LANGUAGE POLICY AND HUMAN DEVELOPMENT

David D. Laitin and Rajesh Ramachandran*

August 2015

Abstract

This paper explores how language policy affects the socio-economic development of nation states through two channels: the individual's exposure to and (in reference to an individual's mother tongue) linguistic distance from the official language. In a cross-country framework the paper first establishes a robust and sizeable negative relationship between an official language that is distant from the local indigenous languages and proxies for human capital and health. To establish this relationship as causal, we instrument language choice with a measure of geographic distance from the origins of writing. Next, using individual level data from India and a set of eleven African countries, we provide micro-empirical support on the two channels - distance from and exposure to the official language - and their implications for educational, health, occupational and wealth outcomes. Finally, we suggest policy implications based on our findings.

JEL: I24, I25, I28, Z18.
Keywords: Language Policy, Institutions, Development.

*Laitin: Department of Political Science, Stanford University, dlaitin@stanford.edu. Ramachandran: Department of Microeconomics and Management, Goethe University, ramachandran@econ.uni-frankfurt.de

1 Introduction

One remnant of the colonial era is its language legacy, with a large majority of post-colonial countries retaining English, French, Portuguese and Spanish as their official languages; and relying on these languages for education and administration.[1] These languages tend not to be the native language of any indigenous group and are typically distant from the languages spoken by the local population.[2] With a distant language serving as a gatekeeper allocating education, jobs, political participation and self-esteem, we explore the consequences of language choice for the economic and human development of post-colonial states.

It is widely acknowledged that language is central to the organization of human society and interpersonal relations. Without this method of communication, no leader could command the resources necessary for an inclusive political system extending beyond family and neighborhood (Weinstein, 1983). The choice of language influences human capital, as it provides those who speak the official language of the state with greater access to economic and political opportunities.

In order to conceptualize the notion of "distant languages", we employ the measure of structural distance between languages based on Ethnologue's (Lewis et al., 2014) language trees. Ours is a weighted measure that calculates the average distance and exposure of the local population's languages from the official language. The theoretical framework advances two channels through which the choice of official language affects socio-economic development,

[1] In the data we define the official language as one in which the constitution or the organic laws of the country have been written. For a general discussion on official language, see Eastman 1983, 37.

[2] Exception is the continent of South and North America, where due to the spread of germs from the old world, nearly the entire local population was decimated. The colonialists in turn settled in these places and hence the former colonial language is also the native language of the majority of the population.

the *distance* from and the *exposure* to the official language.[3] More specifically, we assume that increasing distance and lower exposure results in increasing learning costs and consequently reduces the level of human capital in society. Similarly the use of a distant language increases the cost of acquiring and processing pertinent health information, and acts as a barrier to fostering desirable health behavior, as well in affecting access and quality of health care provided. These differences in physical and mental human capital in turn translate into differences in productivity and wealth.

We demonstrate that the constructed measure of language distance and exposure, in line with our theory, is a statistically significant and economically meaningful correlate of proxies for human capital, health, income and productivity.[4] The pattern of lower distance to the official language, implying higher country wealth and human development, holds both within and across continents.

To better understand the relationship we examine the motivations underlying choice of official language in Sub-Saharan Africa, and provide evidence that the language policy observed today is almost indistinguishable from the one during the colonial period; and hence does not reflect active choices made by the political elite. By studying factors affecting official language choice, we find that it is not past wealth or development levels but in fact possessing a writing tradition that is a key explanatory factor. Using distance from the sites of invention of writing as an instrument for our constructed measure, we show that, like the OLS estimates, the IV estimates are also negative and significant, providing a causal logic linking higher distance from

[3]This second channel is especially relevant to Africa. While teachers in Africa rely on code switching (see Brock-Utne and Holmarsdottir 2003) between official and local languages to better communicate with students, it works against passing national examinations and qualifying for high status jobs.

[4]The proxies used are internationally comparable cognitive test scores, life expectancy, log GDP per capita, log output per worker, and as a composite measure the Human Development Index (HDI).

the official language to lower levels of socio-economic development. The economic magnitude of the estimates is large, and shows that if a country like Zambia were to adopt Mambwe instead of English as its official language, it would move up 44 positions on the HDI ranking and become similar to a country like Paraguay in human development levels.

We next provide empirical support in favor of the two assumptions made under the theoretical framework. Data from the 2005/06 National Family and Health Survey of India (International Institute for Population Sciences (IIPS) and Macro International, 2007) provides evidence for the first channel, viz. that individual level distance to the official language affects various socio-economic outcomes. The data reveal that the distance to the official language of the state in which the individual is resident predicts lower schooling and occupational outcomes. For a Hindi speaker resident in West Bengal, where an Indo-European language Bengali is used, moving to a state using a Dravidian language (e.g. Tamil Nadu) as opposed to moving to another state where an Indo-European language is official (e.g. Uttar Pradesh), would reduce average years of schooling by around one year and decrease the probability of using a mosquito net, of ever having heard about AIDS, or holding a white-collar job by four, nine and three percentage points, respectively. As the identification strategy accounts for state, language group, and time specific trends through the inclusion of fixed effects, as well as a rich set of other controls, we can be reasonably confident that the effects of language distance are being captured.

Evidence on the importance of the exposure channel is evaluated using data from a set of eleven African countries where English is the medium of instruction. It is shown that exposure to English at home is a significant factor in explaining student performance. Using a model with classroom fixed effects and a rich set of pupil controls at the home level, we find that exposure to English increases the probability of reaching the minimum reading level by around ten percentage points; and Math scores increase by around one-fifth of a standard deviation.

A natural question, assuming that reliance on colonial languages constrains human development, is why many post-colonial countries have relied so heavily on colonial languages for

education and administration. Building on ideas presented in Laitin (1994, 2000), a subsequent paper will explain the perpetuation of inefficient language policies. It will focus on potential opposition of language groups whose languages are not officialized. But it will also show how elites, taking advantage of their command of colonial languages, can perpetuate their returns to political power by (over-)emphasizing the costs of transition to an indigenous language policy. Though this paper does not assess the political economy of a shift in language policy, it provides a reasonable estimate of the costs entailed in the reliance on colonial language.

2 The cross-country framework

One institutional factor distinguishing "developed" from many "developing" nations today is their official language. The official language in developed nations is typically one which is spoken and used widely by a majority of the population. To be sure, at the time when the official languages of today's developed states were chosen, they were not universally understood, even in countries as linguistically homogeneous today as France (Weber, 1976) or Japan (Laitin, 1992, 14), but in those countries, there was a core indigenous group fluent in the official language of state. On the other hand, in most developing states today, the official language is often one that is neither indigenous nor spoken by citizens outside of an elite minority.

Sub-Saharan African countries in particular have primarily chosen non-indigenous languages, typically distant from the local language, as official. Relying on current data from Albaugh (2014, 237), for those sub-Saharan countries that are in our dataset, an average of only 18.7 percent of the population could speak the official language of the state. This reaches depths of 4.5 percent for Niger and 5 percent for Guinea and Malawi. And these low cases include countries that were ruled directly (Niger) where the colonial language was the medium of rule and those that were ruled indirectly (Malawi) where indigenous language and cultures were supposedly recognized. To be sure, there is great variation across estimates on what counts

as "speaking" the official language of the state. However, we can surmise that these figures would be lower if the criterion were basic literacy in that language. Secondary education, the key to joining the modern sector in Africa, is almost entirely conducted through the media of non-indigenous languages throughout Africa, with possible exceptions of Somalia (before state collapse) and Mauritania (Albaugh 2014, Appendix A).

The effect of reliance on colonial languages in sub-Saharan Africa, though, a factor highlighted by a small group of educationalists and pedagogues (Alidou et al., 2006), has engendered only a few papers providing systematic quantitative evidence. Eriksson (2014) exploits a language policy change in South Africa, and shows that the provision of two extra years of local language instruction, instead of in English or Afrikaans, had a positive effect on wages, the ability to read and write, on educational attainment, and on the ability to speak English. Ramachandran (2012) using a triple difference-in-differences strategy, finds that introduction of mother tongue schooling for the largest ethnic group in Ethiopia in 1994 resulted in increasing the probability of completing primary schooling and the ability to read a complete sentence by 7 and 28 percent, respectively. Taylor et al. (2013) employ a school fixed-effects model and find that provision of mother tongue instruction in the early grades significantly improves English acquisition, as measured in grades 4, 5 and 6. Laitin et al. (2015) report on a quasi-random introduction of indigenous language medium of instruction in a region of Cameroon. They report that after three years, treated students' overall scores were double those of control students. In the OECD, Dustmann et al. (2010) and Dustmann et al. (2012) highlight the language repertoires of the students as the single most important factor in explaining differences between immigrant and native children's school outcomes.[5] This paper seeks to complement the early

[5]See Angrist and Lavy (1997) on the introduction of Arabic instruction for secondary education in Morocco for findings that do not recommend indigenous language instruction. Our focus in this paper on early education and in places where students do not have easy access outside the school to the language of instruction reduces the relevance to this paper.

controlled studies of language media in schools to assess more generally the costs in human development on the national scale of reliance on official languages that are “distant" from the indigenous languages of post-colonial countries.

2.1 Data and country level measure of distance

For a cross-country estimation of the relationship of linguistic distance to economic outcomes, we need an algorithm to determine distance between any two languages and a measurement strategy to calculate average distance for any population of its language to that of the official language. In order to conceptualize the notion of distances between languages, the measure based on Ethnologue's linguistic tree diagrams is used. The distance between any two languages i and j based on Fearon (2003) is defined as:

$$d_{ij} = 1 - \left(\frac{\#\text{ of common nodes between i and j}}{\frac{1}{2}(\#\text{ of nodes for language i} + \#\text{ of nodes for language j})}\right)^{\lambda}. \quad (1)$$

From Equation 1 we see that if two languages belong to different language families, i.e. the number of common nodes between them is 0, their distance is equal to 1, which by construction is the maximum distance between any two languages. The value of λ determines the relative distance between two languages which belong to the same family compared to two languages that belong to different families. For instance, consider Spanish and Catalan belonging to the Indo-European language family and having seven branches in common.[6] Choosing a value of λ equal to 0.5 would imply the distance between Spanish and Catalan is equal to .116. Choosing a lower λ, such as 0.05, would give greater weight to the similarity in the earlier nodes, and the distance between Spanish and Catalan would fall to 0.012. Of course, if two languages differ at the first node, as would be the case for Spanish and Tamil, whatever the value of λ the distance

[6]The number of nodes before the Spanish and Catalan language are reached starting from an Indo-European language tree are 10 and 8, respectively.

score would remain at 1. As no theoretical basis has been established for choosing the correct value of λ, following Fearon (2003), we fix the value of λ equal to 0.5 in our analysis.[7]

We can now calculate a weighted measure of average distance of a country's population from the official language. The official language/s of the countries included in the regression on Tables III and IV are shown in the excel file accompanying the online Appendix. The data on the number and size of linguistic groups in the country comes from the data of Fearon (2003), which takes into account all linguistic groups that form at least 1% of the population share.[8]

The average distance from the official language (ADOL) for any country i is calculated as:

$$ADOL_i = \sum_{j=1}^{n} P_{ij} d_{jo}, \tag{2}$$

where n are the number of linguistic groups in the country, P_{ij} refers to the population share of group j in country i and d_{jo} refers to the distance of group j from the official language. The coding rules when there is more than one official language depend on whether there is a group associated with an official language in which social and political mobility is possible for monolinguals of that language (e.g. Germans in Switzerland; Afrikaners in South Africa) or whether the group associated with that official language must have proficiency in another official language for full mobility prospects (e.g. Urdu speakers in Pakistan). For the former, language distance equals zero. In case of the latter, language distance equals one-half the distance between their indigenous language and the less prestigious official language plus one-half

[7]We also re-do our analysis using multiple values of λ that have been used in the literature. Our results remain qualitatively very similar and are shown in Table A.1 of the online Appendix.

