian languages.[13] More nodes imply that it is more difficult for a speaker of one language to learn another language and vice versa, and translates into a higher score for linguistic distance. For example, Punjabi and Hindi are adjacent to each other on the family tree, and therefore the linguistic distance between the two is 1. On the other hand, Tamil speakers find it difficult to understand or learn Hindi and vice versa, which is represented by a linguistic distance of 13 on my measure. Table 8 reports the linguistic distance between each pair of Scheduled Indian languages, where the average pairwise distance between

[13] Chiswick and Miller (2005b) and Beenstock, Chiswick, and Repetto (2001) also develop and use measures of linguistic distance. They report the pairwise distance between major international languages and distance between Hebrew and other languages spoken by migrants to Israel, respectively.

the four South Indian languages is 5.8. [14]

Using the data from Table 8, I assign a linguistic distance measure to each minority district based on the pair of languages dominant in the states that the district was assigned to before and after reorganization. I specify the following model where the linguistic distance L_i is interacted with the $minority_i$ dummy in equation 1.[15]

$$y_{it} = \beta_0 + \beta_1 minority_i + \beta_2 minority_i * L_i + \beta_3 \mathbf{Z}_i + \beta_4 \mathbf{X}_{it} + D_t + \epsilon_{it} \qquad (2)$$

In this specification, β_2 estimates the marginal impact of a unit increase in linguistic distance on outcomes in a minority district. I expect that β_2 is negative and statistically significant for economic measures such as educational achievement and participation in non-farm employment.

Table 9 presents OLS estimates of β_2 using the same set of educational and occupational choice outcomes as the previous section . I find strong confirmation for the theoretical prediction that minority language districts have lower economic opportunity, and that outcomes are poorer with increase in linguistic distance between the district's main language and state's majority language. Table 9 reports large differences between minority and majority districts for primary and secondary schooling

[14] Shastry (2010) reports that this measure is strongly correlated with two alternative and logically independent measures of linguistic distance. The first, developed by Shastry (2010) in collaboration with Professor Jay Jasanoff of Harvard University, measures distance based on shared cognates, distance and syntax. The second, based on the Comparative Indo-European Database developed by Dyen, Kruskal, and Black (1997) measures distance as the fraction of words from one language that are cognates of words from the second language. I cannot use the Shastry/Jasanoff measure since it reports the distance only between Hindi and other languages and not for every pairwise combination of languages, nor the Dyen, Kruskal, and Black (1997) measure since it does not include Dravidian languages.

[15] Note that equation (2) does not contain a separate levels term for linguistic distance since the variable is relevant only for minority language districts.

(-6.9% and -7.0%, respectively) which are statistically significant at the 1% level. The matriculation and graduation rates are also lower in minority districts, by 3.9% and 3.8% respectively, but the associated standard errors are large and the point estimates cannot be statistically distinguished from the null. However, the larger and more precisely estimated impact of linguistic distance on primary and middle school education compared to higher education supports the hypothesis that language manifests itself early in the education process when local language teaching is more important.

4.1.3 Test using minority fraction measure

A potential concern with the *minority* variable as defined and used in previous sections is that it does not capture the intra-district mix of languages used. In order to address this concern, I propose an alternative, continuous measure of minority status, *MinorityFraction*, that helps differentiate polarized districts where the state's minority language is spoken by a large fraction of residents from those districts where the number of minority and majority language speakers are evenly matched. An important identifying assumption is that potential economic outcomes do not influence where the speakers of different languages reside. This assumption is justified because native tongue is determined historically, and inter-district migration in India is very low with no discernable economic impact (Munshi and Rosenzweig 2009).

$$MinorityFraction = \frac{OtherLang - StateLang}{OtherLang + StateLang} \tag{3}$$

In this definition, *StateLang* represents the number of speakers of the official language of the state where the district is located. *OtherLang* is the number of speakers of the most popular language spoken in the district other than the state's language. Hence, for minority districts, $OtherLang > StateLang$ and $MinorityFraction > 0$, and vice versa. Additionally, a large positive value for *MinorityFraction* indicates that a large fraction of the population speaks the minority language compared to the state language whereas a small positive value implies that the two languages are spoken by relatively same number of district residents.

