Returns to Speaking English in India*

Suraj Shekhar†

The Pennsylvania State University

sws5487@psu.edu

December 23, 2013

Abstract

The English language is a legacy of British rule in India. Overtime, historical and economic factors have made English the language of commerce and government in India. This paper quantifies the impact of English speaking ability on wages. I show that the returns to fluent English speaking ability are in the range of 10-15 percent. A previous attempt at this (Azam, Chin and Prakash (2013)) points to returns around 30 percent in India. I show that two factors bias their results. One, they don't use enough controls which leads to omitted variable bias. Two, they use wages which are not representative of rural India. In addition, I demonstrate that English speaking skills affect wages by helping one get better paying jobs. Quantile regressions indicate that while returns to schooling are always higher at higher quantiles of wages, the returns to English speaking ability are highest for the middle quantiles in rural areas. Thus, unlike education, English speaking skills don't increase income inequality in rural India.

*Working paper version. Please do not distribute.

†I am deeply grateful to Kala Krishna and Paul Grieco for many helpful comments and discussions.

Contents

1 Introduction **3**

2 Data Features **5**

3 Impact of English Ability on Household Income **6**

4 Individual Wages and English Ability **8**

4.1 Construction of More Inclusive Wages 9

4.2 Impact of English ability on Individual Wages 10

4.2.1 Empirical Estimation 10

4.3 Simultaneous Estimation of English selection and Wages 14

4.4 English Ability and Good Jobs 16

4.5 Complementarity between English ability and Academic ability 18

5 Quantile Regressions **20**

6 Conclusions **21**

A Data **23**

A.1 Variables influencing Household Income 23

A.2 Other Variables Used in the Paper 28

B Additional Results **29**

1 Introduction

The English language is curiously important in India. In most countries the language of government and the corporate world is the native language. However, in India, the official language used in many firms and institutions is English. This is probably due to India's colonial past. The British introduced schools where the medium of instruction was English to produce low level clerks for their administration. Most official work was done in English then and the current civil service exam continues to favour people with good English speaking skills. In addition, the people who were doing really well when India became free (1947) were typically people who had good relations with the British and had probably been educated in English medium schools. These people hired employees who they thought were best qualified, which, more often than not, would be people with similar backgrounds as their own. Add to this the fact that hitherto English was spoken by the ruling class (thus becoming a status symbol) and it's no surprise that English speaking came to be associated with success and good schooling. With time, the globalization of the world and liberalization of the Indian economy (post 1991) gave English speakers another advantage. Multi national companies wanted people who could communicate easily with headquarters. Domestic firms who wanted to interact with these big multi nationals or go into the export market had to hire people who could speak good English. This is a big reason behind major universities in India teaching most of their courses in English. The outcome of all this is that being able to speak good English in India has become a big advantage.

So how much does English speaking ability affect wages in India? Also, how does it affect wages? Is it just that being able to speak good English helps one land a good job? Or is it simply that people with higher ability, good education and wealthy backgrounds get high wages and are more likely to know English so that the relationship between English ability and wages is not causal? In a country as diverse as India, it would also be interesting to know if the returns to English speaking ability are different in urban-rural areas and across gender. The work in this paper is reduced form so I cannot claim to have a model to analyze policy implications. This paper is important because it shows that the effect of English ability on wages is high enough to warrant policy discussions.

There has been some previous work which deals with similar questions for India. Munshi and Rosenzweig (2006) show that students gain (in terms of wages) by going to English medium schools. The returns to these schools turn out to be around 25 percent for both men and women in the year 2000 (This is for a sample from Dadar area of Mumbai). Kapur and Chakraborty (2008) take advantage of a policy change made by the state of West Bengal in India in the year 1983 when they banned English as the medium of instruction in primary schools. The authors compare cohorts using NSS (National Sample Survey) data and claim that switching from English to Bengali hurt that cohort's wages by about 15 percent. Angrist and Lavy (1997) have a similar paper estimating the impact of a change in language of instruction from French to Arabic in Morocco. Another international example is Levinsohn (2007) where he estimates the change in the returns to speaking English in South Africa after their reintegration with the world economy in 1994. These papers also find significant gains from knowing the colonial language. The paper mine is closest to is Azam, Chin & Prakash (2013)(Henceforth referred to as ACP). They consider the impact of English speaking

skills on wages using the IHDS (Indian Human Development Survey) 2005 data. This data set has a self reported variable on English speaking ability which takes one of three discrete values - speak no English, speak some English and speak fluent English. Though they don't claim to have proven a causal relationship, their work points towards returns to fluent English speaking skills being close to 33 percent (on hourly wages). I show that omitted variables bias their results and the wage variable they use is not representative of rural India since it doesn't include family farm income. I construct a better wage variable to correct for this and show that returns are more in the range of 10-16 percent. Although this paper asks some questions similar to the ones asked by ACP, my analysis is very different from theirs. In addition to adding more controls I look at the selection problems more carefully and provide evidence to show that selection into work force and endogeneity of English speaking ability are not biasing my results. Moreover, I show that English speaking skills affect wages by helping one get a higher paying job. Interestingly, English speaking continues to be a significant determinant of wages even after controlling for jobs. This suggests that its impact on wages is not limited to the above mechanism. Further contributions of this paper include showing complementarity between English speaking ability and academic ability in terms of their impact on wages. For example, among males who work, a person who speaks fluent English and did well in the tenth grade exams earns about 32 percent more than one who does not speak English and did not complete tenth grade. Lastly, I run quantile regressions which indicate that schooling and English speaking skills are different in their impact in rural areas. While returns to schooling are always higher at higher quantiles, the return to fluent English speaking ability is mostly flat and marginally higher for middle quantiles. Thus, while schooling increases income inequality in rural India, English ability does not have this distorting effect.

At this point, let me talk about how I deal with some of the obvious problems one faces when estimating the returns to a language in terms of wages. The first one is selection bias. I only have wages for the section of population which works and reports wages. The issue is that this may be a non-random selection of the population. For example, English speakers may have a higher probability of being in this sample. A more serious concern is that people who don't speak English but are still in our sample are special in some unobserved way (like high ability) and would thus not be representative of the non-English speaking group. This would bias our coefficients downwards. In essence, the sample selection issue is one of omitted variables. I will argue that this may not be a big problem since I control for most factors affecting wages. I want to make two additional points and back them with two robustness checks. One, most males (70 percent) in my sample do choose to work. Potentially, we could still have a problem if English speaking males are more likely to work and are therefore over represented in the sample. A simple probit reveals that English ability is not a significant determinant of whether a male works or not. As a robustness check to see that there is little bias from selection into work, I look at a sample of males between the ages of 21 and 60 (instead of 16-65). Close to 80 percent of this sample works. So the sample selection bias must be even lower. The return to fluent English ability here is 14 percent which is very close to the return in the original sample (13 percent). I conduct a further robustness check and look at the impact of fluent English speakers in the household on household income. I have household income for every household so there is no bias from selection into work. The coefficients here are also extremely close

(households with one fluent English speaker earn about 13 percent more than households with no fluent English speakers) to the ones from the individual regression.

The second issue here is that English speaking ability may be endogenous in the wage regression. If there are omitted variables affecting both English ability and wages or if there is reverse causality[1] (people with high wages may have stronger incentives to learn English and improve their job growth rates) then my estimates for returns to English ability may be biased. I run the English selection equation and wages regression simultaneously and allow the errors to be correlated (to control for unobserved variables affecting both). The results indicate approximately zero correlation between the equations suggesting low omitted variable bias.

