Human Capital Response to Globalization: Education and Information Technology in India

Gauri Kartini Shastry*

University of Virginia

October 2008

Abstract

Recent studies have shown that trade liberalization increases skilled wage premiums in developing countries. This result suggests globalization may benefit elite skilled workers relatively more than poor unskilled workers, increasing inequality. This effect may be mitigated, however, if human capital investment responds to new global opportunities. A key question is whether a country with a more elastic human capital supply is better positioned to benefit from globalization. I study how the impact of globalization varies across Indian districts with different costs of skill acquisition. I focus on the cost of learning English, a relevant qualification for high-skilled export jobs. Linguistic diversity in India compels individuals to learn either English or Hindi as a lingua franca. Some districts have lower relative costs of learning English due to linguistic predispositions and psychic costs associated with past nationalistic pressure to adopt Hindi. I demonstrate that districts with a more elastic supply of English skills benefited more from globalization: they experienced greater growth in both information technology jobs and school enrollment. Consistent with this human capital response, they experienced smaller increases in skilled wage premiums.

*Correspondence: shastry@virginia.edu. I am grateful to Shawn Cole, David Cutler, Esther Duflo, Caroline Hoxby, Lakshmi Iyer, Asim Khwaja, Michael Kremer, Sendhil Mullainathan and Rohini Pande for their advice and support, Petia Topalova and David Clingingsmith for help with additional data and all participants of the Labor/Public Finance and Development Economics lunch workshops at Harvard, the IZA/World Bank Conference on Employment and Development 2007 and the Federal Reserve Board of Governors International Finance workshop for their comments. I also thank seminar participants at Wellesley College, the University of Maryland, the Center for Global Development, the Wharton School, the Fletcher School at Tufts, the Anderson School at UCLA, Scripps College, the University of Notre Dame, Mathematica, Brookings Institution and the University of Virginia. In addition, I thank Elias Bruegmann, Filipe Campante, Davin Chor, Quoc-Anh Do, Eyal Dvir, Alex Gelber, Li Han, C. Kirabo Jackson, Michael Katz, Katharine Sims, Erin Strumpf and Daniel Tortorice for helpful conversations. All errors are mine.

1 Introduction

Recent literature has suggested that trade liberalization in poor countries increases skilled wage premiums. It is possible that this increase benefits only elite skilled workers, causing an increase in inequality. Alternatively, if human capital supply responds, globalization's effect on inequality may be mitigated. This is partly because some people acquire the necessary human capital, but also because the increased supply drives down the skilled wage premium over time. I study whether human capital in India responded to the shock of trade reforms by using variation across districts in their ability to take advantage of global opportunities. Some districts have a more elastic supply of English language human capital, which is particularly relevant for export-related jobs. I show that these districts attracted more export-oriented skilled jobs, specifically in information technology (IT)[1], and experienced greater growth in school enrollment. Consistent with this factor supply response, these districts experienced smaller growth in skilled wage premiums.

An important motivation for this paper relates to the consequences of trade. Under a simple Heckscher-Ohlin model with two goods and two countries, the labor-abundant poor country should specialize in unskilled labor intensive industries after liberalization, increasing the demand for unskilled workers and lowering skilled wage premiums. In contrast, several recent studies in Latin America found that trade liberalization increased human capital premiums.[2] There is little evidence on East Asia, but conventional wisdom suggests a smaller effect there.[3] There are multiple explanations for this divergence from standard trade theory.[4] Nevertheless, a key factor in understanding the effects of globalization is the supply response of human capital. If the supply of skilled workers rises (as in East Asia, but less so in Latin

[1] IT refers to both software and business process outsourcing such as call centers and data entry firms.

[2] See Goldberg and Pavcnik (2004) for a review of the literature, which includes, e.g., Hanson and Harrison (1999), Feenstra and Hanson (1997), Feliciano (1993), Cragg and Epelbaum (1996) on Mexico; Robbins, Gonzales and Menendez (1995) on Argentina; Robbins (1995a, 1995b, 1996b) on Chile and Uruguay; Robbins (1996a), Attanasio, Goldberg and Pavcnik (2004) on Colombia; Robbins and Gindling (1997) on Costa Rica.

[3] See Wood (1997) for a survey, Lindert and Williamson (2001) and Wei and Wu (2001).

[4] Two explanations focus on global outsourcing (Feenstra and Hanson 1996, 1997) and complementarities between skill levels (Kremer and Maskin 2006). Others are discussed in Goldberg and Pavcnik (2004).

America[5]), globalization's effect on inequality may be mitigated.

This paper also bears on the debate in many poor countries on whether to encourage people to retain diverse local languages, choose a single native language or promote a global language, such as English. Native language instruction in public schools can strengthen national identity, while instruction in a foreign language may be less accessible for poor households. If the global language is used in white-collar jobs, promoting that language may increase opportunities for the poor.[6] The ability to integrate better with the world economy may bring more trade benefits to places that promote English.[7]

This paper studies how human capital investments responded to Indian trade liberalization in the early 1990s. These reforms, driven by a balance-of-payments crisis, led to an unexpected increase in global opportunities. In addition to reducing trade barriers, reforms removed most restrictions on private and foreign direct investment in telecommunications. While this shock was common across India, historical linguistic forces have created differences in the ability of districts to take advantage of these new opportunities. I examine how the impact of globalization varied with these pre-existing differences.

Specifically, I show that the effect of trade liberalization varies with the elasticity of human capital supply by exploiting exogenous variation in the cost of learning English. There is a serious challenge in identifying a causal effect of lower English learning costs. State governments that care more about global opportunities may advance the teaching of English, but also pursue other policies that promote trade. I use variation driven by historical linguistic diversity that is exogenous to such outward-oriented policies and to reverse causality. In fact, this variation led people to learn English even in 1961, long before trade reforms in the 1990s were contemplated. In India, there is substantial local linguistic

[5]See Attanasio and Szekely (2000), Sanchez-Paramo and Schady (2003).

[6]See Human Development Report (2004), Angrist, Chin and Godoy (2006), Lang and Siniver (2006) and Angrist and Lavy (1997).

[7]Levinsohn (2004) examines how globalization in South Africa increases the returns to English. Using household data from a suburb of Mumbai, India, Munshi and Rosenzweig (2006) find increases in the returns to English and enrollment in English-medium schools from 1980 to 2000.

diversity that motivates individuals to learn a lingua franca.[8] The two choices are English and Hindi. Individuals whose mother tongue is linguistically further from Hindi have a lower relative opportunity cost of learning English: they find Hindi more difficult to learn and they often suffer psychic costs when speaking Hindi, since many feel Hindi was imposed on them. I first show empirically that linguistic distance from Hindi induces more people to choose English and more schools to teach English.

One could worry that variation in linguistic distance from Hindi may be correlated with unobserved regional differences. However, due to the local linguistic diversity, I am able to rely purely on pre-existing within-region variation. The six regions each consist of 2-7 states, each of which is made up of 1-51 districts.[9] I also control for many other potential determinants of IT growth and education and rely on changes over time, comparing post-liberalization to pre-liberalization schooling and wages. Figure 1 maps the within-region variation in linguistic distance to Hindi that I exploit (details are provided in section 5.3).

Using new data on the Indian IT sector, I show that IT firms were more likely to operate in districts with lower costs of English learning. My choice of the IT sector is driven by several factors. This industry grew primarily due to trade liberalization and technological progress; it is difficult to isolate other jobs created by trade reforms. In addition, IT firms almost exclusively hire English speakers; data on other jobs that require English is unavailable.[10] It is important to note that globalization in India consists of more than just IT, e.g. trade in other goods, foreign direct investment, tourism and migration.[11] My IT estimates are likely to underestimate the impact of globalization on English-language opportunities; my estimates on schooling and wages are likely to be driven partly by these other opportunities.

I next provide evidence that districts with native languages linguistically further from

[8] A lingua franca is a language used as a common or commercial language among diverse linguistic groups.

[9] Relying solely on within-state variation tells a similar story, but there is less variation (section 6.3.4).

[10] A search for "English" on an Indian job website (PowerJobs) delivered mostly ads in IT, but also postings for teachers, engineers, receptionists, secretaries, marketing executives and human resources professionals that required English fluency. Many of these jobs are in multinational or exporting firms.

[11] See Topalova (2004, 2005) for the effect of *import* competition on productivity and poverty in India. Edmonds, Pavcnik and Topalova (2007) find that import competition reduces schooling, due to poverty.

Hindi experienced greater increases in school enrollment from 1993 to 2002 relative to pre-existing trends. Using individual-level data, I further find that these districts experienced a smaller increase in the skilled wage premium. The fact that education and skilled wage premiums move in opposite directions is strong evidence that the supply response of human capital shapes the impact of globalization. Demand-driven explanations would predict an increase in both education and the skilled wage premium.

The paper is organized as follows. Section 2 describes trade liberalization and IT in India while section 3 develops a simple theoretical framework. Section 4 describes the Indian linguistic context. In section 5, I discuss the empirical strategy. Section 6 presents the results on IT firm presence and school enrollment growth, as well as discussion of a few robustness checks. Finally, section 7 provides evidence on returns to education and section 8 concludes.

2 Background on trade liberalization and IT

During much of the post-colonial period, India heavily controlled its economy and limited trade through strict investment licensing and import controls. In the late 1970s and 1980s, the government took small steps towards liberalization, but even as late as 1990, numerous tariff and non-tariff obstacles limited trade. Policies in the 1980s raised India's growth rate, but heavy borrowing led to a balance-of-payments crisis in 1991, which resulted in a shift towards an open economy. Reforms ended most import licensing requirements for capital goods and slashed tariffs. The March 1992 Export-Import Policy reduced the number of goods that were banned for export from 185 to 16. From 439 goods subject to some control, the new regime limited only 296 exports. The government also lifted some capital controls and devalued the rupee, further reducing deterrents to trade. Exporters were allowed to sell foreign exchange in the free market or at a lower official price (Panagariya 2004).

Prior to the early 1990s, many services were heavily regulated by the government. State-owned enterprises were large players in sectors such as insurance, banking, telecommunica-

tions and infrastructure. Reforms in the 1990s reduced this level of state intervention. Most importantly, these sectors were opened up to private participation and foreign investment. The 1994 National Telecommunications Policy opened cellular and other telephone services, previously a state monopoly, to both private and foreign investors. Due to technological progress, this policy was revised in 1999 under the New Telecom Policy which further reduced limits on foreign direct investment (FDI) in telecommunications services. FDI for Internet service providers was also allowed with few conditions. FDI in e-commerce was free of all restrictions and foreign equity in software and electronics was granted automatic approval, particularly for IT exporters (Panagariya 2004).

These policy changes led to remarkable growth in Indian exports; annual growth rose 3.3 percentage points in the 1990s from the 1980s. Service exports grew more rapidly than manufacturing exports; even within manufacturing, capital or skilled labor intensive sectors grew faster (Panagariya 2004). Along with technological progress, this policy shift led to the growth of IT services outsourced to India. By 2004, India was the single largest destination for foreign firms seeking to purchase IT services. IT outsourcing accounted for 5% of India's GDP in 2005, forecast to contribute 17% to projected growth to 2010 (Economist 2006). Employment growth has been strong: from 56,000 professionals in 1990-91, the sector employed 813,500 in 2003, implying an annual growth rate of over 20%. In particular, the IT sector increased job opportunities for young, educated workers; the median IT professional is 27.5 years old and 81% have a bachelor's degree (NASSCOM 2004). An entry-level call center job pays on average Rs. 10,000 ($230), considered high for a first job (Economist 2005). The supply of engineers is also found to have influenced the growth of India's software industry (Arora and Gambardella 2004; Arora and Bagde 2007). IT firms are relatively free to locate based on the availability of educated, English-speaking workers, due to their young age, export focus and reliance on foreign capital. While many IT firms are concentrated in a few large cities, young firms have spread out to smaller cities to avoid congestion. Figure 2 maps IT establishments, demonstrating how new firms have spread out across districts.

Trade liberalization in the early 1990s was a shock prompted by external factors; the ad-hoc steps in the 1980s did not reflect a systematic shift towards an open economy. An underlying assumption for my identification strategy is that these reforms were not implemented differently in areas with many English speakers in order to gain from trade in services. This is unlikely since the balance-of-payments crisis that motivated the policy shift and the advances in telecommunications and information technology were unrelated and unexpected.

3 Theoretical framework

Next, I describe a simple model of how the effects of globalization vary with English learning costs. Consider two districts that differ only in this cost (LC and HC for low and high cost, respectively). In the case where most people in either district do not speak English, as in India, a lower cost will generate a higher elasticity of English human capital.[12]

I specify production processes for two goods, one of which is tradeable, and a schooling decision between English- and Hindi-medium instruction. I solve the model before and after trade liberalization to compare the changes in education and the skilled wage premium. After trade reforms, the cost of transporting the traded good falls sufficiently to facilitate its production and export from this economy.[13] English- and Hindi-speaking skilled workers are equally productive in the domestic sector, but only English speakers can produce the globally traded good.[14] Goods travel freely between districts, but workers do not.

Trade reforms cause a positive demand shock for skills, but the impact depends on the supply elasticity of human capital. Identical shocks would cause a greater increase in human capital, but a smaller increase in the wage premium, where the supply is more elastic (figure 3). However, more exporting firms should locate in the low cost LC since it is easier to hire English speakers. The increase in education should still be larger in LC, but the relative change in the wage premium depends on the size of the shocks (figure 4). Formally, I

[12] I do not make an explicit assumption about the elasticity, but this fact aids in the discussion.
[13] I abstract from why India exports goods intensive in skilled labor instead of unskilled labor.
[14] We can think of this good as IT or any good or service requiring English-speaking workers.

vary the difference in demand shocks by changing the importance of a second factor in the production of the traded good that is fixed in the short run, e.g. infrastructure. The more intensive production is in this factor, the more similar the demand shocks. Varying this factor's endowment also predicts how districts with different business environments respond to trade: a pro-business district should see greater growth in both education and the skilled wage premium.

3.1 Production processes for traded and non-traded goods

The non-traded good, Y, is consumed in both districts and produced using

$$Y = \min\left\{\frac{L_Y}{\alpha_L}, \frac{H_Y + E_Y}{\alpha_H}\right\} \qquad where\ \alpha_L > \alpha_H$$

where L_Y, H_Y and E_Y are quantities of unskilled, Hindi- and English-skilled labor and the α's are parameters. Since Hindi- and English-skilled workers are perfect substitutes, firms hire the cheaper workers. Prior to trade liberalization, English and Hindi speakers earn the same wage. The amount of Y produced is determined by the availability of labor.

After trade liberalization, firms who start producing the traded good, X, set up in either district, taking the price of X, p_X, as given. The production function is

$$X = F^{\beta} E_X^{1-\beta} \qquad where\ 0 < \beta < 1$$

where E_X is the amount of English-speaking skilled labor used and F is the exogenous endowment of the fixed factor that earns a return r_F. We can think of F as infrastructure (such as telecommunication networks) that is slow to change, immobile entrepreneurs or even the business environment more generally.

3.2 Schooling decisions

Individuals live for one period and work as unskilled labor or get instantaneous education in English or Hindi and work as skilled labor. P people, born each period, differ in a parameter c_i, distributed uniformly over $[0, 1]$. A second parameter, $\mu_j > 1$, measures the cost of learning English and varies by district j: the low cost district LC has a lower μ_j than the high cost district HC. We can interpret this parameter as the linguistic distance to Hindi of the language spoken in the district: people in LC speak a language further from Hindi.[15] That education is available only in English and Hindi corresponds to the two linguae franca. Studying in Hindi costs $(t_H + c_i)\, w_U$ where t_H is fixed $(0 < t_H < 1)$ and w_U is the unskilled wage. Studying in English costs $\left(t_E + \mu_j c_i\right) w_U$ where t_E is fixed $(0 < t_E < 1)$.

