# Language and Industrialization in Mid-20$^{\text{th}}$ Century India

David Clingingsmith[*]
Department of Economics
Case Western Reserve University

June 2008

**Abstract**

Bilingualism is an important form of human capital in linguistically diverse developing countries such as India, Indonesia, and Kenya. Expansion of economic activities in which communication is relatively important, such as manufacturing and services, provide an incentive to become bilingual, particularly for speakers of minority languages. I use a simple framework to illustrate the relationships between factory employment, bilingualism, and linguistic diversity. I then explore these relationships empirically using a new panel dataset of Indian districts for 1931 and 1961. Instrumental variables estimates show growth of manufacturing employment strongly encouraged bilingualism in mid-20$^{\text{th}}$ century India among minority-language speakers: an additional person became bilingual for every 2.1 manufacturing jobs added. The children of minority-language bilinguals may assimilate to the second language, producing a decline in linguistic heterogeneity. Economists have viewed linguistic heterogeneity as an exogenous determinant of poor economic outcomes, including low economic growth. I find that a one standard deviation increase in manufacturing employment decreased district-level linguistic heterogeneity by a third of a standard deviation, showing linguistic heterogeneity to be endogenous over the medium term.

[*] Email: david.clingingsmith@case.edu. I am grateful for discussions with Leah Platt Boustan, Carola Frydman, Claudia Goldin, Eric Hilt, Lakshmi Iyer, Asim Khwaja, Michael Kremer, Bob Margo, Simona Mkrtschjan, Rohini Pande, Heather Royer, and Jeffrey Williamson. Thanks for useful comments to participants at the 45th Cliometrics Conference, the 36th Annual Conference on South Asia, and seminar audiences at Harvard, UC Davis, University of Toronto, Vanderbilt, Reed College, Case Western, University of British Columbia, and the 2008 Canadian Network for Economic History conference. 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.

## 1 Introduction

Bilingualism is an important form of human capital in many developing countries, where a plethora of languages are often spoken and where many people are linguistic minorities in their local areas. Examples of such countries include India, Indonesia, the Philippines, Nigeria, and Kenya. Increasing gains from trade between individuals, such as those stemming from the expansion of markets and specialization in production, make it more valuable for individuals to be able to communicate with others. Individuals who share a common language face lower transaction costs when exploiting gains from trade with each other, making knowledge of widely-spoken languages advantageous. Development economists have given scant attention to investment in bilingualism, although the migration literature shows that ability in the receiving country's language earns returns in the labor market (Berman *et al.* 2003; Bleakley & Chin 2004; Chiswick & Miller 1995; Dustmann & van Soest 2001).

Bilingualism is further a necessary condition for intergenerational assimilation to a new mother tongue: At least one parent must be bilingual for a child to have a different mother tongue than its parents. Bilingualism is also of interest because it makes possible, though it does not ensure, changes in a population's linguistic composition and heterogeneity over time, referred to by sociolinguists as *language shift* (Fishman 1964; Gal 1978).

Economists have suggested a negative causal relationship between linguistic heterogeneity and economic growth across countries, the provision of public goods across jurisdictions, and trade volume (Alesina *et al.* 1999; Alesina & La Ferrara 2005; Anderson & van Wincoop 2004). This literature has taken linguistic heterogeneity to be exogenous. On the other hand, linguists have documented a long-run consolidation of the 10,000 language spoken in 1500, leading to the extinction of many and the convergence of 70% of the world's population to just 25 languages (Weber 1976; Hill 1978; Krauss 1992; Crystal 1997; Gordon 2005). They have attributed consolidation to industrialization, among other causes

This paper asks whether a structural shift in employment toward more communication intensive sectors, such as manufacturing, spurs individuals to become bilingual. I address this question by looking at the relationship between manufacturing expansion and the growth of bilingualism in India between 1931 and 1961. India is one of the world's most linguistically heterogenous developing countries. The paper attempts to both broaden our understanding of the economics of bilingualism and to show linguistic heterogeneity is endogenous to the process of economic development.

India is an excellent setting in which to study the relationship between manufacturing expansion, bilingualism, and linguistic heterogeneity. Even within small geographic regions such as districts, roughly a quarter of the population were linguistic minorities between 1931 and 1961. Bilingualism increased substantially during this era, particularly among linguistic minorities. There was also a major shifts in the structure of employment toward manufacturing.

I undertake an empirical analysis using a new panel dataset of Indian districts based on the Census of India for the years 1931 and 1961. The dataset includes observations on the six most common languages in each district in each year, allowing me to study the impact of district level changes on languages spoken within the district.

A simple model linking manufacturing jobs and bilingualism makes three main predictions that I explore using the data. First, bilingualism will be greater when manufacturing provides a larger share of jobs or when the wage gap between manufacturing and agriculture is higher. Under the assumption that individuals are more likely to be employed in factories if they can communicate with one another, then, given imperfect sorting, being able to communicate widely increases the chances of getting a factory job. Second, the incentive to become bilingual resulting from the chance to get a high-paying factory job will be larger for individuals whose mother tongue is a minority language. Bilingualism expands the communication potential of someone whose mother tongue is rare more than that of someone whose mother tongue is the majority language. Third, manufacturing growth, through increasing bilingualism, may lead to declining linguistic heterogeneity through assimilation of linguistic minorities.

Causal effects of growth in the district manufacturing share of employment on bilingualism and on linguistic heterogeneity is likely to be confounded by both simultaneity and omitted variables bias. My approach to identification includes language-by-district fixed effects to absorb fixed factors influencing bilingualism at the local level and develops an instrumental variable for the change in the manufacturing share. My instrumental variable is a prediction of what the manufacturing employment share in each district would have been in 1961 if each of its subindustries had grown at the average rate for the rest of the country between 1931 and 1961.

I find that expansion of the manufacturing share of the workforce has strong effects on bilingualism. My instrumental variables regression shows that a one-point increase in the manufacturing share of employment leads to a 1.3 point increase in the bilingual share for minority-language speakers and has no effect on bilingualism for majority-language speakers. These results are consistent with the model. Fixed effects do not change my estimates much, suggesting unobserved fixed factors are not important sources of bias. In absolute terms, my estimates imply that every 2.1 additional manufacturing jobs induces one minority language speaker to become bilingual. Manufacturing employment growth accounts for about 40% of the mean change in bilingualism among minority-language speakers. These are large effects, which may in part reflect the spillovers a new manufacturing job has on other industries and activities.

The district-level bilingual share for a given language might change because people learn a new language or because there is differential migration of bilinguals. In theory, either channel could increase assimilation and lead to language consolidation, though only learning can increase overall bilingualism and provide sustained expansion of communication potential. If migration is important, manufacturing expansion in district $j$ should have negative effects on bilingualism outside of $j$. I find no such effect on bilingualism in either districts adjacent to $j$ or in all other districts outside $j$.

My argument is not ultimately about manufacturing *per se* but about communication-intensive economic activity. My findings thus ought to hold whenever such activities increase, for example, when agricultural laborers increase their share of the agricultural workforce. Agricultural labor is more communication-intensive because these workers are frequently searching for a job. Finding a new job match will be easier, particularly for minority-language speakers, with knowledge of a second widely-spoken language. In fact, a 1-point increase in agricultural labor's share of agricultural employment increases bilingualism among minority-language speakers by 0.46 points.

Given that bilingualism is a precondition for assimilation of minority-language speakers across generations, manufacturing expansion could have a negative effect on linguistic heterogeneity. I measure district-level linguistic heterogeneity using a modified Herfindahl index. I find that a 1-point increase in the manufacturing share of employment decreases linguistic heterogeneity by 1.3 points. Manufacturing expansion reduced district-level linguistic heterogeneity by 5.9 points in the average district between 1931 and 1961. This sizable effect casts into doubt the assumption that linguistic heterogeneity is exogenous to economic performance.

My findings show that bilingualism and linguistic heterogeneity are profoundly affected by structural changes common in the process of economic development. It suggests the need for further micro-level contemporary studies to provide estimates of the returns to bilingualism in linguistically diverse developing countries. Language investment suffers from a network externality (Church & King 1993), suggesting language education could be a particularly fruitful area for policy intervention. My findings also suggest that economic development may be a powerful explanatory factor for the long-run consolidation of languages.

