Industrialization and Bilingualism in India

David Clingingsmith*
Department of Economics
Case Western Reserve University

October 2012

Abstract

Many of the world's poorest people live in linguistically diverse countries where bilingualism is a distinct and important form of human capital. When communication among workers increases productivity, there are economic incentives to learn a second language. I study how the growth of industrial employment increased bilingualism in India between 1931 and 1961. Indian factories were linguistically mixed. I exploit industrial clustering and sectoral demand growth for identification. The effect on bilingualism was strongest in import-competing districts and among local linguistic minorities. Industrialization also increased literacy. Bilingualism was mainly the result of learning, rather than than migration or assimilation. It was not a byproduct of becoming literate. My results shed new light on human capital investment in developing economies and on the long-run evolution of languages and cultures.

JEL codes: J24, J15, O12, I25, N35

Keywords: Language, industrialization, education, culture, identity, ethnicity, India

*Email: david.clingingsmith@case.edu. I am grateful for discussions with Leah Platt Boustan, Carola Frydman, Silke Forbes, Claudia Goldin, Susan Helper, Eric Hilt, Lakshmi Iyer, Asim Khwaja, Michael Kremer, Bob Margo, Rohini Pande, Jim Rebitzer, Heather Royer, Justin Sydnor, Mark Votruba, and Jeffrey Williamson. Thanks for useful comments to participants at the Harvard Economic History Workshop, the Harvard-Hitotsubashi-Warwick conference on Economic Change around the Indian Ocean, the 2008 Canadian Network for Economic History conference, the 36th Annual Conference on South Asia, the 45th Cliometrics Conference, and seminar audiences at UC Davis, University of Toronto, Vanderbilt, Reed College, Case Western Reserve University, and the University of British Columbia. A portion of this project was conducted while I was a fellow of the Project on Justice, Welfare, and Economics at the Weatherhead Center for International Affairs at Harvard University. The Harvard Department of Economics's Warburg Funds supported data entry.

Economists have found the accumulation of human capital to be an important contributor to economic growth (e.g. Mankiw *et al.* 1992). We most often measure human capital with the years of completed schooling or related metrics. From a theoretical perspective, we consider any skill that individuals invest in with the hope of earning a return to be human capital. One such skill is the capacity to communicate with others through speech. This skill may be taken from granted in high income countries where large majorities speak the same mother tongue.[1] Many people in the developing world, however, live in linguistically diverse environments where language differences create real barriers to communication. Consider that India, Indonesia, Nigeria, and the Philippines are each home to hundreds of sizable language communities.[2]

People who cannot speak with those around them may find their prospects for employment or trade limited relative to those who can communicate more widely. Investing in a second language—i.e. becoming bilingual—may confer economic benefits by expanding an individual's capacity to communicate. Despite the large populations in linguistically diverse countries, we are only beginning to learn about the economics of investment in language skills.

This paper asks to what extent the expansion of modern industry in India led individuals to invest in bilingualism. I study the period between 1931 and 1961, which marks the beginning of modern economic growth in the country. Industrial jobs more than doubled during this period, spurred in part be strong increases in tariffs. Most of the new jobs were in larger factories that exploited scale economies through task specialization and mechanization. India is linguistically diverse even within its local labor markets. Large factories mixed workers of different mother tongues in their shops and departments. When workers need to communicate to do their jobs, bilingualism would have been an advantage. Did the expansion of industrial employment in fact lead to increased investment in bilingualism?

A primary advantage of bilingualism in India in this era is the quality of data available. India is the only developing country to have regularly collected data on bilingualism as part of its population census. District-level tabulations for bilingualism were published only in 1931 and

[1] I use "mother tongue" to refer to the first language learned as a child.

[2] I define sizable as having more than $10,000$ speakers.

1961, which forced the choice of period. District-level data allows me to look within states, which set education policies that affected human capital investment and industrial policies that affected industrial growth. I created a new dataset that contains information on district-level economic outcomes and on counts of speakers and bilinguals for the languages spoken within each district. The census tabulated bilingualism for most districts in 1931 and all districts in 1961, which allows me to construct a district-language panel that covers most of the country.[3]

India is linguistically diverse even within the districts, with 25% of the population speaking local minority languages. Data on multiple languages per district allows me to study the growth in bilingualism for mother tongue speakers of majority and minority languages separately. Bilingualism leads to a larger relative expansion in the set with whom one can communicate for minority language speakers than the majority, and I hypothesize that industrial expansion with have a larger impact on this group.

The empirical analysis is centered on estimating how the bilingual share of the population is related to the share of the workforce engaged in industrial jobs. Figure 2 provides an initial view of the data in the form of two maps. The maps show a positive correlation between changes in the industrial employment share and changes in the bilingual share. The principal exception to the pattern is the southernmost part of the Indian peninsula. Note that while the border areas in the north and northeast part of the country show strong increases in both variables, the overall pattern remains even if we exclude them.

My empirical specification differences the district-language observations over time to eliminate a district-language fixed effect. I include state fixed effects after differencing to control for changing state policies.[4] There are several reasons why OLS estimation is unlikely to identify the true effect of industrial growth. First, there is substantial measurement error in the industrial share variable from a change in how the census counted workers. Ths creates a downward bias. Second, literacy is also a human capital investment influenced by the availability of industrial jobs. While

[3] In India the district is the administrative level below the state. The average district in my data is about 70 miles square and had a 1931 population of 1.5 million people.

[4] I discuss how the fixed effects are constructed in light of state boundary changes in Section 3 below.

I have a measure for literacy, it cannot be a used as control because it is endogenous. Literacy is positively correlated with the industrial share and bilingualism, which means that excluding it produces a positive bias in the OLS estimate. Third, other unobservables, such as wage growth, are also correlated with both the outcome and regressor of interest, and so bias.

I create an instrumental variable to provide consistent estimates of the industrial share effect. I collect data on employment shares for 14 industrial sectors, such as textiles and chemicals, for each district in 1931. The instrument is computed by making the counterfactual assumption that in all districts, the 14 sectors grow at their national average rate. This assumption holds the 1931 sectoral structure for each district constant and applies the average rate of employment growth in those sectors to the district. The predicted change in the industrial share that results from applying the counterfactual is the instrument. This approach was pioneered by Bartik (1991) and Blanchard & Katz (1992) and has had several recent applications (Autor & Duggan 2003; Luttmer 2005; Card 2009; Lewis 2011). To compute the instrument for a particular district d, I regress the actual change in the industrial share on the 1931 sectoral shares for the districts other than d. The instrument is an out-of-sample prediction for district d.

Instrumental variables estimation finds that industrial growth has a strong positive effect on bilingualism. A 1-point increase in the industrial share raises the bilingual share by 1.61 points. This effect is much larger than the OLS estimate of 0.55 points. The effect is 2.09 points for speakers of a district's minority languages, which is consistent with the greater potential bilingualism has to increase the set of individuals with whom they can speak.

The large difference between the OLS and IV estimates reflects 1) measurement error as discussed and 2) the source of the identifying variation. IV produces an estimate of the local average treatment effect, or LATE, rather than the average effect (Angrist *et al.* 1996; Angrist & Pischke 2009). However, the large difference also raises the concern of a positive correlation between the instrument and time-varying unobservable determinants of bilingualism.

I provide a check on the exogeneity of the instrument by showing it is not correlated individually or jointly with 11 district characteristics for 1931. Note that the state fixed effects eliminate

concerns about confounding effects of policy changes. Moreover, IV estimation using this instrument is not particularly sensitive to small violations of the exclusion restriction. I show that if the residual correlation between the instrument and unobservables were 0.2, for example, the IV estimate would be too large by 0.27 points.

In my setting, the LATE is the effect of those changes in the industrial share that resulted from national-level growth of the industrial sectors being channeled through the existing pattern of industrial location. National-level growth will tend to matter more for goods that are traded at the national level. Growth in non-tradable industries will mostly reflect changes at the local level. As it happens, industrial growth during the panel was strongly influenced by increased tariffs, which favored the home production of previously imported goods. Further, the 1931 industrial structure will be most strongly influence sectors where factors such as agglomeration economies or proximity to raw materials are important. If production is easily shifted, the preexisting structure matters less.

Is this interpretation of the LATE reflected in districts more exposed to trade? I investigate the role of foreign trade by creating a measure of each district's share in the value net of manufacturing imports. I first assign the net value of imports for each traded industrial good 1928–1931 to one of the 14 industrial sectors. I then allocate the total value for each industrial sector in proportion to the district's share of overall sector employment. I split the districts at the median of the net import value assigned to them. In above-median districts, which I call import competitors, a 1-point increase in the industrial share produces a 1.89-point increase in bilingualism, compared to a 0.50-point increase in the other districts. Note that Indian imports were intensive in technically sophisticated goods such as steel, machine tools, petroleum products, and vehicles. Industrial growth had a larger impact on bilingualism in districts that had more exposure to imports.

Potential bilinguals must choose the language they will learn. India has two *lingua francas*, English and Hindi, that are widely used for communication by people with different mother tongues. The dominant language in a district has a 75% share, which makes it attractive to minorities. I found that the choice of second language differed for dominant and secondary language speakers.

Mother tongue speakers of the dominant language in their district were pushed strongly toward learning Hindi and English and away from the minority languages of their district. Mother tongue speakers of minority languages were pushed most strongly toward English and other languages from the district, with a smaller effect on Hindi.

I consider four channels through which expanded industrial employment could increase the bilingual share of a particular language in a particular district. First, people may decide to invest in learning a second language and become bilingual, which is the channel central to this paper. Second, they may wish to become literate and acquire a second language as part of doing so. Third, speakers of the language may decide to migrate from outside the district to where industry is expanding. Fourth and finally, some parents who are bilingual may decide to teach their children only their second language, causing them to assimilate to the other language.

Industrial expansion did lead to higher literacy. The effect was 1.14 points for each point change in the industrial share. It was stronger in the import-competing districts, but the difference was small—1.17 versus 0.92 points. This suggests that industries differed in their relative demand for bilingualism and literacy.

If bilingualism is a step taken to become literate, then some of the effect of industrial growth on bilingualism will be merely a reflection of investment in literacy. Ideally, we would like to know the effect of industrial growth on bilingualism *conditional* on literacy. Since literacy is endogenous, this would require an additional instrument and the interpretation difficulties associated with two sources of identifying variation. Instead, I explore the sensitivity of the IV estimates to assumptions about the conditional coefficient. Even under the assumption that literacy leads to one-for-one changes in bilingualism, the conditional effect of industrial growth is 0.48 and significantly different from zero. I also show that a 1-point increase in the industrial share raises the number of bilinguals per literate person 13 points. The average growth in the industrial share is 2.9 points. Industrial growth thus increased the number of bilinguals per literate person by approximately 0.38. Overall, literacy rose faster than bilingualism over the panel, leading the number of bilinguals per literate to fall from from 1.47 to 0.55.

I assess migration by considering effects on surrounding districts, which are a likely source of migrants. I take the set of languages spoken in each district and calculate how many speak that language in the geographically adjacent districts. I also calculate how many are bilingual and the bilingual share. I find that industrial growth has a relatively small, statistically insignificant negative effect on the bilingual share in adjacent districts. Level regressions show no effect of industrial growth on the overall size of languages in adjacent districts.

Finally, I find that industrial growth doesn't change the share of the population speaking the majority language or linguistic heterogeneity, which suggests little assimilation is going on. I do find that secondary languages that had higher initial bilingualism had lower population shares 30 years later. This patterns holds for districts between 1931 and 1961 and in state-level data between 1961 and 1991.

By showing that industrial employment growth induces investment in bilingualism, this paper demonstrates that spoken language skill is valued by employers in a linguistically diverse developing economy. The investment response was large even though the jobs concerned were not high skill. The response occurred in an environment where formal education was weak and can't be explained as an epiphenomenon of investment in schooling. My findings suggest that in linguistically diverse developing countries, which are the norm rather than the exception, we further investigate language skill as a type of human capital of potentially independent importance. As I discuss in more detail in the conclusion, the efficiency of second language acquisition begins to fall before children reach school age, so there may be gains to encouraging second language acquisition independent of school.

This paper makes a contribution to several active literatures. It is closely related to recent studies of on the returns to English in India (Munshi & Rosenzweig 2006; Kapur & Chakraborty 2009; Oster & Millett 2011; Shastry 2012; Azam *et al.* 2013). This literature takes the growth of IT and business process outsourcing as its point of departure. I show that bilingualism, including in English, has long been valuable in the larger, lower-skilled industrial sector. This finding is relevant today because the average skill level of Indian workers remains low and the country remains

linguistically fragmented.

An older and larger literature has studied the returns to bilingualism in high-income countries (e.g. Chiswick & Miller 1995; Dustmann & van Soest 2001; Berman *et al.* 2003; Fry & Lowell 2003; Bleakley & Chin 2004; Lang & Siniver 2009). The gist of this literature is that returns tend to be large for immigrants who become bilingual in the primary language of their adopted country and near zero for natives who learn a second language. My study relates to both of these strands. First, linguistic minorities in India tend to be small shares of the local population, meaning the challenge they face is similar to that of immigrants. Second, in contrast to the findings for high-income countries, there is a return to bilingualism for the linguistic majority in India. This difference probably results from greater linguistic diversity in India, which creates the need for a *lingua franca* that is absent in the high-income countries studied in this literature.

Linguistic diversity has been associated with a variety of poor economic outcomes, from low economic growth to low levels of public goods (e.g. Alesina *et al.* 1999, 2003; Alesina & La Ferrara 2005). One root cause for this correlation, among several that have been proposed, may be communication barriers to exchange. Investment in bilingualism induced by industrialization may be an endogenous response to a diverse environment. While I do not find a direct effect on assimilation or linguistic diversity, my results are still consistent with endogenous changes in linguistic diversity over the long run.

The body of this paper contains five sections. Section 1 provides information on the Indian economy that supports the empirical analysis. Section 2 describes the construction of the dataset and provides summary statistics. Section 3 develops a regression model, discusses the challenges of identifying the parameters, and provides an instrumental variables solution. Section 4 presents the empirical analysis. Section 5 discusses the implications of the results.

1 Economic Institutions and Language in Context

I begin the study with a discussion of the historical and institutional contexts in which my empirical analysis is situated. Since my analysis will be conducted at an aggregate level, a primary goal of

this section is to present evidence on lower-level processes. For example, I will show how language fits into the process of industrial recruitment and the conduct of industrial labor. I examine the literature on returns to bilingualism, both in India and more broadly, and how the production of bilingualism and literacy are related. I begin at the aggregate level by characterizing the nature of industrial growth and its relationship with trade.