The 1951 Census reports the top three languages spoken in each district. From this data, I calculate *MinorityFraction* for each district and replace *minority* with this new variable in equations (1) and (2).

$$y_{it} = \beta_0 + \beta_1 MinorityFraction_i + \beta_2 \mathbf{Z}_i + \beta_3 \mathbf{X}_{it} + D_t + \epsilon_{it} \tag{4}$$

In equation (4), β_1 indicates the marginal impact of increasing the share of minority language speakers on outcome variable y_{it}. Table 10 shows that districts with large minority language populations suffer from greater shortfalls in educational attainment. The coefficients on literacy, middle school completion and matriculation are -9.4%, -14.4% and -12.3%, respectively, all of which are statistically different from the null. However, as expected, the impact on college graduation is imprecisely estimated though the point estimate is negative. Concurrently, the fraction of workers in commerce and transportation are also lower (-7.8% and -18.7% and statistically significant at the 5% and 1% level, respectively). Although the coefficient of cultivation as an occupational choice is negative (-7.4%), the associated standard errors are large.

The results in this section show that more polarized districts in 1951, where a smaller fraction spoke the dominant language of the province where the district was located, experienced significantly poorer educational outcomes in the post-Independence period, with lower employment in communication intensive sectors such as commerce and transportation.

4.2 Do minority districts catch up after reorganization?

In this section, I exploit the panel structure of the dataset and the timing of the 1956 reorganization to test for the catch-up hypothesis among minority districts after reassignment using a differences-in-differences framework. The sample used in the test is restricted to those districts for which 1951 census data is available. However, as explained in section 3, this is unlikely to cause selection bias.

$$y_{it} = \beta_0 + \beta_1 minority_i + \beta_2 Post_t + \beta_3 minority_i * Post_t + \beta_4 \mathbf{X}_{it} + \beta_5 \mathbf{Z}_i + D_t + \epsilon_{it} \quad (5)$$

In this specification, y_{it} represents an educational or occupational outcome in census waves from 1951 to 1991. Hence, the coefficient β_3 represents the marginal impact of the 1956 reassignment on minority language districts. The key identifying assumption is that in the absence of the reorganization, there would be no systematic differences in the trend of y_{it} between minority and majority districts. I expect greater change in outcomes (as before, increased enrollment in formal education and shift from agriculture to secondary sector occupations) among minority language districts.

Table 11 shows that reassignment had a large and significant impact (+20.9%

and statistically significant at the 10% level) on basic literacy levels. Even more striking is the impact on middle school completion (+71.6% and statistically significant at the 1% level) and matriculation rates (+67.6% and statistically significant at the 1% level) in minority districts. Although college graduation also increased dramatically, the coefficient (+45.9%) cannot be statistically distinguished from the null.

The results for occupational choice do not follow on expected lines. Employment increased in the agricultural sector (the fraction of cultivators increased by 21.0% and agricultural laborers by 26.7%) but decreased in the non-farm sector (-14.9%). However, all these coefficients are statistically insignificant and cannot be distinguished from the null. One potential explanation for this result is that educational and occupational shifts occur sequentially. Change in language policy affects educational achievement relatively quickly and cheaply since schools only need to change the medium of instruction. However, occupational shift requires workers to finish school and college under the new policy, and only subsequently choose employment in various industries. At the same time, occupational shift also requires creation of new firms in the secondary sector, which was a slow process in India's heavily regulated post-Independence economy.

4.3 Impact of language on income

Finally, I estimate the impact of linguistic differences on income in minority versus majority districts. Estimates of district level per capita income from the National Sample Survey are difficult to use, since it is not representative at the district level. Instead, I use district-level estimates for GDP in 2001 from the Indicus Analytics that are computed by aggregating the volume of production in primary, secondary and tertiary sectors, and multiplied by prevailing prices (Indicus Analytics 2011).

I cannot estimate a fully-specified model of the determinants of income differences since the dataset lacks a reasonably complete set of covariates that determine GDP at the district level.[16] Instead, I take a parsimonious approach and present the differences in income levels between minority and majority districts in Table 12. This table shows that, consistent with greater educational achievement and workforce participation in the secondary and tertiary sectors, majority districts also have higher per capita and per worker income in these sectors. Overall per capita GDP is Rs. 2290 per year higher in majority districts, whereas annual per worker income is Rs. 3870 higher in majority districts. These higher incomes for majority districts are driven by greater secondary and tertiary sector GDP. For majority districts, per capita GDP is Rs. 440 higher in the secondary sector and Rs. 1670 higher in the tertiary sector, but Rs. 130 *lower* in the primary sector.