## 2 Mid 20th Century India

### Language in Indian Factories

An important foundation for my analysis is that workers in Indian factories mix across languages on the job, making bilingualism a valuable skill. Several factory-level studies done by Indian sociologists in the 1950s and 1960s support the idea that speakers of different mother tongues were members of the same work groups and found bilingualism advantageous. Sheth (1968) studied an electrical equipment factory in coastal Gujarat in 1957, and noted that functional units contained workers of several mother tongues and that workers were engaged in "continual interaction" in the course of their work. About 36% of workers were linguistic minorities. A similar pattern held at factories in Bombay, Poona, and south Bihar (Gokhale 1957; Lambert 1963; Vidyarthi 1970). Rice (1958) described interaction among workers in a textile factory in Ahmedabad as taking place in Gujarati, the locally dominant language, with most of the substantial fraction of workers who were mother tongue-Tamil or -Hindi speakers being bilingual.

Some further evidence that factory work rewards the ability to communicate comes from the United States. Around the peak of European mass migration the early 20th century, many industrial enterprises, such as International Harvester, United States Steel, and the Ford Motor Company established schools to teach workers English (Korman 1965). The Ford English School was most prominent among these. Non English-speaking workers were required to graduate from the school to receive incentive payments under Ford's Five Dollar Day profit-sharing scheme (Meyer 1980). By 1919, at least 800 plants offered their workers English classes on site, many in conjunction with the YMCA (Barrett 1992).

The key similarity between the United States and India that makes this comparison worthwhile is that both workforces were linguistically heterogenous and manufacturing in both cases produced specialized jobs, even if firm scale was much larger in the United States.

### Industrialization

The Indian economy showed few signs of advance in the half century before 1920. Between the first all-India census in 1872 and the census of 1921, manufacturing consistently provided jobs to about 10% of the workforce and about 10% of the population lived in cities. There was virtually no growth in per-capita real GDP (Sivasubramonian 2000).

A period of structural change began in the 1920s, at the beginning of the period I study. In the subset of districts in the Indian Union covered by my data, manufacturing employment grew at 2.7% annually between 1931 and 1961, expanding from 7.4% to 11.6% of the workforce[1](Table 1). India also became substantially more urban. In 1961, 19% of India's population lived in cities and towns, up from 12% in 1931.

Indian manufacturing enterprises increased substantially in scale between 1931 and 1961. Large factories, defined as those having more than 20 employees, provided 39.9% of all manufacturing jobs in 1961, more than double their 15.6% share in 1931 (Sivasubramonian 2000; India 1962). Historical studies have suggested that increased task specialization was a major reason for the increase in scale (Roy 1999, 2000). Goldin & Sokoloff (1992) provide quantitative evidence that during early industrialization in the United States, increases in firm size on the order of those we see in India were associated with greater specialization. The shift to larger work groups and greater task specialization increased the communication demands on workers. Labor productivity in large factories grew at a relatively brisk 2.1% annual rate between 1931 and 1947, while small factories actually saw a 1.5% annual decline in labor productivity (Sivasubramonian 2000).

India's external environment and trade policy were important factors driving the structural shift toward manufacturing. India's exports in the early 20th century were primarily agricultural commodities such as tea, wheat, flax, raw cotton, and raw jute. The terms of trade of these export commodities relative to manufactured imports began fall in the late teens (Appleyard 2006).

[1] While this might not seem dramatic by modern standards, it is similar to the 3.1% annual growth U.S. manufacturing employment had between 1849 and 1879 when the industrial revolution took hold (Carter *et al.* 2006).

This negative terms of trade shock favored Indian manufacturing at the expense of agriculture. Additionally, 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 were devolved to the provinces. Thereafter India's central government relied increasingly on import tariffs to raise revenue (Tomlinson 1979). Average import tariffs almost trebled from an average of 4.5% in the teens to 12.3% in the 1920s. Beginning in the mid 1920s, advocates for Indian industries successfully lobbied for protective tariffs. Average tariffs were 23.3% between 1931 and 1961, almost double the level of the 1920s.

### Languages

India is a linguistically diverse country with at least 180 distinct languages. It underwent steady language consolidation during the whole of the 20th century (see Figure 1). Although most of India's languages are concentrated in particular regions of the country, there is still substantial linguistic heterogeneity within small geographic units. The mother tongue of 23% of Indians in my data was a minority language in their district of residence in 1931, rising to 26% in 1961 (Table 1). The average district has two or three minority languages with substantial population shares (Figure 2a).

Between 1931 and 1961, the average bilingualism rate among minority-language mother tongue speakers increased from 28.2% to 43.8% (Table 1). Bilingualism was negatively correlated with the size ranking of a language in its district in 1931. Most of the growth in bilingualism between 1931 and 1961 happened among speakers of medium size minority languages ranked 2 through 4 (Figure 2b). Nearly 80% of minority language mother tongue speakers who were bilinguals chose the majority language of their district as their second language, while the remainder generally chose either English or Hindi. Bilingualism also increased substantially among speakers of local majority languages.

### Literacy and Education

Literacy and formal schooling are two additional forms of human capital investment that were important in India during the period I study. Literacy was expanding rapidly in mid-20th century India. About 24.0% of adults could read in 1961, up from just 9.5% in 1931. Most languages that enter my analysis have written forms, so desire for literacy *per se* probably did not generate substantial second language acquisition. In fact, it is possible that literacy and bilingualism are substitutes.

Bilingualism is related to formal schooling in a more complex way, since schooling may not be available in an individual's mother tongue. Part of the growth of bilingualism in India is likely related to expanded demand for schooling, some of which may be derived from manufacturing growth.

However, it does not seem likely that such a spillover effect is very large. While formal schooling in the vernacular languages of India and in English had been promoted since the 1850s, the primary school completion rate was very low even in 1961. Many children attended primary school for a year or two, perhaps long enough to attain a basic literacy, but few finished. The Census of India did not ask about schooling until 1941 (Srivastava 1972); in 1961, only 7.0% of the population had completed the three to four years that comprised primary school.

## 3 The Economics of Bilingualism and Language Shift

I will now 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 the manufacturing sector. This model is related to previous work by Lang (1986), who developed a language-based theory of discrimination, and Lazear (1999, 2005), who modeled the linguistic assimilation of immigrants.

Interaction with others allows individuals to take advantage of gains from trade of many different types, including working a specialized job in a large firm. Knowledge of a second language expands the network of potential interaction partners. Gains from trade can thus provide individuals with incentives to expand their communication network by becoming bilingual.

Bilingualism in turn enables intergenerational assimilation to a new mother tongue. Bilingual parents choose the language they use in interacting with the children, influencing which languages their children will learn. Parents may discourage or prevent a child's acquisition of a language they consider to be of low value or to be socially stigmatized.

Consider a two-period economy with two production sectors, manufacturing and agriculture.

The economy is populated by $N$ dynasties. Each dynasty 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 dynasty $j$: $U_j = c_j^0 + c_j^1$.