Given this simplified cost structure, I assume that $t_E < t_H$, allowing education in English to be cheaper than in Hindi for some people to ensure that we have some English speakers in autarky.[16] The results are robust to assuming $t_E = t_H$ but the model would then generate no English speakers in autarky. To simplify the algebra, I let $t_E = 0$ and $t_H = t$. Individuals maximize lifetime income. Skilled individuals earn w_H or w_E depending on the language of instruction they choose. Since all skilled workers are equally productive in the Y sector, $w_E \geq w_H$. When solving the individual's problem, I ignore the unrealistic case in which there are no Hindi-skilled workers. Thus, people with low values of c_i get English schooling, those in the middle study in Hindi and those with higher c_i remain unskilled. Letting H and E be the number of Hindi- and English-skilled workers, respectively, I define two additional terms: total education, $ED = H + E$, and the weighted average return to skill, $\widehat{q} = \frac{w_H H + w_E E}{w_U (H+E)}$.

[15] While linguistic distance to Hindi can vary within a district as well, this simplification is not unrealistic since individuals close to Hindi in the low cost district may still be influenced by more English instruction schools and an equilibrium where most people speak English.

[16] While it may seem odd that English education can be cheaper than Hindi in India, the fact that in some states more people speak English than Hindi suggests this is realistic: the absolute cost of studying English may be less than Hindi for some people.

3.3 Characterizing the equilibrium

In equilibrium, all labor markets must clear. Skilled labor market clearing depends on whether the demand for English speakers exceeds their initial supply. Recall that even when $w_E = w_H$, some individuals choose to study in English. If, in equilibrium, the demand for English speakers is less than this initial supply (case A), then $w_E = w_H$ because the remaining English speakers work in the Y sector. The market clearing condition is

$$\alpha_H Y + F w_E^{-\frac{1}{\beta}} \left(p_X \left(1-\beta\right)\right)^{\frac{1}{\beta}} = P\left(\frac{w_H - w_U - t w_U}{w_U}\right) \tag{1}$$

If the demand for English speakers exceeds the natural supply, then $w_E > w_H$ and no English speakers work in the Y industry (case B). The labor market clearing conditions are

$$\alpha_H Y = P\left(\frac{w_H - w_U - t w_U}{w_U} - \frac{w_U t + w_E - w_H}{w_U\left(\mu_j - 1\right)}\right) \tag{2}$$

$$F w_E^{-\frac{1}{\beta}} \left(p_X \left(1-\beta\right)\right)^{\frac{1}{\beta}} = P\frac{w_U t + w_E - w_H}{w_U\left(\mu_j - 1\right)} \tag{3}$$

In both cases, the labor market clearing condition for unskilled workers is

$$\alpha_L Y = P\left(1 - \frac{w_H - w_U - t w_U}{w_U}\right) \tag{4}$$

These labor market clearing conditions, zero profit conditions for each sector and the production function for X close the model. Good Y is the numeraire. The equilibrium without any trade is a special case of A when $F = 0$. Since the demand for English-skilled workers rises after trade liberalization, the wage for English speakers has to rise. Now that fewer English speakers are working in the Y industry, the wage for Hindi-skilled workers rises as well to keep the ratio of skilled to unskilled workers in Y production constant. To compare how these changes differ in districts with different levels of μ_j, we have to first solve the equilibrium in both cases A and B.

Proposition 1 *Case A. If, in equilibrium, $w_E^* = w_H^*$, i.e. the demand for English-skilled workers is less than or equal to the natural supply, then E^* is falling and H^* is rising in the cost of learning English, μ_j. Total education, ED^*, the amount of X produced and the average return to education, $\widehat{q}^*$, are independent of μ_j.*

Proof. All proofs are found in the appendix. ■

If the two districts LC and HC are both in this case, they will have identical wages, production of X and returns to education. They will also have identical total education, but the low cost district will have a higher proportion of English speakers. I next solve case B.

Proposition 2 *Case B. If, in equilibrium, $w_E^* > w_H^*$, i.e. the demand for English-skilled workers is greater than the natural supply, then E^* is falling and H^* is rising in the cost of learning English, μ_j. Total education, ED^*, and the amount of X produced are both falling in μ_j, but the effect of an increase in μ_j on the average return to education, $\widehat{q}^*$, is ambiguous.*

Proposition 3 (in the appendix) provides the necessary and sufficient condition for whether a district is in case A or B. Intuitively, a district is no longer in case A when the demand for English speakers exceeds the natural supply: the high cost district, with fewer initial English speakers, will leave case A at a lower value of F.

The intuition behind the effect of μ_j on returns to education is simple. English-skilled workers are less elastic in HC since the cost of English is higher and most people do not speak English in either district; therefore their wage must rise more. Similarly, the greater relative elasticity of Hindi-skilled workers in HC results in a smaller increase in the Hindi wage. These changes are constrained, however, by the constant skilled-unskilled labor ratio in Y production and the relative size of the demand shocks for English speakers. When X production is intensive in F (a high β), the amount of X that can be produced is more constrained by the amount of F available: the demand for English speakers in the two districts cannot be too different. Thus, the return to education increases by more in HC (figure 4a).

If X production is less intensive in F, the demand shock for skilled labor in LC can be much bigger than in HC, increasing the return to education by more in LC (figure 4b).

I test two predictions of this model, assuming districts in India are in case B (since there is a return to speaking English[17]). First, the district with a lower English learning cost should produce more X and second, school enrollment in the low cost district should grow faster after liberalization. Finally, I provide evidence that the average return to education rises by more in high cost districts. It is straightforward to examine the effect of differences in the endowment of F, the fixed factor (proposition 4 in the appendix). A district with more F would produce more X and experience a larger increase in both education and skilled wage premiums. Thus, the fact that returns to skill move in the opposite direction from educational attainment is evidence against the possibility that these results are driven by improvements in the business environment, such as from state level economic reforms.

4 Linguistic distance from Hindi

4.1 Language of government and media of instruction

The 1961 Census of India documented 1652 mother tongues from five distinct language families, many of which are quite dissimilar. Linguists classify languages such as English and Hindi into the same family (Indo-European), but are unwilling to connect some languages native to India, such as Hindi and Kannada (spoken in Karnataka). As of 1991, 22 languages were native to more than a million people and 114 languages native to 10,000. Much of this diversity is local. One measure of diversity is the probability that two randomly chosen district residents speak different native languages, calculated as one minus the Herfindahl index (the sum of squared population shares of ethnic groups). On average, this probability is 25.6%, ranging from 1% to 89%. A district's primary language is native to 83% of residents on average, ranging from 22% to 100%. Due to this local diversity, many individuals need

[17] See Munshi and Rosenzweig (2006).

to learn a lingua franca (Clingingsmith 2006). Of all multilingual individuals who were not native speakers, 60% chose to learn Hindi and 56% chose to learn English[18]. These are clearly the two linguae franca: at only 6%, Kannada was the next most common second language.

The choice of whether to learn Hindi or English depends on an individual's mother tongue. Someone whose mother tongue is similar to Hindi will find Hindi easier to learn, giving them a greater opportunity cost of learning English, relative to someone whose mother tongue is more different. Historical forces have amplified this tendency. During British occupation, English was established as the language of government and instruction. After Independence in 1947, a nationalist movement to choose an indigenous official language favored Hindi, the native lingua franca. Despite opposition from non Hindi speakers, the official status of Hindi was written into the constitution. This led to riots in non Hindi-speaking areas, the most violent of which occurred in Tamil Nadu in 1963. Speakers of other languages felt at a disadvantage speaking Hindi. In 1967, the central government made Hindi and English joint official languages (Hohenthal 2003): this status reinforces their dominance as the two linguae franca. This history explains greater English literacy among speakers of languages linguistically distant to Hindi. In fact, in some states more people speak English than Hindi.

Over time, this relationship became institutionalized through media of instruction. Early growth in formal education, in the early nineteenth century, was driven by the British who sought to foster an elite class to govern the country (Nurullah and Naik 1949, Kamat 1985). By Independence, missionary societies and princely states had set up many schools in native languages. In 1993, English was still the main medium of tertiary instruction, but there were over 28 media of instruction in urban primary schools. Hindi was most common with 38% of schools and English was second with 9%. At the secondary level, this difference was smaller: instruction was in Hindi in 29% of schools and English in 20%.

I examine the linguistic choices of native Hindi speakers empirically, since these are governed by two opposing forces. While they may not need another lingua franca, English

[18]More than three hundred million people speak Hindi as a first language, while only 180,000 are native English speakers; thus, many more people speak Hindi than English.

allows them to interact with more additional people than any other second language. Since schools often require a second language, we might expect Hindi natives to learn English.

4.2 Measuring linguistic distance

My main measure of linguistic distance from Hindi was developed in consultation with an expert on Indo-European languages, Jay Jasanoff, the Diebold Professor of Indo-European Linguistics and Philology at Harvard University. This measure is based on classifying languages into five degrees of distance from Hindi (see figure 5). For example, Punjabi is one degree from Hindi, while Bengali is three degrees away. From the 1991 Census of India, I calculate a district's linguistic distance from Hindi in two ways: 1) the population-weighted average distance of all native languages and 2) the population share speaking languages at least 3 degrees away ('distant speakers'). Table 2 provides summary statistics.

To ensure that my measure captures linguistic similarity, I calculate two logically independent measures which turn out to be highly correlated with my preferred measure. The first measure is based on language family trees from the Ethnologue database. Figure 6 provides an extract that includes Indo-European languages found in India. I define distance as the number of nodes between two languages: Punjabi is five nodes from Hindi, while Bengali is seven nodes away. I assume there is a node connecting different families. The second measure is the percent of cognates between each language and Hindi from a list of 210 core words.[19] Expert judgments on which words are cognates are from the Dyen et al. (1997) dataset of 95 Indo-European languages.[20] Table 1a provides examples of cognates; table 1b lists the percent of cognates for some languages spoken in India. While Punjabi has 74.5% cognates with Hindi, Bengali has 64.1%. Reassuringly, the correlation between these measures is remarkably high: 0.903 between my preferred measure and the number of

[19] This measure is used in glottochronology, a method to estimate the time of divergence between languages (Swadesh 1972). The formula converting the percent of cognates into a time of divergence is currently out of favor among linguists, but the percent of cognates is still an acceptable measure of similarity.

[20] I assume other languages have 5% of words in common with Hindi. Linguists use 5% as a threshold to determine whether languages are related. For Indo-European languages not in Dyen's list or on Jasanoff's chart, I use the value of the closest language in the tree.

nodes, -0.935 and -0.936 respectively between these and the percent cognates.

5 Identification and empirical specifications

My identification strategy - using within-region variation in the cost of learning English driven by linguistic diversity - depends on two assumptions. First, linguistic distance from Hindi must predict the cost of learning English and second, it must not be correlated with omitted factors that affect schooling or exports conditional on region fixed effects and control variables. If a good measure of English learning costs were available, I would instrument for the cost of English with linguistic distance to Hindi. However, a comprehensive district-level measure of this cost is unavailable, partly due to data limitations and partly because the cost of English is multi-dimensional. We would want to include how many schools teach English and how many adults speak English at the very least, both unavailable at the district level. Therefore, I present reduced form results using the weighted average and percent distant speakers measures of linguistic distance.[21]

In the next subsection, I provide evidence for the first assumption by demonstrating that linguistic distance from Hindi predicts English literacy using data on second languages spoken across India. After that, I describe the specifications to test the effects of globalization and in the following two subsections, I discuss support for the second assumption.

5.1 Linguistic distance and learning English

Using data from the Census of India (1961 and 1991), I estimate

$$E_{lkt} = \alpha_0 + \beta' D_l + \alpha_1' X_{lk} + \gamma_t + \gamma_g + \epsilon_{lkt}$$

[21] In the first robustness check discussed in section 6.3, I proxy for the cost of learning English with the percent of schools teaching in the regional mother tongue, the only data available by district. I then instrument for this proxy with measures of linguistic distance to Hindi. As expected, the results are consistent with the reduced form results.

where E_{lkt} is the percent of native speakers of language l in state k, region g and year t who choose to learn English (conditional on being multilingual), D_l is the distance of language l from Hindi, and γ_t is a year fixed effect. I include region fixed effects, γ_g, (for north, northeast, east, south, west and central India), taking out regional variation in the probability of learning English. X_{lk} includes the share of language l speakers in state k, an indicator for whether language l is the state's primary language and the distance from Hindi of the state's primary language. I weight observations by the number of native speakers and cluster the standard errors by state.

The results confirm the relationship between learning English and linguistic distance to Hindi (table 3). In column 1, I include dummy variables for each distance from Hindi. Languages 1 degree away from Hindi are the omitted group. Two points are worth making. First, individuals at a linguistic distance of zero (native speakers of Hindi and Urdu[22]) tend to learn English. Recall that the prediction on their behavior was ambiguous since they already speak a lingua franca: they have less need of a second language, but English gives them access to more additional people. This creates a non-monotonicity in the relationship between linguistic distance and English learning. In figure 7, I plot the coefficient on each distance from Hindi with a 95% confidence interval to convey this result graphically. The propensity to learn English dips significantly when we move away from native Hindi speakers and then rises as native tongues get further. In my regressions, I account for the non-monotonicity by controlling for the percent of native Hindi speakers and calculating the weighted average linguistic distance only among non Hindi speakers. The percent of distant speakers measure is not affected by this non-monotonicity. I can also exploit this non-monotonicity since districts with more Hindi speakers, all else equal, are likely to have more English speakers.[23]

Second, linguistic distance from Hindi significantly predicts how many multilingual individuals learn English. The p-value at the bottom of column 1 tests whether the dummy

[22] Hindi and Urdu are often considered the same language. Separating them gives similar results.

[23] Hindi speakers are not explicitly in the theoretical model, but we would expect that for them, education in English costs more than in Hindi but the economic return to English outweighs the cost. The predictions would not change if I modelled the non-monotonic relationship between linguistic distance to Hindi and μ_j.

variables are jointly different from zero. The individual dummy variables are not strictly increasing, but the deviation is small. Thus, it is important to examine both the weighted average and percent distant speakers, since the latter does not impose a linear structure.

Columns 2-3 in table 3 test both reduced form measures of linguistic distance. One linguistic degree increases the percent of multilinguals who learn English by 8.9 percentage points. Speaking a language 3 or more degrees from Hindi increases the percent of English speakers by 42 points relative to speakers of languages 1 and 2 degrees away. Columns 3-7 confirm these findings for 1961 and 1991 separately. The relationship between linguistic distance from Hindi and learning English existed even in 1961; this supports the fact that this relationship is exogenous to recent trade reforms.[24] The control variables matter as expected: individuals are more likely to learn English in 1991 and they are more likely to learn English the more other people in their state share their mother tongue (minorities are likely to learn the regional language first). Note that the distance from Hindi of the state's primary language increases an individual's propensity to learn English in general but that of the individual's mother tongue is significant even when we control for this additional measure of distance from Hindi. I reject the hypothesis that speaking any language other than Hindi has a uniform effect on learning English by testing the equality of all linguistic distance fixed effects. Recall that these regressions include region fixed effects: the tendency to learn English is stronger for people further from Hindi even within a region.

The results in table 4 lend further credence to the use of linguistic distance. Columns 1-2 show linguistic distance to Hindi does not influence the decision to become multilingual. The choice to become multilingual is also independent of the distance from Hindi of the state's primary language. Columns 3-4 demonstrate that linguistic distance to Hindi is measured sensibly since individuals further from Hindi are less likely to learn Hindi. All results in tables 3 and 4 are robust to including state fixed effects and clustering by native language, to account for correlated errors between speakers of the same language.

[24] These results are robust to using the overall share of English learners or the number of English speakers as the dependent variable instead.