Differences in per worker income are similarly driven by the secondary and tertiary sector. Table 12 reveals that for majority districts, per worker GDP is Rs. 9500 higher in the secondary sector and Rs. 4460 in the tertiary sector, but only Rs. 620 higher in the primary sector. These calculations suggest that the impact of language on economic outcomes is persistent, not only in terms of education and employment, but also income.

[16] Estimating equation 1 with district per capita GDP as a dependent variable reveals that this model has very low explanatory power ($R^2 < 0.01$), suggesting significant omitted variable bias and a misspecified model.

5 Discussion

The historian Ramachandra Guha has argued that the reorganization of Indian states was a transformative event in the life of a young republic (Guha 2008). It recognized and accommodated the development of a wide array of languages and associated cultural traditions while maintaining a federal and democratic polity. This paper explains how language-based reorganization of state boundaries, which exogenously changed the state language for some districts, had an impact on economic outcomes within those districts.

To explain the impact of language-based reorganization on economic outcomes, I argued that sharing the same language might reduce communication costs, boosting the incentive for education and the economic returns to labor in the communications-intensive secondary sector. In the long run, this may lead to greater trade, productivity and social welfare. I found evidence in support of this mechanism as districts that shifted from states where they were in the linguistic minority to states where they were part of the majority experienced greater educational achievement and an employment shift from agriculture to manufacturing, communications and trade. The effects were largest for primary schooling, as well as for those sectors, such as transportation, that were most likely to be impacted by language. Section 4.2 offered evidence that reassigning a minority district into a state formed on the basis of shared language can reverse the impact of history, and the district can experience disproportionate growth in educational achievement. Finally, section 4.3 compared 2001 per capita and per worker GDP between minority and majority districts, and confirmed that higher education and occupational shift to the secondary and tertiary sectors in majority districts led to higher incomes in those sectors, with stagnant or lower income in the primary sector.

These results should be read with a number of caveats. First, the analysis excludes an investigation of gravity effects by focusing on districts that border states. Second, this paper does not account for bilingualism, multilinguilism and language shift, especially with increasing use of English in the last century.

Nonetheless, this paper has implications on new state formation in India. After 1971, a number of Union Territories (areas administered by the central government) converted to formal statehood. More significantly, three new states – Chhattisgarh, Jharkhand and Uttaranchal – were carved out in 2000 from larger states on the basis of distinct culture of these regions. A number of proposals for separate statehood backed by popular movements remain in active consideration, most notably for a separate Telangana state separated from Andhra Pradesh. The results presented in this paper indicate that language through its impact on commercial growth is a significant driver of economic performance in new states. Therefore, new states formed on the basis of shared economic infrastructure are likely to experience better outcomes.

References

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review 93*(1), 113–132.

Abeyratne, S. (2004). Economic roots of political conflict: The case of Sri Lanka. *World Economy 27*(8), 1295–1314.

Acemoglu, D., S. Johnson, and J. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review 91*(5), 1369–1401.

Acharya, P. (1978). Indigenous vernacular education in pre-British era: Traditions and problems. *Economic and Political Weekly 13*(48), 1981–1988.

Albouy, D. (2008). The wage gap between Francophones and Anglophones: A Canadian perspective, 1970–2000. *Canadian Journal of Economics 41*(4), 1211–1238.

Angrist, J. and V. Lavy (1997). The effect of a change in language of instruction on the returns to schooling in Morocco. *Journal of Labor Economics 15*(1), 48–76.

Annamalai, E. (2004). Medium of power: The question of English in education in India. In J. Tollefson and A. Tsui (Eds.), *Medium of instruction policies: Which agenda? Whose agenda?*, pp. 177–194. Mahwah, NJ: Lawrence Erlbaum Associates.

Arora, A. and S. Bagde (2010). Human capital and the Indian software industry. NBER Working Paper No. 16167.

Arora, S. (1956). The reorganization of Indian states. *Far Eastern Survey 25*(2), 27–30.

Ban, R. and V. Rao (2007). The political construction of caste in South India. mimeo, Available at `http://www.cultureandpublicaction.org/bijupdf/policonscaste2.pdf`.

Banerjee, A. and L. Iyer (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *American Eco-*

nomic Review 95(4), 1190–1213.

Barnes, D. and R. Vanneman (2000). Indian district data, 1961-1991: Machine-readable data file and codebook. Available at http://www.inform.umd.edu/~districts/index.html.

Becker, S., E. Hornung, and L. Woessmann (2011). Education and catch-up in the Industrial Revolution. *American Economic Journal: Macroeconomics 3*(3), 92–126.

