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

Gauri Kartini Shastry*

Wellesley College

This draft: June 2010; First draft: October 2007

Abstract

Recent studies show that trade liberalization increases skilled wage premiums in developing countries. This result suggests globalization benefits skilled workers more than unskilled workers, increasing inequality. However, if human capital investment responds to new global opportunities, this effect may be mitigated. I study how the impact of globalization varies across Indian districts with different costs of skill acquisition. I focus on the relative cost of learning English versus Hindi, since English is useful 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 where it is relatively easier to learn English benefit more from globalization: they experience greater growth in both information technology jobs and school enrollment. Consistent with a larger human capital response, they experience smaller increases in skilled wage premiums. Keywords: education, India, outsourcing. JEL Codes: O, I, F.

*Correspondence: gshastry@wellesley.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, Aimee Chin, Petia Topalova and David Clingingsmith for help with the 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, Maryland, the Center for Global Development, Wharton, Tufts, Anderson, Scripps, Notre Dame, Mathematica, Boston College, Brookings 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 suggests that trade liberalization in poor countries, particularly in Latin America, increases skilled wage premiums, a common indicator of inequality.[1] These results are not predicted by a simple Heckscher-Ohlin model,[2] and there are multiple explanations for this divergence. Some explanations focus on global outsourcing (Feenstra and Hanson 1996, 1997) and complementarities between skill levels (Kremer and Maskin 2006).[3] A key, understudied factor in understanding the effects of globalization, however, is the supply response of human capital. If the supply of skilled workers rises (as in East Asia, but less so in Latin America[4]), this may drive down the skilled wage premium and mitigate globalization's effect on inequality.

I examine this issue in Indian districts with plausibly different supply elasticities. Some districts have a more elastic supply of English language human capital, particularly relevant for service exports. I show that these districts attracted more export-oriented skilled jobs, specifically in information technology (IT),[5] and experienced greater growth in schooling. Consistent with this factor supply response, these districts experienced smaller growth in skilled wage premiums.

I study trade liberalization in India in the early 1990s. Motivated by a balance-of-payments crisis, reforms reduced trade barriers and removed restrictions on private and foreign direct investment. While this shock was common across India, historical linguistic forces created differences in the ability of districts to take advantage of new global opportunities. I examine how the impact of globalization varied with these pre-existing differences.

[1] See Goldberg and Pavcnik (2004) for a review of the literature, which includes 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. There is little evidence on East Asia, but conventional wisdom suggests a smaller effect. See Wood (1997), Lindert and Williamson (2001) and Wei and Wu (2001).

[2] In a two-good, two-country Heckscher-Ohlin model, a labor-abundant poor country should specialize in unskilled labor intensive industries, increasing demand for unskilled workers and lowering wage premiums.

[3] Additional explanations are discussed in Goldberg and Pavcnik (2004).

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

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

Specifically, I show that the effect of liberalization varies with the elasticity of human capital supply by exploiting exogenous variation in the cost of learning English. Some variation is naturally endogenous: state governments that care more about trade opportunities may also promote English instruction. Instead, I use exogenous variation, produced by historical linguistic diversity, that spurred people to learn English even in 1961, long before the 1990s trade reforms. The variation stems from substantial local linguistic diversity that motivates individuals to learn a lingua franca,[6] either English or Hindi. Individuals whose mother tongues are linguistically further from Hindi have a lower *relative* opportunity cost of learning English: they find Hindi more costly to learn and speak, relative to people whose mother tongues are linguistically similar to Hindi. Using several definitions of linguistic distance, I demonstrate that distance from Hindi induces people to learn English and schools to teach English. Because of considerable local linguistic diversity, I am able to rely purely on pre-existing, within-region variation, minimizing concerns that variation in linguistic distance is correlated with unobserved regional differences.[7]

Using new data on the Indian information technology sector, I first show that IT firms were more likely to operate where costs of learning English are lower. My choice of the IT sector is driven by several factors. The industry grew primarily due to trade liberalization and technological progress. Exports make up 82% of the median firm's revenue. In addition, IT firms almost exclusively hire English speakers; data on other jobs that require English is unavailable.[8] I use this industry as a proxy for new global export opportunities.[9]

Next, I demonstrate that districts with languages further from Hindi experienced greater increases in school enrollment from 1993 to 2002 relative to prior trends. These districts also had greater enrollment growth in English than local languages in 2002-2007. Finally, using individual-level data, I show that these districts experienced a smaller increase in

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

[7] The six regions in India each consist of 2-7 states, each of which is made up of 1-51 districts.

[8] 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.

[9] See Topalova (2004, 2005) for the effect of *import* competition on productivity and poverty in India

the skilled wage premium. The fact that education and skilled wage premiums move in opposite directions is evidence that the supply response of human capital substantially shapes the impact of globalization. Solely demand-driven explanations for the increase in skill acquisition would predict a greater increase in both education and the skilled wage premium.

I do not claim that the impacts on education and wages are driven entirely by IT; the IT industry is only a fraction of export-related, English-language opportunities. My results on schooling and wages are likely to be driven by all global opportunities.

This paper contributes to a growing literature on globalization, language and education. 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. Azam, Chin and Prakash (2010) estimate large returns to speaking English in India. Oster and Millett (2010) demonstrate that school enrollment in villages close to where call centers are located increases faster than in other villages in three states in India and that this effect is larger for schools that teach in English. The authors include school fixed effects, but do not explain why the call centers located where they do. Finally, Edmonds, Pavcnik and Topalova (2007) find that *import competition* increases poverty and reduces schooling.

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, school enrollment growth and returns to education and section 7 provides a discussion of robustness checks. Section 8 concludes.

2 Background on trade liberalization and IT

India has traditionally controlled its economy and limited trade through investment licensing and import controls. Since the late 1970s, the government took small steps towards

liberalization, but even as late as 1990, numerous obstacles blocked trade. A balance-of-payments crisis in 1991 precipitated a shift towards an open economy. Reforms ended most import licensing requirements for capital goods and cut tariffs. The March 1992 Export-Import Policy reduced the number of goods banned for export from 185 to 16. From 439 goods subject to some control, the new regime limited only 296 exports. The government relaxed capital controls, particularly for exporters, and devalued the rupee, further reducing deterrents to trade (Panagariya 2004).

Services in particular had been heavily regulated. State-owned enterprises dominated insurance, banking, telecommunications and infrastructure. The reforms in the 1990s opened these sectors 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 reaffirmed in 1999 under the New Telecom Policy which further reduced limits on foreign direct investment (FDI) in telecommunications. FDI for internet service providers was permitted with few limitations. FDI in e-commerce, software and electronics was granted automatic approval, particularly for IT exporters (Panagariya 2004). Note that liberalization was at a national level and reforms did not differ in areas with many English speakers.

These reforms led to remarkable growth in 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, skilled labor intensive sectors grew faster (Panagariya 2004). This policy shift, along with technological progress, led to the growth of IT services outsourced to India. By 2004, India was the single largest destination for foreign firms seeking IT services. IT outsourcing accounted for 5% of India's GDP in 2005 (Economist 2006). Employment growth was strong: from 56,000 professionals in 1990, 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 was 27.5 years old and 81% had a bachelor's degree (NASSCOM 2004). An entry-level call center job

paid on average Rs. 10,000 ($230), considered high for a first job (Economist 2005).

IT firms were free to locate based on the availability of educated, English-speaking workers, due to their export focus and reliance on foreign capital (NASSCOM 2004). Figure 1 maps the location of IT firms from 1992 to 2003. Note that IT establishments do not locate only in the top few metropolitan areas. In fact, many firms have spread to smaller cities. The distribution of entrepreneurs and the supply of engineers also influence where these firms locate (Nanda and Khanna 2010; Arora and Gambardella 2004; Arora and Bagde 2007).

3 Theoretical framework

Consider a country with two districts that differ only in the exogenous cost of learning English relative to Hindi. Workers choose whether to obtain schooling, opting for English or Hindi instruction, and produce both a tradeable and domestic good. English- and Hindi-speaking skilled workers are equally productive in the domestic sector, but only English speakers can produce the traded good. Goods travel freely between districts, but workers do not. Abstracting from why a poor country exports goods intensive in skilled labor, assume that reforms enable the production and export of the traded good.

In a web appendix, I demonstrate that trade reforms cause a positive demand shock for skills and that the impact depends on the relative costs of English and Hindi instruction. The intuition is straightforward. In the case where most people in either district do not speak English, as in India, a lower cost generates a higher elasticity of English human capital.[10] Identical demand shocks would cause a greater increase in human capital where the supply is more elastic (figure 2), but a smaller increase in the wage premium. However, more exporting firms are willing to locate in the "low-cost" district since it is easier to hire English speakers. The increase in education should still be larger in the low-cost district, but the relative change in the premium depends on the size of the demand shocks (figure 3).[11]

[10] I do not make an explicit assumption about the elasticity in the model, but this fact aids in the discussion.