One other problem is that the wage variable available for individuals in this data set is only for wages received outside family farm and family business income. In a country where about 60 percent of the population does not live in urban areas, this might be a serious issue. I try to mitigate this problem by constructing a more inclusive wage variable. [2]

The rest of the paper is organized as follows. In section 2, I talk about the data set and point out some of its special features which make it possible to conduct this analysis. Then, in section 3, I look at the impact of English speaking skills on household income followed by its impact on individual wages in section 4. Finally, I look at quantile regressions in section 5 and then conclude. The data appendix and references are at the back of the paper.

2 Data Features

The most likely reason why this question has not been satisfactorily answered before is the lack of a suitable data set. So first let me point out some special features of this data set(IHDS 2005) which make it possible to answer this question. This data set contains a huge list of variables for a randomly selected population of India including a plethora of consumption, wages and other demographic variables. The ones that are of special interest are:

1. Self reported English speaking ability which could be one of three - Speaks no English, Speaks a little English, Speaks English fluently.

2. The division obtained by the individual in the tenth grade examination (if he reached tenth grade and took this nationwide exam). This takes three values too - Division 1 (high academic ability), Division 2 (medium academic ability), Division 3 (low academic ability). I create a

[1] I expect this to be a small issue since most people would learn their English skills while in school and therefore before starting work. There might be some people who pick up English while working. So we may find that older workers in some jobs speak better English than the younger ones. The experience variable should pick some of this up. In any case, if this problem is bigger than I expect then the simultaneous estimation of English learning and wages would reflect that.

[2] The other issue which might be relevant here is measurement error. English ability is self reported and is therefore prone to bias. There is bound to be some attenuation bias in the coefficients. However, most of the measurement error is probably in the variable indicating that one speaks a little English. I expect much less measurement error when people report speaking fluent English or not speaking English at all

dummy for each of these and use these variables as indicators of ability. All three are zero if the person did not reach tenth grade. I assume that this indicates lowest academic ability.

Of course this data set has its issues too:

1. Only one cross sectional point for the year 2004(133000 individuals in the age group of 16-65).
2. Wages available at the individual level are only for a fraction of the population. Moreover these are wages earned *outside* the family farm and family business. For an agricultural country like India, the exclusion of family farms means that this variable does not measure wages accurately in rural areas.

3 Impact of English Ability on Household Income

I start with the effect of English ability on household income since this information is available for everyone. Thus, here, we don't have bias coming from selection into work. Note that household income refers to total income of the entire household from all possible sources. Since my conjecture is that English ability influences wages, naturally it should have an impact on household income. However, English ability is available at the individual level. To estimate the impact of English skills on household income I count the number of fluent English speakers in a household and the number of household members who speak a little English and use these variables as explanatory variables in the regression with household income as the dependent variable. To make sure that I don't count the same household twice I run the regression using only heads of households from the data. However, unlike Azam, Chin, Prakash (2013), I don't assume that household income depends on characteristics of the head of the household. Instead, I construct household level average variables and use them to capture the nature of the entire household.

Prior to the regression, let's first look at some important variables which could explain household income. Obviously, one needs to control for education levels and ability. I include discrete effects for years of schooling of the highest educated male and female member of the household. I also construct a variable which gives average education of the adults in the household and average division obtained by the adults in the tenth grade examination (to account for ability). Other explanatory variables include multiple controls for place of residence , experience, experience squared, variables to account for social status and social networks of the household, variables which talk about information or how well read the household members are and finally a couple of variables to control for the number of people and number of adults in the household. Please refer to the data appendix at the back of the paper for a detailed description of these variables and the complete model used.

In running the regression I only use households who show a positive income of at least Rs 5000 per year which is very very low. A single individual would have to be earning twice as much to be just on the poverty line. The final regression I run has the following form:

$$log(Income_h) = \beta_0 + \sum_{1}^{8} \beta_i * I((No.\ of\ Fluent\ English\ Speakers)_h = i) + \sum_{9}^{19} \beta_i * I((No.\ of\ Little\ English\ Speakers)_h = i - 8) + {X_h}^T \gamma + \epsilon_h \quad (1)$$

Table 1: Discrete Model Estimation

Variable	**Coefficient** (Std. Err.)
Number of Fluent English speakers=1	0.128** (0.027)
Number of Fluent English speakers=2	0.223** (0.036)
Number of Fluent English speakers=3	0.012 (0.073)
Number of Little English speakers=1	-0.016 (0.018)
Number of Little English speakers=2	0.017 (0.024)
Number of Little English speakers=3	-0.007 (0.060)
Avg. Tenth Grade Division	-0.114** (0.015)
Intercept	9.798** (0.062)

>3 speakers suppressed

Other controls include household size, schooling,location dummys

Full Table in Appendix 2

These results (table 1) show that on an average, household members who speak only a little English don't add to the household income. This could be because of measurement error. I expect much more error when people report speaking a little English as compared to when people say they don't speak English or when they report that they speak fluent English. The first fluent English speaker raises household income by around 13 percent(compared to a household with no fluent English speakers) and the second adds roughly 10 percent more to the aggregate household income.

I realize that some of the explanatory variables in the above regression could be endogenous like number of people in the household (usually poor people in India have more children). Removing this variable does not change the coefficient of the number of English speakers by much. Another possibility is that individuals who grew up in richer households are more likely to be able to speak English and so I may have the relationship backwards! This may indeed be a problem. The first point I want to make in this regard is that most English speakers are earning above average in my sample and therefore make significant contributions to household income. Moreover, I run the same regression again after adding dummy variables for a number of household assets that a household may posses. This should control for the effect of the wealth of a household in determining its current income[3]. Fluent English speakers are still significant determinants of household income(though the effect is smaller than before. The first fluent English speaker adds 9 percent to household income and the second adds another 8 percent). Since wealth of the household is inversely related to the number of people in the household the above exercise also addresses the problem of endogeneity of the latter in a crude way. Also, having the benefit of knowing what is to come in this paper, allows me to assert that English speaking ability significantly affects individual wages and therefore must affect household income.

In their 2013 paper, Azam et al. (2013) conduct a different exercise to estimate impact of English ability on household income. Instead of creating average household characteristics for variables like education, performance in the tenth grade exams, experience etc they use characteristics of the head of the household. Moreover, they do not use many variables like number of adults in the household, information variables,social networks variables, average experience of adults. They find that in households where the head speaks fluent English, the income is 34 percent higher compared to households where the heads don't speak English at all. Apart from the lack of additional controls, the major problem with this analysis is that the entire household income gets attributed to the head of the household. In essence, they use the characteristics of the head to control for the the nature of the entire household. This is a strong assumption to make (and a needless one since we have information about every member). In most cases there will be other members in the household who contribute to the household income. Consider a situation where the head of the household has retired from work and his children are working. The above analysis would still attribute the household income to the head while ignoring the characteristics of the actual earners.

4 Individual Wages and English Ability

While household income is available for everyone in the sample, only a fraction report earning wages individually. Moreover, the wages that are available are annual wages (hourly wages are also in the data) for work done outside the family farm and family business. For an agricultural country like

[3]Household assets could themselves be correlated with current household income. I assume that assets like type of wall of the house (brick, mud, cement, etc.), pressure cooker, TV etc are reflective of the wealth of the household rather than current income. I assume this because these items are unlikely to have been purchased just because the current income is high. If the household is not poor then it would have had these items for a long time. So we can use these assets to control for the wealth of the household when its current members were growing up and acquiring their English skills.