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: $m_D^t + m_S^t + b^t = 1$. 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 manufacturing sector makes use of a more productive technology than the agricultural sector. I take technology to be exogenous. The manufacturing technology requires workers to communicate to take advantage of its superior productivity. Manufacturing 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 manufacturing technology. Otherwise they are only capable of working in agriculture. Common-language firms get access to the manufacturing technology with the exogenous probability $\pi \epsilon (0, 1]$, which reflects how widespread manufacturing is in the economy. Workers in manufacturing each earn the return $w_M$. Workers in firms that do not share a common language or did not get access to the manufacturing technology must use the agricultural technology. Workers in agriculture each earn the return $w_A \leq w_M$. The expected period 0 income of a monolingual $D$ speaker is $p_D^0(\pi w_M + (1 - \pi)w_A) + (1 - p_D^0)w_A$. A parallel expression holds for monolingual $S$ speakers. Bilinguals always form a common-language firm and earn $\pi w_M + (1 - \pi)w_A$. 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, they give 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 $s_j \sim U[0, s]$. Monolingual period 0 workers may also invest in making their child bilingual by paying $c_j \sim U[0, c]$. The two costs reflect language learning ability and are independent. Workers can costlessly borrow against period 1 income 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 the expected income of his or her dynasty 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:

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

A parallel inequality holds for monolingual $D$ speakers. Define $\lambda = \pi(w_M - w_A)$ as the expected increase in return for a worker in a common language firm over a firm without a common language and recall that $m_D^1 = 1 - p_S^1$. Equation 1 can then be rewritten as

$$\lambda m_D^1 \geq c_j. \tag{2}$$

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:

$$q_S = \begin{cases} \frac{\lambda}{c} m_D^1 & \text{if } m_D^1 < \frac{c}{\lambda}, \text{ and} \\ 1 & \text{otherwise;} \end{cases} \qquad q_D = \begin{cases} \frac{\lambda}{c} m_S^1 & \text{if } m_S^1 < \frac{c}{\lambda}, \text{ and} \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

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 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:

$$a_D = \begin{cases} 1 - \frac{\lambda}{s} m_S^1 & \text{if } m_S^1 < \frac{s}{\lambda}, \text{ and} \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

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:

$$p_S^1 = p_S^0 + q_D m_D^0 - a_D b^0 \qquad p_D^1 = p_D^0 + q_S m_S^0. \tag{5}$$

I use equations 3 to 5 to solve for the equilibrium shares $q_S$ and $q_D$ in terms of the initial distribution of language ability. I assume that $q_S, q_D, a_D < 1$. Define $\theta = \frac{\lambda}{c}(1 - \frac{\lambda^2}{c^2} m_D^0 m_S^0 - \frac{\lambda^2}{cs} m_S^0 b^0)^{-1}$. Then

$$q_S = \theta\Big(m_D^0\Big(1 - \frac{\lambda}{c} m_S^0\Big) + b^0\Big(1 - \frac{\lambda}{s} m_S^0\Big)\Big) \quad \text{and} \quad q_D = \theta\Big(m_S^0\Big(1 - \frac{\lambda}{c}(m_D^0 + b^0)\Big)\Big). \tag{6}$$

The shares will be strictly between 0 and 1 as long as the expected wage in a common-language firm is not too large relative to the cost of becoming bilingual or transmitting two languages. I assume that $\lambda$ is sufficiently less than $c$ and $s$ for this to be the case.

### Wages, Manufacturing Prevalence, and Language Learning

Two results flow from equation 6 that link bilingualism to the expected return to being in a common-language firm. These results motivate the empirical analysis the begins in section 4.

**Result 1** *A larger share monolinguals will become bilingual when the expected return to being in a common-language firm is greater.*

This follows from differentiating $q_S$ and $q_D$ with respect to $\lambda$. Under the assumptions that $m_S^0 < \frac{1}{2}$, $\lambda < c$ and $\lambda < s$, we have $\frac{\partial q_S}{\partial \lambda} > 0$. The sign of $\frac{\partial q_D}{\partial \lambda}$ depends on parameters. If $m_D^0 + b^0 < \frac{c}{2\lambda}$, then $\frac{\partial q_D}{\partial \lambda} > 0$.

Note that since $\lambda = \pi(w_M - w_A)$, an increase in either the prevalence of manufacturing or the wage gap increases the expected return to being in a common-language firm.

**Result 2** *The incentive to become bilingual generated by the expected return to being in a common-language firm is larger for the minority $S$ speakers than the majority $D$ speakers.*

Differentiating $q_S$ and $q_D$ as before and using the assumption that $p_S^0 < p_D^0$ gives the result that $\frac{\partial q_S}{\partial \lambda} > \frac{\partial q_D}{\partial \lambda}$. Intuitively, because $D$ speakers are a larger share of the population, the additional return from being able to form a common-language firm with certainty is lower, while the cost $c_j$ of becoming bilingual is fixed. $D$ speakers will thus have a lower communication-based incentive to learn $S$ than $S$ speakers will have to learn $D$.

The model also generates a main result from Lazear's model of linguistic assimilation (Lazear 1999, 2005).

**Result 3** *A larger share of monolingual speakers will become bilingual when they are a smaller share of the population.*

This follows directly from differentiating $q_S$ and $q_D$ with respect to the initial monolingual population shares: $\frac{\partial q_S}{\partial m_S^0} < 0$ and $\frac{\partial q_D}{\partial m_D^0} < 0$.

### Bilingualism, Assimilation, and Language Shift

Language shift within a lineage results when a bilingual parent transmits only their second language to their children. The share of the population knowing $S$ will decline if the number of bilinguals in period 0 who are better off assimilating is greater than the number of $D$ monolinguals who find it worthwhile to learn $S$.

$$\Delta p_S = q_D m_D^0 - a_D b^0 \tag{7}$$

Whether $\Delta p_S$ is positive or negative depends on the parameters.

**Result 4** *Bilingualism among speakers of $S$ may lead to a decline over time in the share of the population speaking $S$.*

Bilingual dynasties may remain bilingual, even if one of their languages has very few or no monolingual speakers and therefore low communication value. There are examples of stable bilingualism of this type, such as in Wales, where the second language has a high cultural value.

Result 4 implies that the growth of manufacturing employment may lead to a decline in linguistic heterogeneity, measured as $h = 1 - \sum_\ell s_\ell^2$, where $s_\ell^2$ is the population share of speakers whose mother tongue is $\ell$.

## 4 Data and Methodology

### Panel Data for Indian Districts

I constructed a panel dataset of Indian districts for the years 1931 and 1961 from the Census of India to test the intuitions developed in the model (India 1933, 1962). The dataset contains information at two levels of aggregation, the district level and the district-by-language level. District-level measures include population, employment, occupational structure, literacy, and urbanization.

I assembled data in this form for all districts in 1931 and 1961. I first selected the six most commonly spoken languages in 1931. I collected the number of speakers and bilinguals of these languages for both 1931 and 1961. I also collected information on employment, urbanization, and literacy for each district and year. In each year, each district has six language-level observations and one observation of employment and other district-level characteristics. The dataset contains 153 districts and covers all of present-day India excluding Uttar Pradesh, Punjab, Himachal Pradesh, and Rajasthan. See Appendix A for further details on how the dataset was constructed.

### Econometric Specification

The model in section 3 links expansion of manufacturing employment to increased bilingualism. It further suggests that this effect should be stronger for district minority languages than majority languages. I test these predictions by estimating an econometric model that measures the effect of a change in the manufacturing share of employment over time on the share of speakers of a language who are bilingual.

$$b_{\ell dt} = \beta_0 + \beta_{m61}(m_{dt} \cdot I_t^{61}) + \beta_m m_{dt} + \beta_{61} I_t^{61} + X'_{\ell dt}\theta + \mu_{\ell d} + \varepsilon_{\ell dt}. \tag{8}$$

The dependent variable $b_{\ell dt}$ is the share of mother tongue speakers of $\ell$ in district $d$ that are bilingual in year $t$. The manufacturing share of employment is $m_{dt}$. The indicator $I_t^{61}$ takes on the value 1 in 1961. The vector $X'_{\ell dt}$ contains controls for the 1931 levels of urbanization, literacy, the workforce participation rate, and the share of the population speaking each language. Fixed effects $\mu_{\ell d}$ allow each language in each district a separate intercept. Unobserved determinants of bilingualism are contained in the residual $\varepsilon_{\ell dt}$. In estimating the model, I cluster standard errors at the district level. I also weight each observation by the number of speakers of language $\ell$ in district $d$ in 1931. This allows me to interpret the coefficients as effects on an average individual.

Our coefficient of interest is $\beta_{m61}$, which measures how the change in the manufacturing share over time affects bilingualism. I estimate equation 8 separately for majority and minority languages, and will refer to the coefficients of interest from these regressions as $\beta^{maj}_{m61}$ and $\beta^{min}_{m61}$. These coefficients are identical to those produced by pooling all the languages together and fully interacting the independent variables with a dummy for minority language: I also present estimates of the differential effect of manufacturing growth on minority languages $\beta^{diff}_{m}$ using this method. Note that there is only one majority language per district, so those estimations will include one fixed effect per district.