[11] Formally, I vary the difference in demand shocks by changing the importance of a second factor in

I test two predictions: First, the district with a lower cost of English should produce more of the traded good and second, school enrollment in the low-cost district should grow faster after liberalization. I also provide evidence that the average return to education rose faster in high-cost districts. This framework allows me to also consider how different business environments respond to trade: a pro-business district should see a greater demand shock and greater growth in both education and the skill premium. That returns to skill and education move in opposite directions help rule out pro-business differences, such as state-level economic reforms, as an explanation for my results.

4 Background on linguistic distance from Hindi

The 1961 Census of India documented speakers of 1652 languages from five language families. Linguists classify languages such as English and Hindi in the same family (Indo-European), but cannot connect many languages native to India, such as Hindi and Kannada (of the Dravidian language family). Much linguistic diversity is local. The probability that two district residents speak different native tongues, calculated as one minus the Herfindahl index (the sum of squared population shares), is 25.6%, but ranges from 1% to 89%. A district's primary language is native to 83% of residents on average, ranging from 22% to 100%. Thus, many people also adopt a lingua franca (Clingingsmith 2008). Of all multilinguals who were not native speakers, 60% chose to learn Hindi and 56% chose English. At only 6%, Kannada was the next most common second language. Of all individuals, 11% speak English (0.02% natively) and 49% speak Hindi (40% natively).

Whether an individual learns Hindi or English depends on the relative costs, which in turn depend on his or her mother tongue. Someone whose mother tongue is similar to Hindi will find Hindi easy 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

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.

tendency. During British occupation, English was established as the language of government and instruction. After Independence in 1947, a nationalist movement chose Hindi, a common but not universal lingua franca, as the "official" language, despite opposition from non-Hindi speakers. This led to riots, the most violent of which occurred in Tamil Nadu in 1963. In 1967, the central government made Hindi and English joint official languages (Hohenthal 2003). This history contributed to the greater English literacy among speakers of languages linguistically distant to Hindi, who saw Hindi as unjustly imposed upon them. In fact, in some states more people speak English than Hindi.

Over time, this relationship was institutionalized through schools. Early growth in formal education in the nineteenth century was in English, driven by the British to foster an elite governing class (Nurullah and Naik 1949, Kamat 1985). By 1993, English was still the main medium of tertiary instruction, but there were over 28 languages of instruction at the primary level. Hindi was most common with 38% of urban schools and English second with 9%. Secondary-level instruction is in Hindi in 29% of schools and English in 20%. Both Hindi and English are taught in schools across the country.[12]

4.1 Measuring linguistic distance

Since there is no universally accepted measure of distance between languages, I calculate three logically independent measures and verify that my results are robust. The first measure of linguistic distance from Hindi was developed in consultation with an expert on Indo-European languages, Jay Jasanoff (personal communication, 2006). It classifies languages into five "degrees" of distance from Hindi (see table 1) based on cognates, grammar and syntax. For example, Punjabi is one degree from Hindi, while Bengali is three degrees away. The second measure is the percent of words from a core list that are cognates of Hindi words.[13] Expert judgments on cognates among Indo-European languages are from Dyen et

[12] Hindi is available as a language of instruction even in non-Hindi speaking districts and English is available even in government schools.

[13] 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

al. (1997). For example, the word "eye" in English is a cognate of the Hindi word, "akh," but not the Bengali word, "cok." Hindi shares 64% of words with Bengali, but only 15% with English.[14] The third measure is based on language family trees from the Ethnologue database (Lewis 2009). I define distance as the number of "nodes" between languages, where a single node connects different families. Reassuringly, these measures are highly correlated: -0.935 between degrees and percent cognates and 0.903 between degrees and nodes.

In what follows, I use degrees to measure a language's distance from Hindi.[15] 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 of languages at least 3 degrees away ('distant speakers'). Panel A in table 2 provides district averages of the percent of speakers at each distance from Hindi.

Note that I proxy English-learning costs as linguistic distance from *Hindi*. One may think the natural proxy is distance from English, but it is the relative costs of learning Hindi and English that should determine which lingua franca one learns. In fact, as distance to English falls, distance to Hindi falls even faster. First, Hindi and English, both Indo-European languages, are more similar to each other than to non-Indo-European languages spoken in India; there is a strong positive correlation between linguistic distance from Hindi and from English (0.9714 for percent cognates). At the same time, languages close to Hindi share on average 66% cognates with Hindi but only 14.6% cognates with English. To be clear, consider two individuals, one speaks another Indo-European language close to Hindi and the other speaks a language very far from Hindi. The first individual would find both Hindi and English easier to learn than the second individual. However, the first individual will find Hindi much easier than English. The second individual finds Hindi and English

of favor among linguists, but the percent of cognates is an acceptable measure of similarity.

[14] I assume non-Indo-European languages share 5% of words in common with Hindi since linguists use 5% as a threshold to determine whether languages are related. For Indo-European languages not in Dyen's list or Jasanoff's classification, I use the value of the closest language on the language tree.

[15] I focus on this measure because its discrete nature allows me to include it in a flexible manner. A common concern about the number of nodes between languages is that two nodes may not be comparable and dealing with a large number of possible distances makes a flexible specification impossible. All the results presented in this paper are robust to using the percent of cognates as the measure of linguistic distance.

roughly equally difficult. So the first individual's relative cost of learning English (cost of English - cost of Hindi) is positive, but the second individual's relative cost is roughly zero.[16] If each person learns one language, the first individual should learn Hindi and the second would be indifferent. Assuming a symmetric distribution around these costs, more people who are like the second individual will learn English and more people who are like the first will learn Hindi.

Also, non-Hindi speakers have historically rebelled against having to learn Hindi, amplifying this tendency. Of course, people can speak more than two languages. Thus, it is an empirical question whether people who speak languages closer to Hindi (relative to people who speak languages far from Hindi) are more likely to learn English (it is easier) or less likely (they only learn Hindi): section 5 demonstrates that the second tendency dominates.

According to this theory, native Hindi speakers should be more likely to learn English than those far from Hindi. For them, the relative cost of Hindi is irrelevant - they already speak Hindi - but their cost of learning English is lower than those farther from Hindi (and thus farther from English). They may not need another lingua franca, however, and could choose to remain monolingual.

5 Does linguistic distance predict who learns English?

My identification strategy relies on within-region variation in the relative cost of learning English driven by linguistic diversity. This strategy rests on two assumptions. First, linguistic distance from Hindi must predict who learns English and second, it must not be correlated with omitted factors that affect schooling or exports conditional on region fixed effects and control variables. In this section, I demonstrate the first and provide evidence for the second.

[16] Using the relative distance (distance from English minus distance from Hindi) gives me similar results since there is little variation in distance from English. This relative distance is greater for those close to Hindi than those far from Hindi. I do not focus on this measure because it is only possible for the percent cognates and not the other linguistic distance measures.

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

$$E_{lkt} = \alpha_0 + \beta' D_l + \alpha_1' X_{lk} + \gamma_t + \gamma_g + \epsilon_{lkt} \tag{1}$$

where E_{lkt} is the percent of native speakers of language l in state k, region g and year t who learn English, conditional on being multilingual,[17] D_l is the distance of language l from Hindi, and γ_t is a year fixed effect. Region fixed effects, γ_g, absorb regional heterogeneity. X_{lk} includes the share of language l speakers in state k, an indicator for 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 (table 3). In column 1, I include dummy variables for each distance. Languages 1 degree away from Hindi are the omitted group. 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 variables are jointly different from zero. The individual dummy variables are not strictly increasing, but the deviation is small. Column 2 in table 3 assumes a linear functional form: One linguistic degree increases the percent of multilinguals who learn English by 8.2 percentage points. Column 3 includes an indicator for sufficiently distant speakers: Speaking a language 3 or more degrees from Hindi increases the percent of English speakers by 37 points relative to speakers of languages 1 and 2 degrees away. Columns 3-7 break these down by year. The relationship between linguistic distance and learning English existed even in 1961, suggesting the relationship is exogenous to recent trade reforms.

As discussed above, individuals at a distance of zero (native speakers of Hindi and Urdu[18]) who choose to become multilingual tend to learn English. This creates a non-monotonicity in the relationship between linguistic distance and learning English. The propensity to learn English dips significantly when we move away from native Hindi speakers and then rises

[17] The results are robust to using the total share of English speakers or the number of English speakers.
[18] Hindi and Urdu are often considered the same language. Separating them gives similar results.

as native tongues get further. I account for this non-monotonicity in my regressions by controlling for the percent of native Hindi speakers and calculating the weighted average linguistic distance only among non Hindi natives. I can also exploit this non-monotonicity since districts with more Hindi natives, all else equal, are likely to have more English learners.

The control variables matter as expected: the propensity to learn English is greater in 1991 and increasing in the share of state residents with the same mother tongue (minorities learn the regional language first). The distance from Hindi of the state's primary language increases the propensity to learn English but that of the individual's mother tongue is still highly significant. 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 the inclusion of region fixed effects: the tendency to learn English is stronger for people further from Hindi even within a region. These results are robust to including state fixed effects and clustering by native language.

This data allows me to further analyze the relationship between linguistic distance and language acquisition (results available upon request). Individuals speaking languages further from Hindi are less likely to learn Hindi, as expected. Reassuringly, linguistic distance to Hindi does not affect whether someone becomes multilingual.