Beenstock, M., B. Chiswick, and G. Repetto (2001). The effect of linguistic distance and country of origin on immigrant language skills: Application to Israel. *International Migration 39*(3), 33–60.

Benhabib, J. and M. Spiegel (1994). The role of human capital in economic development: Evidence from aggregate cross-country data. *Journal of Monetary Economics 34*(2), 143–173.

Benhabib, J. and M. Spiegel (2005). Human capital and technology diffusion. *Handbook of Economic Growth 1*, 935–966.

Carliner, G. (1981). Wage differences by language group and the market for language skills in Canada. *Journal of Human Resources 16*(3), 384–399.

Chakraborty, T. and S. Kapur (2009). English language premium: Evidence from a policy experiment in India. http://www.iza.org/en/papers/Chakraborty08092009.pdf.

Chaudhary, L. (2010). Land revenues, schools and literacy: A historic examination of public and private funding of education. *Indian Economic & Social History Review 47*(2), 179.

Chelmsford (1918). *Report on Indian Constitutional Reforms*. Calcutta, India: Superintendent Government Printing.

Chiswick, B. (1991). Speaking, reading, and earnings among low-skilled immigrants. *Journal of Labor Economics 9*(1), 149–170.

Chiswick, B. and P. Miller (2005a). Do enclaves matter in immigrant adjustment? *City & Community 4*(1), 5–35.

Chiswick, B. and P. Miller (2005b). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development 26*(1), 1–11.

Clingingsmith, D. (2007). Bilingualism, language consolidation, and industrialization in mid-20th century India. Mimeo, Harvard University.

Dar, S. K. (1948). *Report of the Linguistic Provinces Commission*. New Delhi, India: Constituent Assembly of India.

Duflo, E. and R. Pande (2007). Dams. *Quarterly Journal of Economics 122*(2), 601–646.

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

Dyen, I., J. Kruskal, and P. Black (1997). Comparative Indo-European database. Collected by Isidore Dyen. File IE-Rate 1. Available at `http://www.wordgumbo.com/ie/cmp/iedata.txt`.

Evans, S. (2002). Macaulay's Minute revisited: Colonial language policy in nineteenth-century India. *Journal of Multilingual and Multicultural Development 23*(4), 260–281.

Govt. of India (1955). *Report of the States Reorganization Commission*. New Delhi: Manager of Publications.

Guha, R. (2008). *India after Gandhi: The history of the world's largest democracy*. New York: Picador.

Indicus Analytics (2011). *District Gross Domestic Product Series 2001-02 to 2011-12*. New Delhi, India: Indicus Analytics.

Iyer, L. (2010). Direct versus indirect colonial rule in India: Long-term consequences. *Review of Economics and Statistics 92*(2).

Kumar, H. and R. Somanathan (2009). Mapping Indian districts across census years, 1971-2001. *Economic & Political Weekly 10*, 69–73.

Lang, K. (1986). A language theory of discrimination. *Quarterly Journal of Economics 101*(2), 363–382.

Lazear, E. (1999). Culture and language. *Journal of Political Economy 107*(6), 95–126.

Lewis, P. (Ed.) (2009). *Ethnologue: Languages of the world, Sixteenth edition*. Dallas, TX: SIL International. Online version: `http://www.ethnologue.com/`.

Macaulay, T. (1835). Minute by the Hon'ble T. B. Macaulay, dated the 2nd February 1835. In H. Sharp (Ed.), *Selections from Educational Records, Part I (1781-1839)*. Calcutta, India: Superintendent, Government Printing.

Mohanty, P. (2002). British language policy in 19th century India and the Oriya language movement. *Language Policy 1*, 53–73.

Munshi, K. and M. Rosenzweig (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 M. Rosenzweig (2009). Why is mobility in India so low? Social insurance, inequality, and growth. NBER Working Paper No. 14850.

Murphy, K., A. Shleifer, and R. Vishny (1991). The allocation of talent: Implications for growth. *Quarterly Journal of Economics 106*(2), 503–530.

North, D. (1990). *Institutions, institutional change, and economic performance*. Cambridge Univerity Press.

Pande, R. and C. Udry (2006). Institutions and development: A view from below. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Volume 2, pp. 349–411.

Radhakrishnan, S. (1948). Report of the university education commission. `http://www.education.nic.in/cd50years/n/75/7Y/toc.htm`.

Roy, T. (1994). *The economic history of India, 1857–1947*. New Delhi, India: Oxford University Press.

Roy, T. (2002). Economic history and modern India: Redefining the link. *Journal of Economic Perspectives 16*(3), 109–130.

