Common tongue: The impact of language on educational outcomes

http://eprints.exchange.isb.edu/112

Working paper

Indian School of Business

2013

Common tongue: The impact of language on educational outcomes

Tarun Jain*
Indian School of Business

August 20, 2013

Abstract

This paper investigates the impact of official language policies on education using state formation in India. Colonial provinces consisted of some districts where the official language matched the district's language and some where it did not. Linguistically mismatched districts have 18.8% lower literacy rates and 27.6% lower college graduation rates, driven by difficulty in acquiring education due to a different medium of instruction in schools. Educational achievement caught up in mismatched districts after the 1956 reorganization of Indian states on linguistic lines, suggesting that political reorganization can mitigate the impact of mismatched language policies.

Keywords: Language; Medium of instruction; Education; Political reorganization.
JEL Codes: I20, N95, O15, O43.

*Economics and Public Policy Area, Indian School of Business, Hyderabad, India. Email: tj9d@virginia.edu.
This paper benefited from detailed feedback from Tanika Chakraborty, Leora Friedberg, Maitreesh Ghatak, Lakshmi Iyer, Krishna Kumar, Pushkar Maitra, Subha Mani, Prachi Mishra, Kaivan Munshi, John Pepper, Vijayendra Rao, Rohini Somanathan and Ashima Sood. I thank seminar and conference participants at BITS Pilani, Delhi School of Economics, Fordham, Franklin and Marshall, ISB, Monash, UNSW, UVA, ISI Growth and Development Conference, NEUDC and Econometric Society 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

Language policies are associated with a great deal of debate, policy action and conflict. Using uniform languages can lower the cost of communication, facilitate education and expand economic growth. This lesson is quickly learned by new immigrants whose academic and economic success depends critically on mastering the *lingua franca* of their new home. However, attempts to impose official languages can meet considerable resistance, both because language is an important component of identity, and because learning a new language is difficult. For example, adopting Sinhalese as the official language led to a decades long civil war in Sri Lanka not only because of the affront to identity, but also because Tamils who did not speak Sinhalese lost government positions and power. Similarly, separatist movements in Quebec, Belgium and Catalonia are also rooted in resistance to dominant languages.

Do language policies, specifically education in the mother tongue, benefit or hinder long term educational achievement? Language might influence education in a number of ways, not all of which operate in the same direction. First, if the language used in schools is the same as the learner's mother tongue, then educational achievement is greater as residents are more likely to enrol, understand instruction and complete different levels of schooling. For instance, Bleakley and Chin (2004) find that childhood immigrants to the United States from English-speaking countries report significantly greater educational achievement compared to those from non-English backgrounds. Second, language fluency might be key for unlocking future economic opportunities. For example, the returns to post-secondary education declined by half when Morocco switched from French to Arabic as the medium of instruction in schools, primarily because most organized economic activity was conducted in French (Angrist and Lavy 1997).[1]

Language might also influence education through the provision of public schools, although the direction of this effect is not clear. If speakers of the majority language control public spending, they might construct more schools in the areas where the language is spoken because of patronage

[1] Similarly, using a policy change in West Bengal which revoked English as a medium of instruction in government-operated schools, Chakraborty and Bakshi (2013) report that a 1% decrease in the probability of learning English decreases wages by 1.6%.

or shared preferences, leading to higher educational achievement. Conversely, school construction and spending on educational infrastructure might be directed at economically weaker regions. If minority language regions are also economically weaker, then these areas would experience greater improvements in education achievement.[2] Finally, language mismatch might be associated with greater migration if individuals who do not speak the official language of a region relocate to a different region seeking education or economic opportunities in their own language.

The main contribution of this paper is to examine the effect of education in the mother tongue on educational achievement using two large-scale historical events. First, colonial-era provinces were formed without regard to language, which resulted in some areas where schools taught in the same language as the mother tongue of the students, and other areas where they did not. This allows me to compare linguistically matched versus mismatched districts in the colonial era and estimate the impact on post-independence differences in educational achievement. This illustrates the degree to which variation in language explains persistent underachievement in education within countries.

Second, the paper uses the 1956 reorganization of Indian states on strict linguistic lines as a national experiment that had the effect of reversing language policies and aligning the official language with the dominant mother tongue on educational outcomes.[3] The helps to understand whether "fixing" language policies through political reorganization can address persistent development shortfalls.

India is a particularly appropriate setting for studying the impact of language since over 3,000 languages are used in the country, with 18 languages claiming both wide speakership as well as constitutional recognition. The boundaries of modern Indian states correspond to the areas where these languages are used. Public schools offer instruction using the medium of the state language,

[2]Empirical studies on the link between ethno-linguistic diversity and public goods provision are summarized in a survey article by Alesina and Ferrara (2005). In India, Banerjee, Iyer, and Somanathan (2005) examine multiple social cleavages and report lower public goods availability in areas that were directly controlled by the British, had landlord-based tax collection systems, and greater ethnic division.

[3]Ban, Jha, and Rao (2012) study this reorganization in the context of differences in political participation exploiting misalignment at the village level. However, Section 2.3 shows that no districts were misassigned in the 1956 reorganization process.

and regulators, the judiciary and other arms of state governments use the state language for official correspondence with citizens. So most studies that compare a colonial language, especially English, with a vernacular language suffer because the returns from globalization cannot be isolated from pure language effects or because populations who speak these languages are systematically different. In contrast, the setting of this paper permits comparison between multiple sets of major vernacular languages, each with tens of millions of speakers.

There are two major empirical challenges in generating reliable estimates of the relationship between language and education. The primary challenge is that bivariate comparisons between communities that either speak or do not speak the official language might not yield unbiased estimates of the relationship between language and education because of at least two potential sources of endogeneity. First, individuals who do not speak the official language may move to ethnic enclaves where they do not need to learn a new language. Second, communities that recognize the link between speaking a different language and poorer economic outcomes may form their own political units. For example, linguistic minorities in countries such as Spain, Canada and Cameroon launched separatist movements based on language. Thus, unbiased estimates require exogenous matching between languages and communities.

The pattern of British conquest and province formation in the eighteenth century helps generate such estimates. Provincial boundaries in British India were determined either by the sequence of British military conquest, with provinces cobbled together from various districts as imperial rule extended from the coasts into the hinterland, or when the British decided to leave native rulers in place. I argue in Section 2.1 that this process was exogenous to linguistic concerns, leading to the assignment of some districts to provinces where the district's numerically dominant mother tongue language was the same as the official language of the province (henceforth, a "majority" district), and other districts to provinces where it was not (a "minority" district). I hypothesize that if mother tongue instruction facilitates schooling, then historically majority districts should have had better educational outcomes compared to historically minority districts. Insofar that educational achievement persists over generations, minority districts could experience poorer educational outcomes

even till modern times.[4]

In 1955, the States Reorganization Commission recommended forming new states strictly on linguistic lines, a principle which the central government followed while redrawing state boundaries in 1956, 1960, 1966 and 1971 (Govt. of India 1955). After reorganization, new state boundaries consisted of those districts that where the major language was the same as the official language of the state. An immediate policy change in all the new states was extending the official language as the medium of instruction in schools and for all official business. Thus, a natural experiment is set up where each district was classified as either majority or minority before the reorganization by historical accident, and reassigned as majority after the reorganization. I expect that growth rates for each of the measures for educational achievement should be higher in historically 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, this paper finds support for the hypothesis that shared language within a state potentially lowers communication costs and increases educational achievement rates. 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 is conducted in the vernacular. Specifically, the literacy rate in minority districts is estimated to be 18.8% lower than majority districts, whereas the middle school completion rate is 24.4% lower. In contrast, although the fraction of college graduates in minority districts is 27.6% smaller, this coefficient is not robust across specifications. This finding is not surprising since the medium of instruction in most universities is English.

The second major empirical challenge is separating the impact of language from other characteristics of ethnicity and culture that are correlated with language.[5] To ensure that the results reflect the impact of language and not systematic cultural, geopolitical or economic differences between minority and majority districts, I conduct three additional tests. The first test restricts the sample to those minority districts that border majority districts to reduce the potential impact of factors that

[4]For example, Bleakley and Chin (2008) discuss the intergenerational transmission of language skills among immigrants to the United States.