I next explore how distance from Hindi predicts whether schools teach English; I estimate

$$M_{ij} = \alpha_0 + \beta' D_j + \alpha_1 P_j + \alpha_2' Z_j + \gamma_i + \gamma_g + \epsilon_{ij} \tag{2}$$

where M_{ij} is 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 in 1991,[19] and P_j is child population (aged 5 - 19, in millions).[20] Z_j includes 1987 measures of average household wage income, average income for educated individu-

[19] I use a district's distance from Hindi in 1991 since it is more precise than 1961. The 1961 data lists 1652 languages, many of which are difficult to assign a distance from Hindi (in contrast, there are only 114 in 1991 due to prior Census classification). Many districts have been divided since 1961, adding further noise.

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

als, the percent of households with electricity, and the percent of people who: have regular jobs, have graduated from college, have completed high school, are literate, are Muslim, or regularly use a train. 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 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, Uttaranchal, Madhya Pradesh, Chhattisgarh, Haryana, Punjab, Rajasthan, Himachal Pradesh, Jharkhand, Chandigarh and Delhi. In addition, I focus on urban areas and include region and level fixed effects (primary, etc.). The data are described in the appendix.

The results show that linguistic distance from Hindi predicts the percent of schools that teach in English (columns 1-2 of table 4) or teach English as a second language (columns 3-4). An increase in 1 degree in the average distance from Hindi increases the fraction of schools teaching in English by 20 percentage points and the fraction teaching English by 33 points. More Hindi speakers also increases the teaching of English, but only when distance is measured by the percent distant speakers. It is not surprising that not all estimates are significant since the data is aggregated to the state, leaving little within-region variation.

5.1 The exclusion restriction

Next, we must consider whether the variation I exploit is correlated with omitted variables that might bias these results. Most measures of the cost of learning English, e.g. the number of English speakers, are likely to be correlated with unobserved determinants of schooling or trade policies. If a local government cares about access to global opportunities, for example, it may both promote English and provide incentives for FDI. Variation in linguistic distance is unlikely to be related to export-oriented preferences. When using other measures, we might 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. However, we need to verify that there are no other channels through which linguistic distance might correlate with schooling.

One worry about linguistic diversity is that much of the raw variation is geographic. Figure 4 maps the raw variation in percent distant speakers: linguistic distance is not randomly distributed. Indo-European languages have traditionally been spoken in the north, Sino-Tibetan languages in the northeast and Dravidian languages in the south. Since this geographic variation may be correlated with factors that influence schooling or exports, such as agricultural productivity or culture, I include region fixed effects. I also include district control variables, allowing their effects to differ after liberalization. In fact, in the schooling regressions described below, my results are identified off deviations from pre-existing district trends. In figure 5 I demonstrate the residual variation that I use by mapping the residuals from a regression of linguistic distance on region fixed effects and control variables in vector Z_j. This geographically balanced variation is less likely to be correlated with omitted variables. One might wonder where this within-region variation comes from: one important source is historical migration. While recent migration is infrequent, people have migrated across India for millennia, bringing their native languages across present-day boundaries.[21]

One might still worry that communities that speak languages distinct from Hindi differ in ways that might cause bias. Being more forward-looking, for example, could explain the

[21] Some migrants assimilate, but the local diversity shows that these groups retain a separate identity. Linguistic distance to Hindi in 1961 and 1991 are strongly correlated, demonstrating this persistence. On a similar note, Dravidian languages were traditionally considered native to South Asia, but several studies have provided evidence of Dravidian speakers in western and northwestern regions of the subcontinent prior to the tenth century CE (Tyler 1968, McAlpin 1981, 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 a sultan in the Deccan caused some GSB families to migrate down the coast into present-day west and southwest Karnataka (Conlon 1977). 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. The share of the population who speak languages exactly two degrees away from Hindi 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.

tendency to learn English and faster growth in education. The results in table 3 provide evidence against this concern: these groups were more likely to learn English in 1961, before anyone could anticipate trade liberalization in the 1990s. Similar concerns are that these communities migrated to faster growing cities before 1991 or had greater preferences for education. I will rely purely on changes over time when studying enrollment and wages, accounting for time-invariant heterogeneity. Preferences for education, for example, would only bias my results if they changed differentially after reforms. In section 7, I discuss these potential threats to validity in detail and provide empirical evidence against these concerns.

6 Impact of linguistic distance from Hindi

The following section discusses three key results. I first demonstrate that trade liberalization led to more export opportunities in districts far from Hindi. I then show that growth in school enrollment grew faster and that returns to education grew slower in these districts after liberalization. The results suggest that if a district had 2% more English speakers, the probability it receives any IT firms increases by 6% points and school enrollment grows faster, also by a total of 6-7% points over the 9 year interval from 1993 to 2002.

6.1 Information technology

I first test whether export opportunities have grown faster in districts with lower costs of English by studying the IT sector. I estimate

$$IT_{jt} = \alpha_0 + \beta' D_j + \alpha_1' Z_j + \alpha_2' W_j + \gamma_t + \gamma_g + \nu_{jt} \tag{3}$$

where IT_{jt} measures IT presence in district j in year t and D_j measures linguistic distance. Z_j is as in equation (2) and W_j includes other 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. I include year fixed effects and cluster the

standard errors by district. The measures of IT 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.[22] The region fixed effects ensure that my results are driven by within region variation.[23] The data, described in the appendix, contains firm-level employment; I assign employees evenly across branches to estimate employment by district.

Note that if a good measure of English learning costs were available, I would use linguistic distance to Hindi as an instrument. However, a comprehensive district-level measure of this cost is unavailable, because the cost of English is multi-dimensional and data is limited.[24] Therefore, I focus on reduced form results using two measures of linguistic distance: i) the weighted average and ii) percent distant speakers.[25]

Estimating equation (3) reveals a strong positive effect of linguistic distance from Hindi on IT presence (table 5). 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 predicts whether any IT firm establishes a headquarters or branch in a district. An increase in 1 degree from Hindi of the average speaker's mother tongue (three-fourths of a standard deviation, column 1) results in a 3.5% point increase in the probability of any IT presence (the dependent variable mean is 15%); a 20% increase in the percent distant speakers (half a standard deviation, column 2) increases the probability by 6% points. The magnitude of these effects is economically significant: about a fourth of the effect of housing an elite engineering college. Columns 3-4 show that IT headquarters were established earlier in areas linguistically further from Hindi, by approximately one year per 20% distant speakers.[26]

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

[23] In section 7.4, I discuss how these results change when I include state fixed effects or cluster by state; essentially, they tell the same story. I cannot look at changes over time when studying IT because the data does not exist prior to trade reforms; the assumption that IT was negligible pre-1991 is not unrealistic.

[24] For example, the measure would include how many schools teach English and how many adults speak English, both unavailable at the district level.

[25] 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.

[26] Some districts may be unlikely to receive any IT for other reasons. While these reasons are orthogonal to

Linguistic distance also predicts the number of establishments and 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 45%. A back-of-the-envelope calculation suggests that having 20% more distant speakers increases the percent of English speakers by 2% and attracts 0.25 more IT branches (at the mean of 2.5 in 2003) and 270 more employees (at the mean of 600 in 2003).

The results in Panel B of table 5, when I include large cities, are nuanced but not surprising. Being a large city increases all measures of IT presence, even after controlling for 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 the number of establishments is amplified. While having more English speakers has less impact on whether any firm locates in a large city than a small city, it increases the number of establishments.[27]

6.2 Education

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

$$\begin{aligned} \log(S_{ijt}) - \log(S_{ijt-1}) &= \alpha_0 + \beta' D_j \cdot I(t=2002) + \alpha_1 \log(S_{ijt-1}) + \alpha_2' P_{jt} \\ &\quad + \alpha_3' Z_j \cdot I(t=2002) + \alpha_4 B_{jt} + \gamma_i + \gamma_j + \gamma_{gt} + \mu_{ijt} \end{aligned} \tag{4}$$

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 little variation and large standard errors.

[27] Recall that Hindi speakers are as likely to learn English as those with mother tongues far from Hindi. 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. This suggests that the variation in English to which IT firms respond depends more on whether a state is in the Hindi belt, conditional on region, than on the percent of native Hindi speakers.

where S_{ijt} is enrollment in grade i in district j, region g and time t, $I(\cdot)$ is an indicator function, D_j and Z_j are as above[28] and P_{jt} includes log child population and the fraction urban at time t and $t-1$. γ_i is a grade fixed effect. Three years of data gives me one period of growth before liberalization and one after: I can include district fixed effects to control for district trends and region fixed effects interacted with time to allow for region-specific changes in trend. These fixed effects ensure that my results are identified off within-region variation. I use only data from urban areas[29] and cluster by district. I also include a proxy for skilled labor demand growth, B_{jt}, to control for other changes in demand for education. Calculated along the lines of Bartik (1991), B_{jt} 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. Since the data consist of the number of students enrolled, not enrollment rates, I control for population aged 5-19 from the 1991 and 2001 Census.[30] Details are in the appendix. Enrollment increased dramatically, by 32%, between 1993 and 2002.