The Expansion of Indian Industry

India's main industrial sectors in 1931 were textiles, wood products, food processing, and ceramics. Industry made up 8.9% of India's total employment and contributed 13.2% of its GDP (Sivasubramonian 2000). By 1961, industry was 22.1% of GDP and employed 11.8% of the workforce. The overall number of industrial jobs nearly doubled, and about 70% of new jobs were in large-scale industrial enterprises (Sivasubramonian 2000; India 1962).[5] Historical studies have suggested that increased task specialization was a major reason for the increase in industrial scale during this era (Roy 1999, 2000)

India's trade policy was an important factor driving industrial growth from the late 1920s. In 1919, the government of India was given fiscal autonomy from Britain, which meant it could set tariff policy independently. At the same time, rights to land revenue, the main source of income for the central government, were devolved to the provinces. Thereafter India's central government relied increasingly on import tariffs to raise revenue (Tomlinson 1979). Average import tariffs were about 5% from 1900 to 1920, then rose steeply to more than 30% in the early 1930s (Online Appendix Figure A-2). Average tariffs were about 25% between 1931 and 1961. The ensuing substitution of domestically produced goods for technologically advanced imports is consistent with industrial growth being mostly in the large-scale sector.

[5] Industrial statistics divide enterprises into large-scale and small-scale using the threshold of 20 employees without mechanical power or 10 employees with mechanical power.

The Industrial Labor Market, the Industrial Firm, and Bilingualism

How might bilingualism provide an advantage for seekers of these new industrial jobs? First, bilingualism may aid job seekers during the search and recruitment process by enlarging the set of employers they can contact. Second, bilingualism may be valuable in jobs that require communication with other employees.

Since the establishment of the first large factories in the mid-19th century, caste networks have played a central role in connecting industrial firms and employees. Castes are endogamous and hereditary social groups to which most Indians belong. Surveys conducted in the 1950s and 1960s reported 30% to 50% of industrial workers made use of personal contacts, including through caste networks, in getting their jobs (Lambert 1963; Sheth 1968; Holmström 1976). Members of a caste speak the same language, so bilingualism does not play a role in making connections to employers through the caste network.

In their classic studies of the industrial sector in Mumbai, the center of Indian manufacturing, Morris (1965) and Chandravarkar (1994) discuss how labor shortages helped entrench a recruitment system based on caste. The key figure in this system was the *jobber*. The jobber used his contacts among members of his caste in the hinterland to muster labor to the factory in the city. Once there, he supervised the recruits in their jobs. The jobber and his workers shared the common language of their caste (in Mumbai this was typically Marathi), while the jobber also spoke the language of the factory owners (typically Gujarati). A similar system of labor recruitment was found in Calcutta jute mills and the tea plantations of Assam (Roy 2010).

The role of the jobber as a recruiter waned in early 20th century Mumbai as labor became abundant in the city and was taken over by personnel departments (Morris 1965; Chandravarkar 1994; Breman 1999). Personnel departments served as the conduit for referrals, and jobbers no longer selected the workers they would supervise. The jobber's enduring legacy was the establishment of caste connections as a gateway to industrial employment. Munshi & Rosenzweig (2006) found that caste networks and the links they provided to particular occupations continued to influence the occupation and education choices of Maharashtrian children in Mumbai in 2001.

Once the jobber's role in recruitment had ended, the linguistic composition of work groups became less constrained. A number of sociological studies of Indian factories have discussed the use of language in the workplace. Some have collected detailed data on the language, occupation, and work group of factory employees. This work shows that multilingual work groups were the norm rather than the exception, and workers used second-language skills on the job.

The most vivid picture of language use on the factory floor comes from a 1953 study by A.K. Rice of productivity and social organization in an Ahemdabad textile factory.

> Languages are regional and, although Ahmedabad is in the Gujarat, and the common language of all those who work in the industry is Gujarati, it is not uncommon to find three or even four different languages being spoken in the same department of one mill. One one occasion, in a discussion with a group of eight workers, which was being interpreted in three languages, Gujarati, Hindi, and English, it was discovered after half an hour that one worker had not up to that time understood a word that had been said—he came from South India and spoke only Tamil. (Rice 1958)

Bilingualism played a central role in the interaction described. It is easy to see the potential disadvantage that a worker might face by not being able to engage in such discussions, even if they did involve a researcher in this case. Rice goes on to describe how language and caste differences had complicated efforts to improve productivity.

Sheth (1968) studied an electrical factory in a small Gujarati town in 1958. He documented the distribution of mother tongues within its departments and workshops (see Table A-1 in the Online Appendix). The factory had 810 workers. Analyzing his tabulations using a categorical ANOVA (Light & Margolin 1971), I found that 91% of the variation in language spoken is *within* the functional units, rather than across them. In his observations of the factory floor, he noted that employees were in "continuous interaction" with each other. In other words, they needed to talk to do their jobs. Interestingly, caste networks were strong in this factory. A large share of Gujarati-speaking employees, who made up 64% of the total, came from the same caste groups as the owners and top executives. A related study of five factories in Poona during the late 1950s found

that 20%-30% of employees were not native Marathi speakers (Lambert 1963). The percentages were similar across occupational groups and factories. Similar patterns are described by Gokhale (1957) and Vidyarthi (1970).

Moving from the factory to the city level, Holmström (1984) presents data on language and occupation class from a 1979 survey of a random sample of all Bombay industrial workers. The data records five different languages and 36 industrial occupations. I found that 94% of the variation in the primary language spoken is within, rather than across, occupations. There is no segregation of occupations by language groups in the industrial sector as a whole. This finding is particularly interesting as it is these very occupations that caste networks enable the Marathi speaking boys in Munshi & Rosenzweig (2006) to access.

The Returns to Bilingualism

Did bilingualism earn a return in the Indian labor market of the era? Addressing this question completely would require data on wages that do not exist. However, there is a substantial literature on the returns to bilingualism, both within India and beyond. They point to large returns to bilingualism when the second language is a *lingua franca*, such as English or Hindi in India, or the dominant language of the country.

A number of recent studies have found substantial returns to bilingualism in English in India. Using individual-level data from 2005 and conditioning on schooling, Azam *et al.* (2013) find a 34% return to English fluency and a 13% return to knowing a little English compared to knowing none at all. Kapur & Chakraborty (2009) report on a policy intervention in West Bengal in 1983 that removed English instruction from public primary schools. Using variation across cohorts and districts in English exposure, they find a 68% wage premium for English. Shastry (2012) shows that export-oriented IT firms, which rely on English speakers to serve clients in the United States, chose to locate in areas where the cost of learning English relative to Hindi were small. The relative costs were based on predetermined language structure. These areas then showed a response in school enrollment growth.

A more established literature has estimated returns to bilingualism in industrial countries. One strand concerns immigrants learning the dominant language of their adopted home. The effects are generally large. In a study that included Australia, Canada, the United States, and Israel, Chiswick & Miller (1995) found returns to English fluency of 10% to 17% conditional on schooling. A follow-up study on West Germany found an effect of German fluency on wages of 7.3 points per standard deviation of fluency (Dustmann & van Soest 2001). Berman *et al.* (2003) estimated that one-half to three-quarters of the wage convergence for skilled immigrants to Israel came from improved Hebrew. The cost of learning a language increases sharply in adolescence due to biological changes. Bleakley & Chin (2004) use this variation to estimate returns to English for child immigrants to the United States. They find 67% higher wages for those who speak English well rather than poorly, though the difference largely comes through increased schooling.

Another strand addresses returns to bilingualism among the native born in the United States. Fry & Lowell (2003) find that among native speakers of English, there is no additional return to knowing a second language conditional on schooling. However, those monolingual in another language earn 11% less, in line with the estimates of Chiswick & Miller (1995). Chiswick & Miller (1998) found similar results. The situation faced by the native born in the United States is similar to that of dominant-language speakers in India.

National Markets for Goods and Local Markets for Labor

India built a very extensive railway network between 1853 and 1930. The railway penetrated nearly every district and comprised 70,000 km of track. It expanded trade, reduced inter-regional price disparities in major commodities, and created a national market for industrial goods (Donaldson 2010; Burgess & Donaldson 2010).

Interestingly, migration rates remained quite low well into the late 20th century (Cashin & Sahay 1996). At the end of my panel in 1961, only 3.2% of the Indian population were inter-state migrants. Even the substantial economic growth induced by India's 1991 trade liberalization failed to induce substantial cross-district migration (Topalova 2010). Endogamous marriage patterns and

geographic concentration among India's castes are important in explaining low migration (Munshi & Rosenzweig 2009).

Borjas (2003) and subsequent literature highlighted the problem of identifying the impact of labor supply shocks from immigration in the United States by using geographic variation across local labor markets. Labor markets in the United States are well integrated, and local shocks diffuse quickly. If Indian labor markets were as well integrated, it would be impossible to disentangle the migration and learning channels through which industrial employment growth would increase bilingualism in a particular region.

Bilingualism, Literacy, and Education

Bilingualism and literacy are related forms of human capital both in a functional sense and in the way they are acquired. Bilingualism enables face-to-face communication among people with different mother tongues, while literacy enables communication across time and space between people who share a written language. In 1931, 8% of the Indian population were bilingual and 9% were literate. Bilingualism had increased 50% and literacy 300% by 1961.

The census considered a person to be literate or bilingual if they had a basic functional ability India (1933a). A person was considered literate if they could read and respond to a simple letter. They were considered bilingual they regularly used more than one language. Functional literacy and bilingualism can be acquired in school or through independent effort. The population share of primary school completers was only 8.1% as late as 1960 (Barro & Lee 2010). While formal schooling in the vernacular languages of India and in English had been promoted since the 1850s, per-capita spending on primary education and enrollment rates in British India were consistently among the lowest in the world (Chaudhary 2009). Literacy is much higher than primary completion in 1961 (27% versus 8%), which is consistent with basic literacy requiring only a small amount of classroom time and with learning taking place outside the classroom. It isn't possible to directly assess how much of the growth of either literacy or bilingualism resulted from formal schooling and how much resulted from other means because the Census of India did not ask about schooling

independent of literacy or bilingualism until 1941 (Srivastava 1972).

Was bilingualism a byproduct of literacy? All the major languages of India are written, and vernacular newspapers and books were available in most. Generally speaking, then, becoming literate does not *require* becoming bilingual. Becoming literate in English, on the other hand, does require bilingualism, though Chaudhary (2010) reports that only 14% of literates in 1931 knew English. Schools, however, teach primarily in the dominant language of an area, and so we expect there to be some complementarity in the production of literacy and bilingualism, at least for speakers of secondary languages. On the other hand, if the skills can be acquired separately, learners face a trade off about where to invest their effort. As it turns out, increases in bilingualism and literacy are only weakly correlated with $\rho = 0.16$. This, along with low rates of primary completion, points to a decoupling of the production of literacy and bilingualism in this period and a limited role for formal schooling.

2 Data and Summary Statistics

Construction of the Dataset

I constructed a panel dataset of Indian districts for the years 1931 and 1961 from published tables of the Census of India (India 1933a, 1962). Each census was a complete enumeration of the population.

The dataset contains information at the district level on population, employment, occupational structure, literacy, and urbanization. Within each district, there is an observation for each of the languages found in the district that contains the number of speakers and number of speakers who are know an additional language. I refer to such an observation as a district-language. This classification precludes double counting. As I mentioned above, the census used basic functional standards in assessing literacy and bilingualism. A person was considered literate if they could read and respond to a simple letter and bilingual if they regularly used more than one language.

The Censuses recorded the language names reported by respondents. Languages often have

locally specific names, and must be aggregated to consistent categories. The category scheme used in 1961 was finer grained, and the category names were changed in some cases. I matched language categories by hand on a district-by-district basis using *Ethnologue,* a comprehensive global database of languages that includes alternative names and dialects (Gordon 2005).

There were substantial changes in district boundaries between 1931 and 1961. Following India's independence from Britain in 1947, hundreds of sovereign princely states were integrated into the existing colonial administrative framework inherited by India. A reorganization of state boundaries in the 1950s led to further changes in district boundaries. I used maps and a concordance to generate a mapping of all British districts and the princely states that fall within India's 1961 boundaries into consistent geographical units (Singh & Banthia 2004). Implications of the boundary changes for the analysis are discussed in section 3.

I restrict the dataset to districts in which adequate data on bilingualism was reported in 1931. While the census form was standard in that year, the published volumes were prepared at the provincial level and do not always contain the same tables. There is no evidence that the complete data was not collected. Each district has an average of five language observations.

The dataset contains 137 districts and covers present-day India excluding Uttar Pradesh, Punjab, Himachal Pradesh, Rajasthan, and portions of Bihar, which are the states where the bilingualism data were unavailable. Taken together, these excluded states are substantially less urban, more agricultural, less literate, and less bilingual than the others. They are also less linguistically diverse. We therefore need to take care in extrapolating the results to the rest of India.

Characteristics of Districts and Their Languages

Summary statistics for the dataset are presented in Table 1. The average district had a population of about 1.5 million people in 1931 and 2.4 million in 1961. India shows the hallmarks of a developing economy. The share of the population living in cities grew from 13.7% to 19.7% and industrial employment increased from 8.9% to 11.8% of the workforce.

Investment in language skills also increased. Bilingualism rose from 8.0% to 12.1% among

the population as a whole. Literacy expanded substantially even more, from 9.1% to 27.2% of the population.

The first panel of Figure 3 shows the population share of the top five languages in each district. The typical district has a dominant language making up 75% of the population and several large secondary languages with population shares in 5% to 15% range. The growth in the share of the dominant languages mainly came at the expense of the language ranked second. The second panel of Figure 3 plots the bilingual share of speakers and speakers share of population for each district-language observation in the data for each year (N=1,368). There is a great deal of variation in the bilingual share, particularly for the smaller languages at the left side of the figure. A fit to the data from a kernel-weighted polynomial regression shows that on average the bilingual share rises roughly equally for all language sizes between 1931 and 1961.

3 Empirical Specification, OLS Estimates, and Identification

In this section I present my main regression model, show some initial OLS estimates, and factors that may bias them. I then develop an instrumental variables approach to address the biases.

The model measures the effect of industrial employment on bilingualism. Let I_{dt} be the industrial share of total employment in district d at time t and $B_{\ell dt}$ be the share of mother-tongue speakers of language ℓ in district d at time t that are bilingual. The panel structure allows me to eliminate language-district fixed effect by differencing the data over time. My estimating equation is

$$\Delta B_{\ell d} = \alpha + \beta \Delta I_d + s_d + \Delta \varepsilon_{\ell d}. \tag{1}$$

I do not aggregate the language data to the district level to allow for interactions.