I next explore how distance from Hindi of languages spoken in a state predicts whether schools teach English by estimating

$$M_{ij} = \alpha_0 + \beta' D_j + \alpha_1 P_j + \alpha_2' Z_j + \gamma_i + \gamma_g + \epsilon_{ij} \quad (5)$$

where M_{ij} measures language instruction at school level i (primary, upper primary, secondary or higher secondary) in state j, D_j is the linguistic distance from Hindi of languages spoken in state j, and P_j is child population (aged 5 - 19, in millions).[25] Z_j includes average household wage income, average income for educated individuals, the percent of adults with regular jobs, the percent of households with electricity, and the percent of people who: have graduated from college, have completed high school, are literate, are Muslim, or regularly use a train. These are measured for urban areas in 1987. I focus on urban areas since the effects of globalization, such as growth in IT firms, are likely to be concentrated in cities. I include the percent native English speakers, the distance to the closest of the 10 most populous cities and whether the district is on the coast to account for potential trade routes.

The vector Z_j includes the percent native Hindi speakers to account for the non-monotonicity described above. To ensure that these results are not driven by native Hindi populations in the "Hindi Belt" states with high levels of corruption and government inefficiency, I include an indicator variable for the following states: Bihar, Uttar Pradesh (and Uttaranchal), Madhya Pradesh (and Chhattisgarh), Haryana, Punjab, Rajasthan, Himachal Pradesh, Jharkhand, Chandigarh and Delhi. In addition, I include region and level (primary, secondary...) fixed effects. The data appendix describes the sources.

The results show that linguistic distance from Hindi predicts the percent of schools that teach in English (columns 1-2 of table 5) or teach English as a second language (columns 3-4). An increase in 1 degree in the average distance from Hindi increases the percent of schools teaching in English by about 20 percentage points and the percent of schools teaching English by 33 points. More Hindi speakers also increases the teaching of English, but only

[25] These regressions are at the state-level since district-level data is unavailable.

when linguistic distance is measured by the percent distant speakers. It is not surprising that not all of these columns yield a significant result since the data is aggregated to the state level, leaving little within-region variation.

5.2 Empirical specifications

Using this strategy, I first test whether globalization has had a larger impact in districts with lower costs of learning English by studying the IT sector. Growth in IT is one potential benefit of promoting English, since IT firms hire mostly educated English speakers. I estimate

$$IT_{jt} = \alpha_0 + \beta' D_j + \alpha_1' Z_j + \alpha_2' W_j + \gamma_t + \gamma_g + \nu_{jt} \tag{6}$$

where IT_{jt} is a measure of IT presence in district j in year t and D_j measures linguistic distance. Z_j is as in equation (5) and W_j includes other potential predictors of IT firm location such as log population, the number of elite engineering colleges, the distance to the closest airport, and the percent of non-migrant engineers in 1987. As before I cluster the standard errors by district. The measures of IT presence include the existence of any headquarters or branches, the age of the oldest firm, the log number of headquarters and branches and the log number of employees.[26] The data, described in the appendix, contains only firm-level, not branch-level, employment; I assign employees evenly across branches to estimate employment by district. Summary statistics can be found in table 6.

To study how schooling responds in districts with different costs of learning English, I use enrollment from three years (1987,1993, and 2002) to estimate

$$\begin{aligned} \log(S_{ijt}) - \log(S_{ijt-1}) &= \alpha_0 + \beta' D_j \cdot I(t = 2002) + \alpha_1 \log(S_{ijt-1}) + \alpha_2' P_{jt} \\ &\quad + \alpha_3' Z_j \cdot I(t = 2002) + \alpha_4 B_{jt} + \gamma_j + \gamma_t + \gamma_{gt} + \gamma_{it} + \mu_{ijt} \end{aligned} \tag{7}$$

where S_{ijt} is enrollment in grade i in district j, region g and time t, $I(\cdot)$ is an indicator

[26] I add one before taking the natural log to avoid dropping districts with no IT presence.

function, D_j and Z_j are as above[27] and P_{jt} includes log child population at time t and $t-1$. γ_j controls for district trends and γ_{it} is a cohort effect. Region fixed effects are interacted with time, allowing for region-specific changes in trend. As before, I use only data from urban areas[28] and cluster by district. I also include B_{jt}, a proxy for skilled labor demand growth, to control for other factors that may influence demand for education. B_{jt} measures predictable changes in skilled labor demand. This variable, calculated along the lines of Bartik (1991), is an average of national industry employment growth rates weighted by pre-liberalization industrial composition of district employment. B_{jt} should not be correlated with local labor supply shocks but may pick up some of the effect of reforms through national employment growth. Details are in the appendix.

Since the data consist of the number of students enrolled, not enrollment rates, I control for population among 5-19 year-olds from the 1991 and 2001 Census - I cannot use a more precise population for each grade since children start school at different ages and often fail to be promoted. Enrollment increased dramatically, by 32%, between 1993 and 2002 (see table 7), but the increases were greater in the highest grades.

The region fixed effects in specification (6) and the region-specific changes in trend in specification (7) ensure that these results are driven by within-region variation in linguistic distance to Hindi. In section 6.3.4, I discuss how these results change when I include state fixed effects; essentially, they change slightly but tell the same story.

5.3 Origins of linguistic distance variation

It is important to pinpoint the source of within-region variation in linguistic distance. Figure 8 maps the raw variation in percent distant speakers, demonstrating that linguistic distance is not randomly distributed across India. Much of the variation is across region (indicated by thick black lines). Indo-European languages have traditionally been spoken

[27] Since I already include district fixed effects, Z_j is interacted with $I(t = 2002)$, allowing pre-liberalization differences to have a different effect post reform. The results are robust to not including these controls.

[28] The results are robust to using total school enrollment in the district; rural areas show no significant differences in enrollment (see 6.3.3).

in the north, Sino-Tibetan languages are spoken in the northeast while Dravidian languages are found in the south. Since this geographic variation may be correlated with other factors that influence schooling or exports, such as weather, agricultural productivity or culture, I include region fixed effects. I also include district control variables and allow them to have different effects before and after liberalization. In specification (7), the fixed effects and interactions ensure that I am identifying off deviations from a district's own trend and the change in the region's trend in schooling. This strategy rules out many other explanations for these results. Recall that in figure 1 I demonstrated the residual variation that I use by mapping the residuals from a regression of linguistic distance on region fixed effects and controls Z_j. This variation is less likely to be correlated with omitted variables.

This within-region variation is largely due to historical migration. India has had a long, politically fractured history. While recent migration is infrequent, people have migrated across India for millennia, bringing their native languages to other areas. Some have assimilated, but the local linguistic diversity provides evidence of the ability of these groups to retain a separate identity. State boundaries were drawn on linguistic lines in the 1950s, but drawing compact states necessitated including some people of different linguistic groups.

As suggestive evidence of this source of variation, note that a district's linguistic distance to Hindi is highly correlated across the 1991 and 1961 censuses. The correlation between the weighted average measures from 1961 and 1991 is 0.81; for the percent distant speakers, the correlation is 0.88.[29] In addition, there is substantial historical evidence of ethnic groups migrating across present-day state boundaries. For a long time, Dravidian languages were thought to be native to South Asia. Several studies, however, have led some linguists to believe that these languages were first spoken in areas to the northwest of the subcontinent (Tyler 1968, McAlpin 1981). There is also linguistic and cultural evidence of Dravidian

[29] I use data from 1991 to calculate a district's distance from Hindi since it is more precise than data from 1961. The 1961 data documents 1652 languages, many of which cannot be assigned a distance from Hindi (in contrast, there are only 114 in 1991 due to prior classification by the Census). In addition, many districts have been divided since 1961, possibly on linguistic lines, adding further noise. I calculate the correlation by aggregating to 1961 district boundaries and omitting speakers of languages I cannot classify.

speakers in western India even before the tenth century CE (Southworth 2005).

The story of one ethnic group provides a telling example. In the tenth century CE, the Gaud Saraswat Brahmins (GSBs) were concentrated on the western coast of India, particularly in Goa. While there is no direct evidence, GSBs claim to originate from Kashmir. In 1351 CE, unrest due to raiding parties sent by the Bahmani Sultan in the Deccan caused some GSB families to migrate down the coast into present-day west and southwest Karnataka (Conlon 1977). Figure 9 provides a map detailing their migration route. The language still spoken by this group, a dialect of Konkani, is only two degrees from Hindi but the main language in Karnataka is five degrees away. Figure 9 also includes a district map of Karnataka detailing the percent of people who speak languages exactly two degrees from Hindi. The population share in the rest of Karnataka averages 1-4% while in the coastal districts in which the GSBs settled, these languages account for 14-29% of the population. Of course, many others speak languages two degrees from Hindi, but the arrival of this group in 1351 speaks to the persistent effect of historical migration on linguistic diversity today.

5.4 Linguistic distance and omitted factors

Finally, we must check that the within-region variation in linguistic distance to Hindi is not correlated with other omitted factors. Most measures of the cost of learning English, e.g. the number of English speakers, are likely to be correlated with unobserved characteristics that influence schooling or export-oriented policies. If a local government cares about access to global opportunities, it may both promote English and provide incentives for FDI. The residual variation in linguistic distance does not suffer from this problem. Distance to Hindi impacts the cost of learning English in a manner that is orthogonal to preferences of local governments and export-oriented motives before 1990. The results in columns 4-5 of table 3 provide evidence against this concern: the tendency for ethnic groups speaking languages further from Hindi to learn English was evident in 1961, before anyone could anticipate trade liberalization in the 1990s. This supports the assumption that these groups were not more

forward-looking or outward-oriented in the 1980s; it is unlikely they anticipated the trade benefits to learning English.

When using other measures, we may also worry about reverse causality: IT firms often set up English training centers. English-language opportunities will not affect a district's linguistic distance to Hindi. Large movements of people may alter the languages spoken in an area, but migration in India is quite infrequent. According to the 1987 National Sample Survey, only 12.3% of individuals in urban areas had moved in the past five years, only 6.8% had moved from a different district and only 2.4% had moved across state lines. These numbers were even smaller in 1999. Nevertheless, I rely on linguistic distance from Hindi from the 1991 census, before trade reforms were implemented.

Next, I confirm that linguistic distance is not correlated with the supply of schools. If it were, we would worry that districts further from Hindi simply had greater preferences for education. In row (1) of table 8, I present estimates from a regression of the log number of schools on linguistic distance to Hindi, log child population and controls in vector Z_j. Neither measure of linguistic distance predicts the number of schools in 1993 or the number of schools offering courses in science or commerce (results not shown).

Finally, districts that are linguistically distant from Hindi do not, by definition, speak the language spoken by the most Indians. One worry would be that these districts were less integrated with the rest of the country and this may have impacted the evolution of industries or interstate trade. In table 8 rows (2)-(8), I present estimates from a regression of the percent of the labor force employed in specific industries on vector Z_j to demonstrate that linguistic distance has had almost no impact on district industrial evolution.[30]

[30] These results are robust to omitting Z_j, using individual-level regressions or defining the outcome as a population share.

6 Impact of linguistic distance from Hindi

6.1 Information technology

Estimating equation (6) indicates a strong positive effect of linguistic distance from Hindi on IT presence (table 9). In Panel A, I drop the ten most populous cities (as of 1991) since IT firms are likely to locate there regardless of English speaking manpower; in Panel B, I include these cities and an interaction with linguistic distance. The cost of learning English, when measured as the weighted average or percent distant speakers, predicts whether any IT firm establishes a headquarters or branch in a district (columns 1-2). An increase in 1 degree from Hindi of the average speaker's mother tongue (three-fourths of a standard deviation) results in a 4% increase in the probability of any IT presence (the dependent variable mean is 12%); a 20% increase in the percent distant speakers (half a standard deviation) increases the probability by 6%. The magnitude of these effects is economically significant: about a fourth of the effect of housing one of 26 elite engineering colleges. Columns 3-4 show that IT headquarters were established earlier in areas linguistically further from Hindi, by approximately one year per 20% distant speakers.[31]

Linguistic distance from Hindi also predicts the log number of headquarters and branches and log employment (columns 5-8); the coefficients are more significant when using percent distant speakers. Twenty percent more distant speakers increases the number of establishments by 10% and employment by almost 45%. A back-of-the-envelope calculation suggests that having 20% more distant speakers increases the percent of English speakers by about 2%. This would attract 0.06 more IT establishments (at the mean of 0.6) and 63 more IT employees (at the mean of 140) in 2003.

The results in Panel B of table 9, when I include the large cities, are nuanced but not

[31] Some districts may be unlikely to receive any IT for other reasons. While these reasons are orthogonal to linguistic distance, I confirm these results by using firm-level data to focus on districts with any IT between 1995 and 2003. I also use the time variation to study whether firms locate in cities linguistically further from Hindi earlier or whether they branch out to smaller cities that are far from Hindi later due to congestion. The results suggest the latter, but there is insufficient data and variation in time.

surprising. Being a large city increases all measures of IT presence, even after controlling for log population. The effect of linguistic distance on the existence of any IT firm and the age of the oldest IT headquarters is somewhat reduced (the interaction term is negative but not significant), but the effect on how many establishments are located in a city is amplified. While having more English speakers has less influence on whether any firm locates in a large city than a small city, it does increase the number of establishments that locate there.

Recall from section 5 that Hindi speakers are as likely to learn English as individuals whose mother tongues are at least 3 degrees away. While we must be cautious in interpreting these results since the percent of Hindi speakers may be correlated with omitted variables, we should also examine the coefficients on the percent of native Hindi speakers. While being in the Hindi belt is defined independently, it is related to the share of Hindi speakers. Districts with more native Hindi speakers do not attract more IT firms, but being in the Hindi belt has a positive and often significant effect. One possible explanation is that the variation in costs of learning English to which IT firms respond depends more on whether a state is in the Hindi belt than on the percent of native Hindi speakers.

6.2 Education

Estimating equation (7), I find that educational attainment rises more in districts with lower costs of learning English (table 10). Panel A uses the weighted average while panel B uses the percent distant speakers as the measure of linguistic distance. Columns 1-2 pool all levels; columns 3-8 stratify the sample by grade. Both measures of linguistic distance predict an increase in enrollment growth. One degree in average distance to Hindi would increase overall growth by 7% over the 9 year period; an increase of 20% distant language speakers would increase enrollment growth by 6%. At the primary and upper primary levels, the coefficients for girls are larger than for boys, but the increase is comparable at the secondary level. This is partly because enrollment for girls starts from a lower level than boys, particularly for younger children, and the outcome is percent improvement.

The magnitudes are large, but not unrealistic. For the average district, they imply that a district with 2% more English speakers would see urban school enrollment grow by 3500 additional students in primary school, 2700 in upper primary and 1300 in secondary school. Recall that from 1993 to 2002, enrollment on average grew by 13000 (21%) in primary school, 9000 (30%) in upper primary and 15000 (60%) in secondary.

As with the IT results, we may also want to explore what happens to districts with more Hindi speakers over the second time period, since Hindi speakers are more likely to learn English than individuals 1 or 2 degrees away. Districts with more native Hindi speakers ceteris paribus exhibit larger increases in enrollment growth; the effect is similar in magnitude to that of distant speakers. As in table 5, the effect of Hindi speakers is more pronounced when linguistic distance is measured as percent distant speakers. This result is reassuring since it confirms that places with more English speakers saw a bigger increase in school enrollment growth using slightly different variation.

There are two avenues through which job opportunities brought by trade liberalization could increase school enrollment. I focus on human capital responses to returns to education. Another channel is through increased family income: it is unlikely that this channel drives all of these results since the greater job opportunities were concentrated among young adults. This should increase enrollment of children at lower grades by more while the results indicate similar, if not bigger, effects at older ages.

6.3 Robustness checks

This section provides robustness checks for the impact of lower costs of learning English on high-skilled exporting opportunities and school enrollment. The first test is to use the percent of schools teaching in the regional mother tongue as a proxy for the cost of learning English and then to instrument with linguistic distance to Hindi. The second test uses an alternate source of variation in English literacy to find the same effects on the growth of exporting opportunities and school enrollment. Third, I provide the results for a specification

check studying school enrollment in rural areas. Lastly, I discuss whether there is sufficient geographic variation in linguistic distance to Hindi to include state fixed effects. The last section also lists other control variables that do not affect the results such as measures of labor market regulation.