[5]Indeed, a number of studies use linguistic variation as a measure of ethnic cleavages. See Easterly and Levine (1997) and the extensive literature cited in Alesina and Ferrara (2005).

influence educational outcomes, but are unobserved in the data. The results of this test are even stronger than the baseline specification. The second test confirms that language, rather than socio-economic characteristics correlated with language, drive the main results. In this test, I examine the impact of the linguistic distance, a measure of difficulty in understanding the language used in schools in minority districts, on educational achievement. The results show that lower distance is associated with better educational outcomes. Finally, if language is a key determinant of educational achievement, then districts where a larger fraction reports the medium of instruction as its mother tongue should report relatively better educational achievement. The third test confirms this hypothesis and shows that the educational achievement is relatively higher in districts with fewer minority language speakers.

While the precise channel that explains these findings is difficult to pinpoint, the analysis suggests that lower costs of comprehension when instruction is in the same language as the mother tongue is the main factor responsible for the results. Specifically, literacy differences between minority and majority districts are greatest among cohorts who were in school before 1956 rather than after, suggesting a major role for a school-based channel of instruction. Conversely, the occupational structure between minority and majority districts is similar, suggesting that access to language-based business networks and associated economic opportunities are unlikely to motivate differences in educational achievement. The number of schools in minority schools is statistically indistinguishable, indicating that access to schools is not a significant channel. Finally, differences in inter-district migration rates between minority and majority districts are very small and statistically insignificant, ruling out migration as a factor explaining these findings.

The reorganization of states in 1956, which assigned previously minority districts to states where they were part of the linguistic majority, reversed the impact of historical shortcomings in educational achievement. Minority districts experienced greater growth in educational achievement after reorganization as they caught up with majority districts so that the gap in educational achievement was closed by 1991. The basic test comparing minority and majority districts found 67.2% higher matriculation growth rates in previously minority districts. So while linguistic mis-

match has a large and persistent effect on educational achievement, realigning mother tongue with the medium of instruction can alleviate this effect.

2 Historical Background

This section outlines the events that inform the empirical analysis in subsequent sections. Section 2.1 argues that colonial-era provincial boundaries were plausibly independent of the language spoken in the constituent districts, which helps to identify the impact of official language. Section 2.2 examines colonial education policy to determine the medium of instruction in schools located in different provinces. Finally, Section 2.3 outlines the process and outcomes associated with the reorganization of state boundaries in 1956, which helps determine whether changing official language has an impact on educational performance.

2.1 British Conquest of India

Commencing in 1757, when the East India Company gained control over the province of Bengal, British colonial rule in India lasted 190 years. From 1757 to 1857, the Company extended its control over the rest of India. The East India Company's administration ended after the mutiny of 1857 and India was ruled directly as part of the British Empire. In peninsular India, the Company obtained feudal control over the Coromandel coast from the Nawab of Carnatic in 1640. The geographically contiguous areas around the trading post of Fort St. George (later called Madras and now Chennai) formed the Madras Presidency. In Western India, seven islands acquired from Portugal as part of a royal dowry in 1661 became what is now the city of 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.

In addition to direct rule by the British, a number of regions were indirectly ruled through the agency of native kings and princes. The major princely states in peninsular India were Hyderabad, Mysore and Travancore. Iyer (2010) shows that the British were selective about which regions

were annexed for direct rule. Therefore, a key concern for the empirical analysis presented in subsequent sections is whether selection of regions administered directly was correlated with the linguistic characteristics of those areas.[6]

The main events that shaped the boundaries of colonial South India were the Anglo-Mysore Wars between Tipu Sultan of Mysore and the British allied with the Marathas and the Nizam of Hyderabad. Oak and Swamy (2012) describe the process of alliance formation and the pre-war commitments between the British, the Marathas and the Nizam on territorial division, arguing that the commitments were credible and not subject to post-war negotiations. After Tipu's defeat in the Third Anglo-Mysore war, the Marathas regained Dharwar,[7] the Gulbarga region was returned to the Nizam and the British added Malabar, Salem, Bellary and Anantapur to the Madras Presidency.[8] Neither language nor economic factors had a major role to play in determining colonial boundaries. Indeed, Chelmsford (1918) described 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 rulers or administrative convenience

The commission recommended reorganization of states to enable more coherent administration.

[6]Iyer (2010) uses the annexation policy called "Doctrine of Lapse," which specified that a territory was annexed if the ruler died without a natural male heir, as a source of exogenous variation to determine the impact of direct versus indirect British rule. This policy was relevant in North India which had a large number of small states, but less germane in the context of the present study which focuses on South India.

[7]Tipu's father Hyder Ali had captured the fort of Dharwar in 1778 and therefore the Marathas had historical claims on the region.

[8]The British annexed Malabar to prevent Mysore's access to the sea.

> 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. 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 with very few English language teachers or materials. 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).[9]

Educational achievement primarily reflected facility with vernacular languages, with only 14.3% of all literates also able to read and write in English, since in most cases the state or province's major vernacular was the medium of instruction. 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. In Madras Presidency, the medium of instruction was Tamil, except in Andhra districts where Telugu was used (Krishnamurti 1978). In Hyderabad state, Urdu was the language most used in schools as well as official correspondence and judicial transactions even though most of the population spoke Telugu (Reddy 1987). The princely states of Travancore and Mysore used Malayalam and Kannada respectively as the medium of instruction (Tharakan 1984).

At the post-secondary level, the British established universities in Calcutta, Bombay and Madras starting from 1857 (Radhakrishnan 1948). Unlike primary and secondary schools, these universities employed a number of British faculty members and the medium of instruction was English (Annamalai 2004). 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

The nationalist leadership in India before Independence recognized the value of reorganization of states. Although the Indian National Congress, the main nationalist party, endorsed the principle of the linguistic provinces, India's Independence was accompanied by Partition on religious lines

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

which dampened enthusiasm for further division on an ethnic or cultural basis (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, the death of an activist demanding a separate state for Telugu speakers following a hunger strike led to the formation of Andhra Pradesh from the Telugu speaking districts of Madras Province along with the formation of the 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 (the modern states of Andhra Pradesh, Karnataka, Kerala and Tamil Nadu) was reorganized in 1956, West India (Maharashtra and Gujarat) in 1960 and the rest of India in 1965 and 1970 (formation of Haryana, Himachal Pradesh and Nagaland).

Table 2 shows that the commission followed the linguistic majority rule for every district using language data from the 1951 census. Telugu was the most common mother tongue in all districts were assigned to Andhra Pradesh, Kannada in each district assigned to Karnataka, Malayalam in every district to Kerala and Tamil in districts to Tamil Nadu. This rules out both the possibility of bargaining or unobserved district characteristics influencing the reorganization exercise.

3 Data Description

Estimating the impact of language requires district level data on demographic and economic characteristics before and after 1956. The dataset should contain variables that represent 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, religion and caste composition from the 1951 Census of India. This allows me to measure the baseline rates of educational achievement 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 of variables, including educational achievement measures, for 140 districts.[10]

The 1951 and 1961 Census asked respondents about their *mother-tongue*, described in the census forms as the language "first spoken by the individual from the cradle."' In addition, Census forms in all waves defined *literacy* as the "ability to read and write in any language". Since this definition is not specific to literacy in either the mother-tongue or the state's official language, me-

[10]Selective data reporting is not a major concern since the remaining handbooks were destroyed by humidity and pests, and these factors are unlikely to be correlated with economic outcomes in 1951.

chanical differences in literacy levels as a result of redefining literacy when a district is reassigned from one state to another are ruled out.

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 (IMD) provides monthly rainfall readings at the sub-division level.[11] I calculate the mean and variance of the aggregate rainfall in the months of January and July for each census decade and include these four measures in the dataset. Also added are district level indicators for various soil types, especially the fraction of land which is either wasteland or under forests (Department of Land Resources 2000), along with the latitude, longitude and elevation of the district headquarters.

The study is restricted to districts situated in the modern states of South India for four reasons. First, the first wave of reorganization in 1956 took place in South India only. In subsequent waves of reorganization, 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 between 1971 and 2001. Most of these changes were in North India, whereas district boundaries in South India have remained relatively stable over time. Although they offer a method to correct for population totals, correcting for other variables such as educational achievement is not possible.[12] A third issue is that North India contained a large number of small princely states (some just a few square kilometres in area) where education policies are not documented in the literature and difficult to ascertain. In contrast, South India contained only three large princely states (Travancore, Hyderabad and Mysore) where education policies are well documented. Finally, the modern states of Andhra Pradesh, Karnataka, Kerala and Tamil Nadu exhibit significant diversity in major languages (unlike North India where Hindi is

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

[12] The changes in South India between 1951 and 1991 consisted of a small number of districts split into two successor districts. Barnes and Vanneman (2000) consolidate districts up to the parent district which allows all variables of interest to be assigned to the correct district. For example, the acreage under wasteland is available for the split districts in 2000. Adding this acreage for the two split districts and calculating the fraction of wasteland in the consolidated district is straightforward. As a result of consolidation, the sample consists of 67 districts though modern South India consists of 93 districts.

widespread), allowing for cleaner identification of the effects of language on economic outcomes.