The model predicts that manufacturing employment will have a positive effect on bilingualism that will be greater for speakers of minority languages, suggesting the coefficients $\beta^{maj}_{m}$ and $\beta^{min}_{m}$ will be positive and that $\beta^{min}_{m} > \beta^{maj}_{m}$.

### Identification Strategy

I take two steps to mitigate bias in estimating equation 8. First, I include language-by-district fixed effects $\mu_{\ell d}$, which I discuss next. Second, I develop an instrumental variable for the interaction $m_{dt} \cdot I^{61}_{t}$.

#### *Fixed Effects*

Estimates of the impact of manufacturing employment growth on bilingualism will be biased to the extent that fixed district characteristics that promote bilingualism are correlated with manufacturing growth. Coastal districts or those with navigable rivers may promote both interactions among individuals that leads to second-language learning and to more manufacturing through better market access. Similarly, other topographical features such as hills and mountains may hinder both bilingualism and manufacturing.

Further, the ease of learning various second languages depends on how different they are from an individual's mother tongue. We would expect to find fewer members of a linguistic minority becoming bilingualism when there is a great linguistic difference between their language and the majority language. This could introduce an additional source of bias at the language-district level. For example, districts that have been recipients of long-distance migrants over time will likely have minority languages more different from the majority language than districts that have not. These districts may also be more attractive to manufacturing for the same reasons that make them attractive to migrants, such as large cities or a density of transport routes.

Inclusion of the language-district fixed effect $\mu_{\ell d}$ in the estimation of equation 8 means that identification of the coefficient $\beta_{m61}$ comes from variation in the level of bilingualism over time for each language $\ell$ in each district $d$. This fixed effects strategy eliminates sources of bias that are invariant at either the district or district-language level.

Differential economic development across districts presents another threat to identification. It is likely to be correlated with both the expansion of manufacturing and other determinants of bilingualism, such as urbanization and increased education. Since many of these other determinants and manufacturing growth are endogenous, I could control for them even if adequate data were available.

A partial solution to this concern is to estimate a variant of equation 8 that pools minority and majority languages together and includes a district-specific trend in bilingualism $\upsilon_{dt}$ in the specification. This specification allows me to estimate the differential effect of manufacturing growth on bilingualism for minority-language speakers $\beta_m^{diff}$, which result 2 from the model suggests ought to be positive.

District-level variables are absorbed by the trend and are not included separately in the specification:

$$b_{\ell dt} = \beta_0 + \beta_m^{diff}(m_{dt} \cdot I_{\ell d}^{min} \cdot I_t^{61}) + X'_{\ell dt}\theta + \upsilon_{dt} + \mu_{\ell d} + \varepsilon_{\ell dt}. \tag{9}$$

Identification in this specification comes through changes in bilingualism within district-specific language groups relative to overall district changes in bilingualism. Time varying factors that affect *overall* bilingualism at the district level are absorbed by the district trend.

Even estimates with a district-specific trend could still be subject to bias if factors such as urbanization and increasing literacy and education affect language groups differentially. I next outline an instrumental variables approach that provides plausibly exogenous variation in the expansion of manufacturing to identify its effect on bilingualism.

*Instrumental Variable Based on the Manufacturing Mix*

I construct an instrumental variable for the district-level manufacturing employment share in 1961 to mitigate simultaneity and omitted variables bias. The instrument is a predicted value for this share based on national-level employment demand shocks to 10 subindustries within manufacturing. These demand shocks are weighted by the 1931 district-level mix of subindustries within manufacturing, producing an estimate of what the district level manufacturing share would have been in 1961 had all district-level subindustries expanded or contracted their employment shares at the national rate. The instrument therefore isolates component of district-level manufacturing employment growth that resulted from national-level variation in employment demand by subindustry. The approach is related to the shift-share decomposition used by Bartik (1991) and Blanchard & Katz (1992) and to explore the effect of employment shocks on city- and state-level economic variables in the United States. Other empirical analyses using similar approaches include Luttmer (2005), Autor & Duggan (2003), and Bound & Holzer (2000).

I collected data on 1931 manufacturing employment in textiles, wood, metals, ceramics, chemicals, apparel, food processing, vehicles, power, and other manufacturing industries. To create the instrument, I estimate regressions of the 1961 manufacturing share of the workforce $m_{d61}$ on the initial subindustry workforce shares $y_{jd31}$.

$$m_{d61} = \sum_j \mu_j y_{jd31} + \zeta_{jd}. \tag{10}$$

The coefficients $\mu_j$ measure the national average growth rate of employment in industry $j$ relative to the growth rate of overall employment. I use the coefficients from this regression and the district-level subindustry shares for district $d$ to make a predicted share $\widehat{m_{d61}}$ that forms the instrument. See Appendix B for a derivation of equation 10.

The variable $\widehat{m_{d61}}$ will be a valid instrument for both the initial subindustry shares $y_{id}$ and the subindustry growth coefficients $\mu_j$ are uncorrelated with the unobserved determinants of the change in bilingualism. This condition will be satisfied for the $\mu_j$ as long as industries are not concentrated in a small number of districts, in which case national demand growth in the industry would be related to the characteristics of a particular district. Industries are not concentrated, but to be safe I estimate 10 on the districts $\neg d$ and form the instrument as an out-of-sample predicted value.

There are scenarios under which the initial share of workers in a particular subindustry $y_{id}$ could be related to unobserved determinants of bilingualism, leading the exclusion restriction to be violated. The following example illustrates a potential criticism. Suppose subindustries differ in the degree to which they offer year-round employment. Year-round employment may make it more likely a worker will move to the city with his or her family. National employment demand growth in that subindustry may induce relatively more rural to urban migration of complete families in a district that has a large initial employment share in that subindustry. The non-manufacturing workers in those families may be positively selected for bilingualism.

It seems unlikely such a scenario would result in a large effect over the 30-year span of the panel. Further, in the estimations where I include a district-specific trend, this channel would have to affect majority and minority-language speakers differentially to violate the exclusion restriction.

A sample estimation of 10 that includes all districts is shown in Table 4. The estimation has an $R^2$ of 0.91, showing that the initial industrial shares are very good at predicting where manufacturing will expand or contract. Industries that are expanding have a coefficient greater than 1.

I estimate the first stage of an IV estimation of equation 8 by replacing the bilingual share $b_{\ell d}$ with the endogenous variable of interest $m_{dt} \cdot I_t^{61}$ and by including the instruments $\widehat{m_{d61}}$ as regressors. The instrument strongly predicts the district-level manufacturing share in 1961 (Table 5). An F-test of the excluded instrument is 117, well above the critical values that would indicate a weak instrument (Staiger & Stock 1997; Stock & Yogo 2002). Since the instrument is strong as shown below, the bias induced by a violation of the exclusion restriction is likely to be small.

## 5 Main Results

OLS estimation of equations 8 and 9 suggests that an increase in the manufacturing share of employment is correlated with the growth of bilingualism for speakers of minority languages, but not for speakers of majority languages. Table 3 presents the regressions in adjacent panels for minority and majority languages.

My results for minority languages show that a 1-point increase in the manufacturing share over time leads to a 1-point increase in the bilingual share (columns 1-3). This implies a large absolute effect: assuming a stable workforce, one additional minority language speaker becomes bilingual for every 2.7 additional manufacturing jobs added to a district.

Addition of district and district-by-language fixed effects in columns 2 and 3 do not affect the coefficient of interest. This suggests that unobserved time-invariant determinants of bilingualism are not an important source of bias. Interestingly, the cross-sectional correlation between the manufacturing employment share and bilingualism is negative for minority languages and becomes more so when we introduce fixed effects.

Manufacturing expansion has only a small positive effect on bilingualism for majority-language speakers (columns 4 and 5). The effect is not statistically significant. While the formal prediction of the model was that this effect would be positive, it is not surprising that it is small. The majority language is likely the focal language when different mother-tongue groups need to communicate, and hence the burden of learning falls to the minority mother-tongue speakers. About 70% of bilingual minority-language speakers have learned the majority language in their district.