Most estimations will include a fixed effect s_d for the state to which a district belongs to eliminate confounds from state industrial and education policy. Variation in industrial employment occurs at the district level, so I cluster the standard errors by district. I weight the district-language observation by the number of speakers 1931 so that the coefficient β measures the effect of indus-

trial share growth on the average individual.

OLS Estimates

I begin the analysis with OLS estimates of equation 1. For the sake of simplicity, I present an initial estimate with bilingualism aggregated to the district level (Table 2, column 1). A change in the industrial share of employment of 1 percentage point is correlated with a 0.66-point increase in the share of the district population who are bilingual. The correlation is 0.61 in the full district-language data, which I will use for the remainder of the paper (column 2). In interpreting the size of estimates, note that industrial employment share grew by an average 2.9 points between 1931 and 1961. If we assume a homogeneous effect, growth in the industrial share increased bilingualism by about 1.8 points. The overall bilingual share grew by 4.0 points, so industrial growth appears to be an important driver.

State governments undertook effort to improve education and to grow the manufacturing sector, particularly after Indian independence in 1947. Column 3 introduces a fixed effect for the state a district belonged to in 1931. This fixed effect reduces the industrial share coefficient by about 18% to 0.50.

There were also many changes to the 1931 state boundaries over the panel that are related to language. Following popular agitations in the 1950s, the Government of India began to reorganize the states by grouping contiguous districts that shared a common majority language into new states. The goal was to create linguistically homogeneous states. The new organization could have affected both bilingualism and industrial growth. The reorganization placed regions of the 1931 states into different policy environments. For example, the Telugu-speaking region of Madras state were split off to form Andhra Pradesh in 1953. In 1956, the Telugu-speaking region of Hyderabad state was added to Andhra Pradesh. We can say that these two regions each had different policy environments during the last years of the panel than other parts of their 1931 states.

To address this issue, I create a second fixed effect that interacts the district majority language in 1931 with the 1931 state to allow for policy differences across regions due to the states reorga-

nization. Column 4 shows an estimated effect of industrial growth of about 0.55 points, not much different from the 1931 state fixed effects. I will use these fixed effects in most specifications to follow. For the sake of comparison, column 5 shows that the industrial share coefficient is 0.50 when I include fixed effects for the 1961 states. The estimates are also robust to a 1931 $\times$ 1961 state fixed effects (not shown).

Sources of Bias

This section discusses biases in OLS estimates due to measurement error in the industrial share variable and correlation of the change in industrial share with both 1) omitted variables that affect bilingualism and 2) omitted endogenous variables, such as literacy, that are both affected by industrial growth and may themselves increase bilingualism.

The industrial workforce variable suffers from measurement error due to a change in the way it was computed by the census. The 1931 Census divided the workers in each sector into three occupational categories: principal occupation, working dependents, and subsidiary occupation. Working dependents provided assistance to the worker in their job, such as the preparation of materials, though were not otherwise employed. Subsidiary workers had their principal occupation in another sector, which in the case of industry is overwhelmingly agriculture. Overall, 77% of industrial workers fell into the principal occupation category. In 1961, the census abandoned the three categories of workers, and counted only workers and non-workers. I create my industrial employment variables using only the principal occupation data for 1931, which is most comparable to the 1961 categorization. Nevertheless, the change in the industrial share is measured with error, which would attenuate the OLS estimates.

In additional to industrial growth, urbanization, literacy, and income growth are also likely to encourage bilingualism. Urbanization brings populations into contact, creating both the demand for and opportunity to learn new languages. Literacy may be a complement or substitute for bilingualism, in both the acquisition of skill and its use. Income growth provides additional resources to invest in skills. These three processes are likely to be correlated with industrial employment

growth and induce omitted variable bias.

Consider the example of unobserved income growth. Higher wages may facilitate the acquisition of a second language directly through an income effect. Wage growth is likely correlated with increases in the industrial share, where productivity growth tends to be highest. These correlations would lead to an upward bias to the OLS estimate of the industrial share effect as wage growth is unobserved.

Literacy and bilingualism can both be produced through attending school. Reading can also be an important skill in the industrial workplace, particularly in the modern factory sector where rules and procedures are often written down. We would expect an upward bias of the OLS from this factor. Similarly, urbanization brings people who speak different languages into contact, and is also driven by industrial growth, since factories tend to locate in cities. This also produces an upward bias.

The census provides data on literacy and urbanization. However, I cannot use them as controls in estimation because they are themselves outcomes of industrial growth. Literacy is of particular interest, as a related type of human capital, and I will treat it as an outcome in the analysis to follow.

Identification

I address the threats to identification from measurement error and time-varying omitted variables by constructing an instrumental variable for ΔI_d. I take advantage of two aspects of Indian industry to provide a source of exogenous variation. The first is persistence in the location of new industrial jobs. Districts with existing capacity in certain sectors have an advantage in capturing increased demand due to proximity to raw materials, economies associated with existing firm clusters, and other barriers to entry. Take steel production as an example. Proximity to coal and ore supplies confer cost advantages to steel mills. The second factor is the integration of Indian markets through an extensive railway network. Changes in demand for tradable industrial goods can be met by suppliers across the country.

My instrument uses variation in employment growth across 14 industrial sectors that comes from the interaction between national demand growth and the existing sectoral structure of industry at the district level. More concretely, the instrument is a prediction of the change in a district's industrial share under the counterfactual assumption that each of its 14 sectors grew at the overall sectoral average. This type of instrument was pioneered by Bartik (1991) and Blanchard & Katz (1992), and has had several recent applications (Autor & Duggan 2003; Luttmer 2005; Card 2009; Lewis 2011). For example, Card (2009) and Lewis (2011) use the persistence in flows of migrants from particular countries to U.S. cities to form an instrument for changes in the skill mix of workers. My approach differs from this earlier work in using the predicted value from a regression rather than direct computation to produce the instrument. This has the advantage of allowing me to exclude a district's influence on overall sectoral growth when computing its predicted value.

I use a decomposition of industrial growth to derive a regression that produces the instrument as a predicted value. Let the level of industrial employment in district d at time t be E_{dt}, the level of employment in subindustry j at t be Y_{jdt}, and total employment be W_{dt}. We can then express ΔI_d in terms of initial levels and growth rates of the subindustries and total employment:

$$\Delta I_d = \frac{E_{d61}}{W_{d61}} - \frac{E_{d31}}{W_{d31}} = \frac{\sum_j Y_{jd31} g_{jd}}{W_{d31} g_{Wd}} - \frac{\sum_j Y_{jd31}}{W_{d31}}. \tag{2}$$

Let $\mu_{jd} = \frac{g_{jd}}{g_{Wd}} - 1$ measure how much faster or slower subindustry j in district d is growing relative to the overall employment in d. Let y_{jd31} be the share of overall employment in district d in subindustry j in 1931. Then we can rewrite equation 2 as

$$\Delta I_d = \sum_j \mu_{jd} y_{jd31}. \tag{3}$$

The relative growth rate μ_{jd} can be decomposed as $\mu_{jd} = \mu_j + \tilde{\mu}_{jd}$. The component μ_j is the average growth rate of employment in j relative to overall employment. The district-subindustry deviation

from this average is $\tilde{\mu}_{jd}$. We then have

$$\Delta I_d = \sum_j \mu_j y_{jd31} + \sum_j \tilde{\mu}_{jd} y_{jd31}. \tag{4}$$

The first term in equation 4 is the component of the change in the industrial employment share that reflects whether a district's initial complement of industries were relatively fast or slow growers on average. We can write the deviation as a residual and estimate the regression

$$\Delta I_d = \sum_j \delta_j y_{jd31} + \zeta_{jd}. \tag{5}$$

The predicted value from this regression ΔZ_d is just $\sum_j \mu_j y_{jd31}$.

The predicted value ΔZ_d will be a valid instrument if the exclusion restriction holds—subindustry employment shares and predicted values must be uncorrelated with time-varying unobservables. A concern immediately arises for those districts that have a large share of national employment in a particular subindustry. There were six districts that had more than 10% of national employment in at least one subindustry. The estimated coefficient $\hat{\delta}_j$ for the concentrated subindustries will be strongly influenced by the change in industrial employment in those districts and therefore potentially correlated with unobservables. A benefit of using a regression rather than tabulated averages is that I can compute ΔZ_d separately for each district d by 1) estimating equation 5 on the districts $\sim d$ to compute $\widehat{\delta}_{j\sim d}$ and then 2) making an out-of-sample prediction for district d.

To create the instrument I collected district-level data on employment for 14 industrial sectors in 1931. Table 3 shows summary statistics for the sectors. The largest sector is textiles, employing about one-third of all industrial workers. Wood products is the next largest industry, followed by food processing.

First-stage regressions are shown in Table 4. The instrument is strongly correlated with the actual change in the industrial share. The F-statistics of 52.27 on the bivariate correlation and 20.23 on the specification with state fixed effects imply that IV bias will be small. The coefficient on the instrument is 0.73, which implies that the IV estimate will not be sensitive to small violations

of the exclusion restriction. I discuss this further below.

One potential concern with this IV approach is that the instrument's correlation with the industrial share could be picking up cross-sectional characteristics of the districts that influence the location of industry. This would violate the exclusion restriction. Columns 3 and 4 in Table 4 regress the instrument on district characteristics in 1931, controlling for the subindustry shares. All of the coefficients are small and insignificant. The 1931 characteristics are also jointly insignificant. I also provide some sensitivity tests for the exclusion restriction below.

4 Empirical Analysis

This section presents my estimates of the causal effect of industrial employment expansion on bilingualism. After discussing the basic results and sensitivity of the IV estimate, I explore the heterogeneity of the effect by how intensively a district was involved in trade, the particular second language learned, the size of the first language, and the linguistic diversity of the district. Finally, I investigate the mechanism in more detail by comparing the effects of industrial growth on literacy and bilingualism, testing whether migration is an important source of bilinguals, and investigating assimilation. I conclude the section with an analysis of the affects of bilingualism on the evolution of a language community over time.

I present the main instrumental variables results in Table 5. Estimates are presented with and without state fixed effects. The first two columns show that a 1-point increase in the industrial employment share produces a 1.5-point increase in the bilingual share. The estimate is large, both relative to the OLS and in an absolute sense. The estimated effect includes spillovers. For example, a new industrial job may lead to additional demand for transportation and commercial services.

Measurement error, as discussed in the previous section, is one reason why the IV estimate would be larger than the OLS. A second factor is that the IV procedure produces an estimate of the local average treatment effect, or LATE (Imbens & Angrist 1994; Angrist & Pischke 2009). The IV estimator recovers a weighted average estimate of the causal response of each district-language to industrial growth, where the weights are proportional to the first-stage impact on the

district. Even if the OLS estimate were unbiased, to the extent that the responses are heterogenous across district-languages, the IV estimator can produce a different average estimate because it uses different weights.[6]

Recall that the instrument is an estimate of how national-level demand growth in subindustries would affect a district based on its 1931 structure. This variation will have a stronger effect on some districts than others. For goods that do not trade across districts, for example, national demand growth may not matter as much, and for industries where existing locations confer no advantage, the 1931 sectoral structure may not matter much. The R^2 for the first stage in Table 4, column 1 is 0.31, which means that much of the variation in industrial growth is not affected by the instrument.

We thus need to employ caution in comparing the LATE estimate to average changes in bilingualism and industrial growth. For example, if we take overall average industrial employment growth of 2.9 points and multiply it by the LATE of 1.6, we get an increase in bilingualism of 4.4 points, larger than the overall average. This comparison is misleading because the average change in bilingualism and employment growth could be different from the local-average change corresponding to the estimate.

The maps in Figure 2 suggested the correlation between industrial share growth and bilingualism was very strong both in the mountainous and linguistically diverse regions of north and east and in districts that border Pakistan, present-day Bangladesh, Nepal, and Burma. These regions saw the most dislocation during the partition of India in 1947, became sites of military garrisons, and also became more involved in border trade. I check the sensitivity of the estimates by excluding these districts in columns 3 and 4, which show that the estimates are very close in the border and non-border regions.

[6] As a concrete example, I can make a dummy variable that cuts my instrument at the mean and assign the value 1 to districts with above-mean values of the instrument and 0 to the others. This dummy has a strong first stage. In an IV estimation using the dummy, the weighting of the district-language specific industrial-share effects will be different. In fact, this estimation produces an industrial share effect of 0.89, which is much closer to the OLS estimate.

Trade and Import Competition

India increased import tariffs substantially through the 1920s, and they remained high until the end of my panel. The tariffs stimulated growth in import-competing industries. If national growth in industry is substantially driven by import substitution, part of the difference between the OLS and IV estimates may reflect the relative importance of bilingualism in tradable versus non-tradable industrial goods.

I collected data on the value of India's imports and exports of manufactured goods (India 1933b). I matched goods to the industrial sectors that produced them (for exports) or that would produce them had they been made domestically (for imports). Table 3 shows the shares of total export and import value assigned to each sector. Textiles dominate both imports and exports. Leather and chemicals are the second and third largest sources of export value, while metals and foods occupy those slots for imports.

I used this data to construct an indicator of the degree to which each district's manufacturing employment was in sectors in which India was a net importer. For each district and sector, I calculate the share it contributes to national employment in the sector. I then assign the net import value of the sector as a whole to the districts using these shares. I sum this net import value within districts, which gives me a measure of how sensitive a district would be to changes in India's trade policy. I create a dummy variable equal to one for those districts that have an above-median value of net imports. I call these above-median districts import competitors. The dummy variable tells us that a district's manufacturing sectors overlapped relatively strongly with the goods of which India was a net importer. The industrial share increased by 4.3 points for import competitors and only 1.8 points for the others. The import-competing districts were more intensively involved in the production of textiles, processed foods, chemicals, vehicles, and power, and less intensively involved in wood, ceramics, leather, and tailoring.

Both OLS and IV estimates show a much stronger effect of industrial growth in import-competing districts (Table 6). The OLS effect is entirely in the import-competing districts. The IV effect is an imprecisely estimated 0.50 points for non-import competitors and 1.89 for import

competitors. These estimates support the idea that bilingualism is particularly important in the production of tradable goods and that the instrument gives more weight to those districts.

Sensitivity of the IV Estimates

Another factor that might produce IV estimates much larger than OLS is a positive correlation between the instrument and the error term. This is a failure of the exclusion restriction. I will now explore how sensitive the IV estimates are to such positive correlations. If we write out the two IV stages explicitly as

$$\Delta B_{\ell d} = \alpha + \beta \Delta I_d + \gamma \Delta Z_d + \Delta \upsilon_{\ell d} \quad (6)$$

$$\Delta I_d = \theta + \pi \Delta Z_d + \Delta \varphi_{\ell d}, \quad (7)$$

then the exclusion restriction amounts to the assumption that $\gamma = 0$.