India, this is a major restriction. While ACP(2013) have also attempted to quantify the impact of English ability on wages, this section of my paper will differ from theirs in four regards (this also describes how this section is organized):

1. I create a wage variable which will be an approximation to the annual wages of an individual *including* family farm and family business income. Regression results from the original wage variable and this one are compared.

2. Simultaneous estimation of English selection and wages allowing for correlated error terms (to control for unobserved variables affecting both).

3. English speaking ability helps individuals get higher paying jobs. I will use data for young men (between the ages of 16-30) to show that English ability is significant in determining what job an individual gets. Running wages on explanatory variables including job categories shows that the jobs which require English ability pay much better than others.

4. I show that there is complementarity between academic ability and English speaking ability in terms of their impact on annual wages of individuals.

4.1 Construction of More Inclusive Wages

The wage variable available in the data set (at the individual level) describes wages received by an individual outside family farm and family business. Let's call this variable $W0$. For each individual we also have - number of days worked in the family farm in the past year, number of hours worked per day on the family farm, number of days worked per year in the family business, number of hours worked per day in the family business. Thus we can calculate the total number of hours worked by an individual on the farm last year. Similarly, calculate the number of hours worked last year in the family business for each individual. Now we can get the fraction of the aggregate household work hours (sum of all members) contributed by each member and give each one that share of the household farm income and household business income. Of course, the implicit assumption here is that each member has the same productivity.

So the new wage variable($Wages4$) is:

$$\begin{aligned} Wages4_i = & W0_i + (Household\ Farm\ Income)_i * (fraction\ of\ work\ hours\ contributed\ by\ i\ in\ family\ farm) \\ & + (Household\ Business\ Income)_i * (fraction\ of\ work\ hours\ contributed\ by\ i\ in\ family\ business) \end{aligned} \tag{2}$$

Note that no individual gets any part of the family farm or family business income if they have not worked on it. They get the proportional share of their hours spent(as compared to the household) if they do.

4.2 Impact of English ability on Individual Wages

In this section, the primary variables of interest are dummy variables describing whether English speaking ability is fluent or whether the person speaks only a little English. In addition, I have controls for education, ability, location effects, etc which were used earlier in the household level regressions and I have added three other controls to the regression to account for the wealth of the household, marital status and to control for alternative income sources which reduce the incentive to earn more yourself. For example, I use a variable which describes the percentage of household income which is received as remittances from members working elsewhere. Again, details on all the variables used and the full empirical model is described in the data appendix at the back.

4.2.1 Empirical Estimation

In this part, I explain the equations used to estimate the impact of speaking English on $W0$ and on $Wages4$. We can then compare the two regressions and see how the constructed wage variable helps capture aspects which the first regression ignores. The base estimation equation will be of the form :

$$log(WO_j) = \beta_0 + \beta_1 * (Can\ Speak\ Fluent\ English)_j + \beta_2 * (Can\ Speak\ Little\ English)_j + X^T{}_j\gamma + \mu_j \tag{3}$$

Where X_j contains all other controls for individual j.

The above regression is run on males between the ages of 16 and 65 years whose $W0$ is at least Rs 5000/year which as mentioned before is very low. This wage variable does not include own farm income. Therefore, I don't want to include people whose main source of income is farming but they show up in the sample because they do some part time work to supplement their income. Therefore, I include only those males for whom the household farm income is less than 25 percent of the aggregate household income. I also include a dummy variable indicating whether the individual worked for more than 240hrs on his family farm last year to control even better for this defect in $W0$. The same equation is then run with $log(Wages4)$ as the dependent variable [4]

The base regression results (table 2, Model 1) suggests that Ceteris Paribus, speaking a little English leads to an average increase of 6 percent in the $W0$ of a male as compared to a male who does not speak English. However, being able to speak fluent English raises wages by 16 percent compared to the non-speaker. I must point out though that these returns are in no way homogeneous throughout India. I ran the regression separetely on rural and urban areas and it turns out that the return to fluent English in rural areas is around 27 percent while it's only about 6 percent in urban areas. On the other hand, return to high academic ability is very large in urban areas and not significant in rural areas. I will talk more about these patterns in the quantile regression section.

[4] Here I don't restrict the sample to individuals who come from non-farming households. I look at people whose $Wages4 > 5000$.

Table 2: Model Estimations

VARIABLES	(1) Model1 log(W0)	(2) Model2 log(W0)	(3) Model3 log(W0)	(4) Model4 log(W0)
Can Speak Little English	0.0612***	0.0774***	0.107***	0.132***
	(0.0210)	(0.0208)	(0.0223)	(0.0214)
Can Speak Fluent English	0.163***	0.224***	0.275***	0.308***
	(0.0332)	(0.0331)	(0.0338)	(0.0335)
Schooling Years = 10	0.244***	0.312***	0.387***	0.472***
	(0.0431)	(0.0467)	(0.0523)	(0.0470)
Schooling Years = 12 (High School)	0.289***	0.382***	0.470***	0.576***
	(0.0500)	(0.0523)	(0.0557)	(0.0515)
Schooling Years=15 (College)	0.365***	0.522***	0.649***	0.801***
	(0.0529)	(0.0549)	(0.0582)	(0.0534)
High Academic Ability	0.0909*	0.153***	0.184***	0.231***
	(0.0538)	(0.0577)	(0.0620)	(0.0580)
Constant	9.463***	9.615***	9.730***	9.676***
	(0.0576)	(0.0537)	(0.0549)	(0.0556)
Observations	22,812	24,500	24,868	24,868
R-squared	0.602	0.566	0.544	0.527

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Model 2 describes results when I drop controls for household wealth. The coefficient on fluent English ability now jumps to 0.22. This is because household wealth is positively correlated with both English ability and wages of an individual. So when we don't control for it, the coefficient of English ability rises.

Model 3 is essentially model 2 without the controls for social networks and schooling of the most educated female in the household. Unfortunately, nepotism is fairly prevalent in India. Connections with a government official (especially having one in the household) can really improve the chances of finding a good job. A female member of the household being highly educated could be positively correlated with the ability and quality of schooling of males in the household.[5] These will impact English ability and so, as before, the impact of English ability appears larger than it is.

Finally I look at the barest possible model(Model 4) and drop variables reflecting outside income, marital status and information from model 3. This is very close to the model used by ACP(2013). The return to fluent English speaking ability here is about 30 percent. ACP(2013) find that the impact of speaking fluent English on hourly wages(for wages outside family farm and family business income) is about 32 percent. I run my base regression(Model 1) with $log(Hourly\ WO)$ as dependent variable and find the return to be closer to 15 percent. Model 4 shows exaggerated coefficients since it does not control for many variables which might be affecting wages. Some of these variables are correlated with English speaking ability and therefore the coefficient of English

[5]Having a government official in the household may also have similar effects

ability blows up when these are omitted.