I find that educational attainment rises more in districts with lower costs of learning English (table 6). Panel A uses the weighted average while panel B uses the percent distant speakers measure of linguistic distance. Columns 1-2 pool all grades; 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

[28] 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.

[29] The results are robust to using total school enrollment in the district; rural areas show no significant differences in enrollment (see section 7.3).

[30] I cannot use a more precise age group since children start school at different ages and are often held back. Since enrollment and population are all in logs, using enrollment rates calculated as the ratio of enrollment to child population gives me identical point estimates and almost identical standard errors.

Using child population measures from 1993 and 2002 does not impact the results, but the age ranges reported differ between these years.

than for boys, but the difference is not statistically significant.

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.

We may also want to explore growth in districts with more Hindi speakers, since Hindi speakers are more likely to learn English than individuals 1 or 2 degrees away. Districts with more native Hindi speakers exhibit larger increases in enrollment growth; the effect is similar in magnitude to that of distant speakers. As in table 3, 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 new job opportunities could increase schooling. I focus on human capital responses to returns to education. Another channel is through increased family income: it is unlikely that this channel drives my results since the new job opportunities were concentrated among young adults. This should have a larger effect in lower grades while the results indicate similar, if not bigger, effects at older ages.

6.2.1 School enrollment by language of instruction

These results demonstrate that school enrollment responded, but the data does not separate enrollment by language of instruction. I could instead use new data on enrollment by media of instruction from the District Information System for Education (see appendix), but this data is only available after 2002. Not being able to control for pre-liberalization enrollment by language is a problem because I cannot distinguish between pre-existing differences between districts and the impact of growth in export-related jobs.

Nevertheless, I compare enrollment in English to enrollment in Hindi and other local

languages and account for unobserved heterogeneity using district fixed effects. I estimate:

$$\begin{aligned} E_{ljt} &= \alpha_0 + \beta' D_j \cdot \gamma_l + \alpha_1 P_{jt} + \alpha_2' Z_{jl} \\ &\quad + \gamma_l + \gamma_j + \gamma_{gt} + \mu_{ijt} \end{aligned} \tag{5}$$

where E_{ljt} is the fraction of children enrolled in grades 1-8 in language $l \in \{$English, Hindi, and others grouped together$\}$[31] in district j, region g and time t, P_{jt} is child population, $D_j \cdot \gamma_l$ is linguistic distance interacted with an indicator for English instruction, and Z_{jl} is a vector of control variables, interacted with language fixed effects.[32] I also include fixed effects for language and region interacted with year. As before, I include all interactions with percent of Hindi speakers to account for the non-monotonicity. Standard errors are clustered by district.

The results are presented in table 7. The first two columns pool together all years, while columns 3-4 allow the effect to vary by year. An increase in one degree in average distance to Hindi (column 1) increases school enrollment in English by 7.2 percentage points relative to enrollment in Hindi and other languages. An increase in distant language speakers of 20 percentage points (column 20) increases enrollment in English by 4.9 percentage points.

Columns 3-4 allow trends in enrollment to differ by linguistic distance. English enrollment in 2002 was only marginally significantly more responsive to linguistic distance than enrollment in Hindi and other languages, but the interactions with the year dummies are all statistically significant and more or less increasing in magnitude. These results demonstrate that English enrollment in linguistically distant districts grew faster than enrollment in other languages between 2002 and 2007.

[31] Due to data limitations, I have to group together enrollment in all other languages.

[32] Z_{jl} only includes interactions with whether or not a district is in the Hindi Belt, but the results are robust to including interactions with all demographic characteristics of the district used in specification (4).

6.3 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 for skilled wage premiums. As noted above, the theoretical prediction is ambiguous and depends on the relative magnitudes of the demand shocks. I study how the impact of globalization on returns to education varies with linguistic distance by estimating

$$\begin{aligned}\log(wage_n) &= \alpha_0 + \beta_1' D_j \cdot I(t=1999) + \beta_2' D_j \cdot I(t=1999) \cdot HS_n + \beta_3' D_j \cdot I(t=1999) \cdot C_n \\ &+ \alpha_1' D_j \cdot HS_n + \alpha_2' D_j \cdot C_n + \alpha_3 I(t=1999) \cdot HS_n + \alpha_4 I(t=1999) \cdot C_n \\ &+ \alpha_5 HS_n + \alpha_6 C_n + \alpha_7' Y_n + \alpha_8' W_j \cdot I(t=1999) + \gamma_j + \gamma_t + \gamma_{gt} + \mu_n \qquad (6)\end{aligned}$$

where $wage_n$ is weekly wage earnings of individual n in district j in year $t \in \{1987, 1999\}$, HS_n and C_n are indicators for high school and college completion, 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, marital and migration status. 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 coastal 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 have two years of data, allowing me to include district fixed effects, controlling for district wages, and region fixed effects interacted with time which control for region trends. I also cluster the standard errors by district and weight the observations.[33]

Skilled wage premiums rose by less in districts with lower English costs from 1987 to 1999, particularly for high school graduates (see table 8). The coefficients β_2 and β_3 are always negative, but not always significant. The magnitudes are economically significant. The wage premium for high school graduates rises by 5% less over 12 years per degree of linguistic

[33] 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 from 1987.

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.

Note that we should be cautious in interpreting these results. First, since the data do not distinguish between instruction in different languages, I focus on average returns to education. Second, the wage data from the National Sample Surveys is the best available data 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, this provides suggestive evidence that the supply side response of human capital may mitigate rising skilled wage premiums.

7 Threats to validity and robustness checks

This section provides evidence against various threats to validity and a number of robustness checks. I discuss numerous empirical results but omit most tables in consideration of space (all results are available upon request). First, I consider whether other differences between communities that speak different languages could explain my results. An important story to rule out is that people who speak languages distinct from Hindi have different preferences for education. I address this by confirming that linguistic distance is not correlated with the supply of schools. I estimate a regression of the log number of schools in a district on linguistic distance to Hindi, log child population and district control variables. Linguistic distance does not predict the number of schools or the number of schools offering courses in specific fields, such as science, in 1993.

Another way to test preferences for education is to look at pre-trends in school enrollment.

I run the following regression, similar to equation (4), using data from 1987 and 1993:[34]

$$\begin{aligned} \log(S_{ij1993}) - \log(S_{ij1987}) &= \alpha_0 + \beta' D_j + \alpha_1 \log(S_{ij1987}) + \alpha_2' P_j \\ &\quad + \alpha_3' Z_j + \alpha_4 B_j + \gamma_i + \gamma_g + \mu_{ij} \end{aligned} \tag{7}$$

The differences specification accounts for time-invariant heterogeneity across districts.

Results, broken down by grade and gender, are presented in table 9. Linguistic distance is not correlated with pre-trends in school enrollment. Often the coefficient is even negative, although almost always insignificant. The only significant positive result, for primary school boys, is only at 10% and not robust to other linguistic distance measures. The only coefficient significant at 5% is for secondary school boys and is negative. This might cause concern if low enrollment in 1993 allowed for more improvement in the subsequent period, but my estimates control for this trend and allow for convergence by controlling for pre-enrollment.

Another concern is that these communities may have different propensities to migrate. 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 states. These numbers were even smaller in 1999. I rely on linguistic distance from Hindi in 1991, before trade reforms were implemented, which avoids bias from endogenous migration after reforms. In addition, random migration across states before reforms will not bias my coefficients: it will simply bring linguistic distance measures closer to the mean for India. It would only be problematic if migrants had greater preferences for education and speakers of distant languages differentially moved prior to 1991 into districts that will receive export-oriented growth. Given how infrequently people move and the size of these districts, it is unlikely this affected my measure of linguistic diversity.

[34] Data from 1991, when reforms were announced, would be preferable but is not available. Nevertheless, this provides a test of my exclusion restriction, because many reforms were implemented only after 1993 and enrollment is unlikely to have responded so quickly.

The strong correlation between distance to Hindi in 1961 and 1991 (0.81 for the weighted average and 0.88 for percent distant speakers) suggests that the variation is not due to recent migration. Using the National Sample Survey, I can show that migrants tend to be more educated, but linguistic distance is usually not correlated with the likelihood of having migrated, except there is some evidence of *fewer* recent migrants in districts with higher weighted average distance.

Finally, I consider the concern that districts that are linguistically distant from Hindi are less integrated with the rest of the country since they do not, by definition, speak the language spoken by the plurality. This may have impacted the evolution of industries or interstate trade. I test for this by estimating a regression of the percent of workers employed in specific industries, such as manufacturing, tourism or finance, on linguistic distance and Z_j; linguistic distance is not correlated with industrial evolution before 1987.[35]

In the rest of this section, I conduct a few robustness checks. The first uses the percent of schools teaching in the regional mother tongue as a proxy for the cost of learning English, instrumenting with linguistic distance to Hindi. The second check uses an alternate source of variation in English literacy. Third, I examine school enrollment in rural areas and finally, I discuss how the results change when I include state fixed effects and other control variables.