[8]Fearon's (2003) classification of groups, relying on a range of secondary sources, has been recognized in the literature as both principled and objective. See Esteban et al. (2012) for a discussion of the same.

the distance between their language and the more prestigious official language.[9]

The constructed measure of ADOL is distinct from indices of linguistic diversity used in the literature (Greenberg 1956, Alesina et al. 2003, Desmet et al. 2009); while measures of linguistic diversity are concerned with the level of linguistic heterogeneity within a country, our index measures how distant the official language of a country is from the languages spoken within a country. As the choice of official language is not restricted to a set of indigenous languages, countries that are classified as having low levels of linguistic diversity nonetheless maybe linguistically distant from the official language. To see this, consider countries such as Angola, Burundi, Lesotho, Rwanda, Swaziland and Zambia; all have a value of linguistic diversity as measured by the Greenberg index of less than 0.005, however their average distance from the official language is at least 0.50, as all use a non-indigenous imperial language as their official one.[10]

The measure of ADOL is closest in spirit to the peripheral index proposed by Desmet et al. (2005). Their index measures the distance of all peripheral groups to the dominant central group, which in their application referred to the largest linguistic group in the country. Here we extend their application where the official language is the language of the largest linguistic group in the country. In our application, ADOL differs from the original application of the peripheral index when the official language is not the language of the largest ethnic group, such as

[9]In Caribbean countries (Haiti, Jamaica and Guyana) the size of the linguistic groups speaking the official language (French in Haiti and English in Jamaica and Guyana) in the data is estimated to be 95, 98 and 43 percent, respectively. However the correct classification (for a large number of individuals subsumed in this category) of the linguistic background would be "French Creole" in the case of Haiti and "English Creole" in the case of Jamaica and Guyana. The distance here between Creole and the standardized form is taken to be zero whereas in reality there are significant differences. Thus for these countries, the language distance is underestimated.

[10]In fact Angola, Lesotho and Zambia all have the maximum possible distance of 1.

Amharic in Ethiopia, or when a country has adopted a non-indigenous language to act as their official language, as is the case in most post-colonial states in Sub-Saharan Africa and South Asia.[11]

Figure I shows a color coded map of the world depicting the average distance from the official language for the sample of countries included in our study. For illustrative purposes, Table I also provides the average language distance scores for a selected set of ethnic groups and countries.[12]

Insert Table I

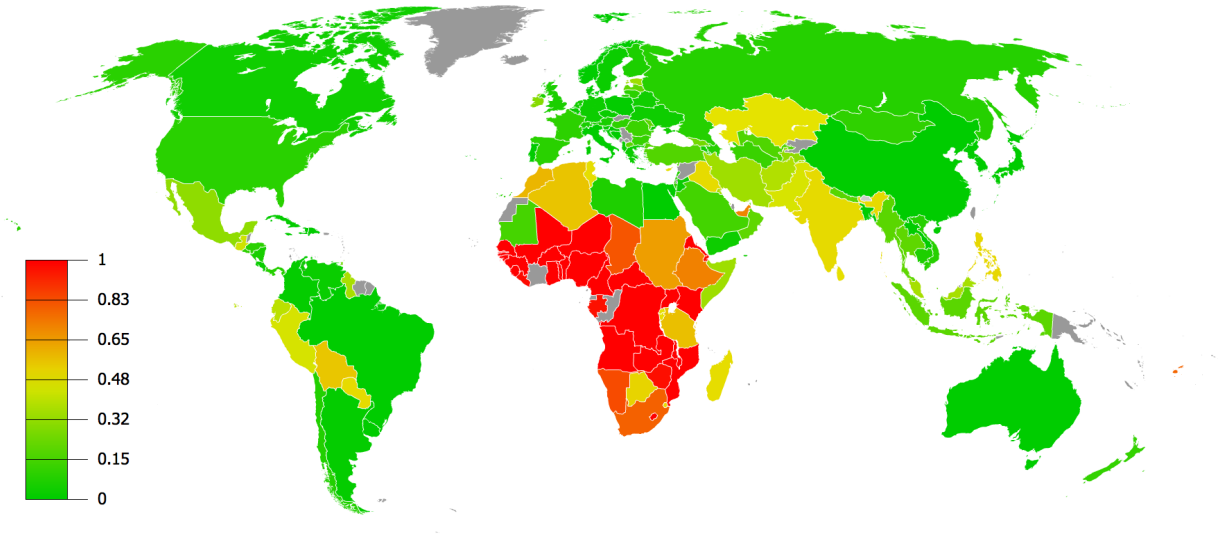

Figure I: **World Distribution of Average Distance from Official Language**

The grey colored areas refer to countries on which information on language distance is not available.

[11]Google Scholar reports 89 citations to the paper that introduced the peripheral index. Ours will be the first that applies it to the case where an official language is not the plurality language in the country.

[12]The following link (http://shar.es/NkqCj) provides an interactive map which shows the average distance from the official language for all countries included in our sample.

Table II in turns shows descriptive statistics for a range for interesting socio-economic variables for the entire sample, as well as by quartiles of language distance. Strikingly, all variables considered are seen to be monotonic with respect to ADOL.

Insert Table II

2.2 Why does the distance from the official language matter?

Outlining a clear theoretical mechanism is essential in order to understand through which channels choice of official language affects socio-economic development. The framework will subsequently guide us in our empirical exercise. It will also enable a theoretically-founded interpretation of the results, as sketched below with a formal exposition provided in the online Appendix. The two main facets of socio-economic development that our theory links to official language choice are *human capital formation* and *health*.

Individuals in our framework are assumed to be utility maximizers and choose the level of human capital and preventive health behavior to maximize their wellbeing. The cost of human capital formation for any individual i is assumed to be a function of their ability, the distance of individual i from the official language of the country, and to the amount of exposure of individual i to the official language.

In our theory, the *first* assumption is the greater the distance of individual i to the official language, the higher the cost of obtaining human capital and participating in the economy. This first assumption implies that all else equal, a native French speaker would face a lower cost of learning Italian than a native German speaker, as Italian is structurally closer to French than German, and hence obtain higher human capital. The *second* assumption states that the greater the exposure to the official language, the lower the costs of obtaining human capital and participation in the economy. The second assumption in turn implies, all else equal, Akan speakers from Ghana would face lower learning and participation costs and obtain higher human capital

due to the use of English as the official language in the United States as compared to in Ghana, as their level of exposure to English would be much higher in the United States.[13]

The health behavior of individuals is assumed to be affected through *two* distinct channels. The first one, directly linked to official language choice, is through language acting as a barrier to health care access or to the comprehension of pertinent health information (Bowen 2001, Djité 2008, Chapter 3, Higgins and Norton 2009, Underwood et al. 2007).[14] Evidence supplied by Translators Without Borders (TWB) provides pertinent information. In a study conducted in Kenya after the outbreak of the Ebola crisis, they were able to gauge the importance of language in comprehending health related information. Randomly sampled rural and urban respondents were first tested on their knowledge regarding channels of Ebola transmission. Before treatment, the average share of correct answers was as low as 8 percent. Subsequently, a random half of the study population was provided information regarding Ebola transmission using posters in Swahili, and the other half using posters in the official language of English. Post-treatment the Swahili-treatment group had on average 92 percent of the answers correct whereas the English-treatment group had only16 percent of the answers correct.[15]

To be sure, governments can use multi-lingual messaging to transmit health education to

[13]Although there is no official language in the Unites States from a legal point of view, by the behavioral criterion stipulated in footnote 1, we can consider English as official in the United States.

[14]Clinical research assumes the immense difficulties of patients not sharing a native language with medical professionals. Recommendations involve (rather expensive) solutions, basically confirming our view of the national health costs of high ADOL. See Bauer and Alegría (2010); and Karliner et al. (2007). Other research shows that the problem of linguistic difference between doctor and patient is exacerbated in emergency departments (see Ramirez et al. 2008). For evidence from Africa on linguistic distance across neighbors and reducing crucial information transmission with identifiable effects on child mortality, see Gomes (2014).

[15]For more details refer to Translators Without Borders (2015).

all their citizens. However, it is costly to transmit technical information in languages that have not been standardized, as would happen if those languages had official status. Indeed, much of the primary health care information, especially in Africa, is only available in the official language.[16] In any event, our claim about the favoring of official language in the health sector is not that it is an absolute impediment to available health care, but that on average, the probability of successful access and treatment is reduced.[17]

The second channel through which language policy affects health behavior is indirect and works through the conduit of human capital. The reasoning being that education matters for the ability of individuals to be able to process and use information regarding best health practices (refer to Dupas (2011, 435-436) and the citations contained therein for an overview on the complementarities between education and health behavior; also refer to De Walque (2007, 2009) on relationship between education, HIV and preventive sexual behavior in Sub-Saharan Africa and De Walque (2010) on the relationship between education and smoking behavior).

It is important to note that our measure of ADOL subsumes both the theoretical concepts of *distance* and *exposure* to the official language. Referring to the formal exposition in the Appendix, the notion of distance from official language is self-evident from equation 2; for the case of exposure, in the cross-country analysis we attribute the distance of other ethnic groups ($i \neq j$) in the country to be a measure of exposure of the ethnic group i to the official language.

[16]Refer to http://translatorswithoutborders.org which highlights lack of availability of health material in local languages as major impediment to combatting recent crises such as drought in the Horn of Africa or Ebola in West Africa. Their site also contains information on initiatives currently being undertaken that help overcome these language barriers.

[17]Refer to, Chang and Emzita (2002), Chantavanich et al. (2002), Drysdale (2004) and Tansey et al. (2010), in the context of South Africa, Namibia, Greater Mekong Sub-region and the Pacific Islands, on the role of lack of local language material as an impediment, and the use of local languages as a key strategy, in checking the growth of HIV incidence among high risk populations such as in the transport industry and among migrant workers.

The intuition here is that the greater the average linguistic distance of all other groups to the official language, the less likely the other groups $j \neq i$ will be relying on the official language in their everyday lives, and therefore the lower the exposure of members of group i to the official language. As the measure takes into account the distance of all ethnic groups, the concepts of group distance and exposure are both captured by the same measure.

2.3 The choice of proxies for our dependent variable

The discussion in the previous section assumes that the choice of official language influences the level of human capital in society by affecting the cost of acquisition. A measure of human capital is thus a natural outcome variable to explore. Since this cannot be measured directly, we need reasonable proxies, and for this we need to address two issues. First, available measures of human capital, such as years of schooling or enrollment rates, mostly capture quantity and not quality, which obscures the variation in the levels of learning that students at the same grade level exhibit across countries. The problem becomes especially pronounced as enrollment levels and years of schooling have sharply risen in developing countries over the past decades, but learning outcomes have either stagnated or even worsened. For instance, in some countries in Sub-Saharan Africa, up to 40 percent of young people who have attended primary school for five years have neither the essential skills to avoid lapsing into illiteracy, nor the minimal qualifications to secure a job (UNECOSOC, 2011). Similarly the latest available round of Demographic and Health Survey (DHS) data from 35 Sub-Saharan African countries shows that 33 percent of the males recorded as having between 4 to 7 years of schooling are still unable to read a complete sentence. This implies that available quantitative measures of human capital might be a poor indicator of actual stock of knowledge, especially for developing countries.

A second issue relates to the time it takes to translate values on average distance to observable changes in levels of human capital. If language choices for post-colonial states were made post World War II, it might take two generations for the effects of this choice to affect standard

outcome variables in a significant way such as output per worker.

Our proposed solution to these problems is to use four distinct measures, each with different advantages, and to show that our results are robust across these different measures, allowing us to combine them for general analysis. For our most direct measure, one that captures the actual level of knowledge (or human capital), we rely on test scores from comparable student achievement tests across countries.[18] Using such a measure however comes at a potential cost. These internationally comparable test scores are available only for 70 countries, and these include only 6 from Sub-Saharan Africa.

As an indirect measure of human capital, here working through the channel of health, we measure life expectancy. As just discussed, we assume that populations with high rates of human capital, controlling for country wealth, are better able to take advantage of modern health resources and communicate successfully with medical staff, thereby improving diagnoses and implementation of remedies. These differences in knowledge (based on test scores) and life expectancy ultimately (albeit slowly) translate into differences in levels of wealth and productivity, as captured by GDP per capita and output per worker, our third and fourth proxies, and both (rather noisy) economic variables that should also be affected by average language distance. Indeed, the down side of using a purely income based measure such as GDP per capita is that it fails to account for the fact that certain countries that are rich in natural resources concentrate income in the hands of a few individuals. Consequently, for such countries, GDP per capita is a poor indicator of the true state of development for the majority of the population. Figure II shows a strong negative relation between ADOL and the four dependent variable of interests.

As noted, none of these four proxies is perfect. Given this lack of a perfect composite

[18]Refer to Hanushek and Woessmann (2012) for further details on how this measure is constructed and how it outperforms traditional measures of human capital in explaining variations in cross-country GDP growth rates.