6.3.1 Proxy for cost of learning English

Using a proxy for the cost of learning English and using linguistic distance from Hindi of languages spoken in a district as an instrument allows me to use a flexible specification in the first stage. The proxies are the percent of schools that teach in the regional mother tongue at the primary or upper primary level. While this is not a comprehensive measure of the cost of learning English, it is the only data on language instruction available by district. Since English is not a regional mother tongue anywhere in India, this variable acts as a lower bound of schools that do not teach in English.

As expected, the results (not shown) from this two stage least squares estimation are consistent with those from the reduced form regressions. The main contribution of this check is to provide some sense of the magnitudes of these effects. If 10% more schools taught in the mother tongue (slightly less than half a standard deviation), a district would be 10% less likely to have any IT presence, the establishment of IT firms would be delayed by 8 months, the number of branches or headquarters would fall by 10% and employment by 50%.[32] Using the percent of primary schools or the percent of upper primary schools that teach in the mother tongue give similar results. The effect of lower English costs on school enrollment growth is also confirmed when using this alternate estimation strategy. The magnitudes of these results are economically significant: a 10% increase in how many schools teach in the mother tongue would reduce enrollment growth by 12% over 9 years.

[32] I control for the percent of Hindi speakers; using this variable as an additional instrument (to exploit the non-monotonicity) provides very similar results.

6.3.2 English speakers in 1961

Another source of variation in English literacy across India is historical variation in the share of English speakers. From the 1961 census, we know the share of people in each district who speak English as either a first, second or third language.[33] While the share of English speakers today suffers from omitted variables bias - it is likely correlated with unobservable characteristics that influence education and IT growth - the share of speakers historically is less correlated with current unobservables. As with linguistic distance to Hindi, we need to worry less about reverse causality since globalization in the 1990s cannot induce a response from language learners years earlier. However, given the history of education in India - most education was in English during British rule - historical English literacy may be correlated with other historical factors that have lasting effects. Nevertheless, while this variation is not exogenous, it is likely to be correlated with different omitted variables than linguistic distance to Hindi. Estimating specifications (6) and (7) using this variation will probably suffer from different biases than the ones that remain in the estimates described above. It is reassuring that these results confirm those found using linguistic distance to Hindi.

First, I re-estimate equation (6) with the four measures of IT presence, substituting the 1961 share of English speakers among bilinguals for linguistic distance to Hindi (panel A of table 11). As before, I drop the ten most populous districts. A 10% increase in the share of English bilinguals (40% of a standard deviation) increases the probability of any IT presence by 17%, advances the establishment of an IT headquarters by an (insignificant) tenth of a year, and increases the number of branches and employment by 3% and 9% respectively. In panel B, I present similar estimates of equation (7). A 10% increase in the share of English speakers in 1961 would have increased enrollment growth by 5%

[33] Recall that we only have this information at the state level from the 1991 census. My 1961 data only consists of the 6 most spoken languages in the district; in future work, I plan to add the other languages.

6.3.3 School enrollment in rural areas

In the results presented above, I focused on school enrollment growth in urban areas since high-skilled exporting opportunities, especially in IT, are likely to be concentrated in cities. In this subsection, I turn my attention to rural areas of the same districts. Note that my measure of linguistic distance to Hindi does not differ in rural and urban areas of the same district. Thus, rural areas that are linguistically distant from Hindi are in the same districts as the urban areas that experienced more IT growth and school enrollment growth.

It is not obvious what we would expect to happen to rural school enrollment. Since these rural areas are closer to the exporting urban areas, we might expect more rural-urban migration in districts with large shares of distant speakers. If families migrated with young children, we would see a negative effect on schooling in rural areas. If, instead, young people migrated after completing their schooling in rural areas, we may see a positive effect on rural education due to spillovers. The results (in table 12) demonstrate that changes in school enrollment trends after 1993 are not correlated with linguistic distance to Hindi in any direction. The point estimates are small in magnitude and of inconsistent signs. In addition, the percent of native Hindi speakers does not affect rural school enrollment trends.

At the same time, this exercise may constitute a check on the validity of my identification assumptions. Some of the alternate explanations for why school enrollment grew faster in urban areas of linguistically distant districts after 1993 should have affected both urban areas and rural areas of those districts. For example, one explanation is that preferences for education may be correlated with linguistic distance to Hindi causing faster growth in certain districts. If this is true, however, we would expect school enrollment trends or changes in trend to differ in the rural areas of these districts as well. The fact that I do not find evidence of differential changes in rural areas supports my identification strategy. Note that this assumes that languages are spread evenly across rural and urban areas, which may not be true; unfortunately, the data needed to confirm this is not available.

6.3.4 Controlling for other variables

In all the results presented in this paper, I control for regional variation. This is especially important to the identification strategy because unobserved regional differences in linguistic distance from Hindi may otherwise bias my results. We may also want to control for variation across states in other unobserved determinants of IT growth or school enrollment such as state-level trade or education policies.

Note that unobserved state variation would only bias my results if it is correlated with linguistic distance to Hindi. From the geographic variation in figure 1, it is clear that a large part of the raw variation in linguistic distance to Hindi is regional; in figure 8, however, after we account for regional variation, much less of the residual variation appears to be across states. While this is reassuring, it is still important to check if my results are robust to including state fixed effects. Relying solely on within-state variation provides similar results (not shown), although some results become insignificant since there is less power.

The first results that used region fixed effects demonstrated how distance from Hindi affects language acquisition for different linguistic groups. These results are fully robust to including state fixed effects. The results in table 9 demonstrating that IT firms locate earlier and more often and grow faster in linguistically distant districts are also robust to isolating within-state variation. The point estimates in table 10 confirming that school enrollment grows faster in districts that speak languages further from Hindi, are not affected systematically (some fall, some rise), but the standard errors double or triple. This is likely due to limited district variation within state.

My preferred specification also controls for a number of district-level demographic and socioeconomic variables prior to trade reforms in India, Z_j. Excluding this vector from the regressions does not affect any of the results. Similarly, the results are robust to adding additional variables to Z_j. For example, it is possible that variation in labor regulation across states in India may affect where IT firms locate. This is unlikely to matter since turnover in these firms is remarkably high with firms raiding each other and many employees migrating

abroad. Nevertheless, I confirm that the results are robust to controlling for pro-worker regulation using data from Besley and Burgess (2004). I do not use this control variable in my main specification since it is not available for all states. In measuring engineering college presence, I count only the locations of the 26 elite engineering colleges, because district-level data on all engineering colleges are not of the same quality. I explored alternative measures of engineering college presence by using the list of accredited engineering programs from the National Board of Accreditation of the All India Council for Technical Education. Each program was affiliated with a college; the program was assigned to a district based on the address of the college. Unfortunately, retrospective data was unavailable, so I only include those colleges that were established prior to 1990. Controlling for the number of engineering colleges with any accredited programs does not alter any of my results.

7 Returns to education

Having demonstrated that districts with greater English literacy experienced greater IT and school enrollment growth after trade liberalization, I now turn to the general equilibrium implications of this human capital response for skilled wage premiums. As noted above, the theoretical prediction is ambiguous and depends on the relative magnitudes of the demand shocks. We must be cautious in interpreting these results for a number of reasons. First, since the data do not distinguish between English-medium education and local language instruction, I focus on average returns to education. Second, the wage data from the National Sample Surveys is the best available data on Indian wages over this time period, but is not particularly suited for this study since the sample affected by export-related jobs is quite small. Researchers are also skeptical of this wage data since it is self-reported and not verifiable; many individuals work in the informal sector. Nevertheless, I provide suggestive

evidence on returns to education, by estimating

$$\begin{aligned}\log(wage_n) &= \alpha_0 + \beta_1' D_j \cdot I(t=1999) + \beta_2' D_j \cdot I(t=1999) \cdot HS_n \\ &+\beta_3' D_j \cdot I(t=1999) \cdot C_n + \alpha_1' D_j \cdot HS_n + \alpha_2' D_j \cdot C_n \\ &+\alpha_3 I(t=1999) \cdot HS_n + \alpha_4 I(t=1999) \cdot C_n \\ &+\alpha_5 HS_n + \alpha_6 C_n + \alpha_7' Y_n + \alpha_8' W_j \cdot I(t=1999) + \gamma_j + \gamma_t + \gamma_{gt} + \mu_n\end{aligned}$$

where $wage_n$ is wage earnings per week of individual n in district j at year $t \in \{1987, 1999\}$, HS_n and C_n are indicators for whether individual n has completed high school or college, respectively and Y_n and W_j contain individual and district characteristics. I only include nonzero wage earners. Y_n includes age, age squared, gender, a dummy for being married and for having migrated. At the district level, W_j includes the percent of native English speakers, whether the state is in the Hindi belt, the distance to the closest big city, whether the district is on the coast and predicted labor demand. To account for the non-monotonicity in linguistic distance, I include the percent of native Hindi speakers interacted with $I(t=1999)$, HS_n, C_n and all triple interactions. I include district fixed effects and region fixed effects interacted with time, cluster the standard errors by district and weight the observations.[34]

I find evidence that skilled wage premiums rose by less in districts with lower English costs from 1987 to 1999, particularly for high school graduates (see table 13). The coefficients β_2 and β_3 are always negative, but not always significant. Note that the magnitudes are economically significant. The wage premium for high school graduates rises by 5% less over 12 years per degree of linguistic distance, relative to a premium of 54% for high school graduates in 1987. Stratifying the sample by age and gender reveals that the results are driven by men and older workers. The results are robust to including state fixed effects.

[34]The results are robust to including the district vector of controls, Z_j, interacted with $I(t=1999)$, but I do not include these in the main specification since the controls are measured in 1987.

8 Conclusion

In this paper, I studied how districts with differing abilities to take advantage of global opportunities responded to the common shock of globalization. I exploited exogenous variation in the cost of learning English, a skill that is particularly relevant for export-related jobs. I first showed that linguistic distance from Hindi predicts whether individuals learn English. One clear benefit of promoting a global language is access to global job opportunities such as in IT. I showed that IT firms were more likely to set up in districts further from Hindi. I next demonstrated that these districts experienced greater increases in school enrollment growth, but smaller growth in the skilled wage premium.

There are two important implications of these results. The first relates to how countries can mitigate any adverse effects of globalization on inequality. During trade liberalization, governments should consider policies to help individuals acquire the skills necessary for global opportunities. The ability to speak English is one such skill. The second implication is the evidence for a long run effect of globalization: factor supply may mitigate the increase in wage inequality brought by trade reforms.

Trade liberalization may also have impacted marriage rates and fertility. IT firms employ more women relative to traditional Indian firms. The male-female ratio among those working was 80:20 in 1987, but 77:23 in software firms and 35:65 in business processing firms (NASSCOM 2004). Anecdotal evidence suggests that women work in call centers between school and getting married, which might increase the age of first marriage and the probability that women continue to work past marriage, potentially impacting fertility rates. In the long run, we might see an impact of future job opportunities on child health. If parents think their daughters are more likely to work before marriage they may invest more in their daughters' health as well as education. The impact on these other measures of development is an important avenue for future research.

9 References

Angrist, J., A. Chin and R. Godoy (2006). "Is Spanish-Only Schooling Responsible for the Puerto Rican Language Gap?," *NBER Working Paper* 12005, National Bureau of Economic Research, Cambridge, MA.

Angrist, J. and V. Lavy (1997). "The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco," *Journal of Labor Economics* 15(1), S48-76.

Arora, A. and A. Gambardella (2004). ″The Globalization of the Software Industry: Perspectives and Opportunities for Developed and Developing Countries." *NBER Working Paper* 10538, National Bureau of Economic Research, Cambridge, MA.

Arora, A. and S. Bagde (2007). "Private investment in human capital and Industrial development: The case of the Indian software industry" (mimeo) Carnegie Mellon University.

Attanasio, O., P. Goldberg and N. Pavcnik (2004). "Trade Reforms and Wage Inequality in Colombia," *Journal of Development Economics* 74, 331-366.

Attanasio, O. and M. Szekely (2000). "Household Saving in East Asia and Latin America: Inequality Demographics and All That", in B. Pleskovic and N. Stern (eds.), Annual World Bank Conference on Development Economics 2000. Washington, DC: World Bank.

Bartik, T. (1991). Who Benefits from State and Local Economic Development Policies? Kalamazoo: W.E. Upjohn Institute for Employment Research.

Besley, T. and R. Burgess (2004). "Can Labor Regulation Hinder Economic Performance? Evidence from India," *The Quarterly Journal of Economics* 119(1), 91-134.

"Busy signals: Too many chiefs, not enough Indians," *The Economist*, September 8, 2005.

"Can India Fly? A Special Report," *The Economist*, June 3-9, 2006.

Clingingsmith, D. (2006). "Bilingualism, Language Shift and Economic Development in India, 1931-1961." (mimeo) Harvard University.

Conlon, F. (1977). *A Caste in a Changing World*. Berkeley, CA: University of California Press.

Cragg, M.I. and M. Epelbaum (1996). "Why Has Wage Dispersion Grown in Mexico? Is It Incidence of Reforms or Growing Demand for Skills?" *Journal of Development Economics* 51, 99-116.

Dyen, I., J. Kruskal and P. Black (1997). FILE IE-DATA1. Available at http://www.ntu. edu.au /education/langs/ielex/HEADPAGE.html.

Edmonds, E., N. Pavcnik and P. Topalova (2007). "Trade Adjustment and Human Capital Investments: Evidence from Indian Tariff Reform." *NBER Working Paper* No. 12884, National Bureau of Economic Research, Cambridge, MA.

Feenstra, R.C. and G. Hanson (1996). "Foreign Investment, Outsourcing and Relative Wages." In R.C. Feenstra, G.M. Grossman and D.A. Irwin, eds., The Political Economy of Trade Policy: Papers in Honor of Jagdish Bhagwati, MIT Press, 89-127.

Feenstra, R.C. and G. Hanson (1997). "Foreign Direct Investment and Relative Wages: Evidence from Mexico's Maquiladoras." *Journal of International Economics*, 42(3), 371-393.

Feliciano, Z. (1993). "Workers and Trade Liberalization: The Impact of Trade Reforms in Mexico on Wages and Employment." (mimeo) Harvard University.

Goldberg, P. and N. Pavcnik (2004). "Trade, Inequality, and Poverty:What Do We Know? Evidence from Recent Trade Liberalization Episodes in Developing Countries," Brookings Trade Forum, Washington, DC: Brookings Institution Press: 223–269.

Hanson, G. and A. Harrison (1999). "Trade, Technology and Wage Inequality in Mexico." *Industrial and Labor Relations Review* 52(2), 271-288.

Hohenthal, A. (2003). "English in India; Loyalty and Attitudes," *Language in India,* **3**, May 5.

Kamat, A. (1985). Education and Social Change in India. Bombay: Somaiya Publications.

Karnik, K., ed. (2002). *Indian IT Software and Services Directory 2002.* National Association of Software and Service Companies, New Delhi.

Kremer, M. and E. Maskin (2006). "Globalization and Inequality." (mimeo) Harvard University.

Lang , K. and E. Siniver (2006). "The Return to English in a Non-English Speaking Country: Russian Immigrants and Native Israelis in Israel," *NBER Working Paper* 12464, National Bureau of Economic Research, Cambridge, MA.

Levinsohn, J. (2004). "Globalization and the Returns to Speaking English in South Africa." *NBER Working Paper* 10985. National Bureau of Economic Research, Cambridge, MA.

Lindert, P. and J. Williamson (2001). "Does Globalization Make the World More Unequal?" *NBER Working Paper* No. 8228, National Bureau of Economic Research, Cambridge, MA.

McAlpin, D. (1981). *Proto-Elamo-Dravidian: The Evidence and its Implications*, Philadelphia, PA: The American Philosophical Society.