The language used by the majority of residents within a district identifies it as either a "majority" or a "minority" district. In a majority district, the district's majority language was the same as the province's official language (also used as the medium of instruction in schools) in the colonial era. In a minority district, the majority language of the district was different from the official medium of school instruction. After reorganization, minority districts were assigned to a state formed on the basis of its language, which was then uniformly used the medium of instruction in schools within the state. Figure 1 shows how each modern district is classified.

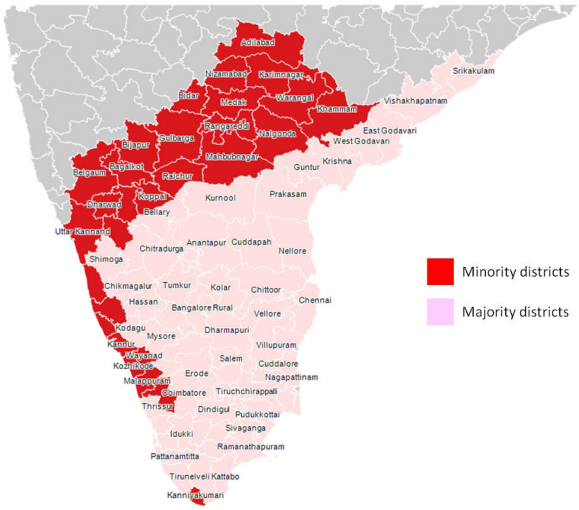


Figure 1: **District assignments in South India**

The dataset yields 335 district-year observations (67 districts each observed five times from 1951 to 1991), although a number of observations are missing for various dependent variables. Table 4 summarizes some of the variables of interest. Among time-invariant characteristics, 22 out of 67 districts are minority districts, 40% have some coastal boundary and nearly 63% of the districts were ruled directly by the British. On average, 13.2% of the district's land was wasteland and 5.6% was under forest cover. The table also reports the mean altitude (1116.3 feet), latitude (13.6 °N) and longitude (77.5 °E) of districts' headquarters. Time varying characteristics reported in Table 4 include the fraction of Scheduled Castes (14.3%) and Scheduled Tribes (3.0%) in the population, the cumulative literacy rate over time (37.0%) and the literacy rate by various cohorts. Younger

cohorts report higher literacy levels which is consistent with the expansion of basic education over time.[13] Completion rates decline for higher education levels, with 25.8% of the population completing primary school, 6% matriculating high school and only 1.2% graduating from college.

4 Empirical Analysis

The objective of the empirical exercise is to estimate the impact of education in the mother tongue on educational outcomes. The classification of a district as a historically minority language or majority language district depends on the exogenous province formation by the British.[14] The subsequent assignment of districts to states in 1956 is on strict linguistic lines. Therefore, the difference in outcomes between minority and majority districts, as well as the difference in growth rates after reorganization identifies the impact of language assignment on educational outcomes.

If the mother tongue of a district is the same as the medium of instruction used in schools, then I expect that more students should be encouraged to enrol in school and achieve basic literacy. I also expect to observe greater rates of completion for subsequent levels of schooling (primary school, middle school, high school and college) both because the supply of students from earlier stages increases and because the demand for schooling increases when students understand instruction better. To test this hypothesis, Section 4.1 estimates a first differences model of the impact of minority status on educational outcomes.

Benhabib and Spiegel (2005) develop a model where countries above a threshold level of human capital "catch-up" with technologically advanced countries once they are able to access technology as well. If language is a "social technology", then districts should experience catch-up growth in educational achievement after reassignment corrects the language mismatch. Section 4.3 tests for whether minority districts experience greater growth in educational achievement rates by estimating a difference-in-difference of the impact of minority versus majority districts on edu-

[13] Cohort-wise literacy available only for census waves from 1961 to 1991 and not for the 1951 census.

[14] Appendix A confirms that the results presented in this section are not sensitive to the influence of direct versus indirect British rule.

cational outcomes before and after reorganization.

4.1 Do minority districts have persistently poorer outcomes?

This section tests whether districts with historical minority language status have persistently poorer educational outcomes. In addition to testing for differences between minority and majority districts, I conduct three robustness checks. The first robustness exercise estimates a first differences model restricting the sample to the set of minority and majority districts that border each other. The second robustness check uses the exogenously determined linguistic distance between the district's majority language and the language used as the medium of instruction in schools. I expect that increasing linguistic distance within minority districts will lead to relatively poorer economic outcomes. Third, I expect that in more polarized districts, where the fraction of minority language speakers is large, outcomes will be relatively poorer compared to districts where the number of speakers of each language are evenly matched.

4.1.1 Test using minority status

This section estimates the difference in outcomes between minority and majority districts. Since pre-assignment data on district-level economic and social characteristics is not available, the key identifying assumption in this model is that the initial assignment of districts as minority or majority districts is not correlated with outcomes, which is justified in Section 2.1. Therefore, I specify the following model.

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

The variable y_{it} represents log of outcomes where I expect systematic differences between minority and majority districts. $minority_i$ is an indicator variable that is 1 if district i is a minority district as defined earlier, and 0 otherwise. Thus, the main 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. The specification does not include district fixed effects because $minority_i$ is constant over time. Hence, I introduce $\mathbf{Z}_i$, which is a vector of time-invariant district characteristics such as the fraction of terrain that is forested or wasteland, a coastal dummy and the altitude of the district headquarters, all of which potentially impact y_{it}. A casual examination of Figure 1 shows that minority districts are geographically clustered in the Northern and Western parts of the peninsula. To capture this aspect of geography, I include the longitude and latitude of the district headquarters. Also included in $\mathbf{Z}_i$ is a dummy variable that indicates whether a district was under direct British rule or ruled indirectly through the agency of a princely state.[15] Therefore, this dummy accounts for factors that are unobserved in the data, such as politics, that might have caused educational provision to be very different across minority and majority districts.[16]

Equation (1) includes a vector of observed time-varying district characteristics $\mathbf{X}_{it}$ that consists of the average and standard deviation in January and July rainfall over the decade and the fraction of residents who are from historically disadvantaged Scheduled Caste and Scheduled Tribe backgrounds.[17] Decade fixed effects ($decade_t$) account for observed and unobserved decade characteristics that affect outcomes for all districts. State fixed effects ($state_i$) control for state characteristics such as attitudes and cultural influences that are constant over time. Finally, unobserved district characteristics are clustered by district.

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 less robust since English is commonly used as the medium of instruction in higher education, mitigating the impact of historical differences in language use.

Column I of Table 5 presents estimates for β_1 with different outcome variables. In this table, the

[15] An alternative is to include a variable that represents the length of British rule. However, since the provincial boundaries of South India were determined well before Wood's Dispatch in 1839, the length of exposure to British education does not vary.

[16] For example, the rulers of Hyderabad, Mysore and Travancore might have had strong political incentives to expand education, partly to forestall direct British rule.

[17] When the outcome variable examines rural residents only, this variable is the fraction of *rural* residents who are from SC or ST backgrounds.

coefficients for literacy rates are negative (-0.208 for total literacy and -0.254 for literacy in rural areas, $p < 0.01$). Since the dependent variable is log transformed, taking the exponential of these coefficients implies that the total and rural literacy rates in minority districts are 18.8% and 22.4% lower than the corresponding rates in majority districts. Similarly, the differences in middle school completion and matriculation rates are, after exponential transformation of the coefficients, even larger (24.4% and 30.4% lower in minority districts than majority districts). In college graduation rates, the fraction of graduates in minority districts is 27.6% smaller than majority districts. The increasing gap as the level of education increases is potentially due to sequentially lower supply of students at the next level. These results can be placed in context of other programs to boost education such as India's school subsidies for girls (National Programme for Education of Girls at the Elementary Level) which increased enrolment by 3 percentage points (Debnath 2013), the Indonesian INPRES school building program which increased schooling by 0.12 to 0.19 years (Duflo 2001), and Mexico's Progresa program which increased enrolment up to 11.1% (Schultz 2004).