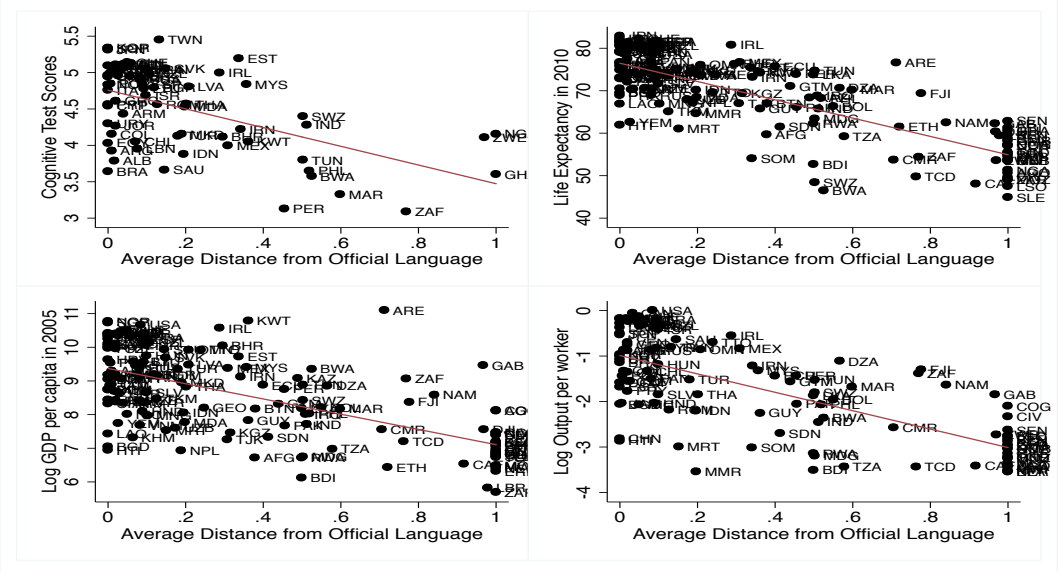

Figure II: **Scatterplot of ADOL and the four socio-economic variables of interest**

measure of socio-economic development we undertake the approach of first presenting our basic regressions with each of the above four dependent variable - *a measure of cognitive skills, life expectancy, log GDP per capita and log output per worker*. After presenting our initial results in support of our thematic framework, we then adopt the strategy of using the standardized score on the Human Development Index (zHDI) as our preferred dependent variable. This index includes health, education, and wealth measures, and is strongly correlated with the four component measures.[19] The rationale of using zHDI as the dependent variable, for robustness exercises and further empirical analysis, is based on the fact that not only does it capture all four dimensions outlined by our theory, albeit imperfectly, but also avoids losing valuable observations.

[19]The correlation between zHDI in 2010 and cognitive test scores, life expectancy, log GDP per capita and log output per worker, are 0.69, 0.89, 0.94 and 0.93, respectively, and all correlations are statistically significant at the 1 percent level.

2.4 Cross country regressions

In order to explore the correlation between the dependent variables of interest and ADOL, we estimate a reduced form regression that takes the form:

$$DV_i = \alpha * ADOL_i + B * X_i + \varepsilon_i, \tag{3}$$

where in all specifications we estimate robust standard errors. The results are shown in Table III, where the DV_i in columns (1) and (2) is a measure of cognitive skills taken from the work of Hanushek and Woessmann (2012). Columns (3) and (4) in turn use life expectancy in 2010 as the dependent variables to explore the effect of ADOL on health. Columns (5) and (6) consider log GDP per capita in 2005 in 2005 constant dollars and finally columns (7) and (8) use log output per worker from the work of Hall and Jones (1999) as a measure of productivity.

X_i refers to a vector of controls and in all 8 specifications shown in Table III, besides our measure of ADOL, we control for three additional confounding factors. First, we control for ethno-linguistic fractionalization (ELF), a measure that takes into account linguistic distance between all ethnic group dyads, and based on Fearon (2003). The concept of ELF and ADOL as explained in section 2.1 are distinct; however, empirically the correlation between the two measures is 0.57 and thus it is important to account for it in a multivariate framework. The choice of the measure of ELF is inspired by the work of Desmet et al. (2009) who show that accounting for distance between groups in diversity measures is important, though once distance is accounted for the choice between the exact nature of the index used - diversity, peripheral heterogeneity or polarization - is empirically irrelevant.[20]

[20]In a companion paper we model the choice of official language in post-colonial states, and show that increasing linguistic diversity increases the probability of retaining the colonial language, and consequently ADOL. Empirically controlling for ADOL turns the coefficient on all standard measures of linguistic diversity close to zero and insignificant, suggesting that most

Second, we include a measure of institutional quality from the Polity-IV data set, quantifying the extent of institutionalized constraints on the decision-making power of chief executives averaged over the years 1960 to 2000.[21] As we are interested in understanding the effects of language policy choices on socio-economic development, the third control we include is the level of log GDP per capita in the year of independence, i.e. before official language choices were instituted and hence account for the previous level of development which were largely unrelated to contemporary language policy choices.[22]

Columns (2), (4), (6) and (8) additionally include continent dummies. The inclusion of continent dummies implies that the coefficient on ADOL is being estimated based on the difference in language distances between countries *within* a continent, and the dependent variable of interest. On the one hand, the inclusion of continent dummies ensures that the effect we are capturing is not being driven by the black box of across continental differences. On the other hand, if our objective is to explain what makes countries in any continent distinct, the inclusion of continent fixed effects by definition will imply that these differences, if they are correlated with the independent variable of interest, are relegated to the black box of fixed effects. As we

of the negative effects attributed to linguistic diversity are mediated through the channel of language choice; we thus provide both theoretical and empirical evidence on a realistic mechanism through which ELF works [Citation removed for review purposes].

[21] Our choice of the measure of institutional quality is guided by theoretical considerations. Refer to Glaeser et al. (2004) for a discussion. However, in the online Appendix we show that the documented correlation is robust to alternative measures of institutions such as the average protection against expropriation risk constructed by the Political Risk Services Group, the index of social infrastructure constructed by Hall and Jones (1999) or the extent of institutionalized democracy as measured by the Polity-IV data set.

[22] As the GDP per capita is not always available at the exact year of independence the closest available date has been used. The Excel file accompanying the online Appendix shows the year of independence and the year from which the GDP data has been used.

later contend (in section 2.7) that geography is a key factor affecting language policy choices, they are consequently correlated with continents. For this reason, the inclusion of continent dummies absorbs a large part of the effect of language distance, though there remains much variance to be explained.

For the dependent variable life expectancy we additionally control for the percentage of people ages 15-49 who are infected with HIV, to ensure that our estimates are not only capturing differences in HIV prevalence rates. For log GDP per capita we control for the availability of natural resources, namely percent of world oil, gold, iron and zinc reserves and number of minerals present in a country. [23]

Insert Table III

In all eight specifications ADOL is seen to be both substantively and statistically an important correlate of the four dependent variables. To have an intuitive understanding of the magnitude of the effect imagine a country such as Ghana switching from using English to Akan, the language of the largest ethnic group, as their official language. This reduces the ADOL from 1 to 0.18, and moves Ghana up 13 and 11 spots in terms of their ranking on cognitive tests scores and life expectancy, respectively, and 21 ranks up in the case of log output per worker.

Table IV in turn considers the standardized value of the HDI in 2010, a composite measure of the facets of socio-economic development outlined by our theory, as the dependent variable. ADOL by itself explains around 55% of the cross-country variation in the HDI, and together with all controls 76% of the cross-country variation in the levels of HDI are accounted for in

[23] We also explore the impact of ADOL on GDP per capita by splitting the sample into countries highly dependent and those not highly dependent on natural resources as a share of a GDP. In line with our theoretical logic, we show (on Table A5 in the online Appendix) that ADOL is much more important in explaining GDP per capita for countries not dependent on natural resources.

the regression. The largest drop in the coefficient occurs between column (4) and (5) when we include continent dummies.

Insert Table IV

Finally, Table V shows that the correlation documented between ADOL and HDI in Table IV cannot be attributed to any particular region of the world. Columns (2) to (6) in Table V drop Africa, Americas, Asia, Europe and Oceania, respectively, and the coefficient on average distance remains, both substantively and statistically, an important correlate of HDI.

Insert Table V

2.5 Theoretically inspired controls and some robustness checks

We now explore other potentially important factors that have been highlighted in the literature as important in explaining cross-country income differences to evaluate the robustness of our results.

Taking into account new insights on deep historical sources of economic performance (Nunn 2009, Ashraf and Galor 2013, Bockstette et al. 2002, and Michalopoulos and Papaioannou 2013), we add a measure of genetic diversity, genetic diversity squared and the index of state antiquity to the specification given by column (5) of Table IV. The results are shown in column (2) of Table VI. The addition of these controls does not affect the precision or magnitude of the coefficient on average distance.

The historical origin of a country's laws has been shown to be correlated to a broad range of economic outcomes (Shleifer et al., 2008). In column (3) of Table VI we additionally control for the legal origin of the countries. As can be seen this control does not affect the precision or magnitude of our estimates.

Insert Table VI

The data on GDP at independence is measured in a common denominator for all countries in our sample. However, given the date of independence between countries varies widely, the same incomes levels in different eras might imply a different stage of development. Alternatively, the timing of independence itself may contain information on a country's wealth. In order to address this concern of comparability across eras, we consider only the sample of countries that gained independence after 1945 and re-estimate Equation 3 for all 5 dependent variables of interest. The results in Table VII show that ADOL is still statistically significant and an economically meaningful predictor of the socio-economic variables considered.[24]

Insert Table VII

We need also to ask how robust our findings are to contemporary changes in the international political economy, from an era of import substitution growth models (where there may have been an advantage to the promotion of indigenous languages) to an era of globalization (where the premium on English would be revealed) (Rodrik, 1990). Perhaps our results supporting the role of languages that are proximate to that of the local populations were appropriate for the first era, but not for the second? We examine this possibility in column (3) of Table VIII, by replacing GDP at independence with zHDI in 1990, and find that the effect remains significant both statistically and substantively in the 1990-2010 period. Globalization, in other words, has not lessened the importance of average distance for human development.

Insert Table VIII

In the online Appendix we conduct a series of robustness tests and show the correlation is robust to additional controls for geography, climate, and alternative measures of ELF and institutions.

[24]The coefficient on ADOL for the dependent variable cognitive test score turns insignificant, as the standard errors increase due to the number of observations reducing to 31. The beta coefficient though is larger than the other 3 explanatory factors considered.

2.6 Methodological concerns

2.6.1 Omitted variable bias

The cross-country framework raises important methodological concerns regarding reverse causality and omitted variable bias (OVB). To quantitatively examine the problem of omitted variable bias we use the test suggested by Oster (2013), which builds upon the methodology of Altonji et al. (2005) that selection on observables can be used to assess the potential bias from unobservables. The results of the test suggest that power of the unobservables would have to be about 2.5 to 10 times stronger relative to the observables, which seems highly unlikely given we explain 75 percent of the cross-country variation in zHDI. The methodological details and results are provided in the online Appendix.

Notwithstanding the quantitative estimate of the extent of OVB, the concern remains that it is not language policy choices, but some other underlying unobservable characteristics that affect both language choices and the socio-economic outcomes. If that were the case, language policy choices would be endogenous in our setting. In this regard, at least with respect to Sub-Saharan Africa, there is good reason to believe that the observed language policy choices strongly mirror the language choices observed during the colonial era, and are hence exogenous.[25]

The objectives of the education policy of the French and British colonialists were identical - train a few elites through the use of the colonial language to help administer the country, and ensure that the masses were docile and controlled through restricting access to secondary and higher education (Bokamba 1984, Fabunmi 2009, Whitehead 2005). The British and French however undertook differing paths to achieve their objectives. In the case of France, a French-only language policy was instituted right from the start of primary schooling, whereas

[25] As can be seen in Table A.8 in the online Appendix, ADOL is a statistically significant correlate of the 4 outcomes variables when we consider only the African continent.

the British adopted a more laissez faire policy and allowed the use of local languages for the initial 1 to 3 years of primary schooling.[26] The fact that less than 3 percent of the population in Sub-Saharan Africa was enrolled in secondary education or higher in 1960 highlights that the policy objective of restricting access to higher education was successfully achieved in both the former British and French colonies.[27]

In line with the colonial-era policy, up until 990, only two former French colonies - Madagascar and Guinea - changed their language policies from colonial times. All others continued with a policy of using only French for all levels of education. On the other hand, the former British colonies also continued with the colonial-era policy of using multiple local languages for a duration of one to three years in primary schooling before switching to the use of English.[28]

Albaugh (2014) makes a compelling case for why language policy in general, and in education in particular, was characterized by policy inertia. Drawing on the works of Tilly and Ardant (1975) and Herbst (2000), she argues that in an environment of low external threat due to stable borders, and income taxation rendered relatively unnecessary due to foreign aid and taxes on primary commodities, African leaders did not have to engage in language planning and ratio-

[26]The two reasons highlighted in the literature for this difference in policy are: (i) the differing roles played by Catholic and Protestant missionaries; and (ii) the differing extent of control exercised by the state. Refer to Albaugh (2014), Michelman (1995), and Whitehead (2005) for details.