If we look at the results from the base (Model 1) $Wages4$ regression[6](close to 40000 males in this sample) then the coefficients of returns to little and fluent English are 4 percent and 13 percent respectively. One would think that this is not a major shift from estimates in model 1, table 2. However, if we run this regression over urban and rural areas separately we begin to see the differences in the two outcomes. While the numbers are similar in urban areas, in rural areas the return to fluent English is about 19 percent - much lower than the 27 percent before. Let's try and understand this through some rough statistics on the approximately 17000 males that are added to the sample when we look at $Wages4$ instead of $W0$[7]. The mean of the *URBAN* dummy (which is 1 if urban and 0 if rural) falls from 0.45 in the $W0$ regression to 0.33 in the $Wages4$ regression. So we are mostly adding to the rural pool. The mean of the *WKFARM* variable (1 if worked on family farm for more than 240 hours last year and zero otherwise) jumps from 0.09 to 0.29 i.e. we are adding a lot of farmers. The mean of the *WKBUSINESS* variable (which is the family business counterpart to *WKFARM*) jumped from 0.02 to 0.17. Thus a lot of businessmen were also added to the sample. The average fluent English speaking ability in the sample that was added is also lower than the sample for the $W0$ regression. So, to sum up, to our original sample we add a set of males in the rural areas who are earning money without being able to speak good English. Moreover, for the farmers and small businessmen who were in the original sample due to some part time work they did, the incomes were under-represented when we ran the wage equation using wages from outside family business and family farm. Naturally then, the returns to speaking English in the $Wages4$ regression are smaller (and perhaps more accurate). Since the additions have been largely to the rural pool, only their returns change. The basic idea is that the wage variable $W0$ is not representative of the rural population in India and thus the regression results using that variable suffers from sample selection.

Consider the coefficients of completing tenth grade, high school and college in the two regressions. The $Wages4$ regression seems to indicate much lower returns. Very similar arguments would work here as well. Farmers can earn money without being well educated. Here is an additional test to show that the presence of these small farmers is making the difference - I ran the $Wages4$ regression after removing those individuals which come from farming households (keep only those individuals for whom the percentage of household income coming from farm income is less than 10). Now the returns to both English speaking ability and schooling are similar across the $W0$ regression and the $Wages4$ regression.

Table 2 tells us that omitted variable bias could be a serious issue in ACP's estimates and table 3 indicates the pitfalls in using wages which don't reflect rural India correctly. However, both these estimation processes were silent on selection bias and endogeneity of English ability. We only observe wages for people who work. This causes two potential issues. One, English speakers may be

[6]Table 3 shows a comparison of the regression outcomes in the base regression (Model 1) if $W0$, $log(Wages4)$ are used as dependent variables respectively.

[7]Of course there is also a small fraction(roughly 3000) of the sample who were already in the $W0$ regression but whose income was underestimated by that measure of wages. However, I believe that the additional 17000 people will be driving the changes in the coefficients simply because they are much more in number than the group whose incomes were underestimated.

Table 3: W0 vs Wages4

VARIABLES	(1) log(W0)	(2) log(Wages4)
Speaks Little English	0.0612***	0.0433**
	(0.0210)	(0.0171)
Speaks Fluent English	0.163***	0.133***
	(0.0332)	(0.0288)
Schooling Years = 10	0.244***	0.151***
	(0.0431)	(0.0351)
Schooling Years = 12 (High School)	0.289***	0.171***
	(0.0500)	(0.0407)
Schooling Years = 15 (College)	0.365***	0.262***
	(0.0529)	(0.0426)
High Academic Ability	0.0909*	0.125***
	(0.0538)	(0.0432)
Medium Academic Ability	0.0325	0.0837**
	(0.0480)	(0.0379)
Low Academic Ability	-0.0774	-0.0259
	(0.0493)	(0.0399)
Constant	9.463***	9.371***
	(0.0576)	(0.0496)
Observations	22,812	39,849
R-squared	0.602	0.493

Robust standard errors in parentheses
Other controls same as Model 1, table 2
*** p<0.01, ** p<0.05, * p<0.1

more likely to work and therefore be over-represented in the sample. Two, there might be people who don't speak English but are still in the work force because they are special in unmeasured ways (like higher ability). This would bias the returns to English ability downwards. The ususal method for addressing this problem is to run a Heckman Selection model. Unfortunately, identification in this model requires an exclusion restriction i.e. a variable which influences the decision to work but not wages.[8] This is very hard to find for males.

I will consider a person to be working if he makes at least Rs 5000 ($log(Wages4) > 8.5$) per year. This is an extremely low cut off. In exchange rate terms it is roughly equivalent to $100 per year. One would have to be earning double of this to be just on the poverty line. I assume that this is approximately what one needs to earn to buy enough food for survival. I make two points to show that the selection into work problem is small for males. One, 70 percent of the males in my sample are working. Two, while English speakers are more likely to work for wages outside of family farm and family business, they are not more or less likely to be in the work force once we look at the more inclusive wage variable $Wages4$. That is to say, English is a significant explanatory variable when we run a probit model where the dependent variable is a dummy for $log(WO) > 8.5$

[8]Identification could also be achieved via parametric assumptions on the joint distribution of the errors but this is usually not convincing

but not if the dependent variable is a dummy for $log(Wages4) > 8.5$. This is because of all the non English speaking workers I add to the sample when I look at $Wages4$. I present two robustness checks in support of little selection bias in my regressions. One, compare the results in section 3 to the ones here. I have household income for every household so there is no bias from selection into work. The returns to English speaking ability are very similar in both regressions. Two, consider the $log(Wages4)$ regression with males who are between the ages of 21 and 60. This is a prime age group for working. Close to 80 percent of this sample works. Yet, the coefficients in this regression are almost exactly the same as before. If there had been any selection bias in the original sample, we would have observed higher returns in the prime sample. Essentially, selection is not a problem for males in India because males are always expected to work.[9] Thus, we can assume that the coefficient of English ability in the $log(Wages4)$ equation in table 3 is free of bias coming from selection into work.

The second issue is that English speaking skills may not be uncorrelated with the error term i.e. unobserved factors like quality of schooling could influence both English speaking ability and wages. If we don't correct for this then the coefficient of English ability would be biased. I estimate the English selection equation and the wage equation simultaneously in the next subsection.

4.3 Simultaneous Estimation of English selection and Wages

Suppose there was an unobserved variable like quality of schooling, which affects both English ability and wages. Then I may be overestimating the effect of English ability. To correct for this possibility, I estimate the English selection equation (ordered probit) and the wage regression(linear regression) simultaneously and allow the errors to be correlated. The correlation between the errors should pick up the effect of unobserved variables affecting the two dependent variables. I assume the errors to be bivariate normal distributed with means zero, standard deviations - σ_e, 1 and correlation coefficient ρ. If I estimate my base regression (Model 2, table 3) and the English equation simultaneously then I have to estimate over 200 variables with 40000 data points for males who work! This may be too time consuming. Instead, I take a random sample of 10000 males and drop the state dummys, social network dummys, variables reflecting information and some variables reflecting household wealth from the regression. I am left with a model which is close to Model 4 (in table 2). 67 coefficients are to be estimated including σ_e and ρ. I use all variables from the wage equation in the English equation and add another variable to the latter to serve as an exclusion restriction. The additional variable is a dummy which takes the value 1 if the most educated male *and* the most educated female in the household have both passed high school. The rationale behind this assumption is twofold. First, let's think about why the education level of the most educated women could be uncorrelated with the wages of the male (who constitutes our observation). Suppose the most educated women is the mother of the male. The education of the mother would be correlated with the education level and academic ability of a male in the household but once I have controlled for both of these (through years of schooling and division obtained in tenth grade) it may not be relevant. Now suppose the most educated women is the spouse of the male. There could be two reasons for the education level

[9] It would have been a completely different story if I was looking at a sample with females!