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 estimate equation (1) on a sample of 37 districts that share a geographical border, but different classifications as minority or majority districts (Figure 2). Since these districts are arguably similar on unobserved characteristics compared to districts far from each other, this strategy helps mitigate omitted variable bias.

Column II of Table 5 shows larger and more precisely estimated differences between historically minority and majority districts in education achievement rates when considering the restricted sample (-28.5% for literacy, -44.8% for middle school completion, -48.2% for matriculation and -50.8% lower for graduation). Figures 5 and 6 show the log of the literacy and college graduation rates separately for every decade from 1951 to 1991.[18] These show that the difference between minority and majority districts persists for every year in the decade, although the difference dimin-

[18]I chose literacy and graduation rates since the complete time series for every wave between 1951 and 1991 is available only for these two measures.

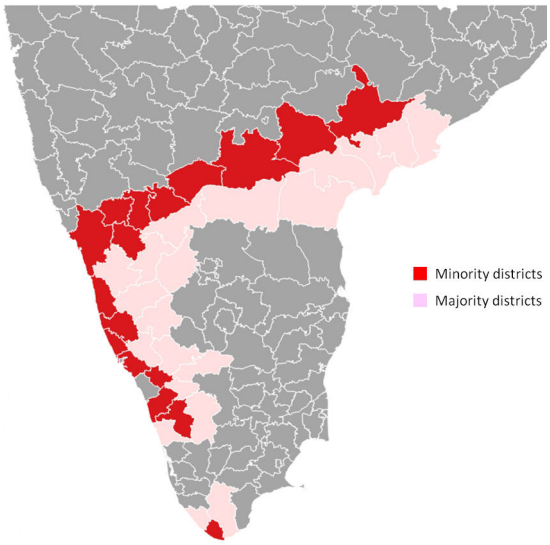


Figure 2: **Sub-sample of neighboring districts**

ishes over time.

Thus, the test of first differences between minority and majority language districts offers evidence that historical mis-assignment on the basis of language has persistent impact on modern educational outcomes. An additional test using a restricted sample of districts bordering each other 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 results presented in the previous section is that language may be correlated with systematic cultural (for instance, access to social capital, or differences in social heterogeneity) differences, so the findings represent cultural rather than linguistic differences between minority and majority districts. To address this concern, I propose a test using a measure of linguistic distance that is logically orthogonal to educational 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.[19] 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, in Figure 3, Tamil and Malayalam are close to each other on the family tree, implying that learning one language is relatively easy for speakers of the other language. This ease is captured by the pairwise linguistic distance of 3 between the two languages. On the other hand, Telugu speakers find it difficult to understand or learn Kannada and vice versa, which is represented by a pairwise linguistic distance of 6.

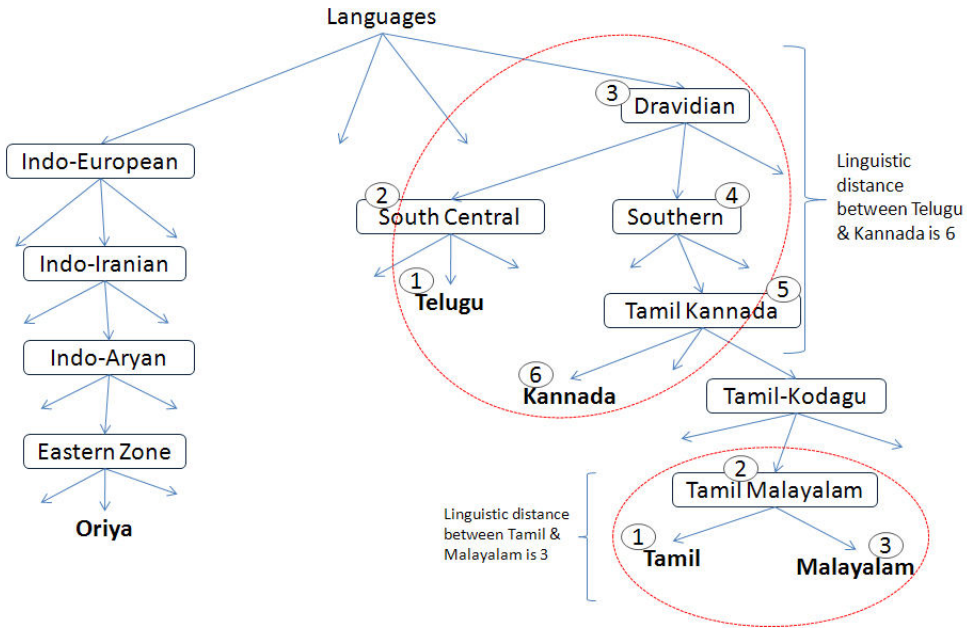


Figure 3: **Construction of linguistic distance**

Table 6 reports the linguistic distance between each pair of Scheduled Indian languages, where the average pairwise distance between the four relevant South Indian languages is 5.83.[20] Using the data from Table 6, I assign a linguistic distance measure to each minority district based on the pair

[19]Chiswick and Miller (2005) and Beenstock, Chiswick, and Repetto (2001) also develop and use measures of linguistic distance to report the pairwise distance between various languages.

[20]Shastry (2012) reports that this measure is strongly correlated with two alternative and logically independent measures of linguistic distance. The first, developed by Shastry (2012) 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. This study cannot use the Shastry (2012) 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.

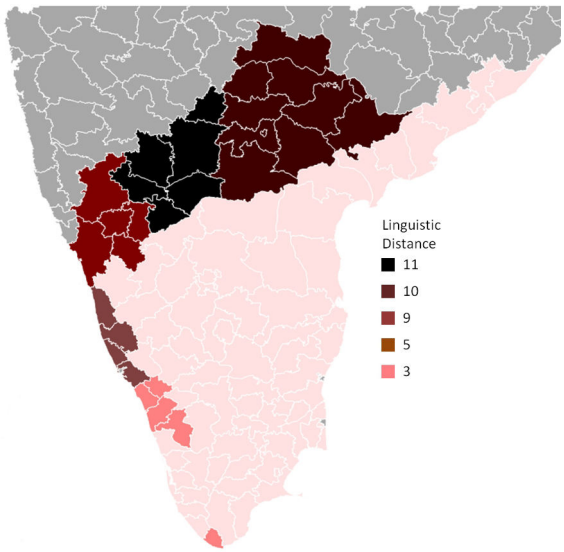


Figure 4: **Linguistic distance by district**

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

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

In this specification, β_2 is the marginal impact of a unit increase in linguistic distance on outcomes in a minority district. I expect $\beta_2 < 0$ if increasing linguistic distance between the mother tongue and the official language makes it more difficult to complete various educational levels.

Table 7 presents OLS estimates of β_2 using the same set of educational outcomes as the previous section. Note that since linguistic distance is constructed by counting nodes on the language tree, β_2 does not have direct economic interpretation. The estimates suggest that economic outcomes are poorer with increase in linguistic distance between the district's mother tongue and historic official language. Table 7 reports that a marginal increase in linguistic distance within minority

districts decreases literacy and middle school completion rates (-0.025 and -0.040, respectively, with both estimates statistically significant). Matriculation rates are also lower in minority districts but the point estimate is very small (-0.001) and cannot be statistically distinguished from the null. Since English is the medium of instruction at the tertiary level, linguistic distance between vernacular languages does not have an impact on graduation rates, as evidenced by the small positive coefficient on college graduation (0.041 and statistically insignificant). Conversely, 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 effects manifest themselves early in the education process when local language teaching is more important.

4.1.3 Test using minority fraction measure

The binary *minority* variable as defined and used in previous sections does not capture the intra-district mix of languages used. I expect that minority districts that are more polarized (i.e. fewer speakers of the official language) will experience poorer educational outcomes than otherwise. I propose a 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 more evenly matched.

$$MinorityFraction = \frac{MotherTongue - OfficialLang}{TotalPopulation} \tag{3}$$

In this definition, *OfficialLang* is the number of speakers of the official language of the pre-reorganization state where the district is located. *MotherTongue* is the number of speakers of the most popular language spoken in the district other than the state's official language. Hence, for minority districts, $MotherTongue > OfficialLang$ 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).

$$y_{it} = \beta_0 + \beta_1 MinorityFraction_i + \beta_2 \mathbf{Z}_i + \beta_3 \mathbf{X}_{it} + decade_t + state_i + \mu_i + \epsilon_{it} \quad (4)$$

In equation (4), β_1 indicates the marginal impact of increasing the share of minority language speakers on outcome variable y_{it}. Table 8 shows that districts with large minority language populations suffer from greater shortfalls in educational attainment. The coefficients associated with literacy, middle school completion and matriculation are -0.111, -0.154 and -0.178, respectively ($p < 0.01$ for all). The coefficient on college graduation is -0.143 ($p < 0.01$). 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. This suggests that language mis-assignment had a persistent impact on educational outcomes.