[27]The percentage enrolled were 3.31 and 2.39 percent for the former British and French colonies, respectively, and the differences are not statistically significant ($t = 0.47$) (Barro and Lee, 2014).

[28]Refer to Albaugh (2014, 62-3) for examples of some experiments in the realm of language policy in education undertaken in the 1960-70s in Sub-Saharan Africa, which she argues were largely symbolic or short-lived.

nalization for state building.[29] The nature of incentives, compared to those that faced European state builders, implied that African leaders did not have to engage in the spread of a standard language to project power and retained the language policy they inherited from their colonial predecessors. Leaders in the face of public pressure to increase access to schooling predictably decided to invest in education to pacify the population, though with little or no interest in actual outcomes. The main challenge to their power came from internal rather than external threats, and therefore patronage was a common resort to maintain power.[30] These internal competitors in turn were concerned with their share of spoils rather than language rights (Cooper 2008). The strongest indication of the continued colonial influence on language policy in Sub-Saharan Africa is that not a single nation in the past 60 years has ever used an indigenous language for secondary or higher education. The available evidence on student outcomes suggests that the language policy today has been as effective as in colonial times in restricting access to a small section of the population and ensuring continuous replenishment in the ranks of the elite, while still separating it from the masses.[31]

The above discussion lends weight to the assertion that language policy choices in Sub-Saharan Africa reflect choices made during the colonial-era. However, one concern that remains is that perhaps countries become independent with entrenched elites having an interest in perpetuating the inefficient policies of the colonial state. For example, consider policies affecting exchange rates (Bates, 1981) or political boundaries (Michalopoulos and Papaioannou,

[29]Refer to Englebert (2009) and Young (1983) for a discussion on the sanctity of the principle of existing sovereign units in postcolonial state system in Africa.

[30]Refer to Francois et al. (2014) for empirical evidence on allocation of political power as a tool of patronage to minimize the probability of revolutions from outsiders and coup threats from insiders.

[31]The Barro and Lee (2014) data for Sub-Saharan Africa, from the year 2010, shows that only 12 percent of the population aged 15 and over has finished secondary schooling, and less than 2.6 percent are enrolled in tertiary education.

2011), that while inefficient, helped perpetuate the rule of post-independence leaders. From this perspective, the causal variable would be the entrenched elite interests rather than any particular policy. On Table A10 in the online Appendix, we rely on the Archigos dataset and use leader duration since independence for all countries as a proxy for entrenched elites.[32] Including leader duration (or duration squared) in our standard regression does not affect the coefficient on ADOL, and we thereby gain confidence that the channel of language policy, over and above the general interests of entrenched elites, is an important factor affecting cross-country development.

2.6.2 Reverse causality

Reverse causality is less troublesome. The measure of language distance is time-invariant, to the extent the composition of ethnic groups remains constant at the country level and language policy choices do not change, and hence are not affected by the levels of socio-economic development directly. The concern regarding endogeneity might still arise as poorer countries plausibly chose more distant language policies, while rich states are able to assimilate minorities thereby reducing average distance. If this is the complete story, all we are observing in our regressions are secondary consequences of weak and poor states vs. strong and rich ones.

Does income determine language choice? In order to answer this, we control for the level of GDP per capita at the time of independence of countries since language policy choices were instituted at the time of independence. Hence if it is difference in income levels rather than language policy choices that is the underlying cause, inclusion of GDP per capita at independence should reduce the magnitude and significance of our coefficient. However as can be seen in column (4) of Table IV, controlling for initial income does not affect the precision and magnitude

[32]The dataset has been accessed at www.rochester.edu/college/faculty/hgoemans/data.htm and the results of the regression are shown in Table A.9 of the online Appendix.

of the coefficient on average distance.[33]

2.7 An instrumental variable approach

To provide evidence that the documented relationship between ADOL and socio-economic development is indeed causal, we now undertake a strategy of using an instrument that is correlated with ADOL but uncorrelated with other country characteristics.

We identify the availability of a written tradition as one of the important factors affecting language policy choices. The rationale being that in the absence of a written language states first need to invest in creating a standardized orthography, vocabulary and modern scientific terminology before a language can be utilized to fulfill the functions of an official language. Thus many states in the face of uncertainty associated with the cost and returns involved in the creation of written language might resort to using the colonial language. The proposed relationship finds strong support when we observationally examine availability of written traditions and choice of official language. Looking across the globe, nearly every country that had a writing script for an indigenous language has adopted at least one indigenous language as at least co-official. This factor can explain the language policy choices observed in Sub-Saharan African. Most Sub-Saharan African countries (with Ethiopia, Tanzania and Liberia as exceptions) did not possess a writing tradition and are characterized by the usage of only the colonial language as the official language. To empirically test whether availability of writing tradition has any explanatory power, we regress our measure of distance from official language on a dummy for having a writing tradition.[34] The results are shown in Table IX.

Insert Table IX

[33] A formal test for equality of the coefficients in columns (3) and (4) of Table IV is not rejected at conventional significance levels ($z = -0.82$).

[34] In the Excel file accompanying the online Appendix is shown the countries coded as one or zero.

The availability of a written tradition is seen to be a statistically significant predictor of ADOL. In columns (2) and (3) we control for log GDP per capita at independence and log population in 1500 (as a proxy for levels of development in the Middle Ages), respectively. The two wealth related factors are not only seen to be statistically insignificant but also their explanatory power is seen to be less by a factor of 40-60 as compared to the hypothesized factor.

The regressions shown in Table IX thus provide support to the assertion that possessing a writing tradition is an important determinant of ADOL. However, the indicator variable cannot be used as an instrument, as states which had a writing tradition, as compared to those which did not, arguably also differ on other important unobservable characteristics which might affect socio-economic development.

Drawing from the work of Diamond (1998), we hence propose using distance from the sites at which writing was independently invented as an instrument for ADOL.[35] He argues that geography was a crucial factor as to why a set of polities - Tonga's maritime proto-empire, the Hawaiian state emerging in the late 18th century, all of the states and chiefdoms of subequatorial Africa and sub-Saharan West Africa, and the largest native North American societies, those of the Mississippi Valley and its tributaries - did not acquire writing before the expansion of Islam and the arrival of the Europeans.

Writing was independently invented in Mesopotamia (Sumer) around 3200 BCE, in China around 1200 BCE, and in Mesoamerica around 600 BCE, and then diffused through trade and exchange to the rest of the world. The rationale for using the distance from the site of invention as an instrument is that the further the distance from the site of invention, the less likely is a country to have obtained the writing tradition through the process of diffusion, and consequently based on the evidence in Table IX will have a higher ADOL. Observe that using the distance from the site of invention as an instrument exploits the exogenous component of the probability

[35] In the online Appendix section A. 3 we use an alternative instrument, applicable to Africa, and document results similar to those shown in Table X.

of having a writing tradition, i.e. geography. The key underlying assumption for it to be a valid instrument is that the distance from these sites of invention should have no independent impact on socio-economic development today, except through the channel of affecting the probability of possessing a writing tradition.

To operationalize the measure we calculate the Great-Circle-Distance, using the Haversine formula, from each of the sites of invention to every other country in our sample. We then take the minimum of the distance from the three sites as the measure of distance from the place of invention of writing. Figure III shows the relationship between the shortest distance from the sites where writing was invented and the ADOL; as hypothesized the distance from official language is seen to be increasing in the distance from where writing was invented. The IV

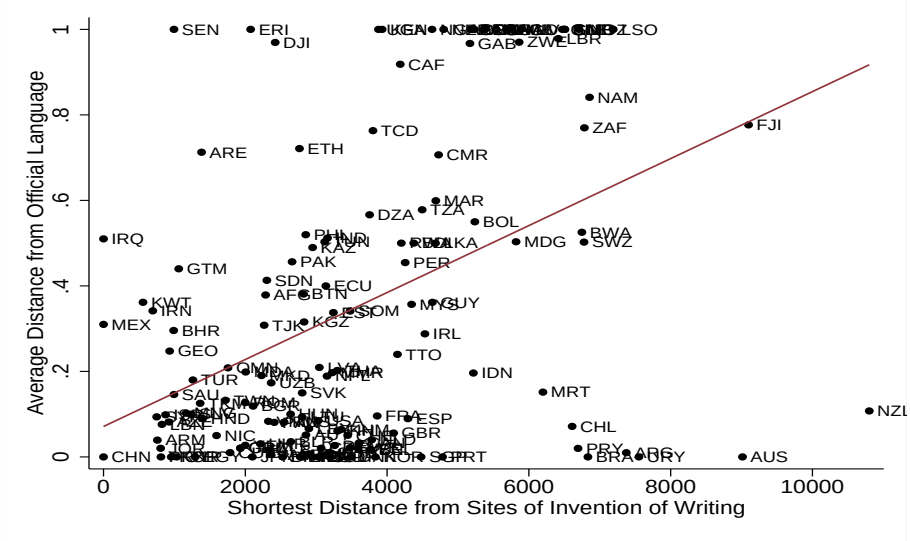

Figure III: **Reduced Form Relationship Between ADOL and Distance from Site of Invention of Writing**

estimates for the five dependent variables of interest are shown in Table X.

Columns (1), (3), (5), (7) and (9) regress cognitive test scores, life expectancy, log GDP per capita, log output per worker and zHDI, respectively, on ADOL instrumented for by the minimum distance from the sites of invention of writing. In Panel (B) the first stage regressions of distance from the sites of invention of writing on ADOL are shown. Inspecting the

F-statistics shows that all, except in column (1), meet or exceed the value of 10, and in most cases are greater than 30, suggesting distance from the site of invention is a strong instrument for ADOL.[36] In panel A are the results of the second stage; we see that the predicted value of ADOL is a statistically significant and economically important predictor of all the socioeconomic variables. The point estimates slightly exceed the OLS estimates in Table III and IV.

Insert Table X

In columns (2), (4), (6), (8) and (10) we additionally add the three controls outlined before in section 2.4 - linguistic diversity accounting for distance, constraints on the executive, and log GDP per capita at independence. We additionally control for an America dummy and the proportion of population of European descent in 1975. The reason is that the majority of the population on the American continent can be classified as either settlers or individuals of mixed race heritage (also known as 'mestizos'), whose mother tongue is a language which the settlers brought along with them. Thus for these countries distance from the site of invention of writing is not an important determinant of ADOL. Again the ADOL is seen to be a statistically significant predictor of the levels of socio-economic development. Results remain stable adding genetic diversity, genetic diversity squared and latitude to these estimations, as shown in tables A.12 of the online Appendix.

A potential concern with the estimates in Table X is that the distance from the sites of invention of writing could be correlated to other factors affecting socioeconomic development. If for instance we were to assume that distance from these earliest sites of invention of writing was responsible not only for acquiring the writing tradition but also a determinant of quality

[36]The F-statistic for the first stage for cognitive test scores takes the value of 2.39, and in the second stage regression ADOL is statistically insignificant. This is not surprising as the test scores are primarily from Europe and America, and hence the instrument does not have much variation leading to an increase in the standard errors. However the magnitude of the coefficient is greater than the one in column (2).

of state institutions and/or governance, then we would be violating the exclusion restriction for our instrument to be valid. In order to assess whether this is a cause for concern we run reduced form regressions of the minimum of the distance from the sites of invention of writing on the three most widely used measures of state institutional capacity and governance - (i) average protection against expropriation risk from the Political Risk Services (PRS) group averaged over the years 1995-05; (ii) social infrastructure combining government anti-diversion policies and openness to international trade from the work of Hall and Jones; and (iii) constraints on the executive from Polity-IV and averaged over the years 1960-2000. The results are shown in Table XI.

Insert Table XI

The distance from the sites of inventions of writing is not a significant correlate of any of the three measures of state institutions or governance, with the F-statistic taking a value of less than one in all three regressions. Thus the IV results confirm the negative relationship between ADOL and socio-economic development estimated by the OLS, and suggest that the OLS estimates may be a lower bound of the true effect of ADOL.

Finally, to gauge the economic magnitude of the IV estimates, again consider Ghana adopting Akan, the language of its largest ethnic group, as its official language instead of English. Such a change would move Ghana 23, 24 and 31 positions up in the ranking of countries on cognitive test scores, life expectancy and log output per worker. Alternatively it would move Ghana from the 7^{th}, 22^{nd} and 21^{st} percentile of the distribution of cognitive test scores, life expectancy and log output per worker to the 38^{th}, 40^{th} and 47^{th} percentile, respectively.