It is easy to show that the bias is $\widehat{\beta} - \beta = \gamma/\pi$. Table 4 tells us that $\hat{\pi} = 0.73$, which means that the bias is approximately 1.36γ. A benefit of having such a strong instrument is that the bias will be small even if the exclusion restriction holds only approximately.

We can see how the IV estimates vary by choosing a set of fixed values for γ, which I will call γ_0, and then estimating

$$\Delta B_{\ell d} - \gamma_0 \Delta Z_d = \alpha + \beta \Delta I_d + \Delta \upsilon_{\ell d}. \quad (8)$$

By doing a separate estimation for each value $\gamma = \gamma_0$ that is of interest, we can evaluate the sensitivity of the IV estimate of β (Conley *et al.* 2012).

Figure 4 plots the IV estimate and confidence intervals for $\gamma_0 \in [-0.3, 0.3]$. The IV estimates are not very sensitive to even moderate amounts of bias. A correlation of ± 0.1 yields estimates of 1.63 and 1.32. The endpoints of this interval represent a very substantial amount of bias, up to half as large as the reduced form OLS coefficient on industrial growth itself, and yet all are within the confidence interval of the actual IV estimate.

Secondary Languages and Heterogeneous Districts

Industrial share growth had a greater impact on bilingualism for speakers of secondary languages (Table 7, column 1). Because they comprise a small share of the population, secondary language speakers have the greatest theoretical potential to expand the pool of others with whom they can communicate. This point can be formalized in a simple model (Online Appendix A1.)

My estimates show a 1-point increase in industrial employment raises the likelihood of being bilingual for an average dominant language mother-tongue speaker by 1.3 points and for an average secondary language mother-tongue speaker by 2.1 points. The difference between these effects is positive with $p = 0.11$.

In more linguistically heterogeneous districts, impediments to communication will generally be higher, while the benefits of learning any particular second language will generally be smaller. The impact of growth in the industrial share in such districts in ambiguous. I divide districts into high and low linguistic heterogeneity groups by the median heterogeneity in 1931. The two groups of districts are similar in terms of initial levels and changes in industrialization, literacy, and urbanization, though we should keep in mind that linguistic heterogeneity may be correlated with unobservables. I estimate differential effects of industrial share growth for high heterogeneity districts in column 2 of Table 7. The point estimates suggest a greater impact in heterogeneous districts, though the estimates are imprecise.

Choice of Second Languages

What languages did people learn as a result of industrial employment growth in India? English and Hindi are the major *lingua francas* of India, meaning they are widely learned by people of different mother tongues as a common language. For speakers of uncommon languages in a locality, the local majority language would likely provide the greatest increase in the probability of being able to speak with the average person.

I have collected additional district-language level data from the 1961 census on the number of bilinguals in English and Hindi as well as bilinguals overall. The census did not tabulate data on

the specific second language spoken at the district level in 1931 (or in any other census year). I therefore conduct the analysis in levels, estimating how the share of speakers of a district-language who are bilingual in English, Hindi, or another language is affected by industrial growth over the prior 30 years, controlling for the overall level of bilingualism for each district-language in 1931. This estimation includes state fixed effects. It differs from the other specifications in this paper by not including district-language fixed effects.

Table 8 shows IV estimates with interactions for minority languages. For speakers of the district dominant language, industrial growth had the strongest effect on learning Hindi, with a coefficient of 1.43, and smaller 0.88 effect on learning English. Learning other second languages was actually decreased by industrial growth by -0.49. In south India, Hindi has long been associated with north Indian dominance and is not widely used as a *lingua franca*. As we would expect, industrial growth leads predominantly to English bilingualism in the south for dominant languages (regression not shown). Speakers of secondary languages had similar effects for all three categories: 0.95 for English, 0.39 for Hindi, and 0.78 for other. Outside of Hindi-majority areas, the dominant language will fall into the other category, which explains the large effect for secondary speakers.

These results relate quite directly to the evolving literature on the returns to English in the IT and business process outsourcing sector in the present day. They show that bilingualism in a *lingua franca* has been an economically important skill not only over the long run but also in the lower-skilled and much larger industrial sector.

Literacy

This section turns to the effect industrial employment expansion had on the investment in literacy and makes comparisons with bilingualism. Literacy expanded from 9% to 27% of the population between 1931 and 1961. Recall that the standard of literacy used by the census was that a person be able to read and respond to a simple letter. The fact that only 8% of the Indian population had completed primary school in 1961 reminds us that learning to read does not require extensive instruction.

I estimate regressions at the district level using first differences and include fixed effects as in the bilingualism regressions. OLS estimation shows a 1-point increase in the industrial share is correlated with 0.54-point increase in the literate share (Table 9, column 1). This is very close in magnitude to the coefficient on bilingualism in the same specification, which was 0.55. The expansion in literacy (18.1 points) was much larger than that of bilingualism (4.1 points), which make industrial growth appear relatively less important as a driver of literacy growth. The IV estimate is 1.14, about twice as large as the OLS (column 2). The difference reflects the same set of factors discussed above for bilingualism. Since the same first-stage is used in both cases, the sensitivity to the exclusion restriction is similar for this estimate.

I show how the effect on literacy varies by import competing status in column 3. Literacy in import competing districts is more affected by industrial growth, but the difference is quite small, 1.17 versus 0.92 points. This difference was stronger for bilingualism, 1.89 versus 0.50. Demand for literacy was relatively similar across industries, whereas bilingualism was more important in the import competing ones. This suggests that there are differences across secotrs in the relative demand for literacy and bilingualism.

Bilingualism and Literacy

How much of the industrial share effect on bilingualism might be a consequence of increased literacy? Is it likely that, as both bilingualism and literacy can be an outputs of a formal education, most of the unconditional effect of the industrial share on bilingualism actually operates through increasing literacy, so that bilingualism is a kind of byproduct of literacy?

Bilingualism is certainly a stepping stone to literacy if one wishes to read a different language than one's mother tongue. However, all languages in my data have scripts and written forms, so bilingualism wasn't strictly necessary for literacy. Chaudhary (2010) reports that only 14% of literates knew English in 1931. Further, given the low level of primary completion in 1961 compared to the level of literacy and the fairly rudimentary literacy standard, it isn't necessarily true that most people who became literate did so through schooling rather than a more informal

arrangement. However, schooling was not offered in all languages in all places, and so some of those desiring education would have learned a second language to do so.

On the other hand, bilingualism and literacy do move together. The coefficient on the change in literacy regressed on change in bilingualism is 0.28. The correlation is even stronger, 0.52, if we look within policy environments by including state fixed effects.[7]

Overall, literacy grew faster than bilingualism. The number of bilinguals per literate fell from 1.47 in 1931 to 0.55 in 1961. My unconditional estimates of the industrial share effect of bilingualism, 1.61, is higher than that on literacy, 1.14, suggest that industrial growth works against the trend. The final two columns of Table 9 show that a 1-point increase in the industrial share increased the number of bilinguals per literate by 13 points. At the mean change in the industrial share of 2.9 points, with the usual caveat about LATE, the IV estimate in column 5 suggests industrial growth increased the ratio by 0.38 bilinguals per literate.

Rigorously estimating the effect of the industrial share on bilingualism conditional on literacy is more difficult. We would need an additional instrument for literacy. Even so, the first stages of the conditional estimates would be different than the unconditional ones, which implies that, following the logic of LATE, the conditional and unconditional estimates might not be strictly comparable.

I take the simpler approach of investigating how the IV estimate changes for fixed values of the coefficient on literacy as a conditioning variable. In the spirit of the sensitivity test conducted above, consider the regression

$$\Delta B_{\ell d} - \theta L_{\ell d} = \alpha + \beta \Delta I_d + s_d + \Delta \varepsilon_{\ell d} \tag{9}$$

where $L_{\ell d}$ is the change in the literate share of the population. How will the estimate of $\hat{\beta}$ change for different fixed values θ_0? If θ_0 is positive, $\hat{\beta}$ will fall.

I estimate equation 9 for various θ_0 and plot the results in Figure 5. Were the entire effect of industrial growth on bilingualism to actually come through literacy, we would need to have

[7] Changes in bilingualism and literacy are not rank-correlated, however.

$\theta_0 > 1.5$. In other words, even if literacy and bilingualism were perfect complements in production, there would still need to be substantial additional spillovers from literacy into bilingualism to drive away the effect of industrial growth. If $\theta = 1$, we have $\hat{\beta} = 0.48$, and if it is the same size as the unconditional correlation with bilingualism, $\theta = 0.52$, then $\hat{\beta} = 1.02$.

Learning, Migration, and Assimilation

Learning, migration, and assimilation are the most plausible channels through which a change in the industrial share would affect bilingualism. Human capital theory says people will learn a second language when the net benefits are high enough. These benefits may motivate bilinguals to move into a district where industrial employment is growing. The effect of the industrial growth could result in part from the sorting of bilinguals across districts. Industrial employment growth may also spur in-migration of monolinguals, which would also affect the bilingual share. Assimilation is also a mechanism through which the bilingual share may change. If parents decide to teach their children only their second language, the children become monolinguals in a different mother tongue group from their parents.

In this section I will present evidence about the roles played by migration and assimilation. I cannot directly measure these channels with the data available. However, the analysis I conduct provides support for the idea that migration and assimilation played small roles in producing the effects I measure.

My approach for the study of migration is to consider the effect of industrial employment growth in district d on the population in the districts that border d that speak the languages spoken in d. For example, if Tamil is spoken in both districts A and B, industrial share growth in A might induce some Tamil speakers in B to move to A. This would change the bilingual share in A due to migration. The magnitude and direction of the change would depend on the share of bilinguals in the migrating Tamil speakers relative to the share of Tamils who are bilinguals in district A.

I first make a list of all the languages spoken in d. For each language on the list that is also spoken in at least one of the adjacent districts, I collect the number of speakers and bilinguals and

compute the changes in the bilingual shares $\Delta B_{\ell d}^{ADJ}$ as well as log changes in the number of speakers and bilinguals. There are 547 such languages. I then analyze how changes in the industrial share affects languages in adjacent districts. Regressions are unweighted. In order to minimize bias from spatially correlated unobservables, I recompute the instrument excluding both the district concerned and its adjacent districts from the prediction regressions.

My analysis suggests that migration from adjacent districts is a small component of the effect of the industrial share on bilingualism. The first piece of evidence is that the effect of industrial growth has a small and statistically insignificant effect on the bilingual share in surrounding districts (Table 10, column 1). Column 2 shows the unweighted IV estimate of the industrial share effect on own-district bilingualism for comparison. At 1.29 points it is smaller than the weighted estimate of 1.61 points.

Even if industrial growth had no effect on the bilingual share in surrounding districts, it might still be drawing bilinguals in a similar proportion to monolinguals. If the movement were large enough and the adjacent districts began with a higher bilingual share, in-migration could still account for part of the industrial share coefficient. I check how the log change in bilinguals and speakers in surrounding districts is affected by the log change in industrial jobs. For consistency, I recompute the instrument using log levels of 1931 sectoral employment. The point estimates in columns 3 and 4 are statistically insignificant, though the relative magnitudes suggest that to the extent in-migration is important it mostly involves bilinguals.

Industrial employment growth may depress the bilingual share of secondary languages by promoting the assimilation of children to the dominant language. Assimilation would then raise the population shares of the dominant languages. The process of assimilation takes at least one generation, and is thus much slower than learning or migration. The secular trend is for the average population share speaking a dominant language to grow from 75% to 77%. I estimate the differential impact of industrial employment growth on the population share speaking the dominant language in Table 10, column 5. The coefficient is small and the sign the opposite predicted by the assimilation hypothesis. Assimilation would also produce a decline in linguistic heterogeneity.

Linguistic heterogeneity is sensitive to changes among the secondary languages, and might be affected independently of the dominant language share if there were assimilation among secondary languages. Column 3 shows no effect of industrial share growth on linguistic heterogeneity, though the estimate is imprecise.

Bilingualism and Assimilation Over Time

While industrial share growth doesn't produce assimilation to larger languages over the span of the panel, it clearly does increase bilingualism. Bilingualism is a precondition for assimilation as parents and children always have at least one language in common. This implies that for children to assimilate, parents must be bilingual. A number of scholars have pointed out that the economic return to the parent's mother tongue is not the only consideration (Grin 1992; Linton 2004; Wickstrom 2005). The social status and political power of the mother tongue community, the use of the language in important cultural activities, and the strength and value of the social network associated with the language are also part of the decision.

If assimilation is occurring, initial bilingualism among speakers of a given language should be negatively correlated with the share of the population speaking that mother tongue later on. Table 11 explores the correlation between bilingualism in 1931 among speakers of a secondary language and the population share speaking that language thirty years later. The first two columns use the 1931–1961 district-language panel. A larger bilingual share in 1931 has a negative correlation with the population share speaking a language thirty years later, conditional on the 1931 population share (column 1). Adding district fixed effects does not alter the correlation (column 2). For the final two columns I use a state-language dataset spanning 1961 to 1991. A negative correlation between initial bilingualism and the share speaking the language appears here as well.

This data does not provide evidence of the mechanism. While it suggests that assimilation is underway, the pattern could be generated simply by differential population growth. If we assume that the estimated coefficient applies to variation in bilingualism generated by industrial expansion, we can roughly estimate the impact. Recall that my IV estimation suggested industrial

share growth increased bilingualism among secondary language speakers by 6.0 points. Applying this estimated change to the -0.05-point follow-on impact of initial bilingualism (column 2), secondary languages in a district that saw average industrial share growth would have a 0.3-point smaller population share 30 years later.

5 Conclusion

I have provided causal estimates showing the large effect industrial employment growth had in increasing bilingualism in mid-20$^{\text{th}}$ India. Additional analysis and findings by other scholars support my contention that individuals were motivated to become bilingual by returns to communication in the new industrial jobs. The effect was larger for speakers of locally less common mother tongues, who would have relatively large gains from bilingualism. My measured LATE is identified by industries that trade goods nationally and have locational persistence. Moreover, the effect is larger in import-competing districts following a large increase in tariffs. Industrial goods affected by this variation tend to be relatively high value and to be produced in more sophisticated plants. While industrial growth increased literacy, I have provided evidence that bilingualism is not an epiphenomenon of the demand for literacy. Migration of bilinguals, as opposed to learning, does not explain the effect.