Shastry, K. (2010). Human capital response to globalization: Education and information technology in India. *Journal of Human Resources forthcoming*.

Simon, J. (1930). *Report of the Indian Statutory Commission*. New Delhi: Swati Publications.

Windhausen, J. (1964). The Vernaculars, 1835-1839: A third medium for Indian education. *Sociology of Education 37*(3), 254–270.

Appendices

A Endogeneity in migration

A possible concern with the district-level, rather than individual-level, analysis presented in previous sections is that language can possibly generate differences in economic outcomes through the channel of inter-district migration. This would bias my estimates if productive individuals in minority districts migrate to majority districts, enroll in schools and higher education and work disproportionately in the non-farm sector.

I address this concern in two ways. First, the empirical literature which relies on district-level analysis in India does not find significant inter-district migration. For example, using data from the National Sample Survey, Duflo and Pande (2007) find that their results are robust to potentially endogenous inter-district migration. Second, I estimate the impact of linguistic differences on rates of migration using the same specifications presented in Section 4.1, and total migration and rural male migrants from the district as dependent variables. Table 13 shows that the coefficients for the minority variable are all small and statistically insignificant. This suggests that districts' historical linguistic status did not have much impact on migration and inter-district migration is unlikely to bias the results presented in the main text.

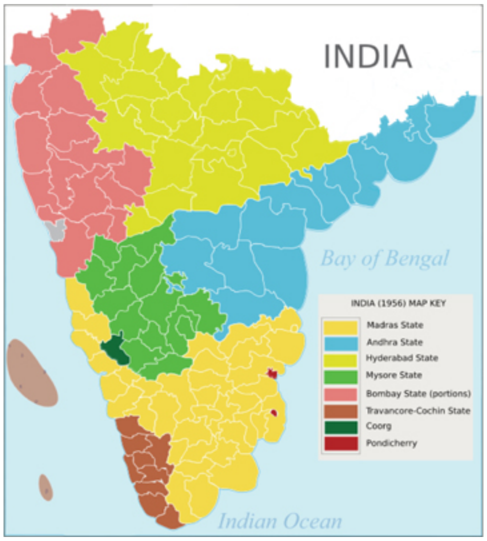

Figure 1: **District assignments in South India before 1956**

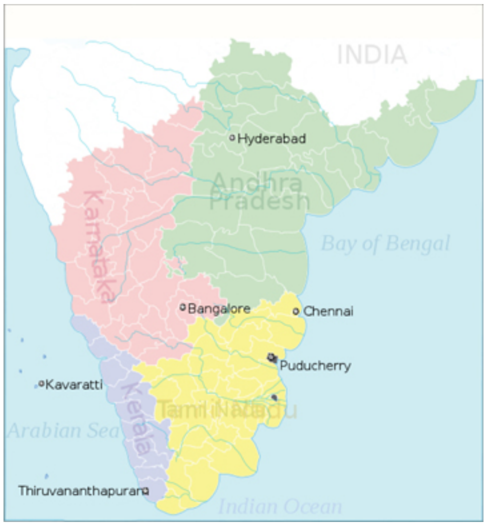

Figure 2: **District assignments in South India after 1956**

Map source: `www.wikipedia.org`

Table 1: **State-wise concentration of languages**

Language	**Main States**	**Fraction of all speakers who reside in the state**
Scheduled Languages that form basis of state		
Assamese	Assam	98.8%
Bengali	West Bengal	82.0%
Gujarati	Gujarat	92.8%
Hindi	Himachal Pradesh, Uttaranchal, Haryana, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Chhattisgarh, Madhya Pradesh	90.2%
Kannada	Karnataka	91.9%
Kashmiri	Jammu and Kashmir	98.2%
Malayalam	Kerala	93.2%
Manipuri	Manipur	86.3%
Marathi	Maharashtra	92.6%
Oriya	Orissa	92.6%
Punjabi	Punjab	77.6%
Tamil	Tamil Nadu	91.8%
Telugu	Andhra Pradesh	86.4%
Scheduled Languages that do not form basis of a state		
Bodo	Assam	96.0%
Dogri	Jammu and Kashmir	96.6%
Konkani	Goa, Maharashtra, Karnataka	88.2%
Maithili	Bihar	97.1%
Nepali	Assam, West Bengal, Sikkim, Uttar Pradesh	76.3%
Santali	Jharkhand, West Bengal	79.2%
Sindhi	Gujarat, Rajasthan	65.8%
Urdu	Uttar Pradesh, Bihar, Maharashtra, Andhra Pradesh, Karnataka	81.0%

Note: *Scheduled languages* are major Indian languages listed in the Eighth schedule of the Constitution. Source: Census of India, 2001.