3 Micro evidence for the theoretical framework - The effect of individual level distance from the official language

Distance for every individual between his/her language and the official language is the first channel through which language policy operates. Our theory holds that high distance from the official language, holding other factors constant, increases learning as well as information acquisition and processing costs for the individual. This increased cost affects human capital formation, knowledge and adoption of best health practices, and in turn these translate into differences in occupational and wealth outcomes.

In order to estimate the effect of distance from the official language on individual outcomes, consider the case of India. Most Indian states use their majority indigenous language up to the end of secondary schooling. Government affairs, administration and courts carry out their functions in the state language and English.[37] The central government in turn operates in Hindi and English, where Hindi is the mother tongue of around 45% of the population. The languages in India come from two distinct language families, the Indo-European and the Dravidian, which provides us with crucial variation at the sub-national level, as the distance within each language family is around 0.29 and across language families, by construction is 1.

3.1 Data

The data come from the Indian National Family Health survey (NFHS 3) of the year 2005-06. We consider the sample of males and females aged 15-54 years to estimate the effect of individual language distance on various socio-economic outcomes of interest. The data provides information on the native language of the respondent, typically a proxy for the language of one's ethnic group even if the respondent has only limited facility in it, and state of residence, which allows us to calculate the language distance for individuals from the official state language. The

[37]The highest court in the land, the Supreme Court, however, operates in English.

data set also provides information on relevant individual characteristics such as age, religion, caste, educational attainment, a wealth index, employment status, nature of occupation, as well as knowledge and adoption of health practices.

3.2 Identification strategy

We estimate the effect of the distance of an individual's native language from the official state language on six variables. The first two are proxies for human capital - (i) years of education (ii) a dummy variable for whether the individual is literate; the next two measure health knowledge and practices - (iii) an indicator variable for whether the individual has ever heard of AIDS; (iv) whether the household uses a mosquito bed net for sleeping[38]; and the final two measure occupation and wealth outcomes - (v) whether the individual holds a white-collar job[39]; (vi) an indicator for whether the individual falls in the top quintile of the income distribution.

Comparing across Indian states indicates large variations in their levels of socio-economic development, which are important to account for in any empirical exercise. Accordingly in all our specifications we account for state fixed effects.[40]

A naive comparison of language distance and socioeconomic outcomes based on native speakers (non-migrants) vs non-native speakers (migrants) resident in the same state fails to account for the fact that natives and migrants might differ along unobservable dimensions which

[38]This information is available only for women and is estimated on the sample of women.

[39]Here we restrict the sample to individuals who are classified as employed and above 35 years of age.

[40]Accounting for state fixed effects implies we are controlling for the number of native speakers that the second-generation migrants are exposed to. However, though the effect of exposure to the state's official language is accounted for, it cannot be retrieved. We are unable to create an exposure indicator at a lower geographical unit, thus allowing for variation among individuals within a state, as the NFHS 3 data does not contain GIS information.

are not accounted for and which might be correlated with language distance.[41] In order to address this concern we restrict ourselves to the sample of individuals who report as having always lived in the same state, or in other words we exclude any first generation migrants. For non-majority language speakers, our data include both members of rooted minority groups (who by rights in the Indian Constitution can receive primary education in their mother tongues) and individuals whose families were migrants in recent generations (who do not have concentrations of their population that would make them eligible for indigenous language instruction in public education). To the extent that rooted minorities are getting the indigenous language instruction that the Constitution affords them, our results seeking to estimate the effect of not receiving mother-tongue education would be an underestimate. On the Constitutional formula for minority language instruction, see Sridhar (1996).

As we observe individuals belonging to the same linguistic groups in states having different official languages, we are able to account for any linguistic-specific group differences through the inclusion of language group fixed effects. In sum, our identification strategy ensures that the estimated effect of language distance is not due to any time invariant state or linguistic group characteristics.

3.3 Results

To estimate the effect of language distance on the dependent variables of interest, the following regression is estimated:

$$O_{ijm} = S_j + \delta_0 * Distance\_State\_Language_{ijm} + \beta_k + L_m + X_{ijm} + \varepsilon_{ijm}, \quad (4)$$

[41] Here we take out of our sample the families that decide to migrate, as this selects for characteristics such as ambition that would confound our results. Our results are stronger if we include first-generation migrants, but that would be an unfair test of our theory.

where O_{ijm} is the outcome of interest for individual i in state j and linguistic group m; and where all individuals report having always been resident in the same state, or in other words are not first-generation migrants. S_j refer to state fixed effects, β_k refer to a set of year of birth dummies and L_m to language group fixed effects. X_{ijm} is a vector of individual level characteristics which include dummies for caste, religion, whether individual lives in a city, town or countryside and the altitude of the primary sampling unit. The coefficient of interest δ_0 captures the effect of distance from the official state language on various socio-economic outcomes, and which according to the theoretical framework should be negative.

Insert Table XII

The results of the estimation exercise are provided in Table XII. The effect on years of education and literacy is calculated using an ordinary least squares regression, whereas the other four dependent variables are estimated using a logit regression, and all six models account for individual sample weights. The Table XII reports the average marginal effect of moving from a language distance of 0.292 to 1 (that is, between language families).

In column (1) and (2) the dependent variables considered are years of education and whether the individual is able to read a complete sentence. The marginal effect shows that the moving from a language distance of 0.29 to 1 decreases the years of education by 0.81 years, and is statistically significant at the 1 percent level. On the other hand, for the dependent variable literacy, the average marginal effect shows that the probability of being literate reduces by 5.9 percentage points moving from a language distance of 0.29 to 1. In other words comparing a Bengali speaker living in Delhi with one in Tamil Nadu shows that the Bengali living in Tamil Nadu would have 0.81 fewer years of education and would be less likely to be literate by a whole 9 percent; after accounting for state and language group specific differences, as well as any time trends.

Columns (3) and (4) use binary indicators for whether the individual has ever heard of HIV, and if the household uses a mosquito net for sleeping as dependent variables. We observe

that the marginal effect of moving from a language distance of 0.29 to 1 reduces the probability of having ever heard about AIDS or the household using a mosquito net for sleeping by 9 and 4.4 percentage points, respectively. Given that the sample average for the binary variable, usage of mosquito nets, is around 40 percent, the estimated marginal increase amounts to a 11 percent increase in the likelihood of using a mosquito net.

Finally columns (5) and (6) consider a binary indicator of whether the individual holds a white-collar job and belongs to the top quintile of the income distribution, respectively. The estimate shows the probability of holding a white collar job and belonging to the top income quintile decreases by 2.5 and 1 percentage point, respectively, when we move from a language distance of 0.29 to 1. Given that on average only 8 percent of individuals hold a white-collar job, the estimated marginal probability amounts to a 31 percent increase in the probability of holding a white-collar job.

The above results confirm the pattern observed in the cross-country data, but are now based on individual level data from India. The individual level data shows that distance from the official language has important implications for human capital (education and health), as well as for occupational and wealth outcomes. The identification strategy ensures that the effect of language distance cannot be attributed to state specific or language group specific differences, time trends, or issues of selection related to migration.

4 Micro evidence for the theoretical framework - the exposure channel

Relying on micro level data, let us now test for the effects of the *exposure* channel. Our evidence comes from countries that participated in the second round of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) program. SACMEQ is a consortium of education ministries, policymakers and researchers that in conjunction with UN-

ESCO's International Institute for Educational Planning (IIEP) collects data on primary schools from eleven African countries.

Consistent with our second assumption, other analysts have conjectured that one of the potentially important reasons for the poor educational outcomes observed on the African continent is not just the fact that the language of instruction is very distant from the native language of the students, but the fact that their exposure to this language remains virtually absent outside the classroom (Brock-Utne 2002, Dutcher 2003). Unrelated directly to education, but still related to the notion of exposure, Lazear (1999) shows that the likelihood that an immigrant will learn English is inversely related to the proportion of the local population that speaks his or her native language. Since everyday family, social and community life is based on the use of their native language or lingua franca, the exposure to the language of instruction is limited. The two forces in combination - use of a non-indigenous language along with limited exposure - imply that learning costs of the official language are high.

4.1 Data

The SACMEQ II round collected data on around 40,000 students, 5,300 teachers and 2,000 school heads from 2000 primary schools.[42] The dataset provides information on standardized student achievement tests in reading and mathematics across the thirteen countries for pupils currently in the 6^{th} grade.[43] The scores are standardized with a mean of 500 and standard deviation of 100. Moreover the standardized scores are provided for essential reading and math tests as well as for a comprehensive math and reading test. The data also provides a categorical

[42]Southern and Eastern Africa Consortium for Monitoring Educational Quality. SACMEQ II Project 2000-2004 [dataset]. Version 4. Harare, SACMEQ [producer], 2004. Paris, International Institute for Educational Planning, UNESCO [distributor], 2010.

[43]We exclude Mozambique, Tanzania and Zanzibar from our sample as the medium of instruction is not English in the 6^{th} grade, and hence have eleven countries in our sample.

indicator which captures whether students meet the minimum and desirable reading levels of SACMEQ. These are the main pupil related outcomes which form the dependent variables of interest. The dataset also provides extensive information on the students' socio-economic background such as parents' education, possessions, housing quality, availability of extra lessons outside the classroom (often referred to as tuitions), support at home for homework, and school absences. It also asks a question regarding usage of the medium of instruction, English, at home, which is divided into the category of never, sometimes and often. The dataset also collects information regarding teachers, headmasters, schooling infrastructure and quality. It also allows us to identify the classroom to which each student belongs. Control variables and descriptive statistics are provided in Table XIII.

Insert Table XIII

The descriptive statistics convey the gravity of the problem facing the educational sector in Africa. About 60% of the students do not reach the minimum reading level. When the bar is fixed at the desirable reading level, about 86% of the students are classified as not reaching that level, and this in spite of vast foreign aid expenditures over the previous decade directed at the educational sector (Devarajan and Fengler, 2013). Obviously, fundamental factors affecting student achievement have not yet been addressed, which directs attention to the exposure channel.

4.2 Identification strategy

To test for exposure, the key independent variable of interest is the frequency with which pupils use English at home. Regarding the usage of English at home, 23% report as never using English at home, 55% report using English sometimes at home and 21% report using English often at home. We construct a binary indicator which takes the value 0 in case the student never uses

English at home and the value 1 if the students use English often or sometimes at home.[44] As all students are Africans in the data their distance to the official language, English, is equidistant and equal to 1.[45] This means there is no need to control for the effect of individual level distance from the official language. The choice of our independent variable, use of English at home, is inspired by the work of Dustmann et al. (2012) who show that the single most important factor in explaining differences between immigrant and native children PISA tests scores in OECD countries is the language spoken at home.

Recall that more than 70-80% of the population in most African countries do not speak the official language and this is especially true for the older generations. It is therefore not surprising that the variable "using English at home" captures a rather small increment in academic success. In the data around 70% of the pupils who do not reach the minimum reading level still claim to use English at home. Given the low level of skills the pupils themselves possess it can be inferred that the exposure to English that takes place even at home is not comparable in quantity or quality in any way to the exposure that language minority students in advanced industrial countries, for instance as immigrants, experience while learning in a majority language. Thus the reported levels of high usage might still be very low in quality and quantity when compared

[44] As we explain below, more than 70 percent of the students who do not reach the minimum reading level still claim to use English at home. Thus we believe the distinction between the categories "sometimes" and "often" is at best tenuous, and prefer to combine them. Using the two categories separately shows the category "sometimes" has a larger effect on achievement than "often", though both have a significant and positive effect.

[45] This is because all African languages belong to non-Indo-European language family trees implying no shared branches and a distance equal to 1. However certain countries such as South Africa and Kenya do have populations which speak languages belonging to the Indo-European language family as their mother tongue (Afrikaans, English). In order to account for this we estimate the effect of exposure to English individually for every country in our sample and show that the results also hold for all countries which have no Indo-European language groups.

to conventional exposure to the medium of instruction in countries where it is spoken by the local population as a native language. That said, given that our measure of exposure captures low quantity and quality of exposure, if it still turns out to be a significant explanatory factor of student performance, this would imply that the estimate should be considered to be the very lower bound of the effects of exposure.

The data identifies the classroom to which each student belongs. We have information on 33,141 students in 4,686 classes across the eleven countries.[46] We are hence able to account for classroom fixed effects in our analysis. Taking classroom fixed effects implies common factors - such as teachers, school infrastructure and other unobservables - which affect student performance at the classroom level are accounted for. We can now estimate the effect of using English at home, which is our proxy for exposure to the medium of instruction, on test scores with class fixed effects and controls at the level of the student's home.