The demand for bilingualism in relatively low-skilled industrial jobs has implications for education policy in linguistically diverse countries. Spoken language differs from other skills taught in schools in significant ways and should be considered a distinct species of human capital for that reason. Children have the highest capacity to learn languages before they are of school age (Johnson & Newport 1989). The earlier the learning begins, the quicker and better the results (Johnson & Newport 1991). Further, at this age learning does not require explicit instruction but happens automatically. This suggests an efficient way to increase bilingualism is to expose children to a second language before they are of school age. My results suggest that there can be a return this skill even for otherwise low-skilled workers.

Another implication follows from the observation that language investment decisions suffer

from a network externality (Church & King 1993). If person i learns a new language and completes a previously suboptimal transaction with person j, the welfare of person j will have increased at no cost to j. This suggests there will be underinvestment in bilingualism. Additionally, changes in economic networks that come from bilingualism can have distributional consequences. Consider monolingual speakers m and n of a secondary language who choose to trade with each other. If m learns the dominant language, he may find a new majority-language trading partner p whom he prefers to n. Social welfare will increase overall, but there will be a loss for n, who must now take a second-best trading partner. Those minority language speakers who can afford investment in a second language are likely to be better off than others with their mother tongue, suggesting that the remaining secondary language speakers will be negatively selected. This may in part explain why linguistic minorities are often of relatively low socioeconomic status.

More broadly, my findings have implications for the economics of cultural diversity. A vast literature in the social sciences has investigated the relationship between racial, ethnolinguistic, and religious differences and economic outcomes (e.g. Akerlof & Kranton 2000; Alesina *et al.* 1999, 2003; Alesina & La Ferrara 2005).

First, measures of group membership and diversity in this literature often take groups to be mutually exclusive. Bilinguals span linguistic groups and may have multiple ethnic identities. In my study, bilingualism rises without much change in the usual measure of linguistic diversity. However, the presence of bilinguals may serve to reduce effective ethnic diversity. Common measures of ethnolinguistic heterogeneity can be modified to take account of bilinguals to see whether their presence mediates the economic and political effects of ethnic diversity.

Second, language, ethnic identity, and culture are intimately related. Learning a new language provides a person not only a functional capacity to communicate, but also grants access to cultural resources such as media, literature, and even religion. It may provide them with a new identity. I have shown that such cultural changes can be spurred by economic forces that make the functional capacity more important.

Where might we see examples of this process more clearly? Consider the formation of Western

European nations. Western Europe was once much more linguistically diverse than it is today. As late as the 1880s, only half of the population of France were mother-tongue French speakers (Weber 1976).[8] A process of linguistic homogenization produced the more uniformly French-speaking France of today. The situation was similar in Britain before the Industrial Revolution.[9] Linguistic homogenization went hand in hand with the development of national cultures and the claims of nations to political sovereignty.

Political scientists have been the most active scholars of this process. Their contributions have naturally focused on the role of the state policy, particularly concerning the language of instruction in schools and official language status (e.g. Laitin 1993; De Swaan 1993; van Parijs 2000). As the birthplace of modern industry, Europe is a promising place to look for long-run links between economic development, language consolidation, and cultural change. The findings in this paper suggest a complementary line of research focused on the economic drivers of cultural change through language acquisition. To cite one example, we know that falling transport costs helped to integrate labor and product markets in 19th century Europe (O'Rourke & Williamson 2000). Understanding the role played by the expansion of markets in the linguistic consolidation of Europe would greatly illuminate our understanding of the interplay between culture, markets, and the state.

[8]Common languages in 19th century France included Occitan, Breton, Norman, Gallo, and Burgundian.
[9]Cornish, Welsh, Scottish Gaelic, Scots, and Manx were important regional languages.

References

Akerlof, George, & Kranton, Rachel. 2000. Economics and Identity. *Quarterly Journal of Economics*, **115**(3), 715–753.

Alesina, Alberto, & La Ferrara, Eliana. 2005. Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, **42**, 762–800.

Alesina, Alberto, Baqir, Reza, & Easterly, William. 1999. Public Goods and Ethnic Divisions. *Quarterly Journal of Economics*, **114**(4), 1243–84.

Alesina, Alberto, Devleeschauwer, Arnaud, Easterly, William, Kurlat, Sergio, & Wacziarg, Romain. 2003. Fractionalization. *Journal of Economic Growth*, **8**, 155–194.

Angrist, JD, & Pischke, JS. 2009. *Mostly Harmless Econometrics*. Princeton University Press.

Angrist, Joshua D., Imbens, Guido W., & Rubin., Donald B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91**(434).

Autor, David H., & Duggan, Mark. 2003. The Rise In The Disability Rolls And The Decline In Unemployment. *Quarterly Journal Of Economics*, **118**(1).

Azam, Mehtabul, Chin, Aimee, & Prakash, Nishith. 2013. The Returns to English-Language Skills in India. *Economic Development and Cultural Change*.

Barro, Robert, & Lee, Jong-Wha. 2010. *A New Data Set of Educational Attainment in the World, 1950-2010*. NBER Working Paper 15902.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute.

Berman, Eli, Lang, Kevin, & Siniver, Erez. 2003. Language-skill complementarity: returns to immigrant language acquisition. *Labor Economics*, **10**, 265–290.

Blanchard, Olivier, & Katz, Lawrence. 1992. Regional Evolutions. *Brookings Papers on Economic Activity*.

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

Borjas, GJ. 2003. The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market. *The Quarterly Journal of Economics*, **118**(4), 1335–1374.

Breman, J. 1999. The Study of Industrial Labour in Post-Colonial India–The Formal Sector: An Introductory Review. *Contributions to Indian Sociology*, **33**(1-2), 1–41.

Burgess, Robin, & Donaldson, Dave. 2010. Can Openness Mitigate the Effects of Weather Fluctuations? Evidence from Indias Famine Era. *American Economic Review Papers and Proceedings*, **100**(2), 449–453.

Card, David. 2009. Immigration and Inequality. *American Economic Review*, **99**(2), 1–21.

Cashin, P, & Sahay, R. 1996. Internal Migration, Center-State Grants, and Economic Growth in the States of India. *Staff papers - International Monetary Fund. International Monetary Fund*, **43**(1), 123–71.

Chandravarkar, Rajnarayan. 1994. *The Origins of Industrial Capitalism in India*. Cambridge University Press.

Chaudhary, L. 2010. Land Revenues, Schools and Literacy: A Historical Examination of Public and Private Funding of Education. *Indian Economic & Social History Review*, **47**(2), 179–204.

Chaudhary, Latika. 2009. Determinants of Primary Schooling in British India. *The Journal of Economic History*, **69**(1), 269–302.

Chiswick, Barry, & Miller, Paul. 1995. The Endogeneity between Language and Earnings: International Analyses. *Journal of Labor Economics*, **13**(2), 246–288.

Chiswick, BR, & Miller, PW. 1998. The Economic Cost to Native-Born Americans of Limited English Language Proficiency. *Pages 413–430 of:* Chiswick, B.R, & Miller, P.W. (eds), *The Economics of Language: International Analyses*.

Church, Jeffrey, & King, Ian. 1993. Bilingualism and Network Externalities. *The Canadian Journal of Economics / Revue Canadienne d'Economique*, **26**(2), 337–345.

Conley, Timothy G, Hansen, Christian B, & Rossi, Peter E. 2012. Plausibly Exogenous. *Review of Economics and Statistics*, **94**(1), 260–272.

De Swaan, a. 1993. The Emergent World Language System: An Introduction. *International Political Science Review*, **14**(3), 219–226.

Donaldson, Dave. 2010. *Railroads of the Raj: Estimating the Impact of Transportation Infrastructure*. NBER Working Paper 16487.

Dustmann, Christian, & van Soest, Arthur. 2001. Language fluency and earnings: estimation with misclassified language indicators. *Review of Economics and Statistics*, **83**(4), 663674.

Fry, Richard, & Lowell, BL. 2003. The Value of Bilingualism in the US Labor Market. *Industrial and Labor Relations Review*, **57**(1), 128–140.

Gokhale, R.G. 1957. *The Bombay Cotton Mill Worker*. Bombay Millowners Association.

Gordon, Raymond G. 2005. *Ethnologue: Languages of the World*. 15th edn. Dallas, Tex: SIL International.

Grin, F. 1992. Towards a Threshold Theory of Minority Language Survival. *Kyklos*, **45**(1), 69–97.

Holmström, Mark. 1976. *South Indian Factory Workers: Their Life and World*. Cambridge University Press.

Holmström, Mark. 1984. *Industry and Inequality: The Social Anthropology of Indian Labour*. Cambridge University Press.

Imbens, Guido, & Angrist, Joshua. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**(2), 467–475.

India. 1933a. *Census of India 1931*. Census Commissioner, Government of India.

India. 1933b. *Statistical Abstract for British India*. London: H.M.S.O.

India. 1962. *Census of India 1961*. Census Commissioner, Government of India.

Johnson, Jacqueline, & Newport, Elissa. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, **21**, 60–99.

Johnson, Jacqueline, & Newport, Elissa. 1991. Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition*, **39**, 21525.

Kapur, Shilpi, & Chakraborty, Tanika. 2009. *English Language Premium: Evidence from a Policy Experiment in India*. IZA Working Paper.

Laitin, D. D. 1993. The Game Theory of Language Regimes. *International Political Science Review*, **14**(3), 227–239.

Lambert, R.D. 1963. *Workers, Factories, and Social Change in India*. Princeton University Press.

Lang, Kevin, & Siniver, Erez. 2009. The Return to English in a Non-English Speaking Country: Russian Immigrants and Native Israelis in Israel. *The B.E. Journal of Economic Analysis and Policy*.

Lewis, Ethan. 2011. Immigration, Skill Mix, and Capital-Skill Complementarity. *Quarterly Journal of Economics*, **126**(2).

Light, RJ, & Margolin, BH. 1971. An Analysis of Variance for Categorical Data. *Journal of the American Statistical Association*, **66**(335), 534–544.

Linton, April. 2004. A Critical Mass Model of Bilingualism among US-born Hispanics. *Social Forces*, **83**(1), 279–314.

Luttmer, Erzo F. P. 2005. Neighbors as Negatives: Relative Earnings and Well-Being. Quarterly Journal Of Economics. *Quarterly Journal Of Economics*, **120**(5), 20–54.

Mankiw, N. Gregory, Romer, David, & Weil, David N. 1992. A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics*, **107**(2), 407–437.

Morris, MD. 1965. *The Emergence of an Industrial Labor Force in India*. University of California Press.

Munshi, Kaivan, & Rosenzweig, Mark. 2006. Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy. *American Economic Review*, **96**(4), 1225–1252.

Munshi, Kaivan, & Rosenzweig, Mark. 2009. *Why is Mobility in India so Low? Social Insurance, Inequality, and Growth*. NBER Working Paper 14850.

O'Rourke, Kevin, & Williamson, Jeffrey. 2000. *Globalization and History*. MIT Press.

Oster, Emily, & Millett, Bryce. 2011. *Do Call Centers Promote School Enrollment? Evidence from India*. Working Paper.

Rice, A.K. 1958. *Productivity and Social Organization: The Ahmedabad Experiment*. Tavistock Publications.

Roy, Tirthankar. 1999. *Traditional Industry in the Economy of Colonial India*. Cambridge: Cambridge University Press.

Roy, Tirthankar. 2000. *The Economic History of India, 1857-1947*. Oxford University Press.

Roy, Tirthankar. 2010. *Company of Kinsman: Enterprise and Community in South Asian History, 1700–1940*. Oxford University Press.

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

Sheth, N.R. 1968. *Social Framework of an Indian Factory*. Manchester University Press.

Singh, R.P., & Banthia, Jayant Kumar. 2004. *India Administrative Atlas, 1872-2001: A Historical Perspective of Evolution of Districts and States in India*. New Delhi: Controller of Publications.

Sivasubramonian, S. 2000. *The National Income of India in the Twentieth Century*. New Delhi: Oxford University Press.

Srivastava, Shyam Chandra. 1972. *Indian Census in Perspective*. New Delhi: Office of the Registrar General.

Tomlinson, B.R. 1979. *The Political Economy of the Raj, 1914-1947 : The Economics of Decolonization in India*. London: Macmillan Press.

Topalova, Petia. 2010. Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India. *American Economic Journal: Applied Economics*, **2**(4), 1–41.

van Parijs, P. 2000. The Ground Floor of the World: On the Socio-Economic Consequences of Linguistic Globalization. *International Political Science Review*, **21**(2), 217–233.

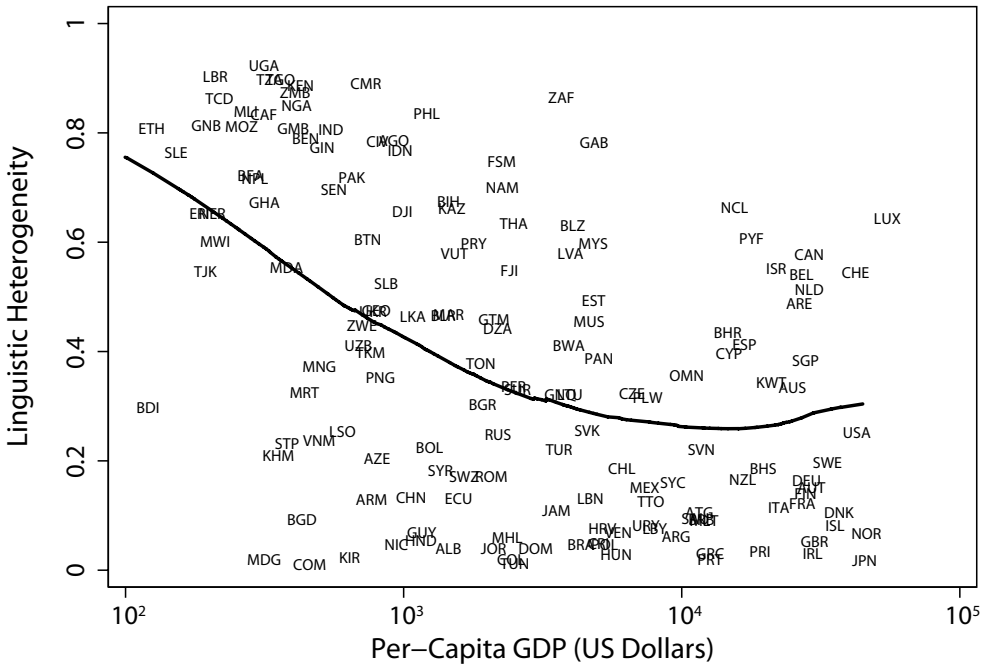
Vidyarthi, L.P. 1970. *Sociocultural Implications of Industrialization in India*. Planning Commission.

Weber, Eugen. 1976. *Peasants into Frenchmen : the Modernization of Rural France, 1870-1914*. Palo Alto: Stanford University Press.