[35] Clingingsmith (2008) demonstrates that growth in manufacturing in the 1930s led to declines in linguistic heterogeneity as minorities learned other languages and claimed new mother tongues. I do not believe this invalidates my strategy. First, despite all the industrial development before 1991, Indian districts still exhibit tremendous linguistic heterogeneity, much of which is geographic. Identity is strongly associated with mother tongue; people may learn other languages, but in the time frame I study, will not claim a new mother tongue. Second, my source of variation is a specific type of linguistic diversity, not simply heterogeneity. Third, the potentially problematic biases are unlikely. This phenomenon would be a concern only if linguistic heterogeneity fell at different rates in different types of districts within a region *and* people of minority languages have different preferences for education. This seems unlikely. There are also extremely few native English speakers (0.02%); we need not worry that many highly able minorities chose to call English their native language. Nevertheless I control for percent native English speakers. If linguistic heterogeneity fell in districts that had more manufacturing in the 1930s, this could possibly affect linguistic distance to Hindi. However, it is most likely people switched to Hindi or to the regional language; neither move will affect my measures since I control for Hindi speakers and focus on within region variation.

7.1 Proxy for cost of learning English

In the first check, I use a proxy for the cost of learning English and instrument for this cost using linguistic distance. The proxies are the percent of schools that teach in the regional mother tongue (the only relevant information available by district). Since English is not a regional mother tongue in India, these proxies are a lower bound on schools not teaching in English. The instruments are the share of district residents who speak languages at each degree of distance from Hindi.

The results from this 2SLS estimation are consistent with the reduced form. The contribution of this exercise is to provide a 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% points less likely to have any IT presence, the establishment of IT firms would be delayed by 8 months, the number of branches would fall by 10% and employment by 50%. The magnitudes for school enrollment growth are also economically significant: a 10% increase in how many schools teach in the mother tongue reduces enrollment growth by 12% over 9 years. Returns to education rise by 2-6% points less over 12 years.

7.2 English speakers in 1961

Another source of variation in English literacy is historical variation in the share of English speakers. The 1961 census includes data on the share of people in each district who speak English as either a first or second language.[36] While the share of English speakers today is endogenous - likely correlated with unobservable determinants of education and IT growth - the historical share of speakers is less likely to be biased. As with linguistic distance to Hindi, we need not worry about reverse causality. 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. While this variation is not exogenous, it is likely to be correlated with different omitted variables than

[36] We have this data only by state in 1991; district data in 1961 only covers the 6 most common languages.

linguistic distance to Hindi. Estimating specifications (3), (4) and (6) using this variation will probably suffer from different biases than any that remain in my main results.

The results are similar to those found using linguistic distance to Hindi. A 10% increase in the share of English bilinguals (40% of a standard deviation) increases the probability of any IT presence by 17% points, 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. A 10% increase in the share of English speakers in 1961 would have increased enrollment growth by 5%. Returns to education are not statistically significant.

7.3 School enrollment in rural areas

I have thus far focused on school enrollment in urban areas since high-skilled exporting opportunities are likely to be concentrated in cities. I now turn to rural areas of the same districts. Census data on languages spoken does not distinguish between rural and urban; rural areas linguistically distant from Hindi are in the same districts as the urban areas that experienced more IT and school enrollment growth.

It is not obvious what we would expect. Since these rural areas are closer to the exporting urban areas, we might see 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 positive spillovers. The results demonstrate that changes in rural enrollment trends after 1993 are not correlated with linguistic distance to Hindi. The point estimates are small in magnitude and of inconsistent sign.

This exercise also constitutes a further check on my exclusion restriction. Many alternate explanations for why school enrollment grew faster in urban areas of linguistically distant districts after 1993 should be relevant for rural areas as well. For example, if preferences for education are correlated with linguistic distance to Hindi, causing faster growth in certain districts, we would expect enrollment trends to differ also in rural areas. The lack of evidence

for differential changes in rural areas supports my identification strategy.[37]

7.4 Controlling for other variables

In the results above, I control for regional variation to eliminate biases from unobserved regional differences. We may also want to control for state variation in unobserved omitted factors such as state-level policies. Unobserved state variation would only cause bias if it is correlated with linguistic distance to Hindi or IT growth, after controlling for regional variation, or school enrollment growth, even after accounting for district pre-trends. Results in table 3 - how distance from Hindi influences language acquisition - and in table 5 - IT firms locate earlier and more often in linguistically distant districts - are fully robust to including state fixed effects. I already account for district trends and region-specific changes in trend in table 6. When I allow for state-specific changes in trend, the standard errors rise and the coefficients fall: I can no longer reject that some coefficients are zero, but I also cannot reject that they are the same as in table 6.[38] Results in table 7 on enrollment by language and table 8 on returns to education are fully robust to state-year fixed effects. Clustering my standard errors by state, instead of district, does not substantially alter my results.

My preferred specification includes district-level demographic and socioeconomic variables prior to trade reforms in India, Z_j. Excluding this vector does not affect any results. Similarly, the results are robust to adding additional variables. For example, variation in labor regulation across states may affect where IT firms locate. This is unlikely since turnover in these firms is remarkably high with firms raiding each other and employees migrating abroad. Nevertheless, I confirm that the results are robust to controlling for pro-worker regulation using data from Besley and Burgess (2004). My main results do not include this variable since it is not available for all states. In measuring engineering college presence, I

[37] This assumes languages are spread evenly across rural and urban areas, which may not be true; unfortunately, the data needed to confirm this is not available.

[38] It is not surprising that the coefficients fall and the standard errors rise. There is less variation within state and therefore less power. In addition, it is not obvious that we should include these additional fixed effects since the signal to noise ratio falls. These regressions would likely be relying on very noisy variation in linguistic distance, exacerbating the attenuation bias.

count only the 26 elite engineering colleges, because district-level data on all engineering colleges are not of the same quality. Another measure is from the list of accredited engineering programs from the National Board of Accreditation of the All India Council for Technical Education. Each program was assigned to a district based on the address of the affiliated college. I only include colleges established prior to 1990. Controlling for this measure does not alter any of my results.

8 Conclusion

In this paper, I demonstrated 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. Next, I showed that IT firms were more likely to set up in districts further from Hindi. Finally, I demonstrated that these districts experienced greater increases in school enrollment growth, but smaller growth in the skilled wage premium.

There are two important implications. The first relates to how countries can mitigate 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 example. The second implication is the evidence for a long run effect of globalization: factor supply may mitigate the increase in wage inequality.

Trade liberalization may also have impacted other development indicators, such as fertility. IT firms employ more women than traditional Indian firms. The male-female ratio among workers 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, potentially impacting age of first marriage and fertility rates. The impact on these other outcomes is an important avenue for future research.

9 References

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.

Azam, M., A. Chin, and N. Prakash (2010). "The Returns to English-Language Skills in India," (mimeo) University of Houston.

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. (2008). "Bilingualism, Language Shift and Economic Development in India, 1931-1961." (mimeo) Case Western Reserve 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.

Jasanoff, J. (2006), Diebold Professor of Indo-European Linguistics and Philology, Harvard University. Personal communication.
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.

Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

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.

Nanda, R. and T. Khanna (2010). "Diasporas and Domestic Entrepreneurs: Evidence from the Indian Software Industry." Journal of Economics and Management Strategy (forthcoming).

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.

Oster, E. and M. B. Millett (2010). "Do Call Centers Promote School Enrollment? Evidence from India." (mimeo) University of Chicago.

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.

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 Data Appendix

10.1 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. Summary statistics are presented in table A1.

10.2 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. Summary statistics are presented in table A2.

10.3 Enrollment by language of instruction in 2002-2007

Beginning in 2002, the Indian government began compiling data on school enrollment by language of instruction as part of a broader effort to collect administrative school-level data (District Information System for Education, DISE). Using school-level surveys among primary and upper primary schools, each district collected data on enrollment in various categories (grade, gender, caste, etc.) and characteristics of the schools, such as the number of teachers, and teacher qualifications. Enrollment is available for the most common languages in each district and then other languages grouped together. I group Hindi and Urdu together and group all other languages together. I use data on child population and the number enrolled by language of instruction, aggregated to the district level. The data is available for 2002 and then annually from 2004 to 2007.

10.4 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.[39] The NSS provides individual-level information on wages

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

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 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.[40] 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.

10.4.1 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.

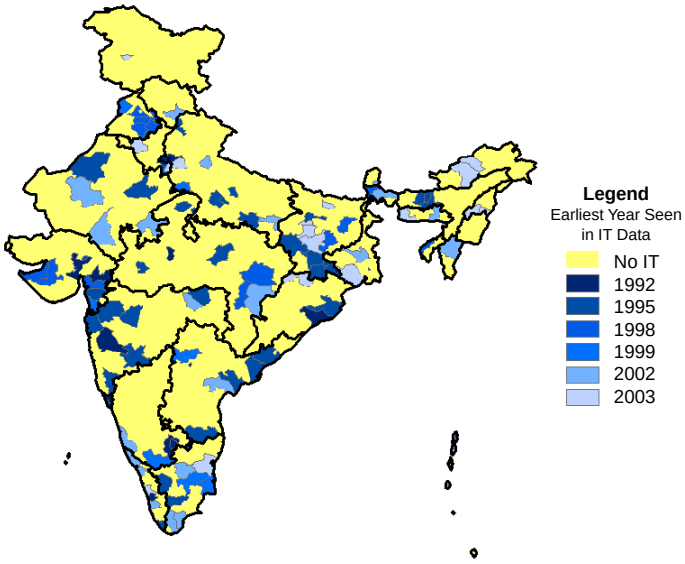
Thus, the proxy is a weighted average of growth rates of national industry employment where the weights are the 1983 share of multi-district employment in each industry. To predict labor demand for educated workers in particular, I further weight these growth rates using the share of employment with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

10.4.2 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.