Mehta, D., ed. (1995). *Indian Software Directory 1995-1996.* New Delhi: National Association of Software and Service Companies.

Mehta, D., ed. (1998). *Indian Software Directory 1998.* New Delhi: National Association of Software and Service Companies.

Mehta, D., ed. (1999). *Indian IT Software and Services Directory 1999-2000.* New Delhi: National Association of Software and Service Companies.

Munshi, K. and M.Rosenzweig (2006). "Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy," *American Economic Review* 96(4), 1225-1252.

NASSCOM, 2004. *Strategic Review 2004.* New Delhi: National Association of Software and Service Companies, 185-194.

Nurullah, S. and J. Naik, 1949. A Student's History of Education in India, 1800-1947. Bombay: Macmillan and Company Limited.

Panagariya, A. (2004). "India's Trade Reform: Progress, Impact and Future Strategy" International Trade 0403004, EconWPA.

Robbins, D. (1995a). "Earnings Dispersion in Chile after Trade Liberalization." Harvard Institute for International Development, Cambridge, MA.

Robbins, D. (1995b). "Trade, Trade Liberalization, and Inequality in Latin America and East Asia: Synthesis of Seven Country Studies." Harvard Institute for International Development, Cambridge, MA.

Robbins, D. (1996a). "Stolper-Samuelson (Lost) in the Tropics: Trade Liberalization and Wages in Colombia 1976–94." Harvard Institute for International Development, Cambridge, MA.

Robbins, D. (1996b). "HOS Hits Facts: Facts Win. Evidence on Trade and Wages in the Developing World." Harvard Institute for International Development, Cambridge, MA.

Robbins, D. and T. Gindling (1997). "Educational Expansion, Trade Liberalisation, and Distribution in Costa Rica." In Albert Berry, ed., Poverty, Economic Reform and Income Distribution in Latin America. Boulder, Colo.: Lynne Rienner Publishers.

Robbins, D., M. Gonzales, and A. Menendez (1995). "Wage Dispersion in Argentina, 1976–93: Trade Liberalization amidst Inflation, Stabilization, and Overvaluation." Harvard Institute for International Development, Cambridge, MA.

Sanchez-Paramo, C. and N. Schady (2003): "Off and Running? Technology, Trade, and the Rising Demand for Skilled Workers in Latin America," World Bank Policy Research Working Paper 3015.Washington, DC: World Bank.

Southworth, F. (2005). *Linguistic Archaeology of South Asia.* New York: RoutledgeCurzon.

Swadesh, M. (1972). "What is glottochronology?" In M. Swadesh, The origin and diversification of languages. London: Routledge & Kegan Paul: 281-284.

Topalova, P. (2004). "Trade Liberalization and Firm Productivity: The Case of India." *IMF Working Paper* 04/28, International Monetary Fund.

Topalova, P. (2005). "Trade Liberalization, Poverty, and Inequality: Evidence from Indian Districts." *NBER Working Paper* 11614, National Bureau of Economic Research, Cambridge, MA.

Tyler, S. (1968). "Dravidian and Uralian: The lexical evidence," *Language* 44, 798-812.

United Nations Development Programme, Human Development Report 2004: Cultural Liberty in Today's Diverse World, New York: Oxford University Press, 2004.

Wei, S. and Y. Wu (2001). "Globalization and Inequality: Evidence from Within China." *NBER Working Paper* 8611, National Bureau of Economic Research, Cambridge, MA.

Wood, A. (1997). "Openness and Wage Inequality in Developing Countries: The Latin American Challenge to East Asian Conventional Wisdom." *World Bank Economic Review* 11(1), 33-57.

10 Theory Appendix

10.1 Proof of Proposition 1

Since $w_E^* = w_H^*$, we know that the supply of English skilled workers does not depend on the wages. To be in this equilibrium, the demand for English skilled labor from X production must be less than or equal to this supply; English speakers not working in the X industry can work in Y production and earn the same wage. From equations (4) and (1), we can show that

$$\frac{\alpha_H}{\alpha_L} P \left(1 - \frac{w_H^* - w_U^* - t w_U^*}{w_U^*}\right) + F \left(\frac{w_H^*}{p_X (1-\beta)}\right)^{-\frac{1}{\beta}} = P \left(\frac{w_H^* - w_U^* - t w_U^*}{w_U^*}\right)$$

Substituting the zero profit condition for Y ($w_H^* = \frac{1}{\alpha_H} - \frac{a_L}{\alpha_H} w_U^*$) into this expression implicitly solves for w_U^*:

$$(2+t)\left(1 + \frac{\alpha_H}{a_L}\right) + \frac{\alpha_L}{\alpha_H} = \frac{1}{w_U^*}\left(\frac{1}{\alpha_H} + \frac{1}{\alpha_L}\right) - \frac{F}{P}\left(\frac{1 - \alpha_L w_U^*}{\alpha_H p_X (1-\beta)}\right)^{-\frac{1}{\beta}} \tag{8}$$

Note that this expression does not depend on μ_j. Thus, $\frac{dw_U^*}{d(\mu_j - 1)} = 0$. The variables, w_H^*, r_F^*, Y^*, X^*, $\widehat{q}^*$, ED^* can be written as functions of w_U^* which also do not depend on μ_j. Comparative statics on E^* and H^* follow easily from the expressions, $E^* = P\frac{t}{\mu_j - 1}$ and $H^* = P\left(\frac{1}{w_U^* \alpha_H} - \frac{\alpha_L}{\alpha_H} - 1 - t - \frac{t}{\mu_j - 1}\right)$.

10.2 Proof of Proposition 2

To be in this equilibrium, the demand for English skilled labor from X production must equal the supply; if there was excess supply, w_E^* would fall to increase firm profits and if there was excess demand, w_E^* would rise to attract additional English workers. From equations (4) and (2), we have

$$\frac{a_L}{\alpha_H} P \left(\frac{w_H^* - w_U^* - t w_U^*}{w_U^*} - \frac{w_U^* t + w_E^* - w_H^*}{w_U^* (\mu_j - 1)}\right) = P \left(1 - \frac{w_H^* - w_U^* - t w_U^*}{w_U^*}\right)$$

Substituting in for w_H^* and solving for w_E^* gives us

$$w_E^* = \frac{1}{\alpha_H} + (\mu_j - 1)\left(\frac{1}{\alpha_L} + \frac{1}{\alpha_H}\right) - w_U^* \left[\frac{a_L}{\alpha_H} + t + (\mu_j - 1)\left[\frac{a_L}{\alpha_H} + \left(\frac{a_H}{\alpha_L} + 1\right)(2+t)\right]\right] = A - w_U^* B$$

Plugging these expressions for w_H^* and w_E^* into equation (3), we get

$$0 = \left(\frac{P}{F}\right)^{-\beta}\left(\frac{1}{w_U^*}\left(\frac{1}{\alpha_L} + \frac{1}{\alpha_H}\right) - \left[\frac{a_L}{\alpha_H} + \left(\frac{a_H}{\alpha_L} + 1\right)(2+t)\right]\right)^{-\beta} - \frac{A - w_U^* B}{p_X (1-\beta)} = G\left(w_U^*; \mu_j\right) \tag{9}$$

Thus,

$$\frac{dw_U^*}{d\left(\mu_j-1\right)}=-\frac{\frac{\delta G}{\delta\left(\mu_j-1\right)}}{\frac{\delta G}{\delta w_U^*}}>0$$

Writing the other variables in terms of w_U^* and differentiate with respect to μ_j-1 is simple. The variables w_E^*, Y^*, H^* are rising in μ_j, while $w_H^*, r_F^*, X^*, E^*, ED^*$ are falling in μ_j. It is similarly straightforward to construct examples where $\widehat{q}^*$ is greater in a district with a higher μ_j and examples where $\widehat{q}^*$ is smaller in the higher cost district.

10.3 Propositions 3 and 4

Proposition 3 *$w_E^*=w_H^*$ holds if and only if*

$$F\leq P\frac{t}{\mu_j-1}\left[\frac{\alpha_L}{\alpha_H}\frac{1}{p_X\left(1-\beta\right)}\left(\frac{1}{\alpha_L}-\frac{\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)}{\left[\left(2+t\right)\left(1+\frac{\alpha_H}{\alpha_L}\right)+\frac{\alpha_L}{\alpha_H}\right]+\frac{t}{\mu_j-1}}\right)\right]^{\frac{1}{\beta}}=\overline{F}\left(\mu_j\right) \tag{10}$$

First, I prove that if $w_E^*=w_H^*$, condition (10) holds. From Proposition 1, we know

$$\frac{F}{P}\left(\frac{w_E^*}{p_X\left(1-\beta\right)}\right)^{-\frac{1}{\beta}}=\frac{F}{P}\left(\frac{w_H^*}{p_X\left(1-\beta\right)}\right)^{-\frac{1}{\beta}}=\frac{F}{P}\left(\frac{1-\alpha_L w_U^*}{\alpha_H p_X\left(1-\beta\right)}\right)^{-\frac{1}{\beta}}\leq\frac{t}{\mu_j-1} \tag{11}$$

From the proof of Proposition 1, equation (8), we can write

$$\frac{F}{P}\left(\frac{1-\alpha_L w_U^*}{\alpha_H p_X\left(1-\beta\right)}\right)^{-\frac{1}{\beta}}=\frac{1}{w_U^*}\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)-\left(2+t\right)\left(1+\frac{\alpha_H}{a_L}\right)-\frac{\alpha_L}{\alpha_H}\leq\frac{t}{\mu_j-1}$$

Solving for w_U^*, we get

$$w_U^*\geq\frac{\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)}{\left(2+t\right)\left(1+\frac{\alpha_H}{a_L}\right)+\frac{\alpha_L}{\alpha_H}+\frac{t}{\mu_j-1}} \tag{12}$$

Solving (11) for F and plugging in (12), we get

$$F\leq P\frac{t}{\mu_j-1}\left[\frac{\alpha_L}{\alpha_H}\frac{1}{p_X\left(1-\beta\right)}\left(\frac{1}{\alpha_L}-\frac{\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)}{\left[\left(2+t\right)\left(1+\frac{\alpha_H}{\alpha_L}\right)+\frac{\alpha_L}{\alpha_H}\right]+\frac{t}{\mu_j-1}}\right)\right]^{\frac{1}{\beta}}$$

Next I prove that if $F\leq\overline{F}\left(\mu_j\right)$, then $w_E^*=w_H^*$ by contradiction. We know that $w_E^*\not<w_H^*$ since English skilled workers can always take jobs as Hindi skilled workers. Suppose $w_E^*>w_H^*$. From condition (3), we know that

$$\begin{aligned}F &= \left(\frac{w_E^*}{p_X\left(1-\beta\right)}\right)^{\frac{1}{\beta}}P\left(\frac{t}{\mu_j-1}+\frac{1}{\left(\mu_j-1\right)}\frac{w_E^*-w_H^*}{w_U^*}\right)\\ &> \left(\frac{w_E^*}{p_X\left(1-\beta\right)}\right)^{\frac{1}{\beta}}P\left(\frac{t}{\mu_j-1}\right)>P\frac{t}{\mu_j-1}\left(\frac{1}{p_X\left(1-\beta\right)}\right)^{\frac{1}{\beta}}\left(\frac{1-\alpha_L w_U^*}{\alpha_H}\right)^{\frac{1}{\beta}}\end{aligned}$$

where the first inequality is due to $w_E^* - w_H^* > 0$ and the second is due to $w_E^* > w_H^* = \frac{1-\alpha_L w_U^*}{\alpha_H}$. Putting this together with $F \leq \overline{F}\left(\mu_j\right)$ and rearranging terms, we can show that

$$\frac{1}{w_U^*}\left(\frac{1}{\alpha_L}+\frac{1}{\alpha_H}\right) < \left[(2+t)\left(1+\frac{\alpha_H}{\alpha_L}\right)+\frac{\alpha_L}{\alpha_H}\right]+\frac{t}{\mu_j-1}$$

However, this contradicts what we know from the proof of Proposition 2, equation (9)

$$\frac{1}{w_U^*}\left(\frac{1}{\alpha_L}+\frac{1}{\alpha_H}\right)-\left[\frac{a_L}{\alpha_H}+\left(\frac{a_H}{\alpha_L}+1\right)(2+t)\right] = \frac{F}{P}\left(\frac{A-w_U^* B}{p_X(1-\beta)}\right)^{-1/\beta} = \frac{w_U^* t + w_E^* - w_H^*}{w_U^*\left(\mu_j-1\right)} > \frac{t}{\left(\mu_j-1\right)}$$

where the second equality is from the fact the English skilled labor market must clear and the inequality is from $w_E^* - w_H^* > 0$. Thus, $w_E^* = w_H^*$.

Proposition 4 *In both case A and case B: the amount of X produced, total education ED^*, and the average return to education, $\widehat{q}^*$, are all increasing in F.*

The proof follows directly from the algebra in the proofs of propositions 1 and 2.

11 Data Appendix

11.1 Information technology

I collected and coded data on IT firms from the National Association of Software and Service Companies (NASSCOM) directories published in 1995, 1998, 1999, 2002 and 2003. These directories contain self-reported firm level data on the location of firm headquarters and branches, and the number of employees. According to NASSCOM, the sample accounts for 95% of industry revenue in most years (Mehta 1995, 1998, 1999; Karnik 2002). While the data is self-reported, firms have no reason to under-report their performance since IT firms receive generous tax exemptions.

11.2 Enrollment and language of instruction in 1993, 2002

Data on school enrollment comes from the Sixth and Seventh All India Educational Surveys (SAIES), conducted by the National Council of Educational Research and Training, which began in September 1993 and September 2002 respectively. The surveys collect school level data on enrollment, facilities, languages taught, courses available, teacher qualifications and other aspects of education. The only data currently available from the 2002 data is enrollment, by grade and gender. Some data from 1993, such as languages taught, are only available at the state level. All data is separated into urban and rural setting.

11.3 Employment, enrollment in 1987 and district-level controls

Data on employment and returns to education are from the National Sample Surveys (NSS) conducted in 1987-1988 and 1999-2000.[35] The NSS provides individual-level information on wages paid in cash and in-kind as well as employment status, industry and occupation codes and observation weights. Employment status includes working in household enterprises (self-employed), as

[35] NSS surveys were conducted in 1983-1984 and 1993-1994, but district identifiers are not available.

a helper in such an enterprise, as a regular salaried/wage employee and as casual day labor. In addition, the data conveys whether individuals are seeking work, attending school or attending to domestic duties. I examine employment in agriculture, manufacturing, wholesale/retail/repair, hotel & restaurant services, transport services, communications (post and courier), financial intermediate/insurance/real estate and other services (education, health care, civil).

This data was also used to calculate 1987 school enrollment and district-level controls. I construct district-level measures of grade school enrollment at the primary, upper primary and secondary levels. The NSS also contains household-level information on household structure, demographics, employment, education, expenditures, migration and assets, from which I calculate district averages of all control variables mentioned above such as household wage income, the percent of working-age adults who are engineers, the percent Muslim, the percent who regularly travel by train and the percent of households that have electricity.