Table 2: **New states in India, 1956-71**

State	**Year established**	**Legislation**
Karnataka, Kerala, Lakshadweep, Andhra Pradesh, Tamil Nadu, Madhya Pradesh	1956	States Reorganisation Act, 1956
Gujarat, Maharashtra	1960	Bombay Reorganization Act, 1960
Nagaland	1963	State ofNagaland Act, 1962
Haryana	1966	Haryana Act, 1966
Himachal Pradesh	1971	State of Himachal Pradesh Act, 1971

Table 3: **Language concentration**

Colonial province (1911)	**Major languages (Share of speakers)**	**HHI (1911)**	**Successor state (2001)**	**Major languages (Share of speakers)**	**HHI (2001)**	**HHI Diff. (2001 - 1911)**
Assam	Bengali (48.0%) Assamese (22.8%)	3082.1	Assam	Assamese (53.1%) Bengali (30.0%)	3791.1	708.9
Bengal	Bengali (92.1%) Hindi (3.8%)	8505.5	West Bengal	Bengali (85.8%) Hindi (7.2%)	7427.8	-1077.7
Bihar and Orissa	Hindi (70.0%) Oriya (14.2%)	5161.6	Bihar	Hindi (73.2%) Maithili (14.3%)	5695.0	533.4
Bombay	Marathi (45.8%) Gujarati (17.0%)	2792.1	Maharashtra	Marathi (72.2%) Hindi (11.2%)	5417.1	2625.0
United Provinces	Hindi (92.1%) Western Hindi (7.6%)	8543.9	Uttar Pradesh	Hindi (91.3%) Urdu (8.0%)	8407.2	-136.7
Central Provinces & Berar	Hindi (43.4%) Marathi (35.0%)	3223.0	Madhya Pradesh	Hindi (94.1%) Marathi (2.6%)	8861.2	5638.3
Madras	Tamil (40.3%) Telugu (38%)	3165.9	Tamil Nadu	Tamil (89.5%) Telugu (5.7%)	8043.8	4877.9
Punjab	Punjabi (59.6%) Western Punjabi (18.5%)	4049.5	Punjab	Punjabi (91.7%) Hindi (7.6%)	8474.0	4424.5
Hyderabad	Telugu (47.6%) Marathi (26.2%)	3219.4	Andhra Pradesh	Telugu (84.7%) Urdu (8.7%)	7267.5	4048.1
Kashmir	Kashmiri (37.8%) Punjabi (23.5%)	2443.3	Jammu & Kashmir	Kashmiri (55.4%) Dogri (22.5%)	3943.3	1499.9
Cochin and Travancore	Malayalam (76.0%) Tamil (20.8%)	6218.5	Kerala	Malayalam (97.2%) Tamil (1.9%)	9449.4	3230.9
Mysore State	Kannada (71.4%) Telugu (15.8%)	5408.6	Karnataka	Kannada (68.5%) Urdu (10.9%)	4901.0	-507.6
Rajputana	Rajasthani (78.8%) Hindi (11.3%)	6377.5	Rajasthan	Hindi (95.5%) Punjabi (2.1%)	9129.5	2752.0

Note: The 1911 provinces of Ajmer-Merwara, Balochistan, Burma, Coorg, NWFP, Andamans and Nicobars, and princely states in Assam, Balochistan, Baroda, Bengal, Bihar and Orissa, Bombay, Central India Agency, Central Provinces, NWFP, Punjab, Sikkim, United Provinces not shown. Source: Census of India, 1911 and 2001.

Table 4: **Summary statistics**

Variable	Obs.	Mean	Std. Dev.
Total Population	370	2072956	1090839
Fraction of Scheduled Castes	292	14.30%	0.053
Education			
Literacy rate	294	37.20%	0.181
Middle school completion rate	226	14.60%	0.103
Matriculation rate	293	6.04%	0.053
College graduation rate	226	1.16%	0.012
Occupation			
Cultivators	319	18.58%	0.156
Agricultural laborers	319	11.87%	0.056
Non-farm workforce	319	17.80%	0.107
Commercial workers	319	3.08%	0.020
Transportation workers	319	1.05%	0.009

Source: Census of India (1951) and Indian District Panel Dataset (1961-1991).