Wickstrom, B.-a. 2005. Can Bilingualism be Dynamically Stable?: A Simple Model of Language Choice. *Rationality and Society*, **17**(1), 81–115.

World Bank. 2012. *World Development Indicators Online*. http://data.worldbank.org/data-catalog/world-development-indicators.

Figure 1: Linguistic Diversity and GDP per Capita



Notes: This graph shows data on linguistic diversity from Ethnologue and GDP per capita in 2000 from the World Development Indicators (Gordon 2005; World Bank 2012). Linguistic heterogeneity is $1 - h$, where h is the Herfindahl index of language shares of population within each country.

Table 1: Summary Statistics

A. Average District Characteristics

	1931	1961	Change
Population (Millions)	1.51	2.37	57.2%
Urban Share	13.7%	19.7%	6.0
Workforce Population Share	36.2%	44.7%	8.4
Workforce Population Share, Males	53.5%	57.4%	4.0
Industrial Employment Share	8.9%	11.8%	2.8
Secondary Language Speakers Share of Population	24.9%	23.3%	-1.6
Bilingual Share of Population	8.0%	12.1%	4.1
Bilingual Share, Secondary Language Speakers	29.7%	33.7%	4.0
Literate Share of Population	9.1%	27.2%	18.1
Bilinguals per Literate Person	1.47	0.56	-0.91
Linguistic Heterogeneity	0.36	0.34	-0.02

B. Breakdown of Bilingualism in 1961 Districts

		Share of Total Bilinguals by Second Language Spoken		
Dominant Mother Tongue of District	Overall Share Bilingual	English	Hindi	Other
Hindi (N=47)				
Hindi Mother Tongue Speakers	4.2%	57.8%		42.2%
Secondary Mother Tongue Speakers	25.7%	7.9%		92.1%
Not Hindi (N=159)				
Dominant Mother Tongue Speakers	5.6%	46.7%	27.7%	25.3%
Secondary Mother Tongue Speakers	41.3%	10.5%	10.3%	74.7%

Notes: Averages are weighted by the district population, except for the bilingual shares, which are weighted by the average number of speakers for each district-language observation. There are 137 district- and 684 district-language observations in the panel dataset described in Panel A. There are 206 districts covering the same territory in 1961 due to boundary changes.

Table 2: OLS Estimates

	Δ Bilingual Share				
	(1)	(2)	(3)	(4)	(5)
Δ Industrial Share Emp.	0.659***	0.610***	0.501***	0.547***	0.495**
	(0.193)	(0.177)	(0.187)	(0.196)	(0.204)
Constant	0.016***	0.022***	0.025***	0.024***	0.025***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Adjusted R-squared	0.170	0.051	0.149	0.170	0.162
N	137	684	684	684	684
Aggregation†	D	DL	DL	DL	DL
Fixed Effects‡	No	No	St31	ML×St31	St61

† D: Observation is a district; DL: Observation is a language within a district.
‡ St31: State the district belonged to in 1931; ML: Majority language of the district in 1931 (i.e. Bengali, Tamil, Hindi, etc.); St61: State the district belonged to in 1961.
Notes: Regressions are weighted by the average number of speakers of the district-language. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 3: Industrial Sectors in 1931

	Dist.-Av. Share Industrial Emp.	Share of Sector, Total Value of Ind. Exports	Ind. Imports
Textiles	0.349	0.708	0.354
Boots and shoes	0.050		0.005
Tailors and Millners	0.049	0.003	0.016
Personal Care Products	0.001		
Leather	0.022	0.105	0.011
Wood products	0.143	0.020	0.005
Furniture	0.002		0.002
Metals and Metalworking	0.063		0.269
Ceramics	0.070		0.017
Chemicals	0.032	0.096	0.091
Food Processing	0.116	0.069	0.135
Vehicles	0.003		0.046
Power	0.002		
Other	0.098		0.050

Notes: The first column presents unweighted averages across all 137 districts in the panel. The second and third columns take the average value between 1927 and 1931 of India's industrial exports and imports and shows the share each subindustry represents of the total.

Table 4: Analysis of the Instrument

	Δ Industrial Share Employment		Pred. Δ Industrial Share Employment	
	(1)	(2)	(3)	(4)
Pred. Δ Industrial Share Employment	0.733***	0.731***		
	(0.101)	(0.162)		
F-Statistic	*52.27*	*20.23*		
Cultivator Emp. Share, 1931			0.005	0.012
			(0.004)	(0.011)
Ag. Labor Emp. Share, 1931			0.007	-0.007
			(0.005)	(0.013)
Urban Share, 1931			-0.002	0.010
			(0.018)	(0.025)
Literate Share, 1931			-0.038	-0.045
			(0.026)	(0.038)
Overall Share Bilingual, 1931			0.009	0.005
			(0.007)	(0.009)
Linguistic Heterogeneity, 1931			0.001	-0.007
			(0.005)	(0.008)
Population Density, 1931			0.000	0.001
			(0.000)	(0.001)
Import Competitor, 1931			-0.002	0.000
			(0.005)	(0.005)
Major City			-0.006	-0.008
			(0.008)	(0.012)
Coastal District			-0.002	0.001
			(0.002)	(0.002)
Border District			-0.002	0.005
			(0.002)	(0.007)
Constant	0.006*	0.019	0.025***	0.021**
	(0.003)	(0.015)	(0.005)	(0.010)
Adjusted R-squared	0.304	0.352		
ML × 1931 State Fixed Effects	N	Y	N	Y
Subindustry Share Controls	N	N	Y	Y
N	137	137	137	137

Notes: Observations are at the district level. Standard errors are corrected for heteroskedasticity. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 5: IV Estimates of Industrial Share Effects on Bilingualism

	Δ Bilingual Share			
	(1)	(2)	(3)	(4)
Δ Industrial Share Emp.	1.479***	1.609***	1.366***	1.482***
	(0.25)	(0.26)	(0.27)	(0.21)
Constant	-0.002	-0.030	-0.005	-0.042*
	(0.01)	(0.02)	(0.01)	(0.02)
N	684	684	580	580
Districts	All	All	Central	Central
ML × 1931 State Fixed Effects	N	Y	N	Y
First-Stage F-Stat	*23.09*	*13.11*	*37.75*	*32.34*

Notes: Observations are at the district-language level and are weighted by the average number of speakers. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 6: Differential Effects by Import Competition

	Δ Bilingual Share	
	(1)	(2)
Δ Industrial Share Emp.	-0.020	0.501
	(0.142)	(0.517)
Δ Industrial Share Emp. × Imp. Comp.	0.725***	1.390**
	(0.257)	(0.597)
Import Competitor, 1931	-0.016**	-0.042***
	(0.007)	(0.013)
Constant	0.035***	-0.008
	(0.005)	(0.021)
ML × 1931 State Fixed Effects	Y	Y
Estimation	OLS	IV
N	684	684
Adjusted R-squared	0.176	
First-Stage F-Stat		8.97
First-Stage F-Stat, interaction		4.98

Notes: Observations are at the district level and weighted by average district population. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 7: Differential Effects by Secondary Language and Heterogeneous Districts

	Δ Bilingual Share	
	(1)	(2)
Δ Industrial Share Emp.	1.333***	1.028
	(0.351)	(0.874)
Δ Ind. Share Emp. X Sec. Lang.	0.757	
	(0.610)	
Δ Ind. Share Emp. X High Ling. Het.		0.655
		(0.842)
Secondary Language	-0.033	
	(0.024)	
High Linguistic Heterogeneity		-0.011
		(0.029)
Constant	-0.018	-0.007
	(0.021)	(0.036)
N	684	684
ML × 1931 State Fixed Effects	Y	Y
Estimation	IV	IV
First-Stage *F*-Stat	*19.54*	*5.43*
First-Stage *F*-Stat, interaction	*19.98*	*11.88*

Notes: Observations are at the district-language level and are weighted by the average number of speakers. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 8: Industrial Share Growth and Bilingualism by Second Language in 1961

	1961 Share of Speakers who were Bilingual in		
	English	Hindi	Other
	(1)	(2)	(3)
Δ Industrial Share Emp.	0.875***	1.428***	-0.493**
	(0.336)	(0.399)	(0.210)
Δ Ind. Share Emp. X Sec. Lang.	0.077	-1.034*	1.269**
	(0.236)	(0.577)	(0.576)
Sec. Lang.	-0.004	0.026	0.046*
	(0.007)	(0.021)	(0.025)
Share of Speakers Bilingual, 1931	-0.090	-0.240*	0.732***
	(0.065)	(0.140)	(0.120)
Share Bilingual × Sec. Lang.	0.080	0.272*	-0.136
	(0.065)	(0.148)	(0.124)
Constant	-0.011	-0.042*	0.016
	(0.014)	(0.023)	(0.010)
N	682	567	684
1961 State Fixed Effects	Y	Y	Y
Estimation	IV	IV	IV
First-Stage F-Stat	10.43	10.43	10.43
First-Stage F-Stat, interaction	46.80	46.80	46.80

Notes: Observations are at the district-language level and are weighted by the average number of speakers. Columns 1 and 2 exclude mother tongue speakers of English and Hindi, respectively. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 9: Industrial Share Effects on Literacy

	Δ Literate Share			Δ Bilinguals per Literate Person	
	(1)	(2)	(3)	(4)	(5)
Δ Industrial Share Emp.	0.538***	1.138***	0.923***	14.810***	13.113**
	(0.126)	(0.326)	(0.298)	(5.342)	(5.275)
Δ Ind. Sh. Emp. × Imp. Comp.			0.246		
			(0.461)		
Import Competitor, 1931			-0.003		
			(0.011)		
Constant	0.138***	0.107***	0.109***	-1.370***	-0.819***
	(0.018)	(0.017)	(0.015)	(0.279)	(0.265)
N	137	137	137	137	137
ML × 1931 State Fixed Effects	Y	Y	Y	Y	Y
Estimation	OLS	IV	IV	OLS	IV
Adjusted R-squared	0.647	0.498	0.506	0.054	0.451
First-Stage *F*-Stat		*10.19*	*6.13*		*10.19*
First-Stage *F*-Stat, interaction			*3.97*		

Notes: Observations are at the district level and weighted by average district population. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 10: Migration and Assimilation

	Δ Bilingual Share Adj. Districts (1)	Δ Bilingual Share (2)	Δ Log Bilinguals Adj. Dists. (3)	Δ Log Speakers Adj. Dists. (4)	Δ Majority Language Share of Pop. (5)	Linguistic Heterogeneity (6)
Δ Industrial Share Emp.	-0.224	1.292***			0.029	0.052
	(0.391)	(0.368)			(0.445)	(0.358)
Δ Industrial Workers			-0.263	0.096		
			(0.429)	(0.220)		
Constant	0.085***	-0.004	1.498***	0.311	-0.008	-0.003
	(0.027)	(0.036)	(0.393)	(0.200)	(0.029)	(0.035)
N	547	684	547	547	137	137
Estimation	IV	IV	IV	IV	IV	
First-Stage F-Stat	*11.36*	*9.30*	*7.88*	*7.88*	*23.79*	*10.19*

Notes: Instrument for columns 1 to 4 additionally excludes adjacent districts in prediction regressions. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 11: Bilingualism and Assimilation of Secondary Languages

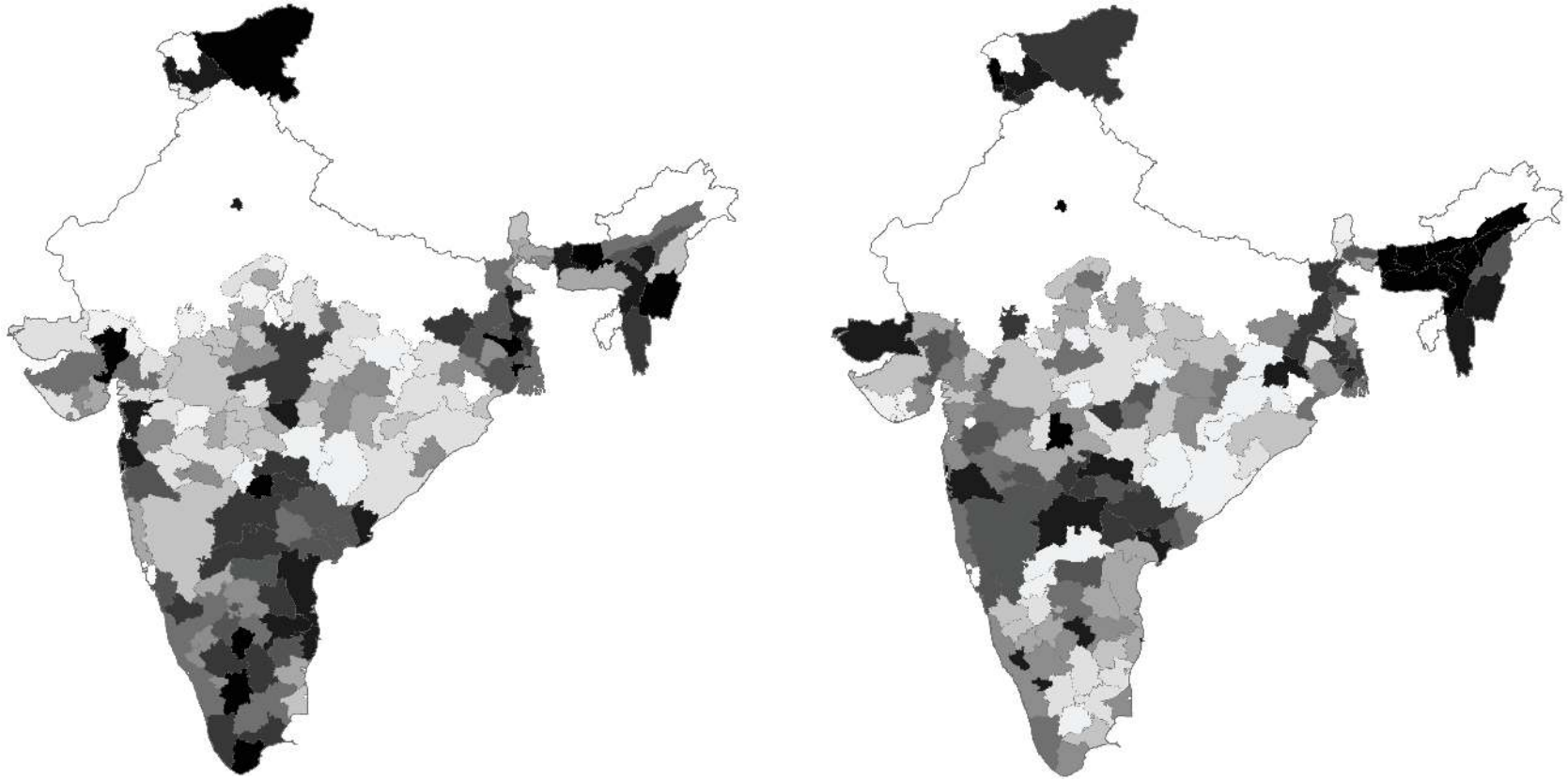
	Final Pop. Share Speaking Language			
	District-Language Panel, 1931–1961		State-Language Panel, 1961–1991	
	(1)	(2)	(3)	(4)
Initial Share of Speakers Bilingual	-0.023**	-0.052***	-0.041**	-0.022**
	(0.011)	(0.017)	(0.021)	(0.009)
Initial Pop. Share Speaking Language	0.845***	0.825***	1.22***	1.20***
	(0.075)	(0.051)	(0.017)	(0.033)
Constant	0.014	0.000	0.011	0.048***
	(0.011)	(0.005)	(0.008)	(0.018)
Fixed Effects	None	District	None	State
Adjusted R-squared	0.832	0.934	0.989	0.995
N	547	547	554	554