11.3.1 Predicted labor demand growth

Using this NSS data and following Bartik (1991), I calculate the proxy for the growth in labor demand for educated workers using the formula

$$\widehat{\varepsilon}_{it}^{E} = \sum_{j=1}^{54} \left(\frac{\widetilde{e}_{-i,j,1983}^{E}}{\widetilde{e}_{-i,j,1983}} \right) \left(\frac{e_{i,j,1983}}{e_{i,1983}} \right) \left(\frac{\widetilde{e}_{-i,j,t} - \widetilde{e}_{-i,j,t-1}}{\widetilde{e}_{-i,j,t-1}} \right)$$

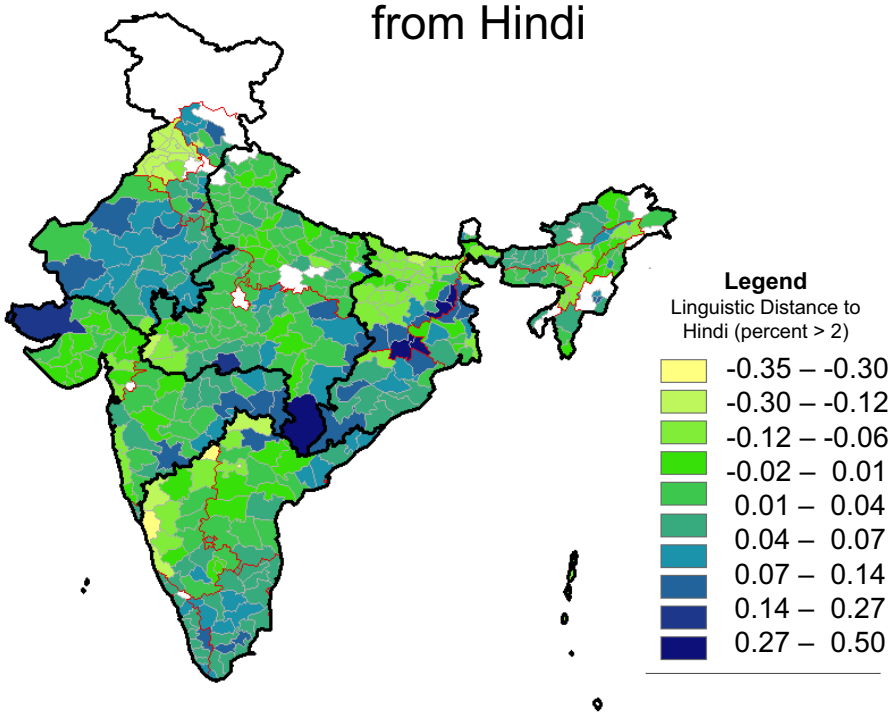
where $\widehat{\varepsilon}_{it}^{E}$ is the predicted skilled labor demand growth from year $t-1$ to t in industry j for area i, e denotes employment and $\widetilde{e}_{-i}$ denotes employment outside area i. The superscript E indicates workers with at least a high school degree. Specifically, $\widetilde{e}_{-i,j,t}$ is national employment outside area i in industry j in year t, $\widetilde{e}_{-i,j,t}^{E}$ is national employment outside area i in industry j for educated workers, and $e_{i,j,t}$ is employment in area i in industry j. NSS data for 1983 does not contain district identifiers so I use a larger aggregation of multiple districts that is smaller than a state. I match 54 2-digit categories from the National Industrial Classification of 1970, 1987 and 1998.

Thus, the proxy is a weighted average of growth rates of national industry employment where the weights are the 1983 share of multi-district employment in each industry. To predict labor demand for educated workers in particular, I further weight these growth rates using the share of employment with at least a high school education.

11.4 Other controls

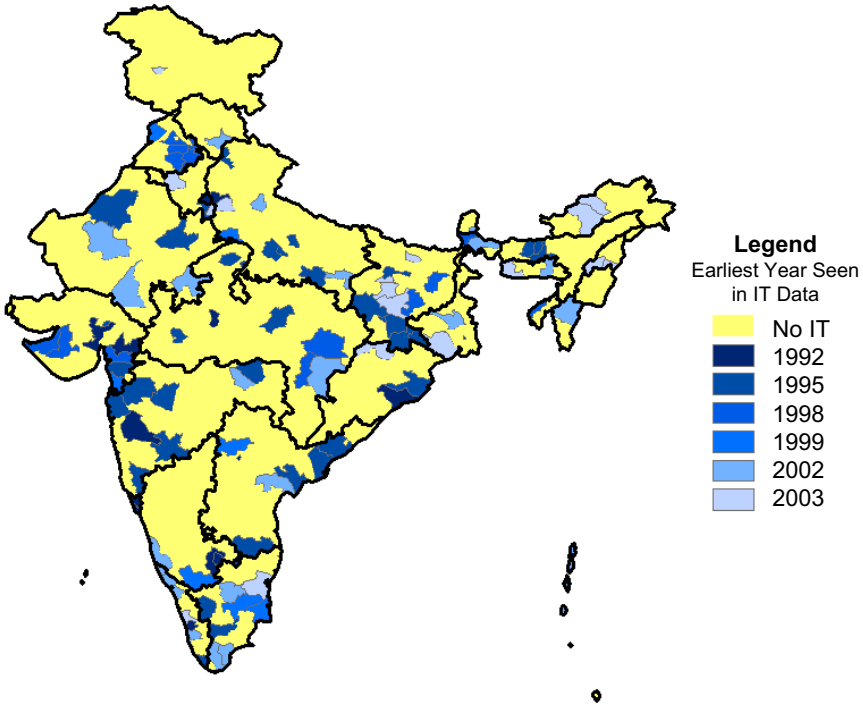
Using latitude and longitude data, I calculate the distance from each district to the closest of the 10 biggest cities in India and to the closest airport operated by the Airports Authority of India. As a measure of elite engineering college presence I count the number of Indian Institutes of Technology and Regional Engineering Colleges (now called the National Institutes of Information Technology) in each district. All of them were established prior to 1990, although some were not given REC/NIIT status until the late 1990s.

Figure 1: Residual Variation in Linguistic Distance from Hindi



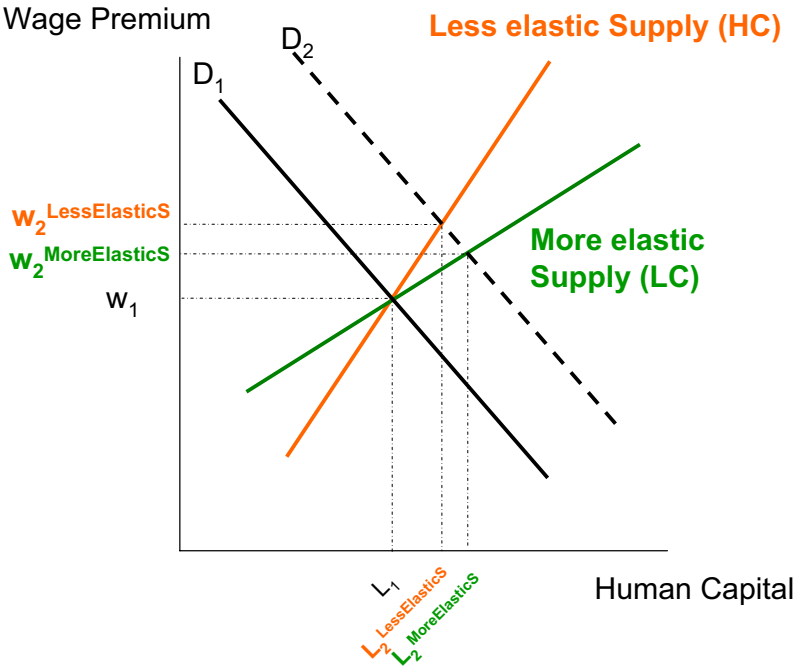
Note: Districts in this map are shaded according to the residual from a regression of linguistic distance to Hindi (the percent of people speaking languages sufficiently distant from Hindi) on region fixed effects and district-level control variables measured prior to 1991. Thick black lines indicate regions, while red lines indicate states.

Figure 2: Growth of IT Industry



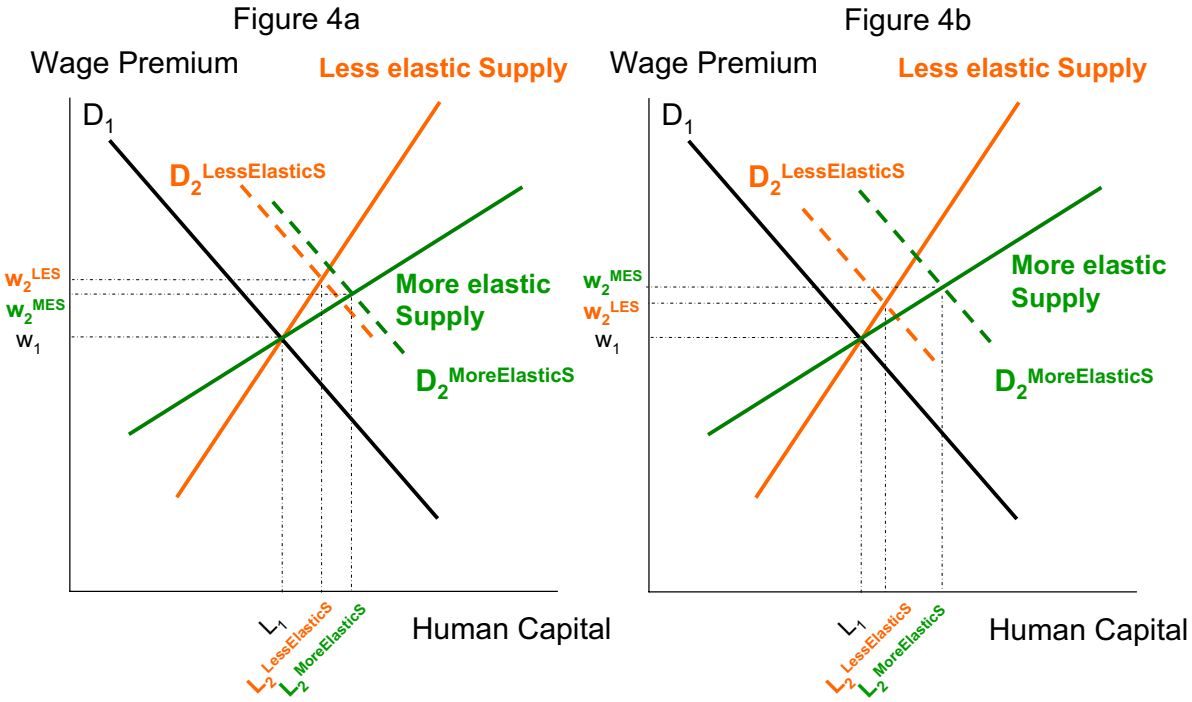
Note: Districts in this map are shaded according to the earliest year an IT establishment from the NASSCOM data is found in the district. Thick black lines indicate regions, while red lines indicate state boundaries.

Figure 3: Effect of Identical Demand Shocks



Note: Identical demand shocks for workers with human capital result in a larger increase in human capital accumulation but a smaller increase in skilled wage premiums in a district with a more elastic supply.

Figure 4: Effect of Different Demand Shocks



Note: Different demand shocks for human capital result in a larger increase in human capital accumulation in a district with a more elastic supply but the relative change in skilled wage premiums is ambiguous.

Figure 5: Chart of Language Distances from Hindi

Note: This figure depicts languages as they get further from Hindi. Source: Jay Jasanoff, Diebold Professor of Indo-European Linguistics and Philology at Harvard University.

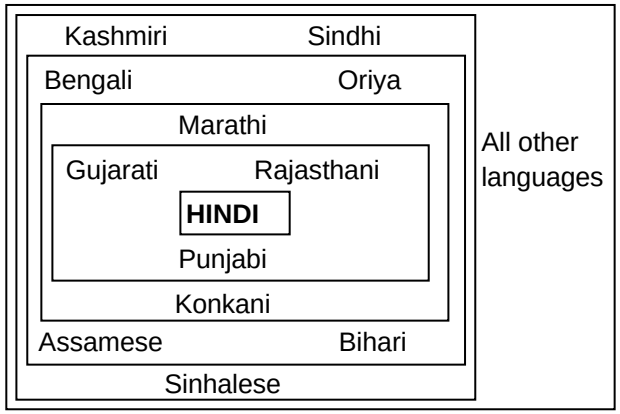
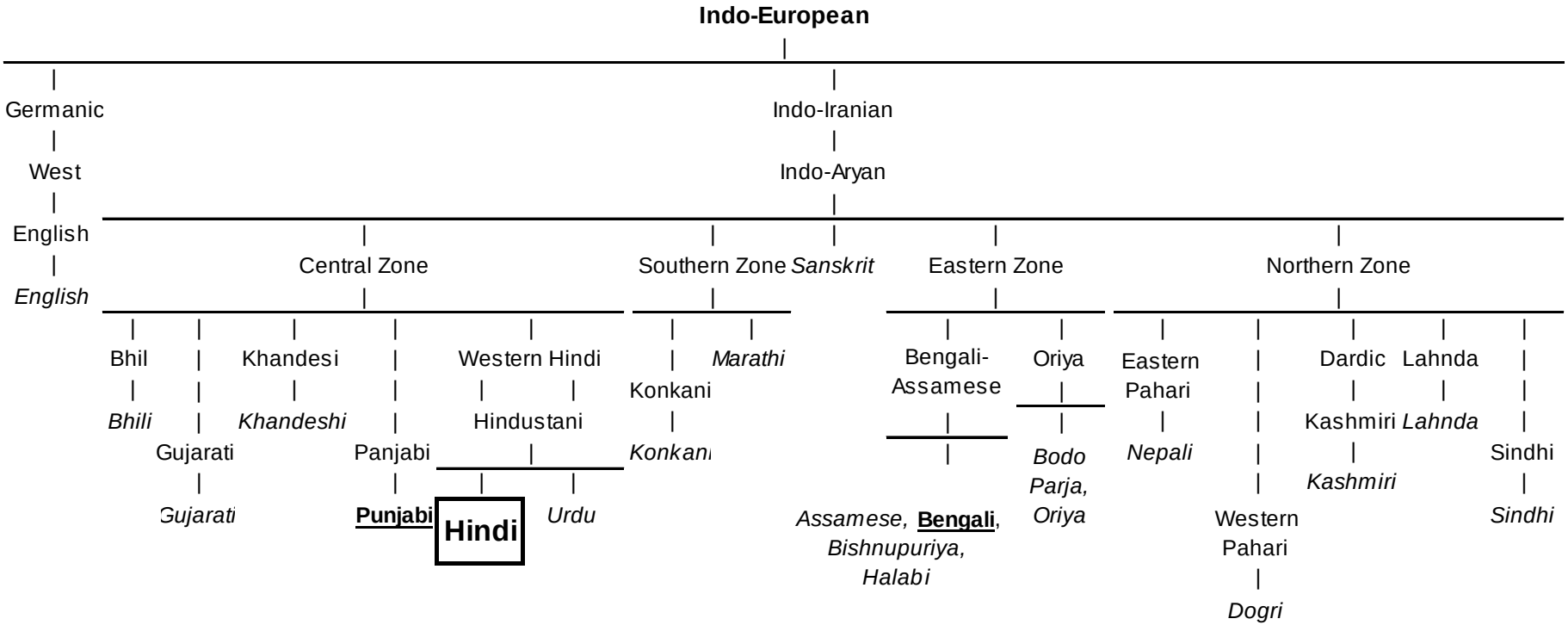
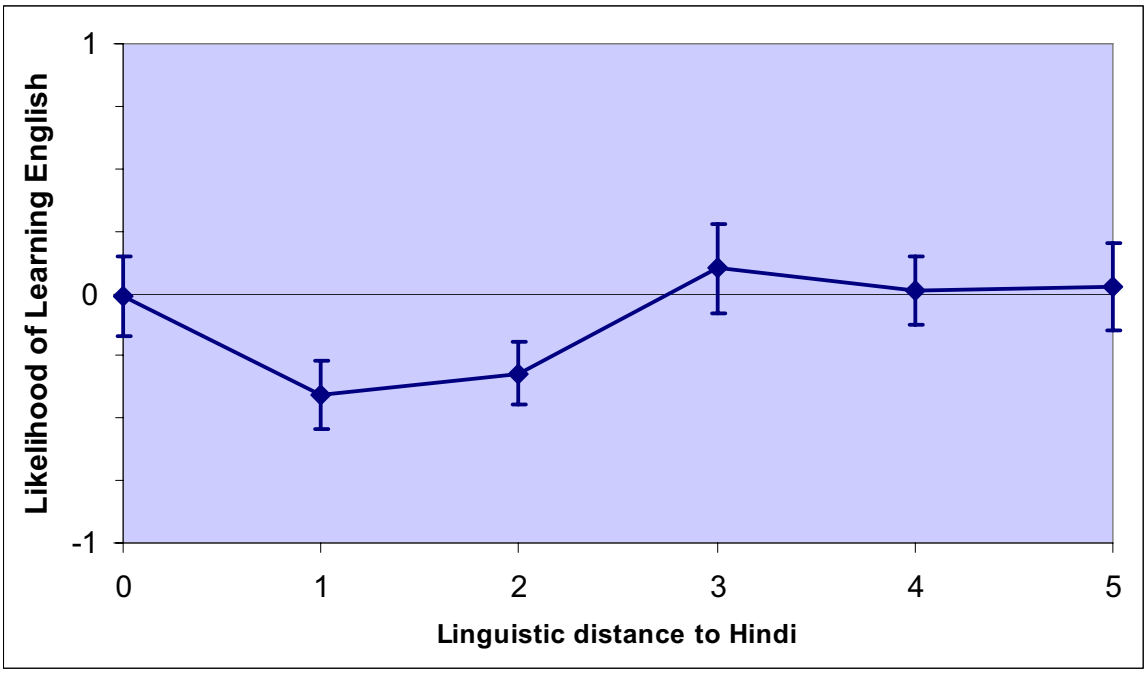