4.3 Results

To estimate the effect of exposure on student achievement we estimate the following reduced form equation:

$$S_{ij} = C_j + \delta_0 * English\_Home_{ij} + \delta_1 * X_{ij} + \varepsilon_{ij}, \tag{5}$$

where S_{ij} refers to the relevant outcome of interest of student i in classroom j. The outcomes considered are the test scores on essential and comprehensive math and reading tests, respectively, for students in the 6^{th} grade.

C_j refers to the classroom fixed effects which accounts for factors at the classroom level which potentially affect student performance. δ_0 is the coefficient of interest and captures the

[46]The number of observations included in the regressions are either 28,349 or 30,952 depending on the dependent variable considered, as some of the control variables are not available for all of the students in the sample.

effect of using English at home on student performance. X_{ij} refers to the student level controls at the family level, which are shown in Table XIII. All regressions are estimated using pupil weights provided by SACMEQ and robust standard errors are estimated.

Insert Table XIV

The results of the estimation exercise are shown in Table XIV. Exposure to English has a positive and statistically significant effect on all six student outcomes considered. The first column considers the essential reading score as the dependent variable. The estimation results suggest that increased exposure to English, captured by frequency of use of English at home, increases the essential reading score by 20 points or $\frac{1}{5}$ of a standard deviation. In column (2), the dependent variable considered is the standardized score on a comprehensive reading test. The results again indicate that exposure to English increases the reading score by 19 points or $\frac{1}{5}$ of a standard deviation.

Columns (3) and (4) consider the essential and comprehensive Math test scores. Exposure to English is seen to have a similar effect to the one on reading scores. It increases the math score on the essential and comprehensive tests by 18.82 and 18.16 points, respectively, amounting to a $\frac{1}{5}$ of a standard deviation in both tests.

The last two columns, (5) and (6), consider the effect of the use of English on reaching the minimum and desirable level of reading. The table reports the average marginal effects of the binary indicator. Use of English at home increases the probability of reaching the minimum and desirable reading level by about 9 and 4 percentage points, respectively.

The fact that even this low (both in terms of quantity and quality) level of exposure that we have isolated has a positive and significant impact on student performance in turn hints at the fact that high levels of official language exposure, a factor missing on the African continent, might play a very important role in increasing human capital.

4.4 Discussion and methodological concerns

One cause for concern is that the indicator of exposure might be correlated to some other omitted home level variable which is driving the results. In Figure A.1 in the online Appendix are plotted the average usage of English at home by socio-economic status and education level of parents. Usage of English is increasing in both the socio-economic status as well as the parents' education level. This suggests that children from better-off households are more likely to use English at home. That said, it should be noted that the coefficient and significance on our coefficient of interest - δ_0, remains remarkably stable even after controlling for a rich set of individual level controls that could affect student performance.[47]

It is often stated that as usage of the indigenous language is very vibrant for home and community affairs in Africa, the use of a foreign language does not really threaten the position or the existence of these indigenous languages. The results here however indicate that this maintenance might very well be at the cost of poor schooling outcomes and the usage of the medium of instruction for home and community affairs might be consequential for reducing learning costs.

5 Conclusion

One of the legacies of colonialism has been the continued use of the former colonial language as the official language in most postcolonial states. Here we theorize that the official language, by acting as a gatekeeper for accessing education, jobs, and elite political networks, imposes costs of participation due to its linguistic distance from popular speech and due as well to the low exposure people have to that official language in everyday life. The question raised by this

[47]No particular country in our sample drives the results. Figures A.2 and A.3 in the online Appendix show the effect of exposure to English on Math and English scores is positive and significant in 10 of the 11 countries.

theoretical orientation is whether a foundation in a language which is proximate in structure and rich in exposure provides a stronger foundation for health and human capital.

In our attempt to test this theory, readers will note that we have analyzed several observable implications of our theory, but without an "interocular traumatic test", that is, one that hits you between the eyes (Putnam et al. (1994, 19) quoting John Tukey). A perfect test would have been if there were a set of countries that were randomly assigned official languages such that we could isolate the average returns to decreased average distance. Alas, the world did not offer us this experiment, nor even a set of countries whose official language were altered in a way that was exogenous to concerns of development.

A less ideal experiment, but still one that is direct, would have been to rely on the digitized Ethnologue map of ethno-linguistic groups within each country, and for each ethnic heartland measure its economic prosperity (proxied by satellite imagery on level of night lighting) and the degree of distance of its language in reference to the official language of the state (Alesina et al., 2014). Alas, this too would not provide valid inferences. Given that our theoretical channel is in human capital formation, we should expect those students who best succeed in gaining human capital will have an incentive to migrate outside the home region in order to get the most prestigious jobs. Therefore the light exposure in the ethnic heartlands would be measuring the impact of the educational system on those who took least advantage of it.

As readers can infer, we abandoned the quest for the ideal experiment, but as we summarize our results below, we hope we have not only exposed the limits of, but also the possibilities for observational data for the study of development.

Using language trees from Ethnologue, a measure of average distance to the official language (ADOL) was constructed for each country. We regress ADOL on four outcome measures that are theorized to be implications of language distance: comparative test scores on internationally comparable exams; life expectancy; per capita GDP; and output per worker. We then combine the elements of these proxies and rely on a standardized score on the Human

Development Index (zHDI). Whether using the proxies or the zHDI, we find a robust negative relationship between ADOL and the development outcomes of interest.

We then address the methodological issues of possible omitted variable bias and endogeneity, and our results hold up. After that, we address the question of causality, and apply several tests. Most important is an instrumental variable estimation. We first show that if there is a major group in the country that has a long written tradition, that country is more likely to have an indigenous language as official. This was the source of the idea for using as our instrument the distance of the country's capital to the nearest spot of the historical origins of writing. Therefore, proximity to the invention of writing is a good predictor of language choice but uncorrelated with standard measures of institutional quality and state strength, and therefore a valid instrument. Our results hold with this IV estimation, giving us confidence that language choice has a causal impact on human capital and health.

Moving to more micro data, we then test for each of our two theoretically derived channels by using data from a set of eleven African countries and India. Exposure to English, which is the medium of instruction in schools in the eleven African countries considered, has a positive effect on student achievement in both math and reading scores. The Indian data shows that the higher the distance between the native language of the individual and the official language of the state he or she is born in, the poorer the human capital and occupational outcomes.

As for policy, in no way does this imply that the colonial languages should be ignored. They are central for international participation and often access to credit from international lenders in one's own country. The first question here is whether a foundation in a language which is proximate in structure and rich in exposure provides a stronger foundation for human capital. The data in this paper suggest the answer is "yes". The second important question that then follows is how best to shift from an inefficient language equilibrium. This paper suggests two possibilities. First, institutionalizing in education and the health sector a language which is proximate in structure to the indigenous language of the population would provide a

stronger foundation for human capital. Second, if this is politically infeasible, efforts to increase exposure to the educational medium is likely to have high returns in human capital. This might be done through adult education campaigns in the official language in the expectation that this would imply more of its use in student homes. Or governments could work to upgrade teacher fluency in the official language, thereby providing better exposure to their students of its standard use. Finally, the employment of professionally trained translators in the health sector would have considerable value. In any event, the data in this paper suggest that future work in development should put more attention on remedying the heavy costs of inefficient language policies in post-colonial states for human development.

References

Albaugh, E. A. (2014). *State-Building and Multilingual Education in Africa*. Cambridge University Press.

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

Alesina, A., S. Michalopoulos, and E. Papaioannou (2014). Ethnic inequality. *Journal of Political Economy*, Forthcoming.

Alidou, H., A. Boly, B. Brock-Utne, Y. S. Diallo, K. Heugh, and H. E. Wolff (2006). Optimizing learning and education in Africa–the language factor. *Paris: ADEA*.

Altonji, J., E. Todd, and C. Taber. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy 113*(01), 151–184.

Angrist, J. D. and V. Lavy (1997). The effect of a change in language of instruction on the returns to schooling in Morocco. *Journal of Labor Economics*, S48–S76.

Ashraf, Q. and O. Galor (2013). The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development. *The American Economic Review 103*(1), 1–46.

Barro, R. and J. Lee (2014). Barro-Lee data set. *Retrieved at http://www. barrolee. com/abgerufen*.

Bates, R. (1981). *Markets and States in Tropical Africa*. Berkeley California: University of California Press.

Bauer, A. M. and M. Alegría (2010). Impact of patient language proficiency and interpreter service use on the quality of psychiatric care: a systematic review. *Psychiatric Services 61*(8), 765–773.

Bockstette, V., A. Chanda, and L. Putterman (2002). States and markets: The advantage of an early start. *Journal of Economic Growth 7*(4), 347–369.

Bokamba, E. (1984). French colonial language policy in Africa and its legacies. *Studies in the Linguistic Sciences 14*(2), 1–35.

Bowen, S. (2001). *Language barriers in access to health care*. Health Canada Ottawa.

Brock-Utne, B. (2002). The most recent developments concerning the debate on language of instruction in Tanzania. University of Oslo, Paper presented to the NETREED conference.

Brock-Utne, B. and H. B. Holmarsdottir (2003). Language policies and practices in Africa–some preliminary results from a research project in Tanzania and South africa. In B. Brock-Utne, Z. Desai, and M. Qorro (Eds.), *Language of Instruction in Tanzania and South Africa*, pp. 80–102. Dar es Salaam, E D Publishers.

Chang, P. and K. Emzita (2002). *Technical assistance for ICT and HIV/AIDS preventive education in the cross-border areas of the Greater Mekong Subregion*, Volume 36648. Asian Development Bank.

Chantavanich, S., A. Beesey, and S. Paul (2002). *Mobility and HIV/AIDS in the Greater Mekong Subregion*. Asian Development Bank.

Cooper, F. (2008). Possibility and constraint: African independence in historical perspective. *The Journal of African History 49*(02), 167–196.

De Walque, D. (2007). How does the impact of an HIV/AIDS information campaign vary with educational attainment? Evidence from rural Uganda. *Journal of Development Economics 84*(2), 686–714.

De Walque, D. (2009). Does education affect HIV status? evidence from five African countries. *The World Bank Economic Review 23*(2), 209–233.

De Walque, D. (2010). Education, information, and smoking decisions evidence from smoking histories in the United States, 1940–2000. *Journal of Human Resources 45*(3), 682–717.

Desmet, K., I. Ortuño Ortín, and S. Weber (2005). Peripheral diversity and redistribution. CEPR Discussion Paper.

Desmet, K., S. Weber, and I. Ortuño-Ortín (2009). Linguistic diversity and redistribution. *Journal of the European Economic Association 7*(6), 1291–1318.

Devarajan, S. and W. Fengler (2013). Africa's economic boom: Why the pessimists and the optimists are both right. *Foreign Affairs 92*, 68.

Diamond, J. M. (1998). *Guns, germs and steel: a short history of everybody for the last 13,000 years*. Random House.

Djité, P. G. (2008). *The sociolinguistics of development in Africa*, Volume 139. Multilingual Matters.

Drysdale, R. (2004). Franco-Australian Pacific regional HIV/AIDS and STI initiative.

Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics 3*(1), 425–449.

Dustmann, C., T. Frattini, and G. Lanzara (2012). Educational achievement of second-generation immigrants: an international comparison. *Economic Policy 27*(69), 143–185.

Dustmann, C., S. Machin, and U. Schönberg (2010). Ethnicity and educational achievement in compulsory schooling. *The Economic Journal 120*(546), F272–F297.

Dutcher, N. (2003). Promise and perils of mother tongue education. Retrieved from http://www.silinternational.org/asia/ldc/plenary_papers/nadine_dutcher.pdf.

Eastman, C. M. (1983). *Language Planning: an Introduction*. San Francisco: Chandler Sharp.

Englebert, P. (2009). *Africa: unity, sovereignty, and sorrow*. Lynne Rienner Publishers Boulder, CO.

Eriksson, K. (2014). Does the language of instruction in primary school affect later labour market outcomes? Evidence from South Africa. *Economic History of Developing Regions 29*(2), 311–335.

Esteban, J., L. Mayoral, and D. Ray (2012). Ethnicity and conflict: An empirical study. *The American Economic Review 102*(4), 1310–1342.

Fabunmi, M. (2009). Historical analysis of educational policy formulation in Nigeria: Implications for educational planning and policy. *International Journal of African & African-American Studies 4*(2).

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth 8*(2), 195–222.

Francois, P., I. Rainer, and F. Trebbi (2014). How is power shared in Africa? *Econometrica*, Forthcoming.

Glaeser, E. L., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2004). Do institutions cause growth? *Journal of Economic Growth 9*(3), 271–303.