Estimates of the differential effect of bilingualism on minority-language speakers $\beta_m^{diff}$ are presented at the bottom of Table 3. The specification in equation 9 with district trend effects, which allow for common district-level shocks to bilingualism, shows a somewhat smaller coefficient estimate than the specification with only language-by-district fixed effects, though the difference is not economically significant. This suggests district-level shocks produce a net upward bias on estimates in the main specification, though it can't be said whether this comes through the level effect on minority or majority languages.

There are two main channels through which the effect of manufacturing expansion might operate. It may lead to second-language learning among speakers within the district, or it may induce a differential migration into the district of bilinguals. I will present evidence later that suggests the learning channel is more important. If the aspect of manufacturing that leads to increased bilingualism is communication intensity, as I have suggested, other sectoral shifts should have effects on bilingualism consistent with the communication intensity of the sector. I will also present evidence showing that this is true for two additional sectors.

While the stability of the estimates under different fixed- and trend- effects specifications increases confidence that the correlation between manufacturing and bilingualism is not driven by some important unobservables, several concerns remain. Simultaneity bias can result from the effect of language on economic outcomes, which is discussed in the literatures on returns to language knowledge and on ethnolinguistic fractionalization. Bilingualism makes communication easier, and therefore districts with a growing bilingual share may attract more manufacturing firms.

Several time-varying aspects of economic development are also simultaneously determined with the manufacturing share of the workforce. Of these, urbanization is probably the source of omitted variables bias of greatest concern in OLS estimates. People interact with a larger set of individuals in an urban environment, which probably leads to a diffusion of languages. Manufacturing became an increasingly urban activity in the years covered by this study. Education and literacy as also likely to be correlated with manufacturing growth. Since literacy is available in many languages and since primary completion rates are so low,

Instrumental variables estimates of the main specification are very similar to the OLS (Table 6). A 1-point change in the manufacturing share of employment leads to a 1.4-point increase in bilingualism among minority-language speakers (p-value $< 0.01$), somewhat larger than the OLS estimate. There is no increase in bilingualism among majority-language speakers. The differential effect of expansion of the manufacturing share on minority-language speakers is 1.1 points. The absolute effect of manufacturing employment on minority speaker bilingualism remains big; 2.1 additional manufacturing jobs in a district results in one additional bilingual among minority-language speakers.

How much of the changes in bilingualism between 1931 and 1961 can my instrumental variables estimates explain? Bilingualism increased by an average 15.6 percentage points among minority-language speakers and an average 5.1 percentage points among majority-language speakers. The population-weighted average manufacturing share of employment increased by 4.2 points between

1931 and 1961 (Table 1). These estimates suggest that manufacturing employment growth accounts for about 40% of the the increase in bilingualism among minority-language speakers and essentially none of the growth among majority-language speakers.

### Learning versus Migration

The growth of manufacturing employment could influence the bilingual share of speakers through three main channels: learning, migration, and fertility. The most interesting of these is learning, because it is most directly related to language consolidation. It increases the size of the group at risk for switching mother tongues. Differential migration does not, though it is nonetheless consistent with the notion that increased communication intensity is the mechanism through which manufacturing growth affects bilingualism. Greater fertility among bilinguals could also increase the population at risk, but seems more likely to be negatively correlated with manufacturing.

I cannot directly test the relative importance of learning and migration as channels through which manufacturing growth increases the bilingual share: Tabulations of the share of language groups and bilinguals by migration status are not available from the Census of India. However, to the extent that manufacturing growth attracts bilinguals to move in to district $d$, we should see a decline in the bilingual share in an adjoining district for a language common to both it and $d$. New employment opportunities should be more attractive to those who must move a short distance.

Table 7 presents instrumental variables estimates of the effect of growth of the manufacturing share in district $d$ on individuals who live outside $d$ but have a mother tongue that is spoken in $d$. All majority languages and virtually all minority languages are also spoken in adjacent districts. Manufacturing growth has no statistically significant effect on bilingualism outside of district $d$, either in adjacent districts or in the rest of India (covered by the data). The only sizable point estimate is for minority languages in adjacent districts is actually positive, though imprecisely estimated.

This test suggests that manufacturing employment growth increases bilingualism through changes within the district in which it occurs, not through spillovers on other districts. While it does not directly prove than language learning is the primary channel, it provides strong support.

### Sectoral Shifts within Agriculture

The general argument motivating my empirical analysis is that sectors of the economy differ in their communication intensity, and that the sectoral shifts would therefore change the demand for communication ability, giving those with a low level of ability, perhaps because their mother tongue is a minority language, an incentive to become bilingual.

Earlier I drew a contrast in terms of communication intensity between manufacturing and cultivation. I argued that the specialization and scale that generate productivity gains in manufacturing makes that sector more communication intensive than cultivation. While communication intensity is difficult to measure, particularly in this context, if other sectoral shifts induce changes in bilingualism consistent it will strengthen the argument in favor of it.

Most agricultural workers in India are owner-operators or tenants. The remainder of the agricultural sector is composed of temporary laborers. Owner-operator and tenant agriculture is typically practiced in small family units. The small size and recruitment among related individuals both mean that the sector is less communication intensive and that matching of workers by mother tongue will be much easier. Larger farms may hire in temporary labor through the labor market, in which matching by mother tongue will be more difficult. Agricultural labor ought to thus be more communication intensive than owner/tenant cultivation.

I collected data on the share of the total agricultural workforce that is agricultural laborers. As my argument about communication intensity suggests, I find the owner/tenant share of employment indeed has a strong and significant negative effect of $-0.32$ points on minority language bilingualism (Table 8, column 1). There is no effect on majority language bilingualism, as we saw with manufacturing growth (column 2). Including district trends shows the differential effect on minority languages to be 0.36 points.

### Linguistic Heterogeneity

The evidence presented so far suggests that growth in the manufacturing share of employment leads to minority-language speakers to learn a second language. Bilingualism is a necessary though not sufficient condition for language consolidation. Further, it is possible that bilingualism resulting specifically from manufacturing employment growth does not lead to language consolidation. For example, it could be that manufacturing employment encourages minority speakers to obtain too shallow a knowledge of a second to allow them to make it the sole language of the next generation.

I measure language consolidation within a region as a decline in linguistic heterogeneity.[2] If large languages grow in relative size at the expense of smaller ones, linguistic heterogeneity will fall. Linguistic heterogeneity is also of independent interest as it has been associated with poor economic performance.

Average linguistic heterogeneity declined between 1931 and 1961 in the area covered by the panel, falling from 0.87 to 0.84. Average district-level linguistic heterogeneity actually increased from 0.30 to 0.35, meaning that languages were becoming less geographically concentrated even as they were consolidating at the national level.

I estimate the effect of expansion in the manufacturing employment share using a version of equation 8 in which there are only district-level regressors. IV estimates show a 1-point increase in the manufacturing share of employment leads to a statistically significant 1.29 point decrease in district-level linguistic heterogeneity (Table 9, column 3). IV estimates are close to the OLS. District-level linguistic heterogeneity increased in India between 1931 and 1961 from 0.30 to 0.35 (Table 1). Manufacturing employment growth actually slowed this trend; in its absence mean linguistic heterogeneity would have climbed to 0.41 by 1961.

It is difficult to form a precise comparison between my district-level results and estimates from the literature of the economic effects of linguistic heterogeneity. The economic variables and the units across which heterogeneity is measured differ. The channels through which linguistic diversity affects economic outcomes may be different at the country and local levels. My estimates imply that the average district-level increase in the manufacturing employment share of 4.2 percentage points leads to a 0.25 standard deviation decrease in linguistic heterogeneity. The best cross-country estimate from Alesina & La Ferrara (2005) is that a one standard deviation decrease in linguistic heterogeneity leads to 0.6% higher annual per-capita GDP growth. Average annual GDP growth in India was only about 0.2% between 1931 and 1961 (Sivasubramonian 2000). Taken at face value, these estimates suggest that the causal channel running from economic development, in the form of the expansion of manufacturing, is relatively large compared to existing estimates of he effect in

[2] Linguistic heterogeneity in region $r$ is $1 - \sum_{\ell \in d} s_{\ell d}^2$, where $s_{\ell d}$ is the share of the population speaking $\ell$ in region $r$.

the other direction.