Figure 6: Family Tree of Indo-European Languages in India



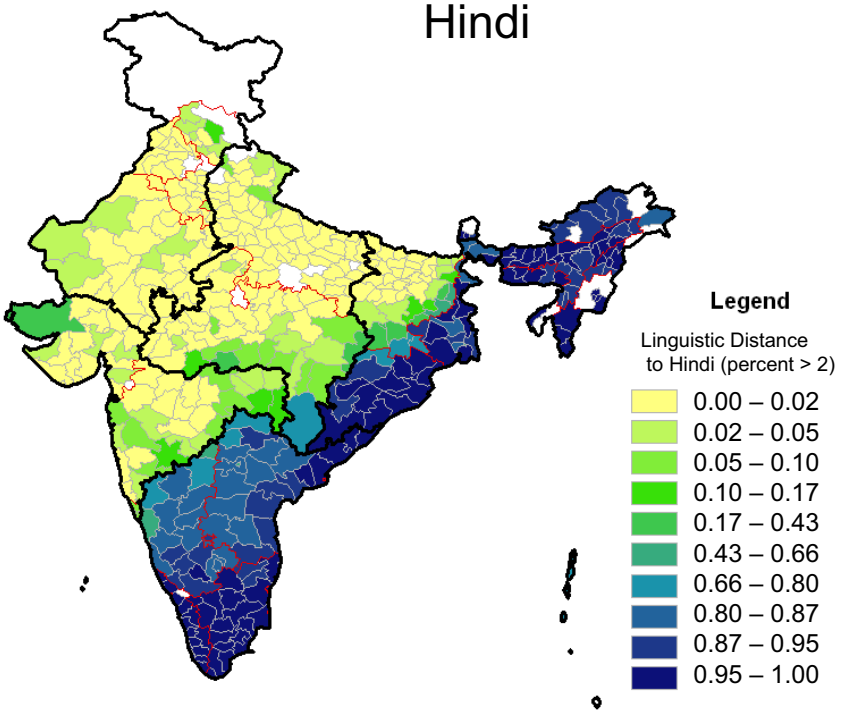
Note: This figure depicts an extract from the family tree of Indo-European languages that are found in India. Trees for the other language families can be found online at www.ethnologue.com.

Figure 7: Linguistic Distance from Hindi and Propensity to Learn English



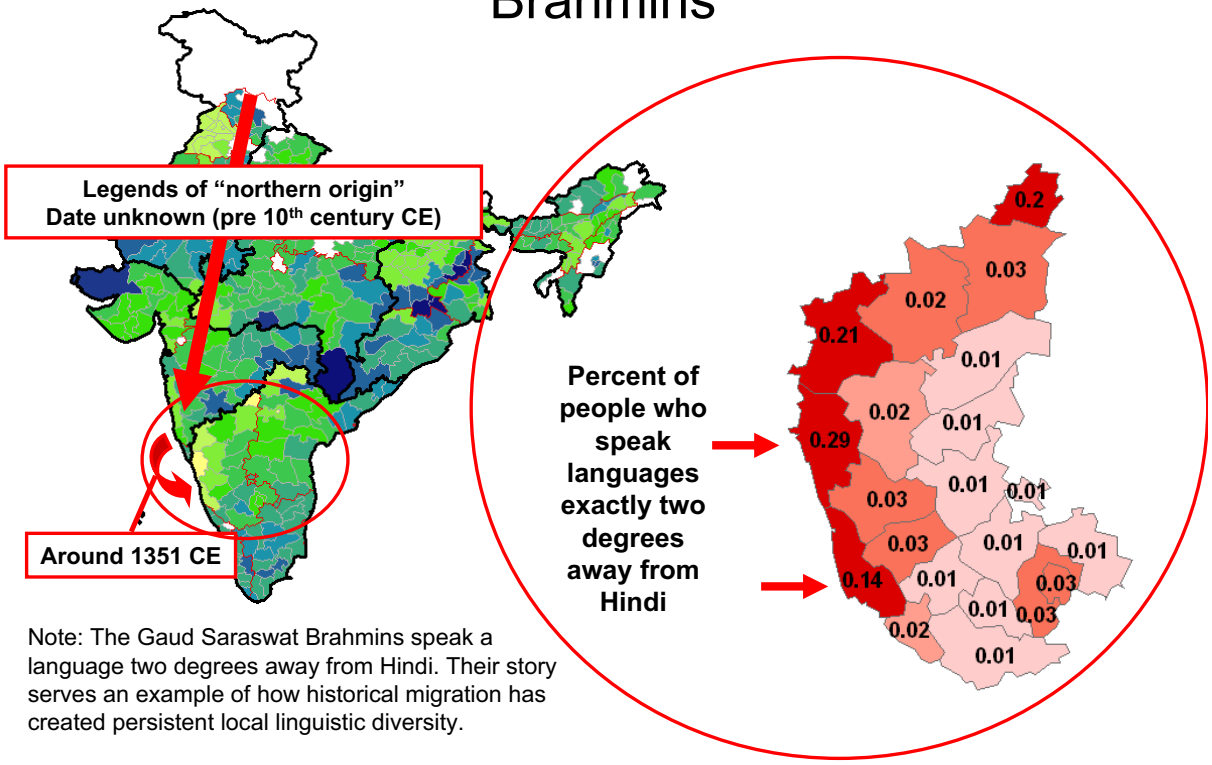
Note: This graph plots the coefficients on each distance from Hindi with a 95% confidence interval from a regression of the percent of multilinguals who learn English on linguistic distance to Hindi (Col 1 in table 3).

Figure 8: Raw Variation in Linguistic Distance from Hindi



Note: Districts in this map are shaded according to the percent of people in a district who speak languages at least three degrees from Hindi. Thick black lines indicate regions, while red lines indicate state boundaries.

Figure 9: Migration Route of Gaud Saraswat Brahmins



Note: The Gaud Saraswat Brahmins speak a language two degrees away from Hindi. Their story serves an example of how historical migration has created persistent local linguistic diversity.

Table 1: Percent Cognates Measure of Linguistic Distance

Panel A: Translations and Cognate Judgments for Sample Words in English, Bengali and Hindi

Meanings			Cognate Judgment		
English	Hindi	Bengali	Hindi - Bengali	Hindi - English	Bengali - English
ALL	SEB, SARA	SOB	**Yes**	*No*	*No*
EYE	AKH	COK	*No*	**Yes**	*No*
FEATHER	PER	PALOK	*No*	**Yes, doubtful**	*No*
GOOD	ECCHA	BHALO	*No*	*No*	*No*
MOTHER	MA	MA	**Yes**	**Yes**	**Yes**
OTHER	DUSRA	ONNO	*No*	*No*	**Yes**
		Percent Cognates	64.10%	14.60%	14.20%

Panel B: Percent Cognates with Hindi

Language	Percent Cognates with Hindi
Bengali	64.1%
English	14.6%
Gujarati	64.6%
Kashmiri	42.4%
Marathi	56.4%
Nepali	64.2%
Punjabi	74.5%

Table 2: Summary Statistics

Variable		Num Obs	Mean	St. Dev.	Min.	Max.
Panel A (at the district level)						
Degree measure of distance from native languages to Hindi		390	3.218	1.364	1.001	4.999
Percent of people who speak languages at distance > 2		390	0.373	0.443	0.000	1.000
Percent of people who speak languages at distance 0 (Hindi/Urdu natives)		390	0.465	0.443	0.000	1.000
Percent of people who speak languages at distance 1		390	0.095	0.265	0.000	0.991
Percent of people who speak languages at distance 2		390	0.067	0.212	0.000	0.958
Percent of people who speak languages at distance 3		390	0.094	0.251	0.000	0.985
Percent of people who speak languages at distance 4		390	0.013	0.069	0.000	0.766
Percent of people who speak languages at distance 5		390	0.265	0.391	0.000	1.000
Percent of people native in English		390	0.00013	0.00049	0.000	0.00670
Percent of urban schools that teach in mother tongue*:	Primary	408	0.889	0.222	0.000	1.078
	Upper primary	408	0.840	0.245	0.000	1.022
Panel B (at the state level, only urban areas)						
Percent of schools that teach English:	Primary	32	0.263	0.164	0.023	0.664
	Upper primary	32	0.310	0.070	0.233	0.511
	Secondary	32	0.337	0.104	0.182	0.615
Percent of schools with English instruction:	Primary	32	0.222	0.237	0.000	1.000
	Upper primary	32	0.321	0.268	0.053	1.000
	Secondary	32	0.380	0.285	0.047	1.000
	Higher secondary	31	0.470	0.295	0.051	1.000

* The percent of schools teaching in the mother tongue can be greater than 1 due to noise in the data.

Table 3: Impact of Linguistic Distance on % of Native Speakers who Learn English

Level of Observation:	State-Mother Tongue						
Dependent Variable:	% of Multilinguals who Learn English						
Sample:	Both Years (1961 and 1991)			1961		1991	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Linguistic distance (0 - 5)		0.082 ***		0.057 **		0.091 ***	
		(0.016)		(0.024)		(0.019)	
Linguistic distance > 2			0.371 ***		0.322 ***		0.395 ***
			(0.053)		(0.067)		(0.066)
D1: Linguistic distance = 0	0.379 ***	0.393 ***	0.359 ***	0.154	0.177 **	0.500 ***	0.449 ***
(Hindi/Urdu speakers)	(0.047)	(0.076)	(0.046)	(0.107)	(0.069)	(0.080)	(0.047)
D2: Linguistic distance = 2	0.046						
	(0.089)						
D3: Linguistic distance = 3	0.391 ***						
	(0.072)						
D4: Linguistic distance = 4	0.376 ***						
	(0.061)						
D5: Linguistic distance = 5	0.385 ***						
	(0.048)						
Year = 1991	0.202 ***	0.203 ***	0.203 ***				
	(0.065)	(0.066)	(0.065)				
Most spoken language in	-0.129	-0.102	-0.089	-0.551 **	-0.568 ***	0.087	0.108
state	(0.137)	(0.164)	(0.116)	(0.255)	(0.216)	(0.130)	(0.084)
Distance to Hindi of most	0.045 ***	0.084 ***	0.122 ***	0.099 ***	0.205 **	0.075 ***	0.076 **
spoken language	(0.014)	(0.010)	(0.034)	(0.029)	(0.097)	(0.008)	(0.032)
Share of native speakers in	0.800 ***	0.790 ***	0.749 ***	1.270 ***	1.270 ***	0.596 ***	0.542 ***
state	(0.176)	(0.189)	(0.151)	(0.305)	(0.265)	(0.143)	(0.118)
Obs (weighted, in millions)	1.3E+09	1.3E+09	1.3E+09	4.2E+08	4.2E+08	8.3E+08	8.3E+08
Observations	1085	1085	1085	537	537	548	548
R-squared	0.778	0.767	0.775	0.640	0.663	0.918	0.924
Test: D1=0, D2=0, D3=0, D4=0, D5=0 (p-value)	0.000						

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by state, are shown in parentheses. All columns include region fixed effects and observations are weighted by number of speakers. In columns 1 and 2, the omitted group has a linguistic distance of 1 degree away from Hindi.

Table 4: Other Dimensions of Language Learning

Level of Observation:	State-Mother Tongue			
Dependent Variable:	% of Native Speakers who are Multilingual		% of Multilinguals who Learn Hindi	
	(1)	**(2)**	**(3)**	**(4)**
Linguistic distance (0 - 5)	-0.008		-0.005	
	(0.009)		(0.011)	
Linguistic distance > 2		-0.017		-0.193 ***
		(0.032)		(0.046)
Linguistic distance = 0	-0.033	-0.019		
(Hindi/Urdu speakers)	(0.032)	(0.030)		
Language = English			0.161	0.251 ***
			(0.051)	(0.063)
Year = 1991	0.086 ***	0.086 ***	0.072	0.071
	(0.015)	(0.015)	(0.045)	(0.045)
Most spoken language in state	-0.092 *	-0.099 *	0.434 ***	0.321 ***
	(0.053)	(0.051)	(0.159)	(0.089)
Distance to Hindi of most	-0.006	-0.012	-0.171 ***	-0.719 ***
spoken language	(0.008)	(0.032)	(0.044)	(0.049)
Share of native speakers in state	-0.281 ***	-0.271 ***	-0.400 **	-0.207 *
	(0.067)	(0.062)	(0.171)	(0.119)
Observations (weighted)	1.3E+09	1.3E+09	6.9E+08	6.9E+08
Observations	1108	1108	1083	1083
R-squared	0.780	0.780	0.850	0.877

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by state, are shown in parentheses. Data is from 1961 and 1991. All columns include region fixed effects and observations are weighted by number of speakers in the state.

Table 5: Impact of Linguistic Distance on Percent of Schools that Teach English

Level of Observation:	State			
Dependent Variable:	% of Schools Teaching In English		% of Schools Teaching English	
	(1)	**(2)**	**(3)**	**(4)**
Linguistic distance	0.198 ***		0.332 ***	
(weighted average)	(0.039)		(0.105)	
Percent distant speakers		0.469		1.052
(>2 degrees)		(0.391)		(1.030)
Percent speakers at 0	-0.126	0.160 ***	-0.186	0.299 **
(Hindi/Urdu speakers)	(0.083)	(0.054)	(0.162)	(0.129)
Native English speakers	1.38	28.03	15.60	57.24
	(14.97)	(20.71)	(36.23)	(35.55)
Hindi belt states	-0.278 **	0.320	-0.181	0.997
	(0.141)	(0.216)	(0.213)	(0.751)
Observations	90	90	119	119
R-squared	0.672	0.576	0.726	0.678
p-value of language distance measures				

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by state, are shown in parentheses. All columns include fixed effects for school level and region. Other controls include child population in 1991, household wage income, average wage income for an educated individual, distance to the closest big city, a dummy for a coastline and the percent of people who: have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity.