Table 5: **Result for education in minority versus majority districts**

Dependent Variable	**OLS Coefficient (β_1)**	**N**	**adj. R^2**
Total literacy rate	**-0.240***** (0.051)	172	0.566
Rural literacy rate	**-0.274***** (0.055)	172	0.580
Middle school completion rate	**-0.249***** (0.074)	129	0.533
Rural middle school completion rate	**-0.305***** (0.081)	129	0.548
Matriculation rate	**-0.239***** (0.078)	172	0.792
Rural matriculation rate	**-0.291***** (0.074)	172	0.860
College graduation rate	**-0.193*** (0.104)	129	0.717
Rural college graduation rate	**-0.258***** (0.086)	129	0.841

Notes: The reported OLS coefficients correspond to β_1 from equation (1). Regression includes decade fixed effects. Standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Data sources: Census of India (1951-1991), Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 6: **Result for occupation in minority versus majority districts**

Dependent Variable	**OLS Coefficient (β_1)**	**N**	**adj. R^2**
Total cultivators	-0.009	172	0.219
	(0.085)		
Total agricultural laborers	0.106	172	0.254
	(0.082)		
Rural agricultural laborers	0.099	172	0.346
	(0.071)		
Total non-farm workforce	**-0.128****	172	0.142
	(0.059)		
Rural non-farm workforce	-0.091	172	0.172
	(0.072)		
Total commercial workers	**-0.166*****	172	0.284
	(0.048)		
Rural commercial workers	**-0.083****	172	0.468
	(0.040)		
Total transportation workers	**-0.329*****	172	0.368
	(0.087)		
Rural transportation workers	**-0.306*****	172	0.560

Notes: The reported OLS coefficients correspond to β_1 from equation (1). Regression includes decade fixed effects. Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Data sources: Census of India (1951-1991), Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 7: **Result for minority versus majority districts (Bordering districts only)**

Dependent Variable	**Coefficient** (β_1)	**N**	**adj.** R^2
Literacy rate	**-0.274***** (0.063)	141	0.797
Middle school completion rate	**-0.434***** (0.116)	109	0.890
Matriculation rate	**-0.323***** (0.106)	140	0.884
College graduation rate	**-0.309**** (0.142)	109	0.880
Cultivators	**0.211**** (0.106)	141	0.589
Agricultural laborers	0.089 (0.108)	141	0.343
Non-farm workforce	**-0.234***** (0.070)	141	0.602
Commercial workers	**-0.200***** (0.063)	141	0.637
Transportation workers	**-0.311***** (0.110)	141	0.436

Notes: The reported OLS coefficients correspond to β_1 from equation (1). Regression includes state and decade fixed effects. Standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Data sources: Census of India 1951-1991, Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 8: **Linguistic distance measure**

	Hindi	**Urdu**	**Gujarati**	**Punjabi**	**Rajasthani**	**Konkani**	**Marathi**	**Assamese**	**Bengali**	**Bihari**	**Oriya**	**Kashmiri**	**Kannada**	**Malayalam**	**Tamil**	**Telugu**
Hindi	0	1	4	4	5	6	5	6	6	6	6	7	11	13	13	10
Urdu	1	0	4	4	5	6	5	6	6	6	6	7	11	13	13	10
Gujarati	4	4	0	3	4	5	4	5	5	5	5	6	10	12	12	9
Punjabi	4	4	3	0	4	5	4	5	5	5	5	6	10	12	12	9
Rajasthani	5	5	4	4	0	6	5	6	6	6	6	7	11	13	13	10
Konkani	6	6	5	5	6	0	2	5	5	5	5	6	10	12	12	9
Marathi	5	5	4	4	5	2	0	4	4	4	4	5	9	11	11	8
Assamese	6	6	5	5	6	5	4	0	1	3	3	6	10	12	12	9
Bengali	6	6	5	5	6	5	4	1	0	3	3	6	10	12	12	9
Bihari	6	6	5	5	6	5	4	3	3	0	3	6	10	12	12	9
Oriya	6	6	5	5	6	5	4	3	3	3	0	6	10	12	12	9
Kashmiri	7	7	6	6	7	6	5	6	6	6	6	0	11	13	13	10
Kannada	11	11	10	10	11	10	9	10	10	10	10	11	0	5	5	6
Malayalam	13	13	12	12	13	12	11	12	12	12	12	13	5	0	3	8
Tamil	13	13	12	12	13	12	11	12	12	12	12	13	5	3	0	8
Telugu	10	10	9	9	10	9	8	9	9	9	9	10	6	8	8	0

Source: `http://www.ethnologue.com` and author's calculations.