of the wife to be uncorrelated with the wages of the husband. One, people in India get married young. This makes it hard for the girl to judge the wage potential of her spouse before getting married. The most she can observe is the education level of her spouse. However, we have already controlled for that variable. Two, in rural India, the education of a woman is not valued highly in the marriage market. Thus, one will see high earning men marrying both educated and uneducated women making the correlation between womans education and husband's wages weak. However, it is a crucial indicator of English ability for males in the household[10] since female education being high could mean that the conversation in the household could sometimes be in English or that English books/newspapers are read in the household. Two, it takes two people to have a conversation so I need at least two people to have the ability to speak in English.[11]

The equations to estimate take the following form:

Wage Equation

$$Log(Wages4_i) = X_i{}^t\beta + I(E_i = 1)\gamma_1 + I(E_i = 2)\gamma_2 + \epsilon_i \tag{4}$$

English Equation

$$E_i{}^* = X_i{}^t\beta + I(Both\ Speak_i)\delta + \mu_i \tag{5}$$

$E_i{}^*$ is the latent variable for English speaking ability. We don't observe this variable. We can only observe E_i which takes one of three values. $E_i = 0$ (don't speak English) if $E_i{}^* < 0$, $E_i = 1$ (speak a little English) if $0 < E_i{}^* < A1$ and $E_i = 2$ (speak fluent English) if $E_i{}^* > A1$. Estimating these equations simultaneously reveals that the correlation between the errors is positive but small (0.12). This is expected. Any variable which increases the probability of speaking good English is almost always going to be positively correlated with wages. Examples for such variables would include ability, school quality and wealth of family. Table 4 compares the key coefficients for the simutaneously estimated model and a standard OLS wage regression with the same regressors as equation 4. The numbers are very similar. Thus, it seems like the controls I am using are enough to get the correct estimates and that there is little bias due to endogeneity of English speaking ability. Note that low correlation between the errors does not mean that unobserved variables such as school quality are not important in determining wages or English speaking ability. This is because variables reflecting wealth, big cities etc are correlated with school quality and so we are partially controlling for the latter. I should also point out that few controls were used in this simultaneous estimation and yet the correlation between errors was not big. This indicates that the correlation between errors in the full model may be even weaker since fewer variables would have been omitted there.

Since the unobserved variable bias is low (low bias from selection into work and low bias from endogeneity of English speaking ability), I will hereafter assume them to be negligible.

[10] An ordered probit shows this clearly.

[11] Being high-school educated is a good indicator being able to speak at least a little English.

Table 4: Simultaneous vs Standard

VARIABLES	(Simultaneous) log(Wages4)	(Standard) log(Wages4)
Speaks Little English	0.11	0.12
	(0.02)	(0.0104)
Speaks Fluent English	0.27	0.30
	(0.04)	(0.0185)
Schooling Years = 15 (College)	0.35	0.39
	(0.03)	(0.0257)
High Academic Ability	0.26	0.17
	(0.04)	(0.0248)
Correlation Coefficient	0.12	
	(0.01)	

Standard errors in parentheses

4.4 English Ability and Good Jobs

In this subsection I indicate a channel through which English speaking skills could affect wages - English abilities helps one get better paying jobs. Thus, a part of the coefficient of English ability in the wages regression is explained by the fact that having good English speaking skills helps one get jobs which pay better. To show the impact on jobs - I take a sample of young males(ages 16-30) and run a multinomial logit regression with categories of jobs as dependent variable and English abilities among the explanatory variables. The assumption here is that these young people have just started working so their explanatory variables may be easily assumed to be what they were when they were applying for a job and thus these variables directly influenced them getting these jobs. Subsequently, I run a wage regression and include job categories among the independent variables. The conjecture is that English abilities would be significant in determining whether you get into some of the job categories and these job categories would pay better and therefore show up with positive and significant coefficients in the wage regression. The other explanatory variables for both regressions are the same as model 1, table 2. The job categories are described by the discrete variable *Job Categ.* It takes the value zero (base case) if the individuals occupation is one of farming/fishing/hunting/logging/carpentering etc. This variable is one if the individual is a teacher/scientist/doctor etc (I expect English to play an important role in getting these jobs and these jobs probably pay more than category 0). Variable is two if the individual is a senior government officer or a senior member of a firm (I don't expect much from the coefficients in this category since this category naturally consists of older people. I am including this category just so that the few young people in this group are not counted among the base group). The variable takes the value three if the occupation is clerks/village officials/ supervisor, etc (I expect English to play some role in getting these jobs but not as much as category 1). Finally the variable takes the value four if the individual is a shopkeeper/salesman/merchant etc (English skills may have a small impact here).

Table 5: Impact of English Ability on Job you get

Variable	**Coefficient** (Std. Err.)
Job Categ : 1	
High Academic Ability	0.983** (0.350)
Speak Fluent English	1.073** (0.204)
Speak Little English	0.494** (0.153)
Job Categ : 3	
High Academic Ability	0.422 (0.280)
Speak Fluent English	0.534** (0.204)
Speak Little English	0.187 (0.136)
Job Categ : 4	
High Academic Ability	0.534† (0.308)
Speak Fluent English	0.448† (0.243)
Speak Little English	0.264* (0.130)

Job Categ 2 suppressed

Other controls include schooling year dummys, location dummys etc

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

The multinomial logit regression (table 5) shows that controlling for everything else, people who are fluent in English are much more likely to be in category 1 than category 0. They are also more likely to be in category 3 or 4 as compared to category zero though the effect for these two categories is not as strong. The second regression (table 6) identifies category 1 and 3 as job categories which pay more(about 35 percent more) than category zero. So our conjecture about English skills helping one get higher paying jobs is true. I should point out that even after controlling for jobs(coarsely through our categories), fluent English continues to have a significant impact on wages. This seems to suggest that while being good at English helps one get a job which pays better, this is not the only route through which English ability influences wages. This could be because having better English helps one get promoted faster within one's job or because the same job category could consist of multiple positions and English speakers are more likely to occupy the higher paying one. As a robustness check to make sure that it's not the coarseness of my constructed job categories which is

getting us this result, I run the wage regression using all possible job codes (99 in number) as part of the explanatory variables. Being able to speak fluent English still turns out to be significant.

Table 6: Impact of Job Category on Wages

Variable	**Coefficient** (Std. Err.)
Job Category 1	0.380** (0.031)
Job Category 3	0.366** (0.023)
Job Category 4	0.020 (0.025)
Speak Little English	0.034† (0.018)
Speak Fluent English	0.078** (0.029)
High Academic Ability	0.093* (0.046)
Medium Academic Ability	0.072† (0.042)
Low Academic Ability	-0.027 (0.043)

Job Categ 2 suppressed
Job Categ 1 is of Teachers, Scietists , doctors etc
Job Categ 3 is of Clerks, village Officials, Supervisors etc
Job Categ 4 is of Merchants, Salesmen, Shopkeepers etc
Other controls include schooling year dummys, location dummys
Significance levels : † : 10% * : 5% ** : 1%

4.5 Complementarity between English ability and Academic ability

Previous regressions confirm that both these abilities significantly affect wages received. Since employers seem to value both these virtues, surely they would be willing to pay even more for people who have both these abilities instead of just one. Consider the base regression (model2, table 3) where instead of academic ability and English ability variables separately we have an interaction term with all possible combinations of the two. The academic ability variable takes four distinct values and the English ability variable takes three distinct values so I have 11 dummys in my regression. For example, 'Fluent English & High Academic Ability' is for a person who speaks fluent English and got first division in the tenth grade examinations. The base here is someone who does not speak English *and* did not reach tenth grade(lowest academic ability).