## 6 Conclusion

I have shown that manufacturing employment growth leads to growing bilingualism among speakers of local minority languages, and that learning, rather than sorting, is the likely mechanism.

Language investment in theory suffers from a network externality (Church & King 1993), suggesting an important role for policy intervention. This externality will be positive when someone decides to become bilingual and negative when a bilingual decides to raise its children as monolinguals. Developing countries' linguistic transitions are likely to proceed more slowly than would be optimal, as those who pay the cost of bilingualism create benefits for other speakers of the languages they learn. Assimilation has the opposite effect; a parent's decision to abandon a language reduces the size of the language in the next generation and negatively affects the remaining speakers. Assimilation decisions may also lead to the loss of important cultural resources.

Further empirical work is needed to both estimate the returns to bilingualism in linguistically diverse developing countries and to understand whether there are economically important network externalities in language ability. If these externalities exist and are substantial, there may be a significant role for government in transferring resources to minority-language speakers to encourage bilingualism and to ease the negative effects of assimilation on those who remain monolingual linguistic minorities.

I have also found that manufacturing employment growth discourages linguistic heterogeneity. This implies that the negative impact of linguistic heterogeneity on economic growth and public goods measured in the literature will be confounded by the process of economic development itself. My results suggest that linguistic diversity ought to be taken as endogenous to the process of economic development.

The rise of new mass communication technologies, such as the Internet, the emergence of India and China as rapid industrializers, and the increasingly global character of production suggests the economic importance of common language will continues to grow. Linguists' prediction of continued consolidation of languages worldwide appear be well founded. In particular, India has seen rapid growth of manufacturing and services in recent decades, and it would not be surprising to see India's stock of languages continue to decline over the next century as the dozen or so major languages gain mother tongue speakers.

## References

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.

Anderson, James, & van Wincoop, Eric. 2004. Trade Costs. *Journal of Economic Literature*, **XLII**, 691–751.

Appleyard, Dennis R. 2006. The Terms of Trade between the United Kingdom and British India, 1858-1947. *Economic Development and Cultural Change*, **54**(3), 635–655.

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

Barrett, James R. 1992. Americanization from the Bottom Up: Immigration and the Remaking of the Working Class in the United States, 1880-1930. *The Journal of American History*, **79**(3), 996–1020.

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.

Bound, John, & Holzer, Harry. 2000. Demand Shifts, Population Adjustments, and Labor Market Outcomes during the 1980s. *Journal of Labor Economics*, **XVIII**, 2054.

Carter, Susan B., Gartner, Scott Sigmund, Haines, Michael R., Olmstead, Alan L., Sutch, Richard, & Wright, Gavin (eds). 2006. *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*. Cambridge University Press.

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

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

Crystal, David. 1997. *Language Death*. West Nyack, NY: Cambridge University Press.

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

Fishman, Joshua. 1964. Language Maintenance and Language Shift as a Field of Inquiry. *Linguistics*, **9**, 32–70.

Gal, Susan. 1978. *Language Shift: Social Determinants of Language Change in Bilingual Austria.* New York: Academic Press.

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

Goldin, Claudia, & Sokoloff, Kenneth. 1992. Women, Children, and Industrialization in the Early Republic: Evidence from the Manufacturing Censuses. *The Journal of Economic History*, **42**(4), 741–774.

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

Hill, Jane H. 1978. Language Death, Language Contact, and Language Evolution. *In:* McCormack, William C., & Wurm, Stephen A. (eds), *Approaches to Language: Anthropological Issues.* The Hague: Mouton.

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

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

Korman, Gerd. 1965. Americanization at the Factory Gate. *Industrial and Labor Relations Review*, **18**(3), 396–419.

Krauss, Michael. 1992. The World's Languages in Crisis. *Language*, **68**(1), 4–10.

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

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.

Luttmer, Erzo F. P. 2005. Neighbors as Negatives: Relative Earnings and Well-Being. Quarterly Journal Of Economics. *Quarterly Journal Of Economics*, **120**(5), 2054.

Meyer, Stephen. 1980. Deindustrialization, Industrialization, and the Indian Economy, c. 1850-1947. *Journal of Social History*, **14**(1), 67–82.

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.

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.

Staiger, D., & Stock, J.H. 1997. Instrumental Variable Regression with Weak Instruments. *Econometrica*, **65**(3), 557–586.

Stock, James H., & Yogo, Motohiro. 2002. *Testing for Weak Instruments in Linear IV Regression.* NBER Working Paper No. T0284.

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

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.

## Appendix

## A Data

I constructed a panel dataset of Indian districts for the years 1931 and 1961 using the published volumes for the Census of India (India 1933, 1962). Many district boundaries changed following India's independence from Britain in 1947, when hundreds of sovereign princely states were integrated into the colonial administrative framework inherited by India. Some British districts were also combined or split up. The census does not always contain sufficient detail in 1961 to use the 1931 district definitions: In constructing the dataset I aggregated geographical units as necessary to form exactly comparable districts based on the equivalence table found in Singh & Banthia (2004). The aggregation produced 244 comparable districts, compared with 339 administrative districts in 1961 and 439 administrative districts, princely states, and territories in 1931. Complete bilingualism data was only collected for 153 of these aggregate districts in 1931. The dataset accordingly excludes the northern states of Uttar Pradesh, Rajasthan, Punjab and Himachal Pradesh. For each district and year, I compiled characteristics of the six languages most commonly spoken in 1931. I thus have six language-level observations per district per year, and one observation of employment and other district characteristics per district per year.

## B Construction of the Instrument

The instrument for my primary explanatory variable $m_{d61}$, the manufacturing employment share in 1961 in district $d$, is constructed by a decomposition.

Let the level of manufacturing employment in district $d$ be $M_d$, the level of employment in manufacturing subindustry $j$ be $M_{jd}$, and total employment be $E_d$. We can then express $m_{d61}$ in terms of initial levels and growth rates:

$$m_{d61} = \frac{g_{Md}M_{d31}}{g_{Ed}E_{d31}} = \frac{\sum_j (g_{Ydj}Y_{dj31})}{g_{Ed}E_{d31}}. \tag{11}$$

Now combine the growth rates of employment to form the relative growth rate of manufacturing subindustry $j$ to overall employment for district $d$, $\mu_{jd} = \frac{g_{Yjd}}{g_{Ed}}$, which allows us to write

$$m_{d61} = \sum_j \mu_{jd}\frac{Y_{dj31}}{E_{d31}} = \sum_j \mu_{jd}y_{jd31}. \tag{12}$$

The relative growth rate $\mu_{jd}$ can be decomposed into a subindustry average $\mu_j$ and a district-level deviation from the average $\tilde{\mu}_{jd}$, so that $\mu_{jd} = \mu_j + \tilde{\mu}_{jd}$. The manufacturing employment share in 1961 can then be decomposed into a component due to the average relative growth of subindustry shares and a component due to district deviations from the subindustry average.

$$m_{d61} = \sum_j \mu_j y_{jd31} + \sum_j \tilde{\mu}_{jd}y_{jd31}. \tag{13}$$

The first term in equation 13 can then be estimated as the predicted value $\widehat{m}_{d61}$ from the regression

$$m_{d61} = \sum_j \mu_j y_{jd31} + \zeta_{jd}, \tag{14}$$

which forms the instrument for the manufacturing employment share in 1961.