Notes: OLS estimates. Observations at the district-language level for columns 1 and 2 and at the language-state level for columns 3 and 4. Observations weighted by the number of average speakers. Standard errors corrected for heteroskedasticity. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Figure 2: Changes in Industrial Employment and Bilingualism by Deciles 1931–1961

(a) Industrial Employment Share

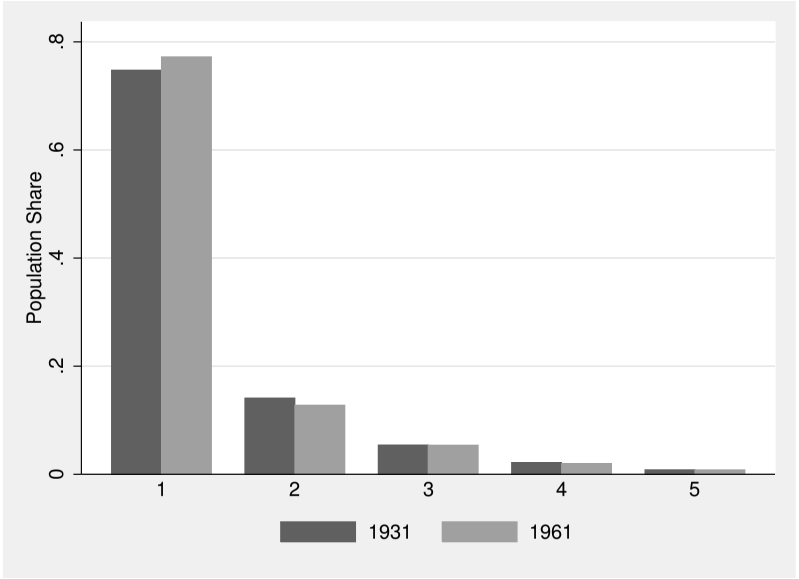
(b) Change in Bilingual Share



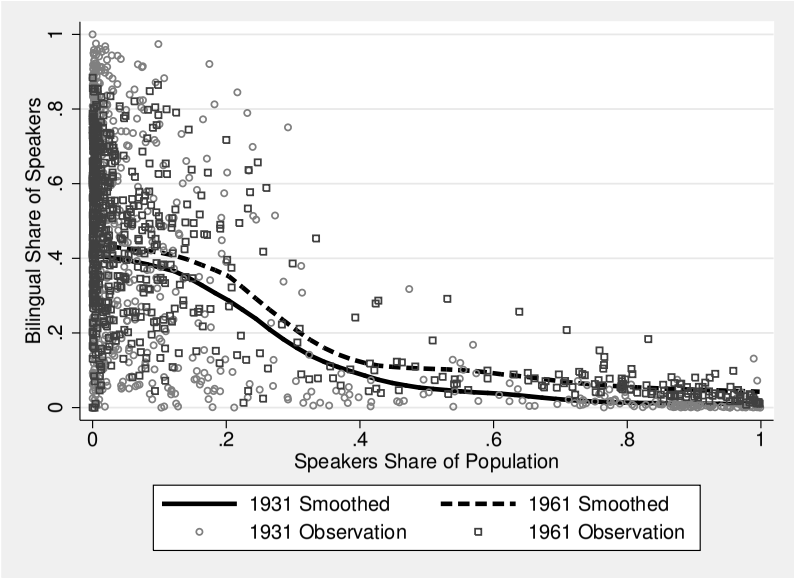
Notes: Maps plot the district-average changes in the industrial share of employment and bilingualism between 1931 and 1961. Light grey represents the first decile and black the tenth decile of the variable.

Figure 3: Language Characteristics

(a) Population Share by District-Level Language Rank

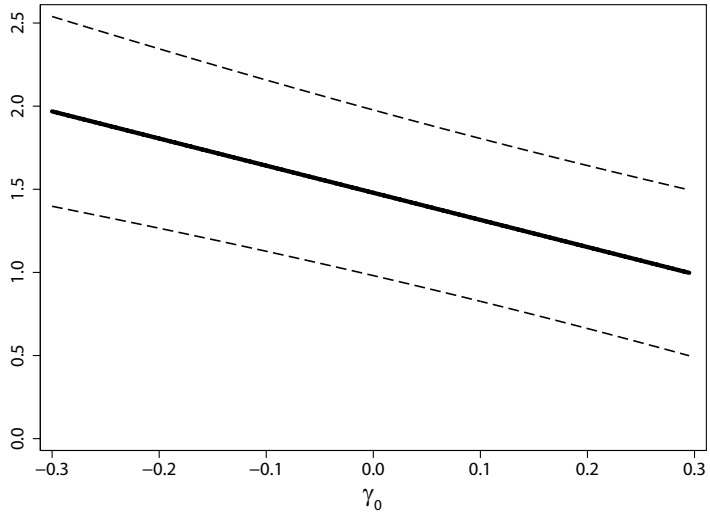


(b) Bilingual Share of Language by its Population Share



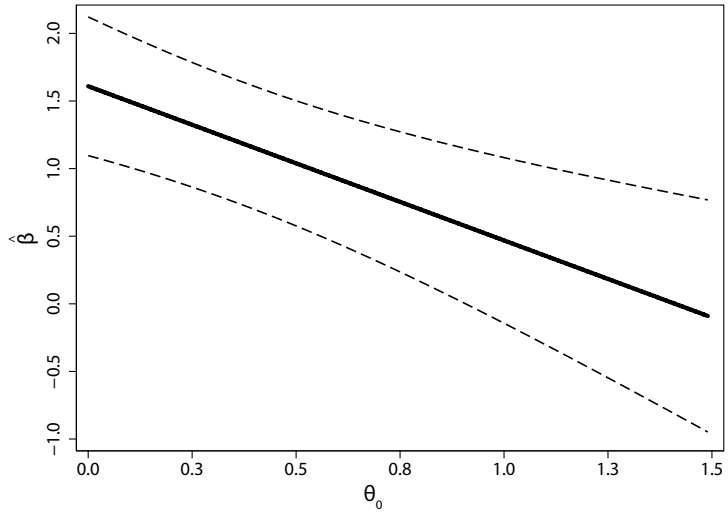
Notes: Panel (a) ranks each languages in a district from 1 to 5 by the number of speakers in 1931. The graph shows the mean share of district population for languages of each rank. Observations are weighted by the district population. N=684. Panel (b) plots data for each district-language observation for 1931 and 1961. The points indicate the bilingual share of speakers and the share of district population. The lines are a fit to the data using a kernel-weighted polynomial regression. Each district-language observation is weighted by the number of speakers in the relevant year. N=1,368.

Figure 4: Sensitivity of IV Estimates to the Exclusion Restriction



Notes: The graph shows how the IV estimate of the industrial share effect on bilingualism varies when the exclusion restriction $\gamma = 0$ is violated. Estimates and confidence intervals come from estimating equation 8 for different fixed values $\gamma = \gamma_0$.

Figure 5: Sensitivity of IV Estimates to the Conditional Effect of Literacy



Notes: The graph shows how the IV estimates of the industrial share on bilingualism are affected by conditioning on literacy for a set of fixed values of the conditional correlation θ. Estimates and confidence intervals come from estimating equation 9 for different fixed values $\theta = \theta_0$.

NOT FOR PUBLICATION

Online Appendix

A1 The Economics of Industrialization and Bilingualism

In this appendix I develop a simple model that links the decision to acquire a second language to imperfect sorting in the labor market and higher productivity enabled by communication in industry. This model builds on earlier work about the implications of language differences for discrimination and immigrant assimilation (Lang 1986; Lazear 1999, 2005).

As we saw in Section 1, workers in Indian industrial firms engage in communication in a common language related to their work. The owners of such firms may prefer to hire only workers who speak a common language. Changes in economic opportunities, such as the expansion of industrial employment, can alter the incentives to become bilingual.

Consider a two-period economy with two production sectors, industry and agriculture. The economy is populated by N families. Each family has one worker alive in each of periods 0 and 1. Workers are endowed with one unit of labor and engage in production in each period. Both sectors produce the same final good, the price of which is normalized to 1. Workers care about overall consumption for their family j: $U_j = c_j^0 + c_j^1$.

Two languages are spoken in the economy. A majority of period 0 workers speak the dominant language D while a minority speak the secondary language S: $p_D^0 > \frac{1}{2} > p_S^0$. Some workers may be bilingual. The population shares of monolingual D and S speakers in period t are m_D^t and m_S^t; the share of bilinguals is b^t. These shares sum to one.

$$m_D^t + m_S^t + b^t = 1. \tag{A1}$$

The period t population shares of everyone able to speak D and S, whether as monolinguals or as bilinguals, are $p_D^t = 1 - m_S^t$ and $p_S^t = 1 - m_D^t$.

The industrial sector makes use of a more productive technology than the agricultural sector. I take technology to be exogenous. The industrial technology requires workers to communicate to take advantage of its superior productivity. Industrial workers must therefore share a common language. Agricultural workers do not need to share a common language.

At the beginning of each period, workers are randomly paired into firms. If members of a firm share a language in common, they are capable of jointly operating the industrial technology. Otherwise they are only capable of working in agriculture. Common-language firms get access to the industrial technology with the exogenous probability $\pi\varepsilon(0,1]$, which reflects how widespread industry is in the economy. Workers in industry each earn the return w_I. Workers in firms that

do not share a common language or did not get access to the industrial technology must use the agricultural technology. Workers in agriculture each earn the return $w_A \leq w_I$.

The expected period 0 income of a monolingual D speaker is $p_D^0(\pi w_I + (1-\pi)w_A) + (1 - p_D^0)w_A$. A parallel expression holds for monolingual S speakers. Bilinguals always form a common-language firm and earn $\pi w_I + (1-\pi)w_A$. While workers in the real world target their job search based on their characteristics rather than getting opportunities randomly, this simple framework captures the intuitively appealing idea that there is a random element that affects match quality.

After workers are matched and produce in period 0, each gives birth to one child and decide how much to invest in its language ability. Period 0 workers may costlessly transmit one of the languages they know to their child. Bilingual workers may transmit both languages by paying the cost $s_j \sim U[0,\bar{s}]$. This cost reflects the delayed speech onset typical of bilingual children. Monolingual period 0 workers may also invest in making their child bilingual by paying $c_j \sim U[0,\bar{c}]$. I assume that $\bar{s}$ lies between $\frac{\bar{c}}{10}$ and $\bar{c}$. Workers can borrow against period 1 income without cost to finance investment if they wish. I assume that N is large enough that workers cannot coordinate investment decisions. Once period 0 workers have made their investment decision, the period ends.

A monolingual S speaker will invest in bilingualism if doing so increases expected family income in period 1. This will be the case if the expected income from forming a common-language firm with certainty, less the cost of bilingualism, is greater than the expected income of a monolingual S speaker in period 1:

$$\pi w_I + (1-\pi)w_A - c_j \quad \geq \quad p_S^1(\pi w_I + (1-\pi)w_A) + (1 - p_S^1)w_A. \tag{A2}$$

A parallel inequality holds for monolingual D speakers. Define $\lambda = \pi(w_I - w_A)$ as the expected increase in return from being in a common language firm and recall that $m_D^1 = 1 - p_S^1$. Equation A2 can then be rewritten as

$$\lambda m_D^1 \geq c_j. \tag{A3}$$

The benefit to a monolingual S speaker from becoming bilingual is the expected increase in return from forming a common language firm multiplied by the probability of matching with someone who only speaks D, in which case bilingualism would enable the formation of a common language firm. The shares q_S and q_D of monolingual S and D speakers for whom the benefits of becoming bilingual outweigh the costs are given by:

$$q_S = \begin{cases} \frac{\lambda}{\bar{c}} m_D^1 & \text{if } m_D^1 < \frac{\bar{c}}{\lambda}, \text{ and} \\ 1 & \text{otherwise;} \end{cases} \qquad q_D = \begin{cases} \frac{\lambda}{\bar{c}} m_S^1 & \text{if } m_S^1 < \frac{\bar{c}}{\lambda}, \text{ and} \\ 1 & \text{otherwise.} \end{cases} \tag{A4}$$

The costs of becoming bilingual ought to outweigh the expected benefits for at least some mono-

linguals, so I assume that $\lambda < \frac{\bar{c}}{2}$.

Bilinguals must decide whether to pass one or both languages to their children. Let the shares of bilinguals that assimilate to become monolingual S and D speakers be a_S and a_D. Assimilating bilinguals will always have higher expected earnings in period 1 if they speak D because $p_D^1 > p_S^1$. Therefore, no bilingual will want its child to become a monolingual S speaker and $a_S = 0$. A bilingual will decide to make its child a monolingual D speaker if the expected additional return from being able to form a common language firm if matched with a monolingual S speaker in period 1 is less than the cost of transmitting S to the child: $\lambda m_S^1 \leq s_j$. This implies that:

$$a_D = \begin{cases} 1 - \frac{\lambda}{\bar{s}} m_S^1 & \text{if } m_S^1 < \frac{\bar{s}}{\lambda}, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{A5}$$

Workers make the investment decision at the end of period 0 anticipating the equilibrium share of workers that will be able to speak S and D in period 1. The period 1 population shares that speak S and D in turn depend on the decisions of the monolingual workers in period 0, given by

$$p_S^1 = p_S^0 + \tilde{q_D} m_D^0 \qquad p_D^1 = p_D^0 + q_S m_S^0, \tag{A6}$$

where $\tilde{q_D} = q_D - a_D(b^0/m_D^0)$. The net increase in the dominant language bilingual share $\tilde{q_D}$ is composed of monolingual D individuals whose children become bilingual less bilinguals whose children assimilate to become monolingual D speakers. The overall increase in bilingualism is $q = \tilde{q_D} m_D^0 + q_S m_S^0$.