Table 7: Complementarity, (Dep Variable - Log(Wages4))

Variable	**Coefficients** (Std. Err.)
No English & High Academic Ability	0.061 (0.060)
No English & Medium Academic Ability	0.088* (0.045)
No English & Low Academic Ability	-0.048 (0.048)
Little English & Lowest Academic Ability	0.027 (0.025)
Little English & High Academic Ability	0.155** (0.052)
Little English & Medium Academic Ability	0.136** (0.043)
Little English & Low Academic Ability	0.025 (0.048)
Fluent English & Lowest Academic Ability	0.059 (0.088)
Fluent English & High Academic Ability	0.327** (0.053)
Fluent English & Medium Academic Ability	0.154** (0.050)
Fluent English & Low Academic Ability	0.168* (0.082)
Schooling Years = 10	0.152** (0.039)
Schooling Years = 12 (High School)	0.169** (0.045)
Schooling years = 15 (College)	0.253** (0.047)

Other controls include location dummys, schooling year dummys, social status and networks

Significance levels : † : 10% * : 5% ** : 1%

This regression (table 7) clearly shows that having both good English speaking skills and high ability is a big advantage. Controlling for everything else, a person who speaks fluent English and

obtained 1st division in the tenth grade exams earns about 32 percent more than a person who does not speak English and did not reach tenth grade. The return for speaking just a little English and still getting 1st division or for speaking fluent English but getting second division is about half of this. So, not surprisingly, both skills are valuable and a person possessing both simultaneously is richly rewarded.

5 Quantile Regressions

In this section I look at the difference between returns to schooling and English ability. Quantile regressions on the entire population seems to suggest that the return to speaking fluent English is about 12 percent at all wage quantiles. However, when we run the quantile regression separately for urban and rural areas the following patterns emerge(table 8) for rural areas [12]:

(a) In rural areas, returns to schooling rises with each quantile. For example, keeping all other things constant, people who do really well (90th percentile) after completing college earn about 40 percent more than the people who do really well (90th percentile) without going to school. However for people at the 10th percentile the difference in wages is about 11 percent.

(b) Returns to fluent English speaking ability are highest for the middle quantiles in rural areas. The returns start at 11 percent for the 10th quantile, rises to 21 percent for the 70th quantile and then falls to 17 percent for the 90th quantile.

(c) While returns to schooling and academic ability are higher in urban areas, the return to English ability is larger in rural areas. This can be seen even without quantile regressions.

Consider (b) first. The middle quantiles of fluent speakers getting higher returns than the upper or lower quantiles seems strange especially in the light of point (a). This could be because having English ability makes a lot of medium skilled jobs available to people. These jobs may not require a high education or ability (like truck driving to the southern parts of India where you may need some English skills to communicate with people). So if you know English and maybe have high school education then you could get these jobs and earn better than the middle quantiles of non English speakers.

As for (c), why are the returns to English ability higher in rural areas? One explanation could be that the supply of English speakers is much higher in urban areas. There are 5 times as many fluent English speakers in urban areas than in rural areas. Moreover, acquiring English skills is much easier in the urban areas. So if the returns were very high in urban areas then most people would have made this investment.

The returns to schooling and academic ability are higher in urban areas since it is very difficult to get jobs without sufficient education in urban areas while in rural areas there are traditional jobs like farming which are just taught by one generation to another and thus can be done without much

[12] Urban in Appendix 2

Table 8: Rural Quantile Regression

Quantile	Speak Little English	Speak Fluent English	High Academic Ability	Complete College
10th Quantile	0.029	0.11*	0.01	0.11**
30th Quantile	0.016	0.11**	-0.02	0.27**
50th Quantile	0.026*	0.17***	0.02	0.31**
70th Quantile	0.015	0.21***	0.04	0.35**
80th Quantile	0.014	0.18***	0.08	0.36**
90th Quantile	0.033	0.17***	0.11*	0.40**

*** p<0.01, ** p<0.05, * p<0.1

Other controls are same as Model 1,table 2

schooling. In addition, once you have good education you can get *really* high paying jobs in urban areas.

Finally for Point (a), The increasing returns to completing college could signal an increasing quality of college. Sadly, there is huge variance in the quality of college education available in India. The people who do really well went to the best colleges and are therefore doing far better than the people who did well without going to school. The people who did just okay after completing college probably didn't go to the best colleges. So they get lower returns to their degree.

6 Conclusions

Being able to speak English is an advantage in India. This could be due to many reasons. It could be a legacy of the British Raj or it could be a consequence of the rapid globalization of the world in the last 50 years. Most likely, it is a result of both these factors and many more. This paper establishes and quantifies a strong relationship between English speaking ability and individual wages. This lays down the foundations for structural work and policy analysis to be done in the future.

(ACP, 2013) attempts to answer similar questions. They find that the returns to English speaking ability could be as high as 33 percent on hourly wages. In this paper I show that the impact on wages is probably more in the range of 10-16 percent. The difference in outcome is a result of omitted variable bias in the former. I have used many more controls in my regressions. I also construct and use a different wage variable to capture the earnings of individuals in rural India a little better.

References

Angrist, J. D. and Lavy, V. (1997). The effect of a change in language of instruction on the returns to schooling in morocco. *Journal of Labor Economics*, pages 48–76.

Azam, M., Chin, A., and Prakash, N. (2013). The returns to english-language skills in india. *Economic Development and Cultural Change*, 61(2):335–367.

Kapur, S. and Chakraborty, T. (2008). English language premium: evidence from a policy experiment in india. *Washington University in St. Louis unpublished paper*.

Levinsohn, J. (2007). Globalization and the returns to speaking english in south africa. In *Globalization and Poverty*, pages 629–646. University of Chicago Press.

Munshi, K. and Rosenzweig, M. (2006). Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy. *The American Economic Review*, pages 1225–1252.

A Data

A.1 Variables influencing Household Income

English - a)*totalengspeakers* (label used in paper - No. of Fluent English speakers)

This is the number of fluent English speakers in the household.

b)*totalengspeakerslittle*(Label used in paper - No. of Little English speakers)

This is the number of people in the household who speak only a little English.

Education - To capture this I use three variables -

a)*HHED5M* -

This represents the number of years of schooling of the highest educated male member of the household. I use discrete effects for each school year since the impact of completing high-school(at grade 12) and that of completing college(schooling level 15) is significantly higher than other school years. Completing tenth grade would also lead to a jump in wages since that grade is of some significance owing to students taking a national level examination in that grade.

b) *HHED5F* -

This represents the number of years of schooling of the highest educated female in the household. This variable could be important in several ways - If the female is working then she could be contributing to income directly, If she is not working but is the mother of a working member - then her having high education could lead to better quality education of her children leading to better wages and even better English ability possibly. Also the mother being better educated may lead to better nutrition for the children growing up which naturally influences future wages. If the highest educated female is not the mother of an earning member of the family but is the wife of one - then higher education of the wife could be correlated to higher ability of the husband (here I am assuming women marry men with higher wage potential!). Similar stories could be thought of for sisters, daughters etc. The reason I stress more on the education levels of the female in the household is that I believe this affects household income in many more ways than the education level of the highest educated male. Again discrete effects of each education level are used in the regression.

c)*avgintegereducationadults*(label used in paper - Avg schooling years adults)-

Apart from the most educated male and female the household could have other earning members and their education levels are important in determining aggregate household income. To account for this - I construct a variable representing average number of years of schooling of the adults in the household(age>21) and round it of to the next integer and use the discrete effects of this variable in the household income regression.