Gomes, J. (2014). The health costs of ethnic distance: Evidence from Sub-Saharan Africa.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*, 109–115.

Hall, R. E. and C. I. Jones (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics 114*(1), 83–116.

Hanushek, E. A. and L. Woessmann (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth 17*(4), 267–321.

Herbst, J. (2000). *States and power in Africa: Comparative lessons in authority and control*. Princeton University Press.

Higgins, C. and B. Norton (2009). *Language and HIV/AIDS*. Multilingual Matters.

International Institute for Population Sciences (IIPS) and Macro International (2007). National family health survey (NFHS-3), 2005-06, India, Volume i. Mumbai.

Karliner, L. S., E. A. Jacobs, A. H. Chen, and S. Mutha (2007). Do professional interpreters improve clinical care for patients with limited english proficiency? a systematic review of the literature. *Health services research 42*(2), 727–754.

Laitin, D. D. (1992). *Language repertoires and state construction in Africa*. Cambridge University Press.

Laitin, D. D. (1994). The tower of babel as a coordination game: Political linguistics in Ghana. *American Political Science Review 88*(03), 622–634.

Laitin, D. D. (2000). Language conflict and violence: The straw that strengthens the camel's back. *European Journal of Sociology 41*(01), 97–137.

Laitin, D. D., R. Ramachandran, and S. L. Walter (2015). Language of instruction and student learning: Evidence from an experimental program in Cameroon.

Lazear, E. P. (1999). Culture and language. *Journal of Political Economy 107*(S6), S95–S126.

Lewis, P., G. Simon, and C. Fennig (2014). *Ethnologue: Languages of the World*. Seventeenth edition. Dallas, Texas: SIL International.

Michalopoulos, S. and E. Papaioannou (2011). The long-run effects of the scramble for Africa. No. w17620. National Bureau of Economic Research.

Michalopoulos, S. and E. Papaioannou (2013). Pre-colonial ethnic institutions and contemporary African development. *Econometrica 81*(1), 113–152.

Michelman, F. (1995). French and British colonial language policies: A comparative view of their impact on African literature. *Research in African Literatures 26*(4), 216–225.

Nunn, N. (2009). The importance of history for economic development. *Annual Review of Economics 1*, 65–92.

Oster, E. (2013). Unobservable selection and coefficient stability: Theory and validation. No. w19054. National Bureau of Economic Research.

Putnam, R. D., R. Leonardi, and R. Y. Nanetti (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton university press.

Ramachandran, R. (2012). Language use in education and primary schooling attainment: evidence from a natural experiment in Ethiopia. *Documents de treball IEB* (34), 1–49.

Ramirez, D., K. G. Engel, and T. S. Tang (2008). Language interpreter utilization in the emergency department setting: a clinical review. *Journal of Health Care for the Poor and Underserved 19*(2), 352–362.

Rodrik, D. (1990). How should structural adjustment programs be designed? *World development 18*(7), 933–947.

Shleifer, A., F. Lopez-de Silanes, and R. La Porta (2008). The economic consequences of legal origins. *Journal of Economic Literature 46*(2), 285–332.

Sridhar, K. K. (1996). Language in education: Minorities and multilingualism in India. *International Review of Education 42*(4), 327–347.

Tansey, E., R. Borland, H. West, et al. (2010). Southern Africa ports as spaces of HIV vulnerability: case studies from South Africa and Namibia. *International maritime health 62*(4), 233–240.

Taylor, S., M. Coetzee, et al. (2013). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach.

Tilly, C. and G. Ardant (1975). *The formation of national states in Western Europe*, Volume 8. Princeton Univ Press.

Translators Without Borders (2015). Does translated health-related information lead to higher comprehension? A study of rural and urban Kenyans.

Underwood, C., E. Serlemitsos, and M. Macwangi (2007). Health communication in multilingual contexts: A study of reading preferences, practices, and proficiencies among literate adults in Zambia. *Journal of health communication 12*(4), 317–337.

UNECOSOC (2011). Background paper - imperative for quality education for all in Africa: Ensuring equity and enhancing teaching quality. Technical report, United Nations Economic and Social Council.

Weber, E. (1976). *Peasants into Frenchmen: the modernization of rural France, 1870-1914*. Stanford University Press.

Weinstein, B. (1983). *The civic tongue: Political consequences of language choices*. Longman New York.

Whitehead, C. (2005). The historiography of British imperial education policy, Part II: Africa and the rest of the colonial empire. *History of Education 34*(4), 441–454.

Young, C. (1983). The temple of ethnicity. *World Politics 35*(04), 652–662.

Table I: **Distance from official language for the three largest ethnic groups for selected countries**

Country	Group Name	Size	Official Language/s	Distance From Offical Language 1	Distance From Offical Language 2
Belarus	Byelorussian	0.78	Belarussian	0	n/a
	Russian	0.13	Belarussian	0.13	n/a
	Poles	0.04	Belarussian	0.34	n/a
Burkina Faso	Mossi	0.5	French	1	n/a
	Western Mande	0.14	French	1	n/a
	Fulani(Peul)	0.1	French	1	n/a
Indonesia	Javanese	0.45	Bahasian	0.181	n/a
	Sunda	0.15	Bahasian	0.212	n/a
	Malays	0.06	Bahasian	0	n/a
Peru	Amerindian	0.45	Spanish	1	n/a
	Mestizo	0.37	Spanish	0	n/a
	White	0.15	Spanish	0	n/a
South Africa	Zulu	0.22	English, Afrikaans	1	1
	Xhosa	0.18	English, Afrikaans	1	1
	Afrikaner	0.09	English, Afrikaans	0.26	0

Note: According to the coding rules, Afrikaans speakers are treated as if Afrikaans were the only official language; hence their value for distance is zero.

Table II: **Socio-economic outcomes by quartiles of language distance**

	Whole Sample	**Quartile 1**	**Quartile 2**	**Quartile 3**	**Quartile 4**
Human Development Index in 2010	0.66 (150, 0.18)	0.78 (38, 0.13)	0.76 (37, 0.13)	0.64 (36, 0.14)	0.46 (39, 0.13)
Log GDP per capita in 2005	8.58 (147 , 1.31)	9.32 (37, 1.03)	9.13 (37, 1.03)	8.47 (35, 1.14)	7.39 (38, 1.05)
Years of Schooling	4.79 (117 , 2.91)	6.04 (33, 2.45)	6.24 (27, 2.96)	4.33 (31, 2.72)	2.22 (26, 1.45)
Institutionalized Democracy Score	3.7 (151, 3.61)	5.57 (38, 4.00)	4.66 (37, 3.75)	3.07 (37, 3.08)	1.54 (39, 1.99)
Life Expectancy in 2010	68.95 (152, 9.84)	75.90 (39, 5.25)	74.65 (37, 5.21)	68.21 (37, 7.98)	57.27 (39, 6.99)
Infant Mortality Rate in 2010	43.01 (152 , 44.67)	18.67 (39, 30.60)	18.28 (37, 18.16)	40.92 (37, 35.04)	92.78 (39, 41.23)
Poverty Headcount (%) under $2 a day.	36.31 (105, 31.47)	17.85 (22, 25.12)	16.62 (21, 19.05)	30.97 (27, 29.37)	63.84 (35, 22.41)

In the parenthesis are provided the number of observations followed by the standard deviation.

Table III: **Regressions of distance on cognitive scores, life expectancy, log GDP per capita and log output per worker**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Cognitive test score	Cognitive test score	Life Expt. in 2010	L. Expt. in 2010	log GDP per capita	log GDP per capita	log Output per worker	log Output per worker
Average distance from official language	-1.000***	-0.848**	-12.79***	-7.316**	-1.381***	-1.290***	-1.570***	-1.089***
	(0.295)	(0.382)	(2.055)	(3.007)	(0.267)	(0.337)	(0.193)	(0.296)
	[-0.439]	[-0.373]	[-0.504]	[-0.288]	[-0.396]	[-0.370]	[-0.570]	[-0.395]
Linguistic fractionalization a/c for distance	0.120	-0.00776	-2.537	-3.663	-0.195	-0.244	0.445	0.231
	(0.292)	(0.380)	(3.024)	(3.270)	(0.432)	(0.404)	(0.318)	(0.300)
	[0.0387]	[-0.00250]	[-0.0559]	[-0.0807]	[-0.0313]	[-0.0390]	[0.0877]	[0.0455]
Executive constraints	0.113***	0.0987***	1.343***	0.788**	0.261***	0.193***	0.173***	0.132***
	(0.0286)	(0.0367)	(0.302)	(0.322)	(0.0455)	(0.0515)	(0.0379)	(0.0449)
	[0.392]	[0.341]	[0.247]	[0.145]	[0.389]	[0.287]	[0.314]	[0.240]
Log GDP per capita at independence	0.116**	0.0476	0.808	0.286	0.374***	0.318**	0.376***	0.312**
	(0.0507)	(0.0487)	(0.616)	(0.634)	(0.115)	(0.127)	(0.115)	(0.124)
	[0.179]	[0.0737]	[0.0640]	[0.0227]	[0.232]	[0.197]	[0.227]	[0.188]
HIV prevalence in 2000			-0.566***	-0.475***				
			(0.114)	(0.116)				
			[-0.323]	[-0.271]				
Natural Resources	No	No	No	No	Yes	Yes	No	No
Continent Dummies	No	Yes	No	Yes	No	Yes	No	Yes
Observations	69	69	106	106	134	134	111	111
R-squared	0.509	0.594	0.754	0.781	0.663	0.685	0.692	0.711

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table IV: **Regressions of distance on zHDI**

	(1)	(2)	(3)	(4)	(5)
Average distance from official language	-1.997***	-2.115***	-1.658***	-1.470***	-1.117***
	(0.131)	(0.157)	(0.167)	(0.160)	(0.260)
	[-0.743]	[-0.787]	[-0.615]	[-0.545]	[-0.415]
Linguistic fractionalization a/c for distance		0.367	0.246	-0.0229	-0.131
		(0.352)	(0.316)	(0.281)	(0.278)
		[0.0760]	[0.0509]	[-0.00473]	[-0.0271]
Executive constraints			0.199***	0.171***	0.127***
			(0.0264)	(0.0237)	(0.0278)
			[0.391]	[0.337]	[0.250]
Log GDP per capita at independence				0.292***	0.243***
				(0.0516)	(0.0554)
				[0.258]	[0.215]
Continent Dummies	No	No	No	No	Yes
Observations	150	150	149	149	149
R-squared	0.552	0.556	0.684	0.742	0.758

*p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table V: **Regressions of distance on zHDI excluding one continent at a time**

	(1)	(2)	(3)	(4)	(5)	(6)
Average distance from official language	-1.117***	-0.888***	-1.138***	-1.143***	-1.175***	-1.078***
	(0.260)	(0.335)	(0.271)	(0.306)	(0.275)	(0.271)
	[-0.415]	[-0.226]	[-0.415]	[-0.427]	[-0.485]	[-0.402]
Linguistic fractionalization a/c for distance	-0.131	-0.0462	-0.340	-0.0980	-0.109	-0.149
	(0.278)	(0.385)	(0.305)	(0.279)	(0.317)	(0.278)
	[-0.0271]	[-0.0124]	[-0.0644]	[-0.0189]	[-0.0258]	[-0.0311]
Executive constraints	0.127***	0.138***	0.125***	0.130***	0.131***	0.127***
	(0.0278)	(0.0325)	(0.0273)	(0.0343)	(0.0348)	(0.0278)
	[0.250]	[0.382]	[0.237]	[0.239]	[0.244]	[0.247]
Log GDP per capita at independence	0.243***	0.184***	0.283***	0.0708	0.382***	0.246***
	(0.0554)	(0.0586)	(0.0579)	(0.0703)	(0.0637)	(0.0554)
	[0.215]	[0.241]	[0.248]	[0.0531]	[0.334]	[0.219]
Continent Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	149	103	124	108	115	146
R-squared	0.758	0.529	0.789	0.829	0.702	0.752

Column (1) considers the entire sample; column (2), (3), (4), (5) and (6) drop Africa, Americas, Asia, Europe and Oceania, respectively. *p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VI: **Robustness tests of regressions of distance on Standardized value of HDI**