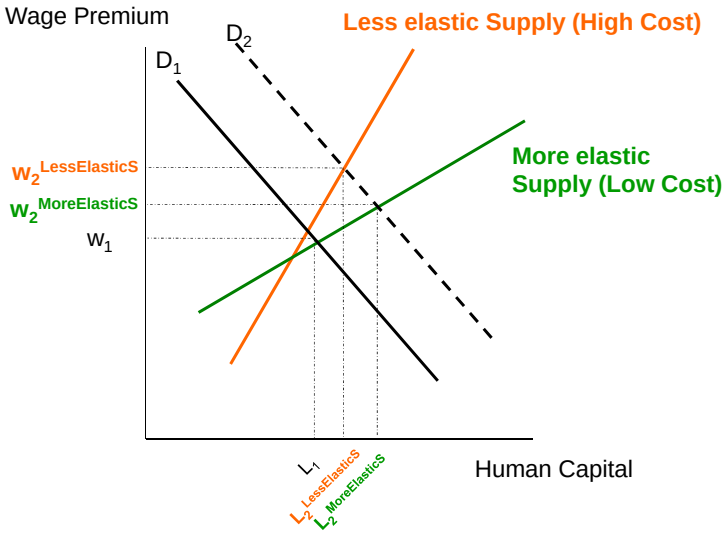
[40] School enrollment in 1987 is used as a control variable. (It is used to calculate a growth rate in specification (4), but is also controlled for on the right hand side). I obtain the same results when using data from the 5th All India Educational Survey from 1986. This is not my preferred data, however, because it is only available at the state level.

Figure 1: Growth of IT Industry



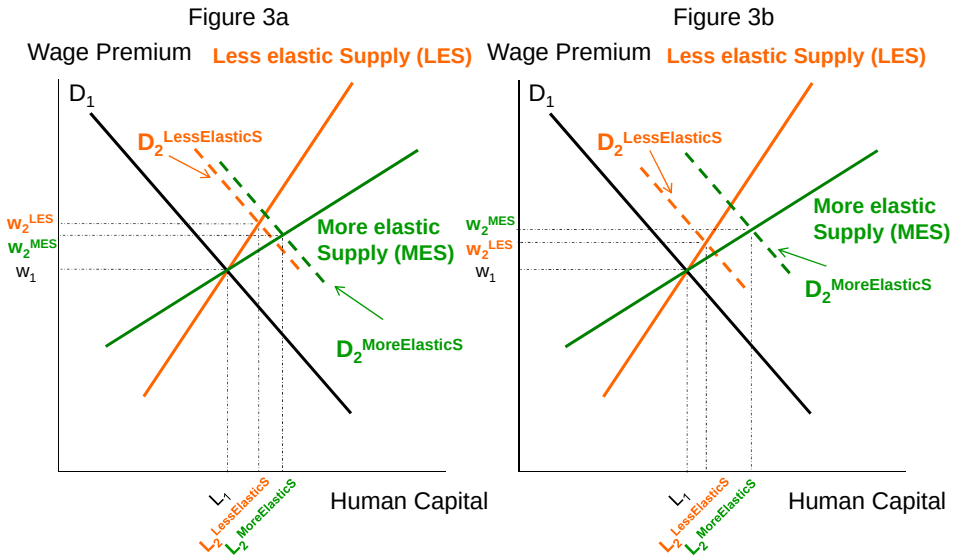
Note: Districts in this map are shaded according to the earliest year an IT establishment from the NASSCOM data is found in the district. Thick black lines indicate states.

Figure 2: Effect of Identical Demand Shocks



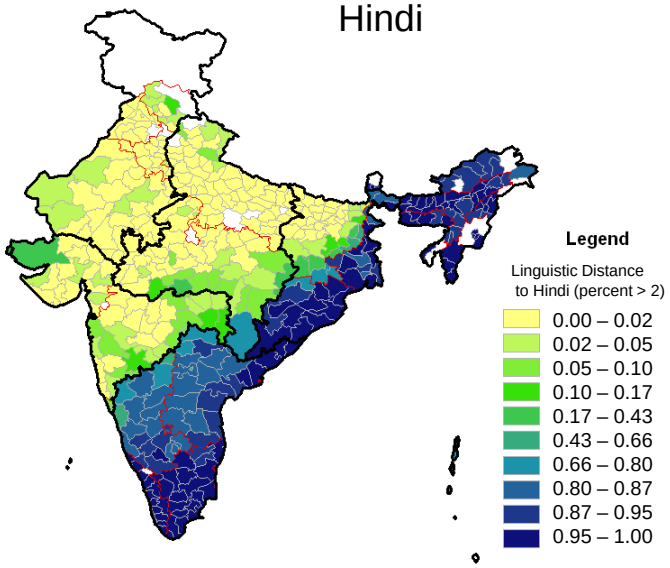
Note: Identical demand shocks for workers with human capital result in a larger increase in human capital accumulation but a smaller increase in skilled wage premiums in a district with a more elastic supply.

Figure 3: Effect of Different Demand Shocks



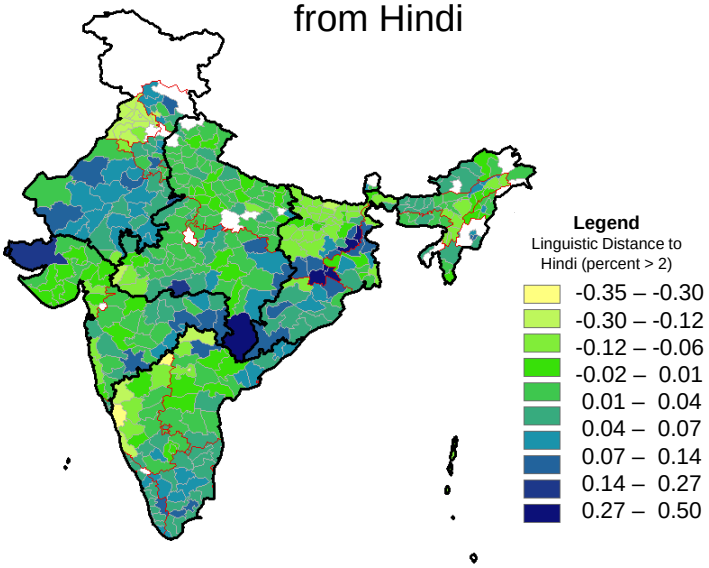
Note: Different demand shocks for human capital result in a larger increase in human capital accumulation in a district with a more elastic supply but the relative change in skilled wage premiums is ambiguous.

Figure 4: 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 lines indicate regions, while thin lines indicate state boundaries.

Figure 5: 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 3 or more degrees from Hindi) on region fixed effects and district-level control variables measured prior to 1991. Thick lines indicate regions, while thin lines indicate states.

Table 1: Measures of Linguistic Distance

Sample Languages	Average Across Languages in 1991			Share of Native Speakers
	Degrees	Percent Cognates	Nodes	
0 Degrees				
Hindi-Urdu	0	100	0	0.456
1 Degree				
Gujarati, Punjabi, Rajasthani	1	67.1	5	0.084
2 Degrees				
Konkani, Marathi	2	56.4	6.5	0.076
3 Degrees				
Assamese, Bengali, Bihari, Oriya	3	64.1	7	0.133
4 Degrees				
Kashmiri, Sindhi, Sinhalese	4	53.3	7.3	0.005
5 Degrees				
All non-Indo European Languages	5	5	12.5	0.244

Sources: Jay Jasanoff (personal communications, 2006), Ethnologue database (2006), Census of India (1991)

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.

** Note that the numbers in the text differ since these are a mean of averages across states, as opposed to a mean across the country.