Ability - As mentioned before, there is a national level exam which students take in the tenth grade. I have a variable indicating whether the student received first division,second division or third division (with first being the best). Of course this variable is only available for people who did complete tenth grade! I will assume the division to be 4 for people who did not reach tenth grade. The implicit assumption being that the people who did not reach tenth grade are either of low ability or have ability but cant utilize it owing to their low education and therefore their impact on household income will be similar to that of one with low ability. I construct a variable which represents the average tenth grade division among the adults in the household(*avgtenthgradediv*, label used in paper - Avg tenth grade division). Obviously a lower average represents higher ability here. Also - as discussed before - *HHED5F* acts as an additional control on ability of household members

Experience - It would be ideal to have the job experience of the household members. Short of it - I construct an experience variable which is essentially (age-16) under the assumption that people start working as early as 16y (especially true in the rural areas). I then construct *avgexperienceadults* which, as the name suggests is the average experience of the adults in the household. Following Mincer and several other papers which have established this- I use both average experience and average experience square(*avgexperienceadultssq*) in the regression.

Place of Residence - I use three variables to account for this effect.

a)*URBAN*
This variable is a dummy which tells me if the area the household is in is classified as urban(takes value 1) or rural area according to the 2001 census of India. Clearly - being in an urban area would give access to many more jobs and probably more lucrative job opportunities. However since it's also likely to be populated with many English speakers the returns to English ability in urban areas may or may not be significant.

b)*METRO*
I also use discrete dummys for 6 of the largest cities in India with this variable.

c)*STATEID*
Both English abilities and job opportunities are affected by which State of India the household is in. Southern parts of India have a much higher density of English speakers and some of these states are much more prosperous too. This leads to better jobs. I use dummys for 33 states (including Union Territories) in the regression.

Social Status and Networks - I use three variables to account for the effect of the historical social status and the current social networks of the household on household income.

a) *GROUPS8*

The data set reports this variable which categorizes a person into one of 8 categories - Brahmans(high cast and traditionally well read and at least moderately wealthy), Other High

Castes, Scheduled Castes (SC)(traditionally badly treated in the society and poor), Scheduled tribes(Similar too SC), Muslims, Jains and Sikhs(traditionally well off),Christians and OBC(other backward castes). Though we would like to believe that the cast system etc are things of the past in India and that the religion or caste of a person has no effect on the opportunities he gets and the choices he makes, unfortunately in many parts of India(especially rural!) these continue to play a significant role and my conjecture is that they affect the wages strongly. Through education levels - they could also influence English ability.

b) *govtrelationinfamily*

I construct this variable and it takes 4 distinct values. 0 - If the individual does not know any government employees or officials, 1- if the individual knows but that person is not in the family, 2- if that person is in the family but not in the household, 3- if the known government worker is in the household, 4 - if that is the individual himself. Note that here the government employee could be a military personnel. This variable is of importance in India. Government employees are powerful people and they can easily help their family members find jobs. Also since they are usually better off and better educated than the average person it is quite possible that having a govt employee in your family affects the English ability and education quality of the children.

c)*medrelationinfamily*

As above I construct this variable representing medical networks of the family and this also has similar four distinct categories. I suspect these networks are more useful in rural areas where information about disease and nutrition are limited and having a medical personnel in the family may significantly improve the health and well being of the family members (especially girls who may otherwise face discrimination and receive far worse medical attention than her male siblings

Information - I capture the role of information in influencing household income in two ways

a)*menreadnews* -
A variable indicating how often do the men in the household read newspapers. This might be important since reading newspapers could make people more aware of new opportunities and technology. It takes three different values - 0 - Never, 1-Sometimes, 2- Regularly

b) *womenreadnews* -
A variable indicating whether women in the household read newspapers regularly. Apart from the direct effect as above this variable also indicates the progressive(or non-progressive as the case may be) attitude of the household which may be directly correlated to income. It takes three different values - 0 - Never, 1-Sometimes, 2- Regularly

The number of members - I use two variables to capture the age distribution of the household.

a)*NPERSONS* - Indicating total number of people in the household.

b)$NADULTS$ - Indicating number of people above the age of 21 in the household.These two variable might be important in determining household income since they describe the number of possible wage earners and the number of possible dependents. Also, since they might give us some indication of the resource constraint faced by the household when the children were growing up - number of people in the family may be inversely related to English speaking ability.

Complete Model used for Discrete Effects model in Household level regression

$$\text{logincome} = \beta_0 + \sum_1^8 \beta_i * I(totalengspeakers = i) + \sum_9^{19} \beta_i * I(totalengspeakerslittle = i - 8) +$$

$$\sum_{20}^{34} \beta_i * I(avgintegereducationadults = i - 19) + \beta_{35} * avgtenthgradediv + \beta_{36} * avgexperienceadults +$$

$$\beta_{37} * avgexperienceadultssq + \beta_{38} * NPERSONS + \beta_{39} * NADULTS +$$

$$\beta_{40} * URBAN + \sum_{41}^{55} \beta_i * I(HHED5M = i - 40)$$

$$+ \sum_{56}^{70} \beta_i * I(HHED5F = i - 55) + \sum_{71}^{74} \beta_i * I(govtrelationinfamily = i - 70)$$

$$+ \sum_{75}^{78} \beta_i * I(medrelationinfamily = i - 74) + \beta_{79} * menreadnews + \beta_{80} * womenreadnews$$

$$+ \sum_{81}^{86} \beta_i * I(metro = i - 80) + \sum_{87}^{93} \beta_i * I(GROUPS8 = i - 86) +$$

$$\sum_{94}^{126} \beta_i * I(STATEID = i - 93)$$

Variables affecting wages at the Individual Level

1. **English Ability**

ed7fluent,ed7little (label used in paper - Can Speak Fluent English, Can Speak Little English) - dummy variables indicating if a person reports himself to be able to speak fluent English or just a little English. The base here is being able to not speak English at all.

2. **Education**

ED5 - (Label used in paper - Schooling years) - with discrete effects for each school year since it's quite clear that the impact of completing High School(12 years of schooling), or College(15 years of schooling) or even tenth grade is much higher than others.

3. **Ability**

I use -

a)*ed9I,ed9II,ed9III*(labels used in paper - High Academic Ability, Medium Academic Ability, Low academic Ability)

Each being a dummy indicating what if the person received first division,second division or third division in the tenth grade examination respectively. Note that if the person did not reach tenth grade - all these variables will take value zero. Thus the interpretation of the coefficients on these variables should be - Compared to someone who did not reach tenth grade(lowest Academic Ability), how much more does reaching tenth grade and getting one of these divisions get you in terms of wages.

b) *HHED5F* - As Described before in the household level regression.

4. *Experience* and *experiencesq*

5. **Place of Residence** - All three variables used before.

6. **Social Status and Networks** - All three variables used in the household regression.

7. **Information** - Both variables used before in the Household regressions.

8. *NPERSONS*- As used before.