4.2 Channels

This section examines various mechanisms through which language might affect educational achievement. Although it is difficult to pinpoint the precise channel with the available data, I find that factors within the school, such as the medium of education, are responsible for language affecting educational achievement while simultaneously ruling out the effects on occupational choice, provision of public schools and inter-district migration.

Figure 7 shows the impact of language on literacy rates among various age cohorts. All four tests described earlier consistently show that the negative impact of language mismatch on literacy is largest among older age groups, which is not surprising since these cohorts are more likely to have been in school before 1956. Therefore, school-based channels such as the impact of medium of instruction on students' ability to understand the material are likely to explain the differences in

educational achievement between minority and majority districts.

A second mechanism through which language might influence educational attainment is through access to new occupations and business networks. If communication intensive occupations (such as those in the secondary and tertiary sectors) require knowing the majority language of the state or province, then the returns to education will be greater for individuals who live in majority districts. However, regardless of the empirical specification, Table 10 shows no significant differences in the occupational structure between the minority and majority districts. This suggests that districts' historical language status did not persistently affect occupational structure. Hence, access to communication-intensive occupations or business networks is unlikely to be a major channel to motivate greater investments in education.

The third mechanism that I examine is the impact of public investments, especially school building, on educational attainment. State governments, which control primary and secondary education in India, might invest in minority language areas with relatively lower educational achievement to spur improvements in schooling completion rates. Alternatively, if representatives from majority language districts are more likely to form the government, they may reward constituents with more educational infrastructure. Table 10 reports differences in the presence of various school types in minority versus majority districts. Although all coefficients are negative, suggesting fewer schools in minority language districts, these cannot be statistically differentiated from the null in the various empirical tests. Thus, while I cannot conclusively rule out the role of public investments as an explanation for differences in educational achievement, the empirical evidence for this channel is weak.

Finally, I examine the impact of inter-district migration. Language policies might impact educational outcomes if individuals in minority districts who are more motivated to study migrate to majority districts, enrol in schools and complete increasing levels of education. To estimate the impact of this channel, I use the same specifications presented in Section 4.1. Along with total migration, I also include the male migration rate as the dependent variable since men are more likely to migrate in search of work than entire families. Table 11 shows that the coefficients

for the minority variable are all small and statistically insignificant which is consistent with other empirical studies that do not find significant inter-district migration in India over this period (Munshi and Rosenzweig 2009). This suggests that districts' historical linguistic status did not have much impact on migration and inter-district migration is unlikely channel through which language influences educational outcomes.

4.3 Do minority districts catch up after reorganization?

This section exploits 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. Only those districts for which 1951 census data is available are included in the sample.

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

As before, y_{it} represents an educational outcome as measured in each census wave from 1951 to 1991. $Post_t$ is an indicator variable that is 0 if the year is 1951 and 1 otherwise. 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 increase in enrolment and completion of formal education among minority language districts.

Column I in Table 12 shows that reassignment had a large and significant impact on middle school completion and matriculation rates, in which minority districts experienced growth rates of 73.2% and 67.2% greater than majority districts, respectively. Literacy and college graduation growth rates also were also higher by 21.1% and 29.6%, respectively, though neither can be statistically distinguished from the null. Figure 5 shows that the difference in literacy between minority and majority districts is large before reorganization and diminishing after 1956, suggesting that alignment of a district's mother tongue with the state's official language is positively associated

with enrolment in schooling. However, Figure 6 does not show much change in college graduation rates between 1951 and 1961, and minority and majority districts converge on this measure only in later decades. This is not surprising, since changing the medium of instruction in 1956 is unlikely to change the supply of potential college students by 1961. Instead, increases in schooling in the 1950s and 1960s as a result of the reorganization of states will impact university-level education by 1970 at the earliest.

Column II in Table 12 reports coefficients associated with a specification where $minority_i$ is interacted with $Post_t$ and L_i. I find that in addition to the middle school and matriculation rates, the coefficient associated with the literacy is also significant. This suggests that districts where the mother tongue was linguistically distant from the official language before 1956 benefitted more from reorganization, and literacy improved drastically as a result of instruction in the mother tongue. The coefficient associated with university graduation rates remains insignificant.

Finally, Column III reports the coefficient associated with *MinorityFraction* interacted with $Post_t$. The important difference from Columns I and II is that the coefficient on college graduation rates is also significant at the 10% level in Column III. Although this suggests that minority districts with a relatively large fraction of residents who do not speak the official language might have experienced greater university enrolment and completion rates, this coefficient is fragile across specifications. This is not surprising since the medium of instruction in most universities remained English both before and after the reorganization, instead of switching to the official language.

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 not only demonstrated that colonial-era provinces which mixed together linguistic minority and majority areas lead to differences in long term educational

outcomes, but also that language-based reorganization of state boundaries might help remedy these differences.

The magnitude of educational differences between districts that do and do not speak the official language is large, indicating that historical factors can dominate government policies or private remedies aimed at alleviating educational shortfalls. Nonetheless, catch up by minority districts after reorganization suggests that historical mismatches can be remedied through large scale political changes.

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 Telangana state separated from Andhra Pradesh. The results presented in this paper indicate that new states formed on the basis of shared language might experience better educational outcomes.

References

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

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*(2), 155–194.

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

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.

Ban, R., S. Jha, and V. Rao (2012). Who has voice in a deliberative democracy? Evidence from transcripts of village parliaments in south India. *Journal of Development Economics 99*(2),

428–438.

Banerjee, A. and L. Iyer (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *American Economic Review 95*(4), 1190–1213.

Banerjee, A., L. Iyer, and R. Somanathan (2005). History, social divisions, and public goods in rural India. *Journal of the European Economic Association 3*(2-3), 639–647.

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

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

Bleakley, H. and A. Chin (2004). Language skills and earnings: Evidence from childhood immigrants. *Review of Economics and Statistics 86*(2), 267–298.

Bleakley, H. and A. Chin (2008). What holds back the second generation? The intergenerational transmission of language human capital among immigrants. *Journal of Human Resources 43*(2), 267–298.

Chakraborty, T. and S. Bakshi (2013). English language premium: Evidence from a policy experiment in India. Mimeo.

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

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

Debnath, S. (2013). School subsidy for girls and gender gap in enrolment. Mimeo.

Department of Land Resources (2000). *Wasteland Atlas of India*. New Delhi, India: Ministry of Rural Development, Government of India.

Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review 91*(4), 795–813.

Dyen, I., J. Kruskal, and P. Black (1997). Comparative Indo-European database. Collected by Isidore Dyen. File IE-Rate 1.

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

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.

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

Krishnamurti, B. (1978). Language planning and development: The case of Telugu. *Contributions to Asian Studies 2*, 37–56.

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

Lewis, P. (Ed.) (2009). *Ethnologue: Languages of the world, Sixteenth edition*. Dallas, TX: SIL International.

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

Oak, M. and A. Swamy (2012). Myopia or strategic behavior? Indian regimes and the East India Company in late eighteenth century India. *Explorations in Economic History 49*(3), 352–366.

Radhakrishnan, S. (1948). *Report of the university education commission*. New Delhi: Government of India Press.

Reddy, V. (1987). *Economic history of Hyderabad State: Warangal Suba, 1911-1950*. New Delhi, India: Gyan Books.

Schultz, T. (2004). School subsidies for the poor: Evaluating the Mexican Progresa poverty program. *Journal of Development Economics 74*(1), 199 – 250.

Shastry, K. (2012). Human capital response to globalization: Education and information technology in India. *Journal of Human Resources 47*(2), 287–330.

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

Tharakan, P. (1984). Socio-economic factors in educational development: Case of nineteenth century Travancore. *Economic & Political Weekly 19*(45), 1913–1928.