Table 1: Population-Weighted District-Level Summary Characteristics

| | 1931 | 1961 | Change<br>1931–1961 |
|---|---|---|---|
| Employment Rate | 0.461 | 0.439 | -0.022 |
| Manufacturing Share of Emp. | 0.074 | 0.116 | 0.042 |
| Urban Share | 0.118 | 0.193 | 0.075 |
| | | | |
| Bilingual Share of Population | 0.069 | 0.096 | 0.027 |
| Linguistic Heterogeneity | 0.299 | 0.347 | 0.048 |
| Literate Share of Population | 0.069 | 0.261 | 0.192 |
| | | | |
| *Mother Tongue* | | | |
| *Speakers of Majority Language* | | | |
| Population Share | 0.773 | 0.739 | -0.034 |
| Share Bilingual | 0.015 | 0.067 | 0.052 |
| | | | |
| *Mother Tongue* | | | |
| *Speakers of Minority Languages* | | | |
| Population Share | 0.227 | 0.261 | 0.034 |
| Bilingual Share | 0.282 | 0.438 | 0.156 |

Notes: All summary statistics are population-weighted averages across districts. Employment rate is the share of the population who are in the workforce. Urban share is the share of the population who live in urban areas. Bilingual share is the share of the district population who are bilinguals. Linguistic heterogeneity is the district-level measure of linguistic heterogeneity $h_d = 1 - \sum_{\ell \epsilon d} s_{\ell d}^2$, where $s_{\ell d}$ is the population share of speakers whose mother tongue is $\ell$ in district $d$. Population share of speakers of the majority language is the district population share that speaks the majority language of the district. Share bilingual for majority-language speakers is the share of speakers of majority languages in a district who are bilingual. Population share of speakers of minority languages is district population share that speaks a minority language of the district. Share bilingual for minority-language speakers is the share of speakers of minority languages in a district who are bilingual.
Source: Census of India

Table 2: All-India Industrial Mix in 1931 and 1961

| | 1931 | 1961 |
|---|---|---|
| Manufacturing Share of Workforce | 0.077 | 0.093 |
| | | |
| *Share of* | | |
| *Manufacturing Workers in* | | |
| Textiles | 0.345 | 0.311 |
| Wood | 0.131 | 0.123 |
| Metals | 0.060 | 0.068 |
| Ceramics | 0.086 | 0.070 |
| Chemicals | 0.053 | 0.036 |
| Apparel | 0.142 | 0.110 |
| Food Processing | 0.119 | 0.165 |
| Vehicles | 0.002 | 0.033 |
| Power | 0.002 | 0.013 |
| Other | 0.060 | 0.071 |

Notes: Equivalent subindustries for 1931 and 1961 were created using the 1901–1961 mapping of occupation codes in the 1961 census. This table includes workers in all states of post-Independence India and is thus not strictly comparable to Table 1.

Source: Census of India

Table 3: Manufacturing Employment and Bilingualism: Weighted OLS

| | Bilingual Share of Speakers | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | Minority Languages | | | Majority Language | |
| Manufacturing Share of Emp. × 1961 | 1.037* | 1.038*** | 1.060*** | 0.142 | 0.003 |
| | (0.61) | (0.25) | (0.34) | (0.12) | (0.07) |
| Year 1961 | 0.302* | 0.105 | 0.087 | -0.012 | -0.022 |
| | (0.16) | (0.08) | (0.11) | (0.02) | (0.06) |
| Manufacturing Share of Emp. | -1.222** | -2.381*** | -2.406*** | -0.019 | -0.178 |
| | (0.53) | (0.48) | (0.66) | (0.03) | (0.13) |
| Fixed Effects | None | District | District X Language | None | District |
| $R^2$ | 0.228 | 0.661 | 0.679 | 0.196 | 0.447 |
| N | 1349 | 1349 | 1349 | 308 | 308 |
| $\beta_{m61}^{diff}$ | | | 1.057*** | | |
| | | | (0.34) | | |
| $\beta_{m61}^{diff}$, District Trend Effects | | | 0.837** | | |
| | | | (0.40) | | |

Notes: All regressions include controls for initial levels of urbanization, literacy, workforce share of population, and language share of population. Observations are at the language-district level and are weighted by number of speakers in 1931. The differential effect of a change in manufacturing employment on minority-language speakers is $\beta_{m61}^{diff}$ and is estimated by pooling the majority and minority language data. Standard errors are 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 4: Instrument using 1931 Manufacturing Employment Shares: Sample Regression

| | 1961 Mfg. Share |
|---|---|
| Textiles | 1.128*** |
| | (0.21) |
| Apparel | 1.297* |
| | (0.50) |
| Wood | 0.495 |
| | (0.55) |
| Metal | 0.387 |
| | (0.48) |
| Ceramics | 1.496 |
| | (1.09) |
| Chemicals | 0.348 |
| | (0.93) |
| Food | 2.079** |
| | (0.64) |
| Vehicles | 25.274* |
| | (10.49) |
| Power | 3.728 |
| | (2.50) |
| Other | 3.979*** |
| | (0.71) |
| $R^2$ | 0.908 |
| N | 155 |

Notes: Observations at the district level. 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 5: First Stage for Industrial Mix Instrument

| | Manufacturing Share of Employment × 1961 |
|---|---|
| Predicted Manufacturing Share of Emp. × 1961 | 0.693*** |
| | (0.06) |
| Year 1961 | 0.040** |
| | (0.02) |
| Manufacturing Share of Emp. | 0.900*** |
| | (0.05) |
| $R^2$ | 0.969 |
| N | 308 |
| *F-Test of Instrument* | *117.14* |
| *Shea's Partial $R^2$* | *0.79* |

Notes: Observations are at the district level. Regression includes district fixed effects and controls for initial levels of urbanization, literacy, and workforce share of population. Standard errors are 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: Manufacturing Employment and Bilingualism: Weighted IV Estimates

| | Bilingual Share of Speakers | |
|---|---|---|
| | (1) | (2) |
| | Minority Languages | Majority Language |
| Manufacturing Share of Emp. × 1961 | 1.350*** | -0.045 |
| | (0.49) | (0.07) |
| Constant | 0.053 | -0.017 |
| | (0.12) | (0.06) |
| Manufacturing Share of Emp. | -2.669*** | -0.136 |
| | (0.71) | (0.12) |
| $R^2$ | 0.673 | 0.446 |
| N | 1345 | 308 |
| $\beta^{diff}_{m61}$ | 1.335*** | |
| | (0.50) | |
| $\beta^{diff}_{m61}$, District Trend Effects | 1.099** | |
| | (0.48) | |

Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Initial level controls include urbanization, literacy, workforce share of population, and language share of population. Standard errors are 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: Manufacturing Employment and Bilingualism Outside District $d$: Weighted IV Estimates

| | Bilingual Share of Speakers | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | Districts Adjacent to $d$ | | | All Districts Other Than $d$ | | |
| Languages | All | Minority | Majority | All | Minority | Majority |
| Manufacturing Share of Emp. × 1961 | 0.046 | 0.140 | 0.007 | -0.006 | -0.030 | 0.095 |
| | (0.12) | (0.37) | (0.06) | (0.03) | (0.05) | (0.06) |
| Year 1961 | 0.104*** | 0.124** | 0.060** | 0.041*** | 0.041*** | 0.055*** |
| | (0.03) | (0.06) | (0.02) | (0.00) | (0.01) | (0.02) |
| Manufacturing Share of Emp. | -0.368* | -0.895 | -0.064 | -0.040 | -0.018 | -0.146* |
| | (0.19) | (0.55) | (0.08) | (0.04) | (0.07) | (0.08) |
| $R^2$ | 0.762 | 0.728 | 0.617 | 0.701 | 0.697 | 0.759 |
| N | 1378 | 1091 | 287 | 1636 | 1328 | 308 |

Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Bilingual share of speakers is the ratio between the number of mother tongue speakers of language $\ell$ in district $d$ who can speak a second language to the total number of mother tongue speakers of language $\ell$ in district $d$. Initial level controls include urbanization, literacy, workforce share of population, and language share of population. Standard errors are 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: Agricultural Labor and Bilingualism: Weighted OLS Estimates

| | Bilingual Share of Speakers | |
|---|---|---|
| | (1) | (2) |
| | Minority Languages | Majority Language |
| Agricultural Laborers Share × 1961 | 0.460*** | 0.017 |
| | (0.14) | (0.03) |
| Year 1961 | 0.177 | -0.083 |
| | (0.14) | (0.08) |
| Agricultural Laborers Share of Ag. Emp. | -0.064 | -0.051 |
| | (0.13) | (0.04) |
| R-squared | 0.708 | 0.620 |
| N | 1237 | 281 |
| $\beta_{m61}^{diff}$ | 0.443*** | |
| | (0.15) | |
| $\beta_{m61}^{diff}$, District Trend Effects | 0.358*** | |
| | (0.21) | |