	(1)	(2)	(3)
Average distance from official language	-1.117***	-1.029***	-1.148***
	(0.260)	(0.328)	(0.328)
	[-0.415]	[-0.381]	[-0.428]
Linguistic fractionalization a/c for distance	-0.131	-0.286	-0.232
	(0.278)	(0.297)	(0.292)
	[-0.0271]	[-0.0593]	[-0.0485]
Executive constraints	0.127***	0.142***	0.0892**
	(0.0278)	(0.0299)	(0.0374)
	[0.250]	[0.280]	[0.174]
Log GDP per capita at independence	0.243***	0.257***	0.332***
	(0.0554)	(0.0676)	(0.0766)
	[0.215]	[0.214]	[0.260]
Predicted genetic diversity		38.40	23.04
(ancestry adjusted)		(66.14)	(66.66)
		[1.029]	[0.628]
Predicted genetic diversity squared		-29.95	-18.93
(ancestry adjusted)		(46.91)	(47.07)
		[-1.137]	[-0.731]
State Antiquity Index		0.577**	0.403
		(0.285)	(0.288)
		[0.137]	[0.0977]
Legal Origins	No	No	Yes
Continent Dummies	Yes	Yes	Yes
Observations	149	136	130
R-squared	0.758	0.777	0.793

*p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VII: **Regressions of distance on cognitive scores, life expectancy, log GDP per capita, log output per worker and zHDI in 2010 - Sample of countries independent post-1945**

	(1)	(2)	(3)	(4)	(5)
	Cognitive test score	Life Expt. in 2010	log GDP per capita	log Output per worker	zHDI in2010
Average distance from official language	-0.608	-10.10***	-0.705*	-0.597**	-0.928***
	(0.474)	(3.409)	(0.398)	(0.295)	(0.323)
	[-0.339]	[-0.399]	[-0.229]	[-0.247]	[-0.373]
Linguistic fractionalization a/c for distance	0.0427	-1.930	-0.532	-0.464	-0.423
	(0.587)	(3.366)	(0.534)	(0.282)	(0.337)
	[0.0161]	[-0.0402]	[-0.0920]	[-0.109]	[-0.0890]
Executive constraints	0.0611	0.740	0.190**	0.0566	0.116***
	(0.0423)	(0.445)	(0.0782)	(0.0399)	(0.0329)
	[0.234]	[0.137]	[0.277]	[0.103]	[0.219]
Log GDP per capita at independence	0.0598	2.706***	0.693***	0.859***	0.488***
	(0.116)	(0.616)	(0.154)	(0.107)	(0.0626)
	[0.117]	[0.277]	[0.541]	[0.698]	[0.511]
Natural Resources	No	No	Yes	No	No
Continent Dummies	Yes	Yes	Yes	Yes	Yes
Observations	31	93	79	63	91
R-squared	0.501	0.744	0.650	0.786	0.788

The dependent variables in columns (1), (2), (3), (4) and (5) are cognitive scores, life expectancy in 2010, log GDP per capita in 2005, log output per worker from the work and zHDI in 2010, respectively. $*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VIII: **Regressions of distance on Standardized value of HDI in 1990 and 2010**

	(1)	(2)	(3)
	zHDI in 2010	zHDI in 1990	zHDI in 2010
Average distance from official language	-1.117***	-0.802***	-0.381***
	(0.260)	(0.289)	(0.125)
	[-0.415]	[-0.291]	[-0.140]
Linguistic fractionalization a/c for distance	-0.131	-0.214	0.0261
	(0.278)	(0.338)	(0.145)
	[-0.0271]	[-0.0446]	[0.00554]
Executive constraints	0.127***	0.158***	-0.00590
	(0.0278)	(0.0366)	(0.0131)
	[0.250]	[0.311]	[-0.0118]
Log GDP per capita at independence	0.243***	0.287***	
	(0.0554)	(0.0714)	
	[0.215]	[0.232]	
Standardized Value of HDI in year 1990			0.855***
			(0.0337)
			[0.869]
Continent Dummies	Yes	Yes	Yes
Observations	149	121	121
R-squared	0.758	0.712	0.955

In column (1) and (3) the dependent variable is zHDI in 2010; in column (2) it is zHDI in 1990. *p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table IX: **Factors affecting average distance from official language**

	(1)	(2)	(3)	(4)	(5)
Dummy for whether country has a written tradition	-0.708***	-0.708***	-0.711***	-0.715***	-0.391***
	(0.0386)	(0.0414)	(0.0386)	(0.0417)	(0.0970)
	[-0.814]	[-0.815]	[-0.817]	[-0.822]	[-0.450]
Log GDP per capita at independence		0.000455		0.00597	
		(0.0226)		(0.0243)	
		[0.00108]		[0.0142]	
Log Population in 1500 CE			0.00462	0.00577	
			(0.00926)	(0.00976)	
			[0.0229]	[0.0286]	
Continent Dummies	No	No	No	No	Yes
Observations	152	152	151	151	152
R-squared	0.663	0.663	0.665	0.666	0.740

*p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table X: **IV Regressions of distance on cognitive scores, life expectancy, log GDP per capita, log output per worker and zHDI in 2010**

	(1) Cognitive test score	(2) Cognitive test score	(3) Life Expt. in 2010	(4) L. Expt. in 2010	(5) log GDP per capita	(6) log GDP per capita	(7) log Output per worker	(8) log Output per worke	(9) zHDI in2010	(10) zHDI in2010
Panel A: Two-Stage Least Squares										
Average distance from official language	-1.49	-1.28**	-24.5***	-25.8***	-1.64***	-1.26**	-1.42***	-1.58***	-1.58***	-1.38***
	(1.29)	(0.56)	(3.06)	(3.35)	(0.56)	(0.53)	(0.44)	(0.41)	(0.36)	(0.31)
	[-0.64]	[-0.56]	[-0.92]	[-0.96]	[-0.47]	[-0.36]	[-0.51]	[-0.57]	[-0.59]	[-0.51]
Linguistic fractionalization a/c for distance		0.16		10.2***		0.054		0.55		0.050
		(0.37)		(3.65)		(0.56)		(0.42)		(0.33)
		[0.052]		[0.21]		[0.0085]		[0.11]		[0.010]
Executive constraints		0.078**		0.58*		0.18***		0.11***		0.13***
		(0.030)		(0.34)		(0.051)		(0.041)		(0.032)
		[0.27]		[0.11]		[0.27]		[0.20]		[0.24]
Log GDP per capita at independence		0.037		1.15*		0.40***		0.31***		0.24***
		(0.059)		(0.62)		(0.092)		(0.096)		(0.057)
		[0.058]		[0.100]		[0.27]		[0.18]		[0.21]
% of European descent in 1975		0.0014		0.0032		0.0045*		0.0053**		0.0041**
		(0.0018)		(0.018)		(0.0027)		(0.0022)		(0.0017)
		[0.12]		[0.014]		[0.15]		[0.20]		[0.17]
America		-0.55***		0.24		-0.042		-0.095		-0.00031
		(0.15)		(1.38)		(0.20)		(0.16)		(0.13)
		[-0.33]		[0.0091]		[-0.012]		[-0.036]		[-0.00012]
Observations	70	66	152	139	147	135	112	110	150	137
R-squared	0.287	0.610	0.630	0.707	0.374	0.638	0.487	0.716	0.528	0.761
Panel B: First-Stage for ADOL										
Distance from Site of Invention of Writing	0.000021	0.000042***	0.000079***	0.000074***	0.000076***	0.000070***	0.000083***	0.000072***	0.000078***	0.000073***
	(0.000014)	(0.000012)	(0.000014)	(9.8e-06)	(0.000014)	(9.7e-06)	(0.000016)	(0.000011)	(0.000014)	(9.9e-06)
	[0.18]	[0.35]	[0.43]	[0.39]	[0.41]	[0.38]	[0.43]	[0.36]	[0.43]	[0.39]
Linguistic fractionalization a/c for distance		0.47***		0.71***		0.73***		0.66***		0.70***
		(0.13)		(0.096)		(0.096)		(0.11)		(0.098)
		[0.35]		[0.39]		[0.40]		[0.36]		[0.39]
Executive constraints		-0.0075		-0.027**		-0.031**		-0.031**		-0.028**
		(0.015)		(0.013)		(0.013)		(0.014)		(0.013)
		[-0.059]		[-0.14]		[-0.16]		[-0.15]		[-0.14]
Log GDP per capita at independence		0.0041		0.011		0.0046		-0.0023		0.010
		(0.030)		(0.025)		(0.024)		(0.036)		(0.025)
		[0.015]		[0.025]		[0.011]		[-0.0038]		[0.024]
% of European descent in 1975		-0.0025***		-0.0030***		-0.0029***		-0.0034***		-0.0030***
		(0.00066)		(0.00062)		(0.00062)		(0.00071)		(0.00063)
		[-0.48]		[-0.35]		[-0.33]		[-0.36]		[-0.35]
America		-0.11		-0.13**		-0.13**		-0.14***		-0.14***
		(0.072)		(0.051)		(0.051)		(0.054)		(0.052)
		[-0.15]		[-0.14]		[-0.13]		[-0.15]		[-0.14]
Observations	70	66	152	139	147	135	112	110	150	137
R-squared	0.034	0.495	0.185	0.671	0.172	0.683	0.188	0.692	0.181	0.671
F-Stat	2.39	9.63	34.0	44.9	30.1	46.1	25.4	38.6	32.6	44.2

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table XI: **IV Falsification test - Regressions of distance from sites of invention of writing on development outcomes**

	(1)	(2)	(3)
	Average Protection against Expropriation Risk	Social Infrastructure	Constraints on the Executive
Distance from Site of Invention of Writing	-1.8e-06	-9.4e-06	0.000060
	(7.4e-06)	(0.000011)	(0.000080)
	[-0.021]	[-0.080]	[0.062]
Observations	127	112	149
R-squared	0.000	0.006	0.004
F-Statistic	0.057	0.71	0.57

*p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table XII: **Marginal probability estimates of language distance from official state language on socio-economic outcomes**

	(1)	(2)	(3)	(4)	(5)	(6)
	Years of Education	Indicator for Literacy	Indicator for having heard about AIDS	Indicator for using mosquite net for sleeping	Indicator for white-collar job	Indicator for belonging to top-income quintile
Distance from State Language	-0.81*** (0.00)	-0.059*** (0.00)	-0.09*** (0.00)	-0.043*** (0.00)	-0.025*** (0.00)	-0.009*** (0.00)
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Language Group Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year of Birth Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Backward Group Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Place of Residence Indicator	Yes	Yes	Yes	Yes	Yes	Yes
Religion Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Altitude in Metres	Yes	Yes	Yes	Yes	Yes	Yes
Observations	76476	76354	76471	34094	18249	76476
Sample Average for the Depedent Variable	6.82	0.66	0.77	0.40	0.08	0.31

*p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis .

Table XIII: **Educational data for East and Southern Africa: Selected Descriptive statistics**

Variable	Observations	Mean	Standard Deviation	Min	Max
Essential Reading Score	33141	492.80	106.22	5.72	1061.83
Comprehensive Reading Score	33141	492.14	101.48	5.72	1061.83
Essential Math Score	32908	492.46	106.98	.432	1143.5
Comprehensive Math Score	32908	492.83	105.00	.432	1200.43
Proportion With Minimum Reading Level	33141	.39	.49	0	1
Proportion With Desirable Reading Level	33141	.14	.35	0	1
Socioeconomic Index	33141	7.02	3.31	1	15
Age	33141	13.52	1.86	9.59	25.5
Male	33141	0.5	0.5	0	1
Whether Repeated Grade	33141	0.49	0.5	0	1
Mean Years of Education of Parents	33141	3.50	1.36	1	6
Poss. of Exercise Books	33141	0.06	0.24	0	1
Poss. of Pencils	33141	0.16	0.37	0	1
Poss. of cattle	33141	7.74	27.74	0	500
Poss. of Sheep	33141	2.44	15.14	0	500
Home Interest	33141	10.65	2.17	5	15
Extra Lessons Outside the Classroom	33141	0.61	0.49	0	1
Pupil Abseentism Problem	33031	0.05	0.22	0	1
Regularity of meals	32674	10.82	1.83	3	12
Home Quality	33141	10.14	3.22	4	16
Homework assistance Maths 1	33141	2.27	0.67	1	3
Homework assistance Reading 1	28809	2.6	0.6	1	3

Table XIV: **Effect of exposure to English on student achievement**

	(1)	(2)	(3)	(4)	(5)	(6)
Use of English at home	19.67***	18.93***	18.82***	18.16***	.085***	.041***
	(1.18)	(1.12)	(1.20)	(1.18)	(0.007)	(0.004)
Classroom Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Level Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	28349	28349	30952	30952	28349	28349

The dependent variables in columns (1) and (2) are the essential and comprehensive reading score; columns (3) and (4) are the essential and comprehensive math score; columns (5) and (6) the dependent variable is a binary indicator of whether the student reaches the minimum and desirable reading level. The list of individual level controls is shown in Table XIII. *p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.