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

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 state	-0.129	-0.102	-0.089	-0.551 **	-0.568 ***	0.087	0.108
	(0.137)	(0.164)	(0.116)	(0.255)	(0.216)	(0.130)	(0.084)
Distance to Hindi of most spoken language	0.045 ***	0.084 ***	0.122 ***	0.099 ***	0.205 **	0.075 ***	0.076 **
	(0.014)	(0.010)	(0.034)	(0.029)	(0.097)	(0.008)	(0.032)
Share of native speakers in state	0.800 ***	0.790 ***	0.749 ***	1.270 ***	1.270 ***	0.596 ***	0.542 ***
	(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						
Dependent var. mean (weighted)	0.55	0.55	0.55	0.41	0.41	0.62	0.62

Note: This table displays estimates of equation (1). Observations are at the state-mother tongue level. The dependent variable is the percent of native speakers who choose to learn English (conditional on being multilingual). The primary independent variables are measures of the distance between the native language and Hindi. All columns include an indicator for whether the native language is the state's primary language, the share of native speakers in the state, the distance from Hindi of the state's primary language and region fixed effects. Columns 1-4 include year fixed effects, while columns 5-6 (7-8) include only observations from 1961 (1991). In column 1, the omitted group has a linguistic distance of 1 degree away from Hindi. Observations are weighted by the number of native speakers in the state. All columns exclude observations with zero multilinguals and observations where the native language is English; these account for an extremely small number of individuals. Robust standard errors, clustered by state, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 4: Impact of Linguistic Distance on Percent of Schools that Teach English

Dependent Variable:	% of Schools Teaching		% of Schools Teaching	
	(1)	**(2)**	**(3)**	**(4)**
Linguistic distance	0.336 ***		0.199 ***	
(weighted average)	(0.105)		(0.038)	
Percent distant speakers		1.063		0.478
(>2 degrees)		(1.029)		(0.392)
Percent speakers at 0	-0.194	0.298 **	-0.128	0.161 ***
(Hindi/Urdu speakers)	(0.161)	(0.129)	(0.083)	(0.054)
Hindi belt states	-0.196	0.998	-0.282 **	0.323
	(0.212)	(0.752)	(0.141)	(0.218)
Observations	119	119	90	90
R-squared	0.727	0.678	0.672	0.576
Dependent var. mean	0.340	0.340	0.303	0.303

Note: This table displays estimates of equation (2). Observations are at the state-school type (primary, upper primary, secondary or higher secondary) level. The dependent variable is the percent of schools in the state that teach in English (columns 1 and 2) or teach English as a second language (columns 3 and 4). The primary independent variables are measures of the distance from Hindi of languages spoken in the state; the first row presents coefficients of the weighted average across all languages spoken, while the second row presents coefficients of the percent of speakers at least 3 degrees away from Hindi. All columns include fixed effects for region and school type, the percent of native English speakers, the percent of native Hindi/Urdu speakers, 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 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. Robust standard errors, clustered by state, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 5: 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	Weighted average	Percent distant	Weighted average	Percent distant	Weighted average	Percent distant
	(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
Mean of dependent variable	0.153	0.153	1.560	1.560	0.253	0.253	0.818	0.818

Note: This table displays estimates of equation (3). Observations are at the district-year level and includes data from 1995, 1998, 1999, 2002 and 2003. The dependent variable is an indicator for the existence of any IT headquarters or branch in the district (columns 1-2), the number of years an IT firm has been headquartered in the district (columns 3-4), the log of the number of headquarters and branches (columns 5-6) and the log of the number of employees per branch (columns 7-8). The primary independent variable is a measure of the distance from Hindi of languages spoken in the district; odd columns use the weighted average across all languages spoken, while even columns use the percent of speakers at least 3 degrees away from Hindi. In Panel A, I drop observations for the ten most populous cities in India (as of 1987) and in Panel B, I include them but also include interactions. All columns include fixed effects for region and year. Other controls include log of population, the percent of native English speakers, the percent of native Hindi/Urdu speakers, a dummy for a Hindi belt state, household wage income, average wage income for an educated individual, distance to the closest big city, distance to closest airport, the number of elite engineering colleges, a dummy for a coastline and the percent of people who: have regular jobs, have college degrees or secondary school degrees or engineering degrees, are literate, are Muslim, ride a train and the percent of households with electricity. Robust standard errors, clustered by district, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 6: Impact of Linguistic Distance on Grade School Enrollment

Level of Observation:	District-Year-Grade							
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	0.066 ***	0.070 ***	0.040 *	0.020	0.066 ***	0.049 **	0.117 **	0.120 **
	(0.024)	(0.024)	(0.020)	(0.019)	(0.022)	(0.022)	(0.051)	(0.050)
Post * Percent speakers at 0	0.151 *	0.045	0.228 ***	0.192 **	0.192	0.109	-0.025	-0.114
(Hindi/Urdu speakers)	(0.088)	(0.074)	(0.079)	(0.076)	(0.133)	(0.080)	(0.146)	(0.150)
Post * Hindi belt states	-0.062	-0.010	-0.279 ***	-0.289 ***	-0.452 ***	-0.335 ***	0.283	0.503 ***
	(0.096)	(0.084)	(0.083)	(0.077)	(0.106)	(0.080)	(0.178)	(0.195)
Observations	8509	8660	3605	3635	2112	2157	2792	2868
R-squared	0.745	0.706	0.930	0.923	0.942	0.931	0.794	0.753
Panel B: Percent distant speakers								
Post * Linguistic distance	0.348 **	0.296 *	0.324 **	0.268 **	0.539 ***	0.328 **	0.283	0.306
	(0.175)	(0.174)	(0.163)	(0.133)	(0.178)	(0.136)	(0.290)	(0.381)
Post * Percent speakers at 0	0.302 ***	0.190 ***	0.343 ***	0.274 ***	0.374 ***	0.241 ***	0.179	0.090
(Hindi/Urdu speakers)	(0.089)	(0.070)	(0.097)	(0.090)	(0.145)	(0.090)	(0.125)	(0.121)
Post * Hindi belt states	0.131	0.151	-0.092	-0.131	-0.140	-0.150	0.422 **	0.659 **
	(0.141)	(0.148)	(0.109)	(0.089)	(0.127)	(0.113)	(0.199)	(0.307)
Observations	8509	8660	3605	3635	2112	2157	2792	2868
R-squared	0.745	0.706	0.930	0.924	0.942	0.931	0.794	0.753
Mean enrollment in 1993	4817	5685	6066	6807	5184	6103	2867	3926

Note: This table displays estimates of equation (4). Observations are at the district-year-grade level. The dependent variable is the log change in the number of children enrolled in the grade in the district from 1987 to 1993 (time period 'pre') or 1993 to 2002 (time period 'post'). The primary independent variable is a measure of the distance from Hindi of languages spoken in the district interacted with an indicator variable for the time period post reforms; panel A uses the weighted average across all languages spoken, while panel B uses the percent of speakers at least 3 degrees away from Hindi. All columns include fixed effects for district, timeperiod and grade and region interacted with post. Other controls include log enrollment in the pre year, log child population in pre and post years, urbanization in pre and post years, a proxy for labor demand growth and post interacted with: the percent of native English speakers, the percent of native Hindi/Urdu speakers, a dummy for a Hindi belt state, 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 regular jobs, have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity. Robust standard errors, clustered by district, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 7: Impact of Linguistic Distance on Enrollment by Language of Instruction (English, Hindi/Urdu or Others)

Level of Observation:	District-Year-Language			
Dependent variable:	Enrollment / Child Population			
Linguistic distance measure:	Weighted average	Percent distant speakers	Weighted average	Percent distant speakers
	(1)	**(2)**	**(3)**	**(4)**
English * Linguistic Distance	0.072 ***	0.243 ***	0.026 *	0.033
	(0.014)	(0.059)	(0.014)	(0.063)
-- * 2004			0.035 **	0.163 **
			(0.015)	(0.064)
-- * 2005			0.058 ***	0.274 ***
			(0.016)	(0.067)
-- * 2006			0.059 ***	0.274 ***
			(0.017)	(0.068)
-- * 2007			0.060 ***	0.251 ***
			(0.017)	(0.069)
English * Percent speakers at 0 (Hindi/Urdu)	-0.106 *	0.034	0.049	0.077
	(0.058)	(0.055)	(0.062)	(0.067)
p-values from F-tests: all English * Linguistic Distance Variables (p-value)			0.000	0.000
Observations	5057	5057	5057	5057
R-squared	0.229	0.226	0.234	0.232

Note: This table displays estimates of equation (5). Observations are at the district-year-language of instruction level. The dependent variable is enrollment of children in grades 1-8 divided by the relevant child population (from the same datasource). The primary independent variable is a measure of the distance from Hindi of languages spoken in the district interacted with an indicator variable for English as the medium of instruction; odd columns use the weighted average across all languages spoken, while even columns use the percent of speakers at least 3 degrees away from Hindi. All columns include fixed effects for district, language and region interacted with year. Other controls include child population, and interactions of language fixed effects with: the percent of native English speakers, the percent of native Hindi/Urdu speakers, and a dummy variable for whether a district is in the Hindi Belt. All two way interactions for the triple interactions are also included in columns 3 and 4. Robust standard errors, clustered by district, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 8: 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	0.014	0.016	-0.016	-0.023	0.026	0.022	0.078	-0.249	0.104	-0.024
	(0.024)	(0.024)	(0.055)	(0.030)	(0.027)	(0.105)	(0.108)	(0.276)	(0.161)	(0.112)
-- * High school	-0.046 **	-0.048 ***	-0.048	-0.034	-0.052 ***	-0.100	-0.124	-0.135	-0.057	-0.100
	(0.019)	(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	-0.036 *	-0.004	-0.002	-0.024	-0.048	-0.014
	(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	-0.038	-0.028	-0.011	0.070	-0.082	-0.039	-0.008	-0.146	0.044	-0.089
(Hindi/Urdu)	(0.066)	(0.069)	(0.186)	(0.097)	(0.072)	(0.072)	(0.074)	(0.173)	(0.103)	(0.077)
-- * High school	0.051	0.021	0.255	-0.112	0.128 *	0.032	-0.015	0.231	-0.109	0.113
	(0.062)	(0.063)	(0.221)	(0.086)	(0.073)	(0.083)	(0.085)	(0.248)	(0.100)	(0.096)
-- * College	0.006	0.001	0.000	0.012	0.039	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 ***	1.047 ***	1.035 ***	1.146 ***	0.976 ***	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	0.075	0.108	-0.026	-0.033	0.123	-0.024	0.018	-0.119	-0.121 *	0.004
	(0.079)	(0.079)	(0.191)	(0.087)	(0.085)	(0.067)	(0.068)	(0.143)	(0.066)	(0.074)
-- * College	0.166 **	0.171 **	0.109	-0.032	0.221 **	0.056	0.076	-0.063	-0.113	0.104
	(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 ***	1.040 ***	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	71255	58631	12624	22196	49059	71255	58631	12624	22196	49059
R-squared	0.670	0.667	0.686	0.648	0.671	0.670	0.667	0.686	0.648	0.671