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

A Impact of direct versus indirect British rule

A concern in the empirical analysis presented in Section 4.1 is whether the pattern of British conquest correlated with linguistic characteristics. If the British selected the districts with the greatest economic potential which was also correlated with linguistic status, then the impact of this selection would be misrepresented as the impact of language on educational outcomes.[21] One strategy for addressing this concern is to add an interaction term $minority_i * British_i$ to equation (1) to separate the impact of direct British rule from linguistic characteristics. This term controls for the impact of direct British rule in minority districts. As before, the specification also includes a term for direct British control in the vector $\mathbf{Z}_i$.

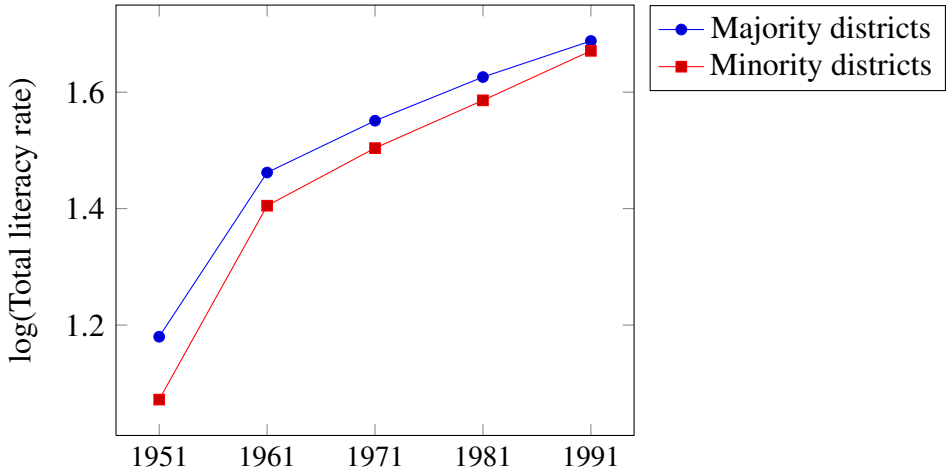
$$y_{it} = \beta_0 + \beta_1 minority_i + \beta_2 minority_i * British_i + \beta_3 \mathbf{Z}_i + \beta_4 \mathbf{X}_{it} + decade_t + state_i + \mu_i + \epsilon_{it} \quad (6)$$

Table 13 shows estimates of the coefficients associated with $minority_i$, $minority_i * British_i$ and $British_i$. The results show that language, rather than direct British rule, is the main determinant

[21] Although I control for a number of georgaphical and demographic variables that might impact education, other factors such as the presence of business communities or transportation links are unobserved in the data.

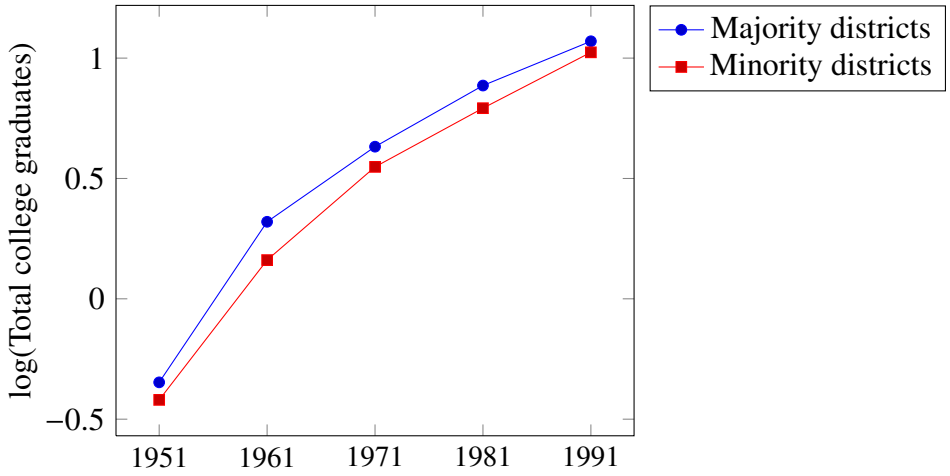
of educational outcomes. The coefficients associated with $minority_i$ are comparable to those presented in Table 5. The coefficients associated with $British_i$ are negative (although all statistically indistinguishable from the null), consistent with Iyer (2010) who found that princely states made greater investments in public education. Most coefficients associated with $minority_i * British_i$ are positive, which perhaps suggests that the British made greater educational investments in minority language districts than princely states. However, these coefficients are also statistically insignificant, leading to the conclusion that the difference on the basis of language alone had a significant influence on educational achievement.

Figure 5: **Total literacy by year**



Notes: This figure shows the log of total literacy in minority and majority districts over time. Source: Census of India 1951-1991.

Figure 6: **College graduation rate by year**



Notes: This figure shows the log of college graduates in minority and majority districts over time. Source: Census of India 1951-1991.

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%

Notes: This table shows the fraction of speakers of each major Indian language who reside in the state formed on the basis of that language. *Scheduled languages* are major Indian languages listed in the Eighth schedule of the Constitution. Source: Census of India 2001.

Table 2: **Verification of language-based reorganization**

State	Major languages			State	Major languages		
Andhra Pradesh	**Telugu**	**Urdu**	**Rajasthani**	**Tamil Nadu**	**Tamil**	**Telugu**	**Kannada**
Srikakulam	1895114	11966	5	Madras	961743	234379	15866
Visakhapatnam	1991689	12803	58	Chingleput	1628012	397088	6537
E. Godavari	2270664	20766	249	N. Arcot	2309896	340313	24316
W. Godavari	1647308	22979	634	S. Arcot	2537541	179893	6459
Krishna	1679530	76655	8706	Salem	2589194	488885	189387
Guntur	2359100	164474	7512	Coimbatore	2138996	653070	267196
Nellore	1641350	108934	30	Nilgiri	120831	26469	107492
Chittoor	1225248	N.A.	6760	Madurai	2216038	417447	150706
Cuddapah	1006250	N.A.	2692	Tiruchirapalli	2587107	249696	14621
Anantapur	1145948	N.A.	19791	Thanjavur	2876292	57419	5137
Kurnool	1294510	N.A.	6401	Ramanathapuram	1819079	195750	44241
Mahbubnagar	1200793	123187	60929	Tirunelveli	2272394	158763	8177
Hyderabad	1041773	571422	32037	Kanyakumari	667528	1463	242
Medak	882177	123715	20243				
Nizamabad	634981	80777	24873	**Karnataka**	**Kannada**	**Telugu**	**Urdu**
Adilabad	483094	64992	32770	Bangalore	1049419	378319	N.A.
Karimnagar	1341214	58284	10792	Tumkur	908688	137540	N.A.
Warangal	1129675	66046	99547	Chitradurga	628600	138465	N.A.
Khammam	636394	39707	51896	Kolar	610974	238604	N.A.
Nalgonda	1121609	52971	69530	Bellary	554499	96147	N.A.
				Mysore	1199303	53047	N.A.
Kerala	**Malayalam**	**Tamil**	**Konkani**	Mandya	667528	11657	N.A.
Cannore	1139740	3404	8498	S. Kanara	272983	2045	15847
Kozhikode	1782135	5322	9	Coorg	80410	3927	3818
Palghat	1269160	102827	1	Hassan	614079	17904	N.A.
Trichur	1165772	17206	3197	Shimoga	492808	26441	N.A.
Ernakulam	1232939	9148	13314	Chikmagalur	297246	14949	N.A.
Kottayam	1046815	112308	461	Belgaum	1085195	14974	120155
Alleppey	1288937	4380	5317	Bijapur	1177809	13428	128455
Quilon	1293174	35176	156	N. Kanara	283024	3313	36986
Trivandrum	966787	53106	32	Dharwar	1282604	22927	178546
				Gulbarga	784387	122992	240202
				Bidar	200689	81251	94562
				Raichur	717485	108472	94879

Notes:This table verifies that the allocation of districts to states in the 1956 reogranization was strictly on linguistic lines. "N.A." indicates data not available.
Source: Census of India 1951.