I now solve for the shares of monolingual D and S speakers who become bilingual, $q_D(m_D^0, m_S^0, \lambda, \bar{c}, \bar{s})$ and $q_D(m_D^0, m_S^0, \lambda, \bar{c}, \bar{s})$, which will use to derive Results 1 and 2. Using the equation A1 and the definitions of p_D^t and p_S^t, equation A6 may be rewritten as

$$m_D^1 = m_D^0 - q_D m_D^0 + a_D(1 - m_S^0 - m_D^0) \qquad m_S^1 = m_S^0 + q_S m_S^0. \tag{A7}$$

Substituting in equations A4 and A5 and solving for q_D and q_S yields

$$q_D = \frac{\lambda \bar{s} m_S^0 (\bar{c} + \lambda(m_S^0 - 1))}{\left(\bar{c}^2 \bar{s} + \bar{c}\lambda^2 m_s^0 (m_d^0 + m_s^0 - 1) - \lambda^2 m_d^0 m_s^0 \bar{s}\right)} \tag{A8}$$

$$q_S = \frac{\lambda(\bar{s}(1 - m_s^0 - m_d^0 m_s^0) - \bar{c}(\lambda m_S^0 (1 - m_D^0 - m_S^0))}{\left(\bar{c}^2 \bar{s} + \bar{c}\lambda^2 m_s^0 (m_d^0 + m_s^0 - 1) - \lambda^2 m_d^0 m_s^0 \bar{s}\right)}. \tag{A9}$$

The main comparative statics of interest are the response of the equilibrium shares to changes in the expected extra return from forming a common-language firm $\lambda = \pi(w_M - w_A)$. I will now differentiate the equations for q_D and q_S with respect to λ and sign the derivatives. Define $\theta =$

$\left(\bar{c}^2\bar{s}+\bar{c}\lambda^2 m_s^0(m_d^0+m_s^0-1)-\lambda^2 m_d^0 m_s^0\bar{s}\right)^{-1}$.

$$\frac{\partial q_S}{\partial \lambda} = \theta^2\bar{c}m_s^0\bar{s}\left(\bar{c}^2\bar{s}-\bar{c}\lambda(\lambda m_s^0(m_d^0+m_s^0-1)-2(m_s^0-1)\bar{s})+\lambda^2 m_d^0 m_s^0\bar{s}\right) \tag{A10}$$

$$\begin{aligned}\frac{\partial q_D}{\partial \lambda} = {} & \theta^2\bar{c}\bar{s}(\bar{c}^2(2\lambda m_s^0(m_d^0+m_s^0-1)-m_s^0\bar{s}+\bar{s}) \\ & +\bar{c}\lambda m_s^0(\lambda(m_s^0-1)(m_d^0+m_s^0-1)-2m_d^0\bar{s}) \\ & -\lambda^2 m_d^0(m_s^0-1)m_s^0\bar{s})\end{aligned} \tag{A11}$$

Under the assumptions I have made about the parameters, both $\frac{\partial q_S}{\partial \lambda} > 0$ and $\frac{\partial q_D}{\partial \lambda} > 0$. Differentiating $\tilde{q_D} = q_D - a_D(b^0/m_D^0)$ with respect to λ using $\frac{\partial q_D}{\partial \lambda} > 0$ yields

$$\frac{\partial \tilde{q_D}}{\partial \lambda} = \left(1+\frac{b^0}{m_D^0}\frac{\bar{c}}{\bar{s}}\right)\frac{\partial q_D}{\partial \lambda} > 0. \tag{A12}$$

A comparison of the derivatives A10 and A11 shows that $\frac{\partial q_S}{\partial \lambda} > \frac{\partial \tilde{q_D}}{\partial \lambda}$.

This derivation establishes the first two results.

Result 1. *The increase in net bilingualism will be higher when industry is more prevalent and/or the wage gap is higher. In other words,* $\frac{\partial q}{\partial \lambda} > 0$.

Result 2. *The expected return to being in a common-language firm has a positive effect on net bilingualism for both dominant and secondary language speakers. The effect is larger for secondary than dominant languages.* $\frac{\partial q_S}{\partial \lambda} > \frac{\partial \tilde{q_D}}{\partial \lambda} > 0$

The effect is larger for S monolinguals because they form a smaller share of the initial population, making the additional return from being able to form a common-language firm higher than for D monolinguals. The expected cost of bilingualism is common to both groups.

The final result concerns the differential effect of industrial jobs by how linguistically heterogenous the economy is. Linguistic heterogeneity measures the likelihood any two randomly selected individuals have no language in common. In this two-language setting, linguistic heterogeneity is increasing in m_S^0 and decreasing in m_D^0 and b^0. In a more heterogeneous economy, bilingualism is more valuable to both D and S monolinguals.

Result 3. *The return to bilingualism will have a larger effect on bilingualism when linguistic heterogeneity is higher.*

I define linguistic heterogeneity as the probability that two randomly selected individuals in period 0 do not have a language in common.

$$h = 1-(m_D^0)^2-(m_S^0)^2-m_D^0 b^0-m_S^0 b^0. \tag{A13}$$

Because we have only two languages, h ranges from zero to $\frac{1}{2}$. Using the definition $m_D^0 + m_S^0 + b^0 = 1$, I rewrite h as

$$h = 1 - m_S^0 - m_D^0 + 2m_S^0 m_D^0. \tag{A14}$$

I solve equation A14 for m_S^0 and inserting it into $q = q_D m_D^0 + q_S m_S^0 - a_D b^0$. Making use of $m_D^0 + m_S^0 + b^0 = 1$ along with a rearrangement of equations A4 and A5 yields

$$q = m_D^0\, q_D(m_D^0, h) + \frac{m_D^0 + h - 1}{2m_D^0 - 1} q_S(m_D^0, h) - \left(1 - \frac{\bar{c}}{\bar{s}}\, q_D(m_D^0, h)\right)\left(1 - m_D^0 - \frac{m_D^0 + h - 1}{2m_D^0 - 1}\right) \tag{A15}$$

The cross partial derivative $\frac{\partial q}{\partial \lambda \partial h}$ shows how impact of the expected return to bilingualism on the total change in bilingualism is affected by linguistic heterogeneity. $\frac{\partial q}{\partial \lambda \partial h} > 0$ is true under the previously outlined parameter ranges if the number of initial bilinguals is less than 20%. As λ falls relative to $\bar{c}$ or $\bar{s}$ rises relative to $\bar{c}$, the inequality is true even if there are more initial bilinguals. In the Indian districts studied in the paper, bilinguals were initially about 8% of the population.

A2 Appendix Tables and Figures

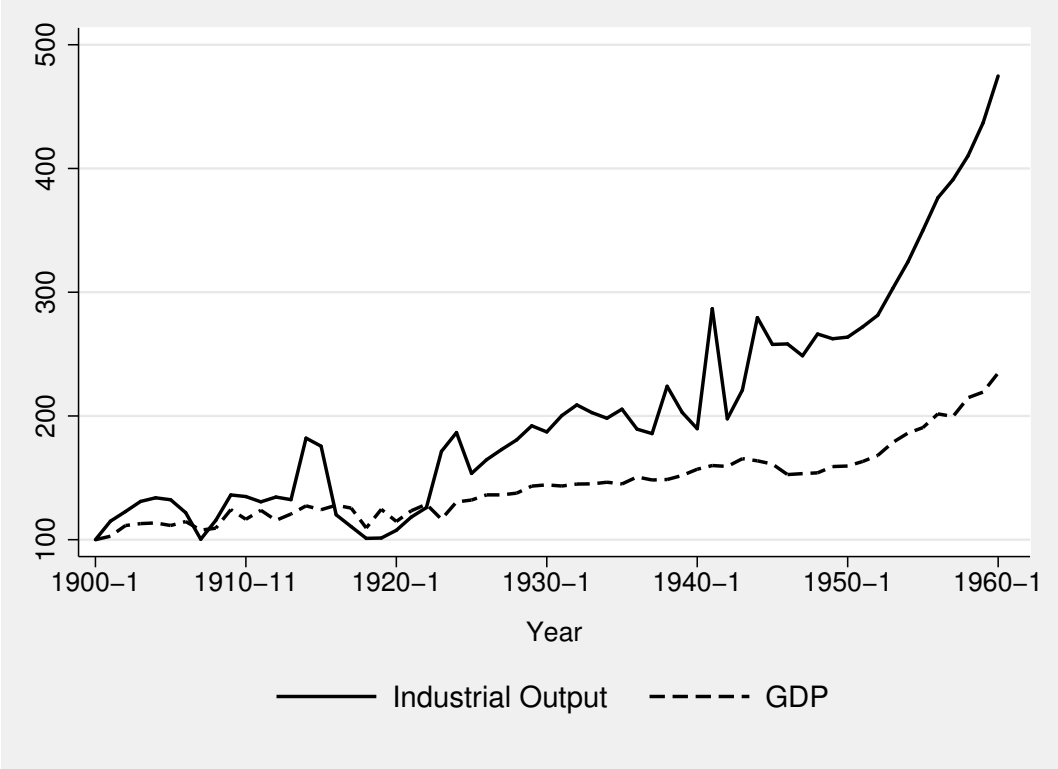
Table A-1: "Oriental" Factory Employees by Department and Mother Tongue†

		Percentage Whose Mother Tongue is			
	N	Gujarati	Marathi	Hindi	Other
Pattern Shop	26	87.5	12.5		
Foundry	135	44.5	26.5	29.0	
Machine Shop	159	80.0	15.0	4.0	1.0
Electric Shop	95	55.2	36.0	2.6	6.2
Switchgear Shop	79	70.0	28.5		1.5
Welding Shop	22	63.5	20.0	16.5	
Fitting Shop	25	68.0	20.0	8.0	4.0
Testing Shop	16	100.0			
Repair and Maintenance Shop	58	76.2	13.8	5.0	5.0
Packing and Painting Shop	28	84.7	10.3	5.0	
General Party	32	81.4	6.2	6.2	6.2
Inspection Dept.	5	100.0			
Personnel Dept.	47	57.5			
Chief Engineer and Assistants	4	50.0			50.0
Costing Dept.	4	100.0			
Stores Dept.	9	100.0			
Time and Methods Dept.	10	100.0			
Scheduling Dept.	6	93.8	6.2		
Designs Dept.	8	81.2	18.8		
Sales Dept.	24	63.0	33.0	4.0	4.0
Purchase and Office	24	62.5	25.0	12.5	12.5
OVERALL	810	63.6	19.6	10.4	6.4

†Variance in mother tongue is 91% within and 9% across departments.

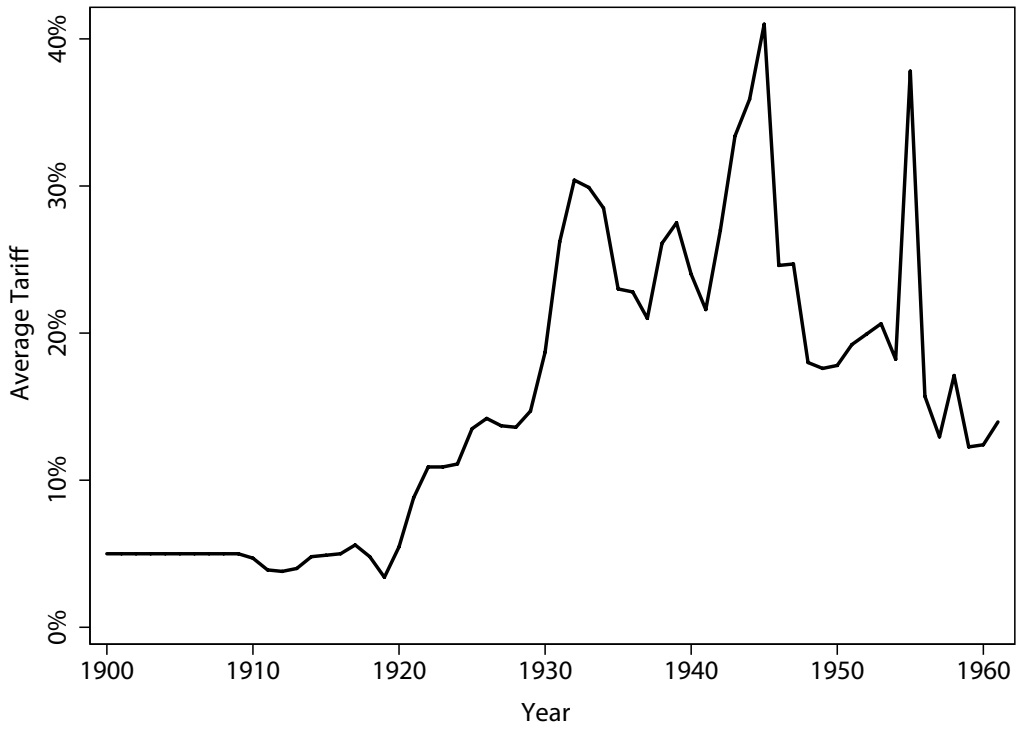
Notes: Data are from Sheth (1968)'s study of the "Oriental" electrical equipment factory in 1957. The factory was located in a small town in Gujarat.

Figure A-1: Real Industrial Output and GDP in India 1900–1961



Notes: Data from Sivasubramonian (2000). The series are deflated to 1938-9 prices. They represent undivided India from 1900-1947 and independent India from 1947-1960. The 1900-1947 series are indexed to 100 in 1900-1. The 1947-1961 series are indexed to the 1900-1947 series in 1947.

Figure A-2: Average Tariffs in India 1900—1960



Notes: This graph shows average tariffs in India based on the trade database developed by Blattman, Hwang, and Williamson for Blattman *et al.* (2007).

References

Blattman, Christopher, Hwang, Jason, & Williamson, Jeffrey G. 2007. Winners and losers in the commodity lottery: The impact of terms of trade growth and volatility in the Periphery 1870?1939. *Journal of Development Economics*, **82**(1), 156–179.

Lang, Kevin. 1986. A Language Theory of Discrimination. *Quarterly Journal of Economics*, **101**(2), 363–382.

Lazear, Edward. 1999. Culture and Language. *Journal of Political Economy*, **107**(6), S95–S126.

Lazear, Edward. 2005. *The Slow Assimilation of Mexicans in the United States*. Unpublished Ms.

Sheth, N.R. 1968. *Social Framework of an Indian Factory*. Manchester University Press.

Sivasubramonian, S. 2000. *The National Income of India in the Twentieth Century*. New Delhi: Oxford University Press.