Table 6: Summary Statistics on IT Presence Across Districts

Year of Data	Number of Disticts with IT (Out of 409)	Average Across All Districts		Average Across IT Districts	
		Number of HQ or Branches	Employees	Number of HQ or Branches	Employees
1995	47	1.36	120	11.83	1043
1998	47	2.25	279	19.60	2428
1999	54	2.48	348	18.76	2634
2002	76	3.24	559	17.46	3006
2003	72	2.54	604	14.40	3430

Table 7: Summary Statistics on Grade School Enrollment, only urban areas

Grade	Mean in 1993	Standard Deviation in 1993	Mean in 2002	Standard Deviation in 2002	Class size / Class size in 1st grade 1993	Class size / Class size in 1st grade 2002	% Growth since 1993
Grade 1	14698	22274	16948	24595			15%
Grade 2	12337	19817	14915	22793	84%	88%	21%
Grade 3	11872	19591	14406	22455	81%	85%	21%
Grade 4	11118	18543	13689	21336	76%	81%	23%
Grade 5	11060	18965	14117	22058	75%	83%	28%
Grade 6	11362	18554	14068	23283	77%	83%	24%
Grade 7	10211	16239	13316	20667	69%	79%	30%
Grade 8	9686	15239	13136	19854	66%	78%	36%
Grade 9	9345	13666	12290	17445	64%	73%	32%
Grade 10	7492	10560	10643	13841	51%	63%	42%
Grade 11	4372	6425	9036	11559	30%	53%	107%
Grade 12	4000	5986	8077	9878	27%	48%	102%
						Overall:	32%

Table 8: Impact of Linguistic Distance on Other Variables

Linguistic distance measure:	Weighted average		Percent distant speakers (>2 degrees)	
	(1)		**(2)**	
(Log) Number of Schools	0.020		-0.022	
N=1450, an observation is a district-school level	(0.025)		(0.184)	
Financing, Insurance, Real Estate, Computer	0.001		0.031	
Related Activities, R & D, Other Business Activities	(0.003)		(0.043)	
Hotels and Restaurants	-0.002		0.007	
	(0.003)		(0.019)	
Agriculture, Hunting, Forestry and Fishing	0.000		-0.103	*
	(0.010)		(0.062)	
Manufacting	0.0003		0.087	
	(0.009)		(0.078)	
Wholesale, Retail and Repair	-0.018	**	-0.014	
	(0.008)		(0.060)	
Transportation (Land, Water, Air) and Related	0.003		-0.035	
Services	(0.007)		(0.036)	
Communications (Post, Courier,	0.0007		-0.006	
Telecommunications)	(0.001)		(0.005)	
Other Services (Public Service, Education, Health,	-0.002		-0.037	
Sanitary, Community Services)	(0.009)		(0.069)	

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, are shown in parentheses. The dependent variable is the percent of the labor force employed in each industry. Other controls include the percent of native English speakers, the percent of native Hindi/Urdu speakrs, a dummy for a Hindi belt state, child population in 1991, household wage income, average wage income for an educated individual, distance to the closest big city, a dummy for a coastline, a Bartik control (in industry regressions, see data appendix) percent of people in the district who have regular jobs, have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity and fixed effects for region. N = 382 (industry regressions), 1450 (number of schools).

Table 9: Impact of Linguistic Distance on Growth of IT presence

Level of Observation:	District-Year							
Dependent variable:	Any HQ or branch		Years of IT HQ presence		(Log) Number of HQ & Branches		(Log) Number of Employees per Branch	
Linguistic distance measure:	Weighted average	Percent distant speakers	Weighted average	Percent distant speakers	Weighted average	Percent distant speakers	Weighted average	Percent distant speakers
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
Linguistic distance	0.035 **	0.299 **	0.346	5.113 **	0.025	0.509 ***	0.155 *	2.251 ***
	(0.017)	(0.124)	(0.271)	(2.400)	(0.026)	(0.158)	(0.092)	(0.794)
Percent speakers at 0	-0.206 ***	-0.102	-1.202	0.183	-0.152	-0.020	-0.975 **	-0.333
(Hindi/Urdu speakers)	(0.072)	(0.073)	(0.950)	(0.596)	(0.114)	(0.104)	(0.417)	(0.334)
Hindi belt states	0.090	0.266 **	1.132	4.369 *	0.047	0.358 ***	0.506	1.866 **
	(0.071)	(0.106)	(1.115)	(2.449)	(0.100)	(0.137)	(0.423)	(0.733)
Number of IITs and NIITs	0.187 ***	0.183 ***	1.329	1.351	0.184	0.174	0.750 **	0.709 **
	(0.066)	(0.066)	(1.339)	(1.330)	(0.117)	(0.116)	(0.367)	(0.351)
Observations	1845	1845	1701	1701	1845	1845	1845	1845
R-squared	0.38	0.38	0.22	0.23	0.31	0.32	0.38	0.38
Panel B								
Linguistic distance	0.035 **	0.263 **	0.443	5.255 *	0.033	0.575 ***	0.168 *	2.221 ***
	(0.017)	(0.123)	(0.285)	(3.003)	(0.027)	(0.176)	(0.094)	(0.817)
Big city * Linguistic	-0.054	-0.225	-1.742	2.170	0.351 **	1.482 *	0.409	1.720
distance	(0.043)	(0.186)	(2.267)	(12.029)	(0.174)	(0.771)	(0.400)	(1.759)
Big city	0.497 ***	0.451 ***	23.996 **	18.283 **	1.653 *	2.074 **	3.386	3.978 **
	(0.156)	(0.121)	(10.862)	(7.697)	(0.956)	(0.841)	(2.188)	(1.864)
Percent speakers at 0	-0.199 ***	-0.102	-1.396	0.128	-0.166	-0.003	-0.973 **	-0.304
(Hindi/Urdu speakers)	(0.073)	(0.074)	(1.042)	(0.632)	(0.118)	(0.107)	(0.428)	(0.341)
Big city * Percent speakers at 0	0.026	-0.033	9.506	11.289	-0.404	0.046	-0.490	0.054
(Hindi/Urdu speakers)	(0.154)	(0.166)	(10.500)	(12.038)	(1.578)	(1.616)	(2.378)	(2.638)
Hindi belt states	0.080	0.231 **	0.217	3.550	0.057	0.411 ***	0.495	1.821 **
	(0.071)	(0.104)	(1.584)	(3.209)	(0.103)	(0.151)	(0.430)	(0.750)
Number of IITs and NIITs	0.159 ***	0.152 ***	-0.110	0.031	0.056	0.037	0.420 **	0.353 **
	(0.061)	(0.061)	(1.740)	(1.692)	(0.125)	(0.129)	(0.363)	(0.356)
Observations	1895	1895	1746	1746	1895	1895	1895	1895
R-squared	0.47	0.47	0.67	0.67	0.69	0.69	0.58	0.58

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, are in parentheses. All regressions include year and region fixed effects. Other controls include the log of population, percent of native English speakers, household wage income, average wage income for an educated individual, distance to the closest big city, distance to closest airport, a dummy for a coastline, the percent of people in the district who have regular jobs, have college degrees, secondary school or engineering degrees, are literate, are Muslim, ride a train and the percent of households with electricity. In Panel A, I drop observations for the ten most populous cities in India (as of 1987).

Table 10: Impact of Linguistic Distance on Grade School Enrollment

Level of Observation:	District-Year							
Sample:	All Grades		Primary (Grades 1-5)		Upper Primary (Grades 6-8)		Secondary (Grades 9-12)	
	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Panel A: Weighted average								
Post * Linguistic distance	0.065 ***	0.075 ***	0.032 *	0.012	0.061 ***	0.043 **	0.125 **	0.134 ***
	(0.024)	(0.025)	(0.019)	(0.017)	(0.019)	(0.020)	(0.049)	(0.049)
Post * Percent speakers at 0	0.127	0.006	0.205 ***	0.174 **	0.186	0.095	-0.002	-0.106
(Hindi/Urdu speakers)	(0.095)	(0.076)	(0.074)	(0.072)	(0.123)	(0.072)	(0.140)	(0.147)
Post * Hindi belt states	-0.005	0.027	-0.281 ***	-0.295 ***	-0.459 ***	-0.335 ***	0.371 **	0.532 ***
	(0.102)	(0.089)	(0.081)	(0.075)	(0.097)	(0.073)	(0.183)	(0.199)
Observations	8509	8660	3605	3635	2112	2157	2792	2868
R-squared	0.902	0.884	0.978	0.978	0.984	0.982	0.871	0.847
Panel B: Percent distant speakers								
Post * Linguistic distance	0.322 *	0.304 *	0.333 **	0.272 **	0.533 ***	0.325 ***	0.261	0.266
	(0.185)	(0.185)	(0.158)	(0.130)	(0.164)	(0.122)	(0.287)	(0.379)
Post * Percent speakers at 0	0.274 ***	0.160 **	0.315 ***	0.247 ***	0.363 ***	0.220 ***	0.210 *	0.111
(Hindi/Urdu speakers)	(0.096)	(0.071)	(0.091)	(0.086)	(0.133)	(0.082)	(0.122)	(0.118)
Post * Hindi belt states	0.172	0.191	-0.087	-0.131	-0.150	-0.150	0.494 **	0.658 **
	(0.149)	(0.160)	(0.106)	(0.088)	(0.117)	(0.102)	(0.200)	(0.311)
Observations	8509	8660	3605	3635	2112	2157	2792	2868
R-squared	0.902	0.884	0.978	0.978	0.984	0.982	0.870	0.846

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, are in parentheses. All regressions include district, timeperiod, gender and grade level fixed effects, and both region fixed effects and grade level fixed effects interacted with post. Other controls include log enrollment in the pre year, log child population in pre and post years, a proxy for labor demand growth, and post interacted with: the percent of native English speakers, household wage income, averaged wage income for an educated individual, distance to the closest big city, a dummy for a coastline, the percent of people in the district who have regular jobs, have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity.

Table 11: Impact of 1961 English Bilingualism

Level of Observation:	District-Year			
Dependent variable:	Any HQ or branch	Years of IT HQ presence	Number of HQ and Branches	Number of Employees per Branch
	(1)	**(2)**	**(3)**	**(4)**
Panel A: IT Growth				
English among bilinguals 196	0.169 **	0.971	0.318 **	0.868 *
	(0.076)	(0.986)	(0.161)	(0.483)
Observations	1335	1208	1335	1335
R-squared	0.3	0.22	0.27	0.31
Sample (all grades):	All	Girls	Boys	
	(1)	**(2)**	**(3)**	
Panel B: School Enrollment				
Post * English among bilinguals 1961	0.250 ***	0.268 ***	0.228 ***	
	(0.076)	(0.078)	(0.076)	
Observations	12929	6413	6516	
R-squared	0.929	0.929	0.930	

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, are in parentheses. Control variables and fixed effects are as listed in table 9 for panel A and table 11 for panel B.

Table 12: Impact of Linguistic Distance on Rural Grade School Enrollment

Level of Observation:	District-Year					
Linguistic distance measure:	Weighted average			Percent distant speakers		
Sample (all grades):	All	Girls	Boys	All	Girls	Boys
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Post * Linguistic distance	0.01	0.03	-0.0004	-0.09	-0.10	0.01
	(0.03)	(0.03)	(0.02)	(0.15)	(0.18)	(0.12)
Post * Percent speakers at 0	0.09	0.14	0.05	0.07	0.14	0.05
(Hindi/Urdu speakers)	(0.08)	(0.09)	(0.06)	(0.08)	(0.11)	(0.07)
Post * Hindi belt states	-0.22 **	-0.35 ***	-0.13	-0.27 **	-0.40 ***	-0.12
	(0.10)	(0.12)	(0.08)	(0.12)	(0.14)	(0.10)
Observations	17298	8537	8761	17298	8537	8761
R-squared	0.89	0.90	0.89	0.89	0.90	0.89

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, are in parentheses. All regressions include district, timeperiod, gender and grade level fixed effects, and both region fixed effects and grade level fixed effects interacted with post. Other controls include log enrollment in the pre year, log child population in pre and post years, a proxy for labor demand growth, and post interacted with: the percent of native English speakers, household wage income, averaged wage income for an educated individual, distance to the closest big city, a dummy for a coastline, the percent of people in the district who have regular jobs, have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity.

Table 13: Impact of Linguistic Distance on Wages and Returns to Education

Level of Observation:	Individual									
Linguistic distance measure:	Weighted average					Percent distant speakers				
Sample:	All	Men	Women	Age < 30	Age > 29	All	Men	Women	Age < 30	Age > 29
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Post * Linguistic distance	0.015	0.016	-0.016	-0.023	0.027	0.022	0.078	-0.249	0.104	-0.024
	(0.024)	(0.024)	(0.055)	(0.030)	(0.027)	(0.105)	(0.109)	(0.276)	(0.161)	(0.112)
-- * High school	-0.047 **	-0.049 ***	-0.048	-0.034	-0.053 ***	-0.100	-0.124	-0.135	-0.057	-0.101
	(0.018)	(0.018)	(0.049)	(0.022)	(0.020)	(0.079)	(0.080)	(0.185)	(0.087)	(0.086)
-- * College	-0.032 *	-0.028	-0.050	-0.031	-0.037 *	-0.002	-0.001	-0.024	-0.048	-0.012
	(0.018)	(0.017)	(0.042)	(0.029)	(0.021)	(0.073)	(0.071)	(0.167)	(0.110)	(0.091)
Post * Percent speakers at 0	-0.037	-0.027	-0.011	0.070	-0.080	-0.039	-0.008	-0.146	0.044	-0.088
(Hindi/Urdu)	(0.066)	(0.069)	(0.186)	(0.097)	(0.072)	(0.072)	(0.073)	(0.173)	(0.103)	(0.077)
-- * High school	0.051	0.020	0.255	-0.112	0.125 *	0.032	-0.016	0.231	-0.109	0.111
	(0.062)	(0.063)	(0.221)	(0.086)	(0.073)	(0.083)	(0.085)	(0.248)	(0.100)	(0.096)
-- * College	0.004	0.000	0.000	0.012	0.036	0.037	0.027	0.062	0.015	0.065
	(0.054)	(0.058)	(0.147)	(0.119)	(0.066)	(0.073)	(0.076)	(0.184)	(0.135)	(0.093)
Post	1.024 ***	1.046 ***	1.035 ***	1.146 ***	0.975 ***	1.057 ***	1.042 ***	1.201 ***	1.061 ***	1.069 ***
	(0.091)	(0.088)	(0.202)	(0.142)	(0.094)	(0.103)	(0.106)	(0.252)	(0.158)	(0.104)
-- * High school	0.077	0.110	-0.026	-0.033	0.128	-0.024	0.019	-0.119	-0.121 *	0.004
	(0.079)	(0.078)	(0.191)	(0.087)	(0.085)	(0.067)	(0.068)	(0.143)	(0.066)	(0.074)
-- * College	0.167 **	0.172 **	0.109	-0.032	0.222 **	0.054	0.074	-0.063	-0.113	0.102
	(0.072)	(0.071)	(0.165)	(0.110)	(0.088)	(0.060)	(0.059)	(0.135)	(0.077)	(0.079)
High school	0.539 ***	0.428 ***	1.177 ***	0.439 ***	0.590 ***	0.638 ***	0.499 ***	1.276 ***	0.480 ***	0.734 ***
	(0.060)	(0.057)	(0.107)	(0.068)	(0.068)	(0.054)	(0.050)	(0.084)	(0.053)	(0.064)
College	1.018 ***	0.891 ***	1.585 ***	0.969 ***	1.039 ***	1.121 ***	0.964 ***	1.728 ***	1.012 ***	1.181 ***
	(0.052)	(0.051)	(0.120)	(0.074)	(0.062)	(0.040)	(0.040)	(0.107)	(0.056)	(0.048)
Obs. (in millions, weighted)	67.2	55.6	11.6	21.6	45.6	67.2	55.6	11.6	21.6	45.6
Observations	71292	58665	12627	22205	49087	71292	58665	12627	22205	49087
R-squared	0.671	0.667	0.686	0.648	0.671	0.671	0.667	0.686	0.648	0.671

* 10%, ** 5%, *** 1%. Robust standard errors, clustered by district, in parentheses. Individual-level controls include age, age squared, married, male, whether the individual has ever moved, high school and college both interacted with linguistic distance and percent native Hindi speakers. District-level controls include a proxy for labor demand growth and post interacted with: the percent of native English speakers, distance to the closest big city, a dummy for a coastline and being in a Hindi belt state. All regressions include districts fixed effects, region fixed effects interacted with post and a dummy variable for whether the individual is self-employed.