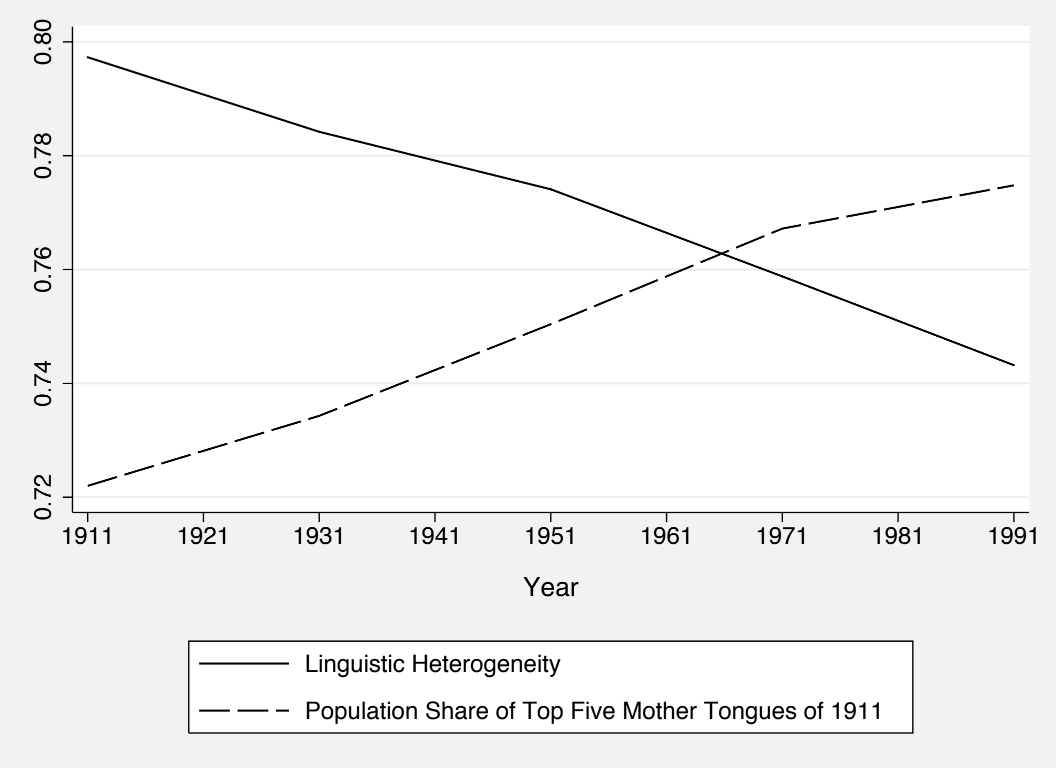
Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Initial level controls include urbanization, literacy, workforce share of population, and language share of population. Standard errors are corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 9: The Effects of Manufacturing Employment on Linguistic Heterogeneity

| | Linguistic Heterogeneity | |
|---|---|---|
| | (1) | (2) |
| Estimation | OLS | IV |
| Manufacturing Share of Emp. × 1961 | -1.109*** | -1.292*** |
| | (0.32) | (0.34) |
| Year 1961 | 0.256*** | 0.275*** |
| | (0.09) | (0.08) |
| Manufacturing Share of Emp. | 2.097*** | 2.271*** |
| | (0.52) | (0.53) |
| $R^2$ | 0.788 | 0.787 |
| N | 308 | 308 |

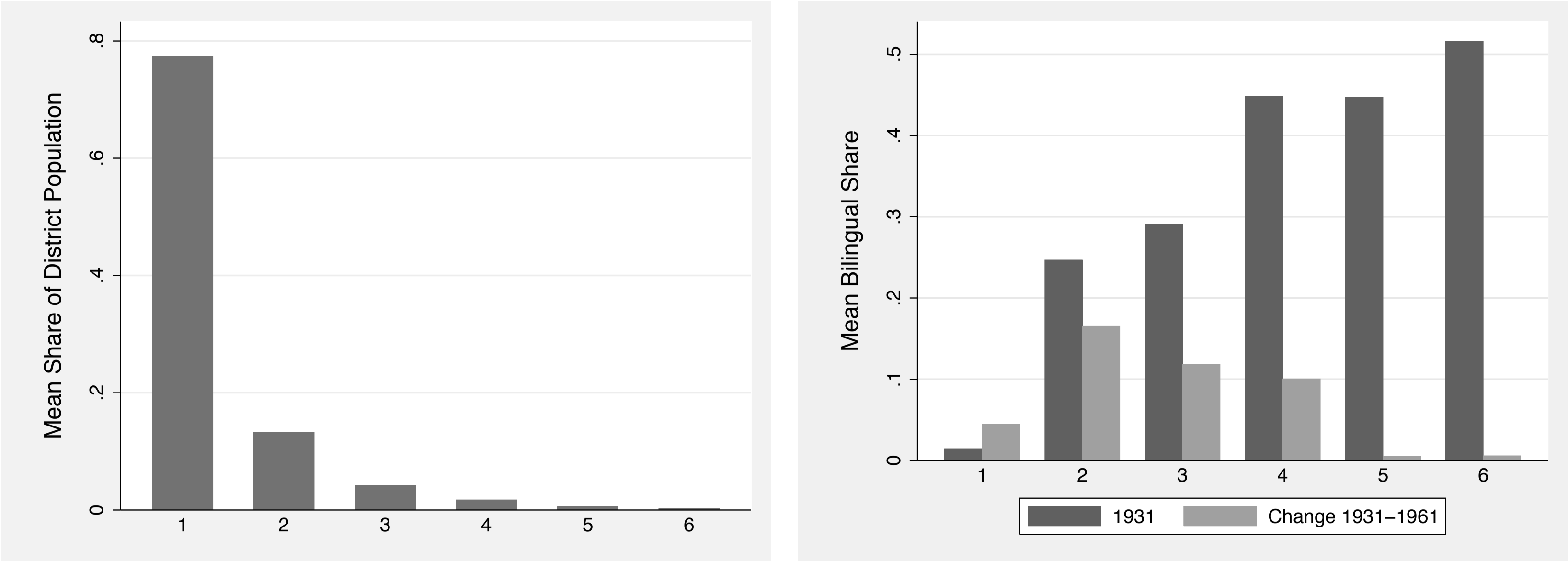
Notes: Each observation is weighted by the population in the district. Standard errors are robust to heteroskedasticity. Linguistic heterogeneity in district $d$ is $h_d = 1 - \sum_\ell s_{\ell d}^2$, where $s_{\ell d}$ is the district population share of speakers whose mother tongue is $\ell$. Regressions include controls for initial levels of urbanization, literacy, workforce share of population, and language share of population. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$

Figure 1: Linguistic Consolidation in India, 1911–1991



This graph shows two measures of linguistic consolidation for the post-Independence territory of India for the years 1911 to 1991. Linguistic heterogeneity is defined as $h = 1 - \sum_{\ell} s_{\ell}^2$, where $s_{\ell}$ is the population share whose mother tongue is $\ell$. It measures the probability that two randomly selected individuals in the population will have different mother tongues. The population share speaking the top five mother tongues of 1911 measures the share of the population speaking Hindi, Marathi, Bengali, Tamil, and Telugu, which were the five largest mother tongues in 1911. Based on tables in the the 1961 and 1991 Census of India.

Figure 2: District-Level Characteristics of Languages by their District-Level Ranks 1–6



(a) Mean share of district population speaking language for each rank

(b) Mean district-level share of speakers of language bilingual in 1931 and change 1931–1961 for each rank

Notes: I ordered languages by their number of mother-tongue speakers within each district for 1931, assigning them a rank of 1 through 6. The right-hand graph shows the mean population share of languages by rank. The left hand graph shows the share of mother-tongue speakers who were bilingual in 1931 by the mother tongue's rank, and the change in bilingualism between 1931 and 1961. The left-hand graph is weighted by the number of speakers of the mother tongue of each rank in 1931. We can therefore interpret the graphs as showing information for the average person whose mother tongue has a given district-level rank.