Table 3: **District allocation**

Region	Pre-1956 state	Post-1956 state	Districts (Modern names)	Pre-Post 1956 language	Classification
Telangana	Hyderabad	Andhra Pradesh	Adilabad, Karimnagar, Khammam, Mahbubnagar, Medak, Nalgonda, Nizamabad, Rangareddy, Warangal	Urdu to Telugu	Minority
Andhra	Madras	Andhra Pradesh	Srikakulam, Vizianagaram, Visakhapatnam, Nellore, East Godavari, West Godavari, Krishna, Prakasam, Guntur, Anantapur, Chittoor, Kadapa, Kurnool	Telugu to Telugu	Majority
Mysore	Mysore	Karnataka	Chamrajnagar, Mysore, Kolar, Mandya, Tumkur, Bangalore Rural, Bangalore Urban, Bellary, Haveri, Shivamogga, Chickamagaluru, Chitradurga, Hassan	Kannada to Kannada	Majority
Coorg	Coorg	Karnataka	Kodagu	Tulu to Kannada	Minority
Coastal Karnataka	Madras	Karnataka	Dakshin Kannada, Udupi, Uttar Kannada	Tamil to Kannada	Minority
Gulbarga	Bombay	Karnataka	Bidar, Bijapur, Gulbarga, Koppal, Raichur	Marathi to Kannada	Minority
Belgaum	Bombay	Karnataka	Bagalkot, Belgaum, Dharwar, Gadag	Marathi to Kannada	Minority
Travancore	Travancore	Kerala	Kottayam, Idukki, Alappuzha, Pathanamthitta, Kollam, Ernakulam, Thiruvananthapuram	Malayalam to Malayalam	Majority
Malabar	Madras	Kerala	Kasaragod, Kannur, Kozhikode, Wayanad, Malappuram, Palakkad	Tamil to Malayalam	Minority
Tamil Nadu	Madras	Tamil Nadu	All modern districts	Tamil to Tamil	Majority
Kanyakumari	Travancore	Tamil Nadu	Kanyakumari	Malayalam to Tamil	Minority

Notes: This table shows the allocation of each 1951 district before and after 1956 and the classification as a minority or majority district. Source: Census of India, 1951 - 1991 and various Acts of Parliament specifying district assignment in states' reorganization.

Table 4: **Summary statistics**

Variable	**Obs.**	**Mean**	**Std. Dev.**
Minority	335	32.8%	
Andhra Pradesh	335	32.8%	
Kerala	335	17.9%	
Karnataka	335	28.4%	
Tamil Nadu	335	20.9%	
Coastal	335	40.0%	
Direct British rule	335	62.7%	
Wasteland	335	13.2%	0.089
Forest land	335	5.6%	0.055
Altitude (feet)	335	1116.3	1278.021
Latitude (degrees)	335	13.6	3.413
Longitude (degrees)	335	77.5	2.485
Scheduled Caste	294	14.3%	0.053
Scheduled Tribe	291	3.0%	0.042
Rainfall July (Mean)	335	2593.3	2402.625
Rainfall January (Mean)	335	89.2	80.303
Rainfall July (Std dev)	335	831.4	675.452
Rainfall January (Std dev)	335	120.8	99.570
Literates (5+ years)	296	37.0%	0.182
Literates (5 to 9 years)	268	35.7%	0.147
Literates (10 to 14 years)	268	61.2%	0.216
Literates (15 to 19 years)	268	57.5%	0.212
Literates (20 to 24 years)	268	52.8%	0.211
Literates (25 to 34 years)	268	46.2%	0.211
Literates (35+ years)	268	34.9%	0.173
Primary school completion	268	25.8%	0.158
Middle school completion	228	14.4%	0.104
Matriculates	295	6.0%	0.053
Graduates	228	1.2%	0.012

This table shows the summary statistics of the final dataset. Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Sources: Census of India 1951-1991, Indian Meteorological Department and Ministry of Rural Development (Government of India).

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

	I: Full sample			II: Bordering districts only		
Dependent Variable	**Coefficient**	**N**	**adj. R-sq**	**Coefficient**	**N**	**adj. R-sq**
Total literacy rate	**-0.208*****	292	0.869	**-0.336*****	128	0.903
	(0.0586)			(0.0911)		
Rural literacy rate	**-0.254*****	290	0.887	**-0.339*****	128	0.894
	(0.0578)			(0.0896)		
Middle school completion rate	**-0.280*****	224	0.903	**-0.595*****	98	0.946
	(0.0907)			(0.102)		
Rural Middle school completion rate	**-0.460*****	222	0.921	**-0.693*****	98	0.926
	(0.0947)			(0.133)		
Matriculation rate	**-0.362*****	291	0.879	**-0.658*****	128	0.937
	(0.110)			(0.131)		
Rural Matriculation rate	**-0.487*****	289	0.932	**-0.698*****	128	0.937
	(0.0942)			(0.134)		
College graduation rate	**-0.323****	224	0.859	**-0.710*****	98	0.917
	(0.146)			(0.202)		
Rural college graduation rate	**-0.455*****	222	0.931	**-0.718*****	98	0.948
	(0.112)			(0.161)		

Notes: This table reports OLS coefficients corresponding to β_1 from equation (1) which estimates the impact of minority language status on various measures of educational achievement. Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department and Ministry of Rural Development (Government of India).

Table 6: **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

This table shows the pairwise linguistic distance between major Indian languages. Source: `http://www.ethnologue.com` and author's calculations.

Table 7: **Results for educational outcomes by linguistic distance**

Dependent Variable	**Coefficient**	**Std. error**	**Obs.**	**adj. R-sq**
Literacy rate	**-0.025****	(0.012)	292	0.870
Middle school completion rate	**-0.040***	(0.020)	224	0.909
Matriculation rate	-0.001	(0.025)	291	0.885
College graduation rate	-0.041	(0.034)	224	0.869

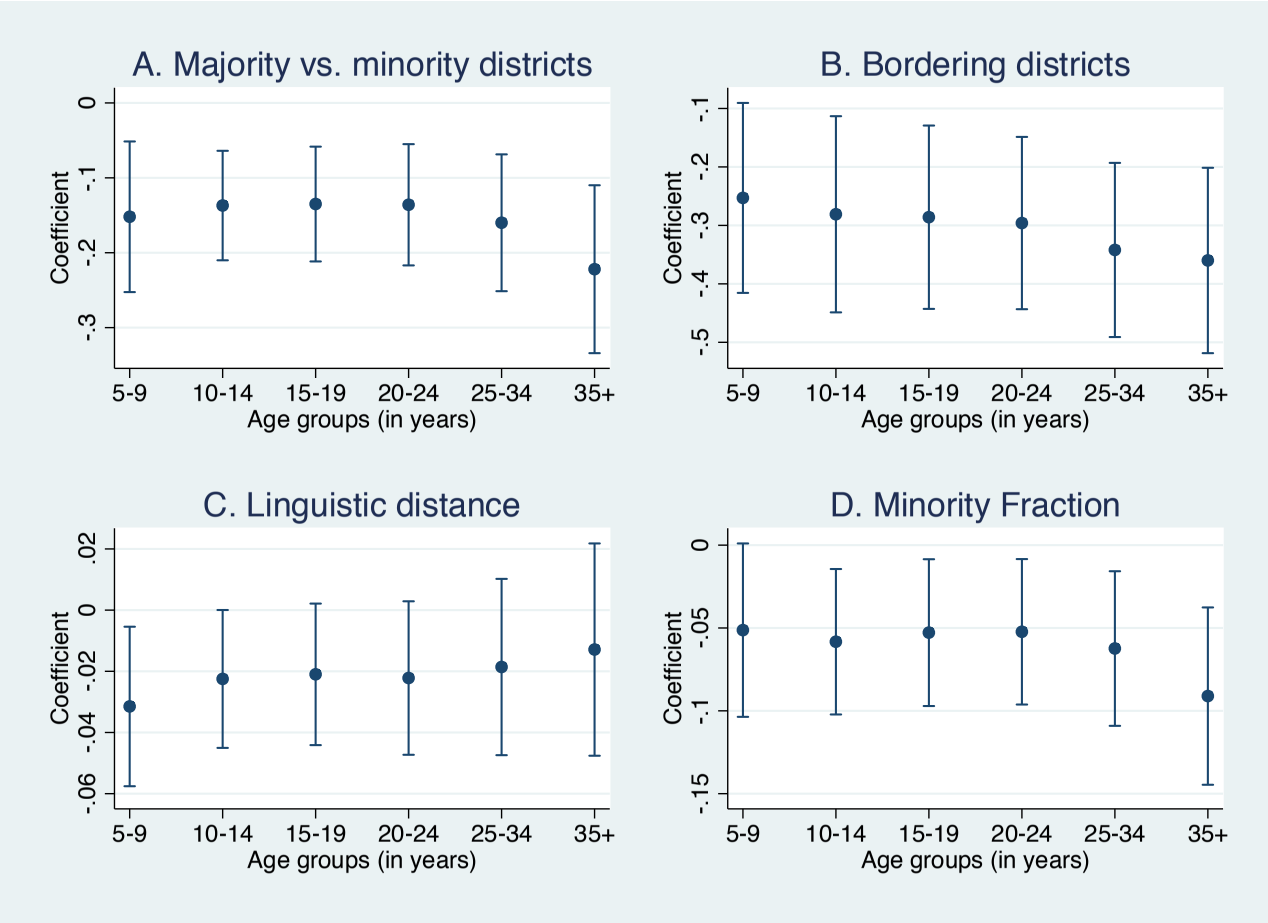
Notes: This table reports OLS coefficients corresponding to β_2 from equation (2) which estimates the impact of linguistic distance between the district's mother tongue and the language used in schools among minority districts on various measures of educational achievement. Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Table 8: **Result for education outcomes using *MinorityFraction***