Table 9: **Results for education and occupational outcomes by linguistic distance**

Dependent Variable	**OLS Coefficient (β_2)**	**N**	**adj. R^2**
Literacy rate	**-0.069*****	172	0.614
	(0.015)		
Middle school completion rate	**-0.070*****	129	0.566
	(0.015)		
Matriculation rate	-0.039	172	0.794
	(0.024)		
College graduation rate	-0.038	129	0.718
	(0.032)		
Cultivators	**0.113*****	172	0.304
	(0.025)		
Agricultural laborers	0.008	172	0.250
	(0.0254)		
Non-farm workforce	0.026	172	0.147
	(0.018)		
Commercial workers	-0.000	172	0.280
	(0.015)		
Transportation workers	-0.015	172	0.366
	(0.027)		

Notes: The reported coefficients correspond to β_2 from equation (2). Regression includes decade fixed effects. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Data sources: Census of India 1951-1991. Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 10: **Result for education and occupational choice outcomes using *MinorityFraction***

Dependent Variable	**Coefficient** (β_1)	**N**	R^2
Literacy rate	**-0.094*****	194	0.693
	(0.033)		
Middle school completion rate	**-0.144*****	150	0.872
	(0.054)		
Matriculation rate	**-0.123****	193	0.855
	(0.052)		
College graduation rate	-0.097	150	0.855
	(0.070)		
Cultivators	-0.074	194	0.540
	(0.049)		
Agricultural laborers	-0.021	194	0.226
	(0.049)		
Non-farm workforce	0.001	194	0.478
	(0.036)		
Commercial workers	**-0.078****	194	0.565
	(0.030)		
Transportation workers	**-0.187*****	194	0.348
	(0.054)		

Notes: The reported coefficients correspond to β_1 in equation (4). Regression includes decade fixed effects. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Data sources: Census of India 1951-1991. Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 11: **Result for education and occupational outcomes before and after reorganization**

‘			
Dependent Variable	**OLS Coefficient (β_3)**	**N**	**adj. R^2**
‘			
Literacy rate	**0.209***	194	0.797
‘	(0.124)		
Middle school completion rate	**0.716*****	150	0.898
	(0.224)		
Matriculation rate	**0.676*****	193	0.871
	(0.231)		
College graduation rate	0.459	150	0.862
	(0.314)		
‘			
Cultivators	0.021	194	0.579
	(0.216)		
Agricultural laborers	0.267	194	0.268
	(0.220)		
Non-farm workforce	-0.149	194	0.529
	(0.156)		
Commercial workers	-0.033	194	0.593
	(0.133)		
Transportation workers	-0.019	194	0.351
	(0.247)		

Notes: The reported coefficients correspond to β_3 in equation (5). Regression includes decade fixed effects. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Data sources: Census of India 1951-1991. Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.

Table 12: **Differences in income between minority and majority districts**

Dependent Variable	**Majority**	**Minority**	**Difference**
Total per capita GDP	22.45	20.16	2.29
Primary sector per capita GDP	5.91	6.03	-0.13
Secondary sector per capita GDP	4.92	4.48	0.44
Tertiary sector per capita GDP	11.78	10.11	**1.67***
Total per worker GDP	58.42	54.55	3.87
Primary sector per worker GDP	32.94	32.32	0.62
Secondary sector per worker GDP	71.26	61.76	9.50
Tertiary sector per worker GDP	100.13	95.66	4.46

Notes: Per capita and per worker GDP in Rs. '000. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Data sources: Indicus Analytics dataset on district GDP of India, 2001-02.

Table 13: **Inter-district migration**

Dependent Variable	**Coefficient**	**N**	**adj.** R^2
Minority vs Majority			
Total migration rate	-0.005	129	0.201
	(0.107)		
Rural male migration rate	0.011	129	0.197
	(0.139)		
Bordering Districts			
Total migration rate	-0.120	93	0.172
	(0.159)		
Rural male migration rate	-0.212	93	0.225
	(0.213)		
Linguistic Distance			
Total migration rate	-0.427	51	0.597
	(0.390)		
Rural male migration rate	-0.280	51	0.463
	(0.559)		
MinorityFraction			
Total migration rate	0.042	129	0.204
	(0.065)		
Rural male migration rate	0.076	129	0.201
	(0.085)		

Notes: Regressions include decade fixed effects. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Data sources: Census of India 1951-1991. Indian Meteorological Department, and World Bank India Agriculture and Climate Data Set.