Note: This table displays estimates of equation (6). Observations are at the individual level using data from 1987 (pre) and 1999 (post). The dependent variable is the log weekly wage and the primary independent variables are measures of the distance from Hindi of languages spoken in the district interacted with fixed effects for high school completion and college completion; odd columns use the weighted average across all languages spoken, while even columns use the percent of speakers at least 3 degrees away from Hindi. All columns include fixed effects for district, year and region interacted with post. Individual-level controls include age, age squared, married, male, whether the individual has ever moved, whether the individual is self-employed and high school and college also interacted with post and the 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. Robust standard errors, clustered by district, are shown in parentheses. *** 1%, ** 5%, * 10%

Table 9: Impact of Linguistic Distance on Grade School Enrollment in 1993

Level of Observation:	District-Grade							
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								
Linguistic distance	-0.013	-0.030	0.035	0.038 *	0.005	-0.012	-0.101	-0.124 **
	(0.029)	(0.030)	(0.022)	(0.023)	(0.025)	(0.026)	(0.062)	(0.059)
Percent speakers at 0	-0.067	0.266 ***	0.051	0.124	-0.198	0.159 *	-0.104	0.542 ***
(Hindi/Urdu speakers)	(0.099)	(0.100)	(0.089)	(0.098)	(0.128)	(0.084)	(0.165)	(0.183)
Hindi belt states	-0.363 ***	-0.469 ***	-0.393 ***	-0.486 ***	-0.044	-0.133	-0.558 ***	-0.724 ***
	(0.111)	(0.116)	(0.092)	(0.095)	(0.109)	(0.084)	(0.214)	(0.245)
Observations	4105	4256	1770	1800	1011	1056	1324	1400
R-squared	0.678	0.606	0.831	0.783	0.860	0.824	0.610	0.554
Panel B: Percent distant speakers								
Linguistic distance	0.073	0.103	0.187	0.225	0.122	0.172	-0.076	-0.135
	(0.198)	(0.206)	(0.187)	(0.186)	(0.175)	(0.135)	(0.337)	(0.421)
Percent speakers at 0	-0.067	0.254 **	0.132	0.217 **	-0.167	0.185 **	-0.240	0.371 **
(Hindi/Urdu speakers)	(0.101)	(0.099)	(0.090)	(0.100)	(0.142)	(0.092)	(0.165)	(0.181)
Hindi belt states	-0.315 **	-0.396 **	-0.290 *	-0.361 **	0.030	-0.026	-0.577 **	-0.769 **
	(0.151)	(0.167)	(0.151)	(0.151)	(0.125)	(0.105)	(0.239)	(0.335)
Observations	4105	4256	1770	1800	1011	1056	1324	1400
R-squared	0.678	0.606	0.831	0.783	0.860	0.824	0.608	0.551
Mean enrollment in 1993	4817	5685	6066	6807	5184	6103	2867	3926

Note: This table displays estimates of equation (7). Observations are at the district-grade level. The dependent variable is the log change in the number of children enrolled in the grade in the district from 1987 to 1993. The primary independent variable is a measure of the distance from Hindi of languages spoken in the district; panel A uses the weighted average across all languages spoken, while panel B uses the percent of speakers at least 3 degrees away from Hindi. All columns include fixed effects for region and grade. Other controls include log enrollment in the 1987, the percent of native English speakers, the percent of native Hindi/Urdu speakers, a dummy for a Hindi belt state, log child population in pre and post years, urbanization in pre and post years, a proxy for labor demand growth, 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 regular jobs, have college degrees or secondary school degrees, are literate, are Muslim, ride a train and the percent of households with electricity. Robust standard errors, clustered by district, are shown in parentheses. *** 1%, ** 5%, * 10%

Table A1: 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 A2: 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%

11 Theory Appendix (NOT FOR PUBLICATION)

11.1 Production processes for traded and non-traded goods

In this appendix I set up the model that provides the implications tested in the paper. First, I specify production processes for two goods, one of which is tradeable. 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, it becomes possible to produce and export the traded good X. Firms 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 (telecommunication networks) that is slow to change, immobile entrepreneurs or the business environment more generally.

11.2 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.[41] 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.[42] 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

[41] 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.

[42] 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.

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, $\widehat{q} = \frac{w_H H + w_E E}{w_U (H+E)}$.

11.3 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

$$\alpha_H Y + F w_E^{-\frac{1}{\beta}} \left(p_X (1-\beta)\right)^{\frac{1}{\beta}} = P\left(\frac{w_H - w_U - t w_U}{w_U}\right) \tag{8}$$

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

$$\alpha_H Y = P\left(\frac{w_H - w_U - t w_U}{w_U} - \frac{w_U t + w_E - w_H}{w_U\left(\mu_j - 1\right)}\right) \tag{9}$$

$$F w_E^{-\frac{1}{\beta}} \left(p_X (1-\beta)\right)^{\frac{1}{\beta}} = P\frac{w_U t + w_E - w_H}{w_U\left(\mu_j - 1\right)} \tag{10}$$

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

$$\alpha_L Y = P\left(1 - \frac{w_H - w_U - t w_U}{w_U}\right) \tag{11}$$

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. 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 (11) and (8), 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^*:

$$(2+t)\left(1+\frac{\alpha_H}{a_L}\right)+\frac{\alpha_L}{\alpha_H}=\frac{1}{w_U^*}\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)-\frac{F}{P}\left(\frac{1-\alpha_L w_U^*}{\alpha_H p_X(1-\beta)}\right)^{-\frac{1}{\beta}} \tag{12}$$

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 $H^* = P\left(\frac{1}{w_U^*\alpha_H} - \frac{\alpha_L}{\alpha_H} - 1 - t - \frac{t}{\mu_j-1}\right)$. ■

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.*

Proof. 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 (11) and (9):

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

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 (10), we get

$$0=\left(\frac{P}{F}\right)^{-\beta}\left(\frac{1}{w_U^*}\left(\frac{1}{\alpha_L}+\frac{1}{\alpha_H}\right)-\left[\frac{a_L}{\alpha_H}+\left(\frac{a_H}{\alpha_L}+1\right)(2+t)\right]\right)^{-\beta}-\frac{A-w_U^*B}{p_X(1-\beta)}=G\left(w_U^*;\mu_j\right) \tag{13}$$

Thus,

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

Writing the other variables in terms of w_U^* and differentiate with respect to $\mu_j - 1$ 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. ■

Proposition 3 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.

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

$$F \leq P\frac{t}{\mu_j - 1}\left[\frac{\alpha_L}{\alpha_H}\frac{1}{p_X(1-\beta)}\left(\frac{1}{\alpha_L} - \frac{\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)}{\left[(2+t)\left(1+\frac{\alpha_H}{\alpha_L}\right)+\frac{\alpha_L}{\alpha_H}\right]+\frac{t}{\mu_j-1}}\right)\right]^{\frac{1}{\beta}} = \overline{F}\left(\mu_j\right) \tag{14}$$

Proof. First, I prove that if $w_E^* = w_H^*$, condition (14) holds. From Proposition 1, we know

$$\frac{F}{P}\left(\frac{w_E^*}{p_X(1-\beta)}\right)^{-\frac{1}{\beta}} = \frac{F}{P}\left(\frac{w_H^*}{p_X(1-\beta)}\right)^{-\frac{1}{\beta}} = \frac{F}{P}\left(\frac{1-\alpha_L w_U^*}{\alpha_H p_X(1-\beta)}\right)^{-\frac{1}{\beta}} \leq \frac{t}{\mu_j - 1} \tag{15}$$

From the proof of Proposition 1, equation (12), we can write

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

Solving for w_U^*, we get

$$w_U^* \geq \frac{\left(\frac{1}{\alpha_H}+\frac{1}{\alpha_L}\right)}{(2+t)\left(1+\frac{\alpha_H}{a_L}\right)+\frac{\alpha_L}{\alpha_H}+\frac{t}{\mu_j-1}} \tag{16}$$

Solving (15) for F and plugging in (16), we get

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

■

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 (10), we know that

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

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 (13)

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

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.*

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

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 3a). 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 3b).

I test two predictions of this model, assuming districts in India are in case B (since there is a return to speaking English[43]). 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.

[43] See Munshi and Rosenzweig (2006).