Dependent Variable	**Coefficient**	**Std. error**	**Obs.**	**adj. R-sq**
Literacy rate	**-0.111*****	(0.033)	264	0.867
Middle school completion rate	**-0.154*****	(0.051)	203	0.907
Matriculation rate	**-0.178*****	(0.058)	263	0.888
College graduation rate	**-0.143***	(0.076)	203	0.881

Notes: This table reports OLS coefficients corresponding to β_1 in equation (4) which estimates the effect of increasing fraction of speakers whose mother tongue is different than the official language on educational achievement. Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Figure 7: **Literacy rate differences by age group**



Notes: The reported coefficients in (A) correspond to β_1 in equation (1), in (B) correspond to β_1 in equation (1), in (C) correspond to β_2 in equation (2), and in (D) correspond to β_1 in equation (4), with the literacy rates for the respective age groups as the dependent variable.

Table 9: **Occupational choice**

Dependent Variable	Coefficient	Std error	Obs.	adj. R-sq
Minority vs. Majority				
All farm workers	-0.097	(0.136)	291	0.341
Land-owning cultivators	**-0.281***	(0.151)	291	0.493
Agricultural labor	0.159	(0.139)	291	0.263
Manufacturing	-0.125	(0.138)	291	0.604
Commerce	-0.093	(0.060)	291	0.600
Transport and communication	-0.139	(0.092)	291	0.415
Minority x Linguistic distance				
All farm workers	0.007	(0.050)	287	0.337
Land-owning cultivators	-0.003	(0.050)	287	0.491
Agricultural labor	0.052	(0.056)	287	0.264
Manufacturing	0.026	(0.039)	287	0.610
Commerce	0.032	(0.024)	287	0.609
Transport and communication	0.055	(0.033)	287	0.429
MinorityFraction				
All farm workers	-0.085	(0.099)	264	0.343
Land-owning cultivators	-0.181	(0.107)	264	0.505
Agricultural labor	0.038	(0.115)	264	0.248
Manufacturing	-0.005	(0.096)	264	0.630
Commerce	-0.032	(0.060)	264	0.617
Transport and communication	-0.075	(0.082)	264	0.470

Notes: This table reports the impact of language on occupational structure. The dependent variables represent the log of the fraction of the population in each occupational category. OLS coefficients correspond to β_1 from equation (1) under "Minority vs. Majority", β_2 from equation (2) under "Minority x Linguistic distance" and β_1 from equation (4) under "MinorityFraction". Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Table 10: **Provision of schools**

Dependent Variable	Coefficient	Std error	Obs.	adj. R-sq
Minority vs. Majority				
Primary schools	-106.8	(73.52)	107	0.373
Middle schools	-16.54	(10.69)	107	0.629
High schools	**-9.913****	(3.806)	107	0.296
Junior colleges	-0.0420	(0.947)	107	0.434
Colleges	-0.154	(0.328)	107	0.242
Minority x Linguistic distance				
Primary schools	-40.03	(31.99)	106	0.381
Middle schools	-11.20	(7.257)	106	0.636
High schools	-0.891	(2.425)	106	0.280
Junior colleges	-0.463	(0.549)	106	0.432
Colleges	-0.0626	(0.125)	106	0.234
MinorityFraction				
Primary schools	-56.56	(74.22)	96	0.411
Middle schools	-6.907	(11.15)	96	0.695
High schools	**-6.179****	(2.473)	96	0.415
Junior colleges	-0.416	(0.786)	96	0.502
Colleges	-0.161	(0.206)	96	0.301

Notes: This table reports the impact of language on the number of schools or colleges per million residents. The reported OLS coefficients correspond to β_1 from equation (1) under "Minority vs. Majority", β_2 from equation (2) under "Minority x Linguistic distance" and β_1 from equation (4) under "MinorityFraction". Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Table 11: **Inter-district migration**

Dependent Variable	Coefficient	Obs.	adj. R-sq
Minority vs. Majority			
Total migration rate	-0.040	268	0.355
	(0.067)		
Male migration rate	-0.170	268	0.299
	(0.106)		
Minority x Linguistic distance			
Total migration rate	0.018	264	0.379
	(0.016)		
Male migration rate	0.022	264	0.303
	(0.029)		
MinorityFraction			
Total migration rate	-0.006	241	0.367
	(0.041)		
Male migration rate	-0.067	241	0.297
	(0.069)		

Notes: This table reports the impact of language on inter-district migration. The reported OLS coefficients correspond to β_1 from equation (1) under "Minority vs. Majority", β_2 from equation (2) under "Minority x Linguistic distance" and β_1 from equation (4) under "MinorityFraction". Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Table 12: **Result for education outcomes before and after reorganization**

	I: Minority x Post			II: Minority x Ling Dist x Post			III: MinorityFraction x Post		
Dependent Variable	**Coefficient**	**Obs.**	**adj. R-sq**	**Coefficient**	**Obs.**	**adj. R-sq**	**Coefficient**	**Obs.**	**adj. R-sq**
Literacy rate	0.192	292	0.871	**0.0196***	292	0.871	**0.127***	264	0.869
	(0.117)			(0.011)			(0.070)		
Middle school completion rate	**0.549***	224	0.907	**0.056***	224	0.908	**0.424****	203	0.914
	(0.324)			(0.032)			(0.176)		
Matriculation rate	**0.514****	291	0.881	**0.0551****	291	0.882	**0.451*****	263	0.894
	(0.240)			(0.0240)			(0.123)		
College graduation rate	0.259	224	0.859	0.0381	224	0.863	**0.327***	203	0.883
	(0.297)			(0.0293)			(0.165)		

Notes: This table reports the impact of language on educational achievement before and after the 1956 reorganization of Indian states. The reported coefficients correspond to β_3 in equation (5). Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).

Table 13: **Result for impact of direct vs. indirect British rule**

Dependent Variable	**Minority**	**Minority x British**	**British**	**Obs.**	**adj. R-sq**
Total literacy rate (5+ years)	**-0.338*** **	0.274	-0.0922	291	0.877
	(0.117)	(0.196)	(0.088)		
Child literacy rate (5 to 9 years)	**-0.255** **	0.196	-0.0798	268	0.817
	(0.099)	(0.165)	(0.078)		
Adult (35+ years) literacy rate	**-0.358** **	0.260	-0.114	268	0.829
	(0.144)	(0.230)	(0.104)		
Primary school completion rate	**-0.273***	0.192	-0.056	268	0.861
	(0.134)	(0.226)	(0.109)		
Middle school completion rate	**-0.405** **	0.286	-0.132	223	0.902
	(0.173)	(0.273)	(0.133)		
Matriculation rate	**-0.370***	0.086	-0.104	290	0.877
	(0.205)	(0.326)	(0.170)		
Diploma	-0.171	-0.133	-0.109	201	0.747
	(0.272)	(0.499)	(0.246)		
Graduation rate	-0.262	-0.062	-0.046	223	0.854
	(0.275)	(0.487)	(0.241)		

Notes: This table checks if the impact of language on educational achievement is affected by direct or indirect British rule. The tables estimates reported OLS coefficients correspond to β_1 from equation (6). Each observation is a district-year, and the sample is pooled over the 1951, 1961, 1971, 1981 and 1991 census waves. Regression includes decade and post-reorganization state fixed effects. Robust standard errors in parentheses are clustered at the district level. *** implies significance at the 0.01 level, ** 0.05, * 0.10. Sources: Census of India 1951-1991, Indian Meteorological Department, Ministry of Rural Development (Government of India).