9. **How Wealthy the Household Is**:

This is directly related to the wages of the individual. However since this variable is strongly correlated to your possible jobs and your English ability I feel this should be included in some way. I don't want to use Household income since the correlation with individual wages may be very strong. Instead I do the following - the data set has a list of 24 household assets that households could have. I use dummys for some of these (or just the number of these assets a household has) to account for this effect. These assets range from - Cycles to Television to Cell Phones to computers to type of wall of the house they live in. For a lot of these household assets, one can assume that these may have been present before the individual got his current job and are therefore less correlated with their present wages. This would be especially true for young individuals. In the case of older people it would be a much stronger assumption to make though! The variables are called *CG(i)* and *HQ4dummy* for different consumer goods and types of House walls respectively. The base for house walls is Grass,thatch and this is compared with Mud/unburnt Bricks, plastic,wood,burnt bricks,metal sheets, stone,cement, concrete.

10. **Income from outside**:

If the household gets remittances from abroad or gets pensions then the individual may be working part time just to supplement the income. So the wages for these individuals may be low due to this reason. Therefore,in the wage regression, I use the variable *contincremitbyincome* which describes the percentage of household income which is sent from members working elsewhere. I also use *contincotherbyincome* which describes the percentage of household income which the household

receives without working actively like pensions, rent from properties etc. I expect these variables to be negatively correlated to own wages.

11. **Marital Status**:

Studies have shown positive correlation between getting married and wages or having children and wages. The link to English ability could be weak though.

Base Regression for Individual Level regression

$log(WO) = \beta_0 + \beta_1 * ed7fluent + \beta_2 * ed7little +$

$\sum_3^{17} \beta_i * I(ED5 = i - 2) + \beta_{19} * ed9I + \beta_{20} * ed9II +$

$\beta_{21} * ed9III + \sum_{22}^{36} \beta_i * I(HHED5F = i - 21)$

$+\beta_{37} * experience + \beta_{38} * experiencesq$

$+\beta_{39} * URBAN + \sum_{40}^{54} \beta_i * I(metro = i - 39) + \sum_{55}^{87} \beta_i * I(STATEID = i - 54)$

$+ \sum_{88}^{94} \beta_i * I(GROUPS8 = i - 87) + \sum_{95}^{97} \beta_i * I(medrelationinfamily = i - 94) +$

$+ \sum_{98}^{101} \beta_i * I(govtrelationinfamily = i - 97) + \sum_{102}^{103} \beta_i * I(menreadnews = i - 101)$

$+ \sum_{104}^{105} \beta_i * I(womenreadnews = i - 103) + \beta_{106} * NPERSONS + \sum_{107}^{131} \beta_i * I(CG(i)$

$+ \sum_{133}^{140} \beta_i * I(HQ4dummy = i - 132) + \beta_{141} * contincremitbyincome + \beta_{142} * contincotherbyincome$

$+\beta_{143} * maritalstatus$

A.2 Other Variables Used in the Paper

1. *WKFARM* - Dummy Variable which takes the value 1 if the person has worked on the family farm for more than 240 hours last year.

2. *WKBUSINESS* - Dummy Variable which takes the value 1 if the person has worked on the family business for more than 240 hours last year.

3. *Job Categ* - Takes the value zero (base case) if the individual's occupation is one of farming,fishing,hunting,logging,carpentering etc. The value for this variable is one if the individual is a teacher,scientist,doctor etc. Variable is two if the individual is a senior govt officer or a senior member of a firm. The variable takes the value three if the occupation is clerks,village officials, supervisor etc. Finally the variable takes the value 4 if the individual is a shopkeeper,salesman,merchant etc.

4. *Work or Not* - Dummy Variable which takes the value 1 if the individual chooses to join the

work force and earns at least Rs 5000 per year.

5. *NCHILDREN* - Number of children in the household.

6. *NADULTS* - Number of adults in the household.

7. *Early Marriage* - Dummy variable which takes the value 1 if the person got married before the age of 21.

8. *speakathome* - Dummy variable which is 1 if both the highest educated female and male in the household have completed school.

9. *englishstates* - Dummy variable which is 1 if it's a state of India which has a high number of English speakers historically.

B Additional Results

Full Results for Table 2

Table 9: Discrete Model Estimation

Variable	**Coefficient** (Std. Err.)
No. of Fluent English Speakers=1	0.128** (0.027)
No. of Fluent English Speakers=2	0.223** (0.036)
No. of Fluent English Speakers=3	0.012 (0.073)
No. of Fluent English Speakers=4	0.169† (0.099)
No. of Fluent English Speakers=5	0.014 (0.250)
No. of Fluent English Speakers=6	0.048 (0.223)
No. of Fluent English Speakers=7	0.469** (0.168)
No. of Little English Speakers=1	-0.016 (0.018)

Continued on next page...

... table 9 continued

Variable	Coefficient (Std. Err.)
No. of Little English Speakers=2	0.017 (0.024)
No. of Little English Speakers=3	-0.007 (0.060)
No. of Little English Speakers=4	-0.068 (0.072)
No. of Little English Speakers=5	-0.089 (0.083)
No. of Little English Speakers=6	-0.006 (0.113)
No. of Little English Speakers=7	0.393 (0.306)
No. of Little English Speakers=8	-0.750† (0.390)
No. of Little English Speakers=9	-1.078** (0.086)
No. of Little English Speakers=11	1.471** (0.088)
Avg Tenth Grade Division	-0.114** (0.015)
Intercept	9.798** (0.062)

Urban Quantile Regression Results

Table 10: Quantile Regression - Urban

Variable	Coefficient (Std. Err.)
Equation 1 : q15	
Speaks Little English	-0.005 (0.027)

Continued on next page...

... table 10 continued

Variable	**Coefficient** (Std. Err.)
Speaks Fluent English	0.048 (0.039)
High Academic ability	0.196** (0.062)
Schooling Years=10	0.077* (0.033)
Schooling Years=12 (High School)	0.057 (0.085)
Schooling Years=15 (College)	0.263** (0.072)
Equation 2 : q30	
Speaks Little English	-0.011 (0.018)
Speaks Fluent English	0.061† (0.036)
High Academic ability	0.127** (0.047)
Schooling Years=10	0.108** (0.028)
Schooling Years=12 (High School)	0.200** (0.068)
Schooling Years=15 (College)	0.412** (0.053)
Equation 3 : q50	
Speaks Little English	0.027* (0.013)
Speaks Fluent English	0.092** (0.028)
High Academic ability	0.161** (0.037)
Schooling Years=10	0.135** (0.027)
Schooling Years=12 (High School)	0.205** (0.073)

Continued on next page...

... table 10 continued

Variable	**Coefficient** (Std. Err.)
Schooling Years=15 (College)	0.341** (0.034)
Equation 4 : q70	
Speaks Little English	0.031† (0.019)
Speaks Fluent English	0.120** (0.035)
High Academic ability	0.149** (0.043)
Schooling Years=10	0.136** (0.031)
Schooling Years=12 (High School)	0.260** (0.064)
ed16	0.335** (0.035)
Equation 5 : q80	
Speaks Little English	0.024 (0.023)
Speaks Fluent English	0.096** (0.037)
High Academic ability	0.126** (0.048)
Schooling Years=10	0.118** (0.033)
Schooling Years=12 (High School)	0.258** (0.071)
Schooling Years=15 (College)	0.357** (0.041)
Equation 6 : q90	
Speaks Little English	0.021 (0.030)
Speaks Fluent English	0.079† (0.041)

Continued on next page...

... table 10 continued

Variable	**Coefficient** (Std. Err.)
High Academic ability	0.114* (0.046)
Schooling Years=10	0.090* (0.042)
Schooling Years=12 (High School)	0.224** (0.084)
Schooling Years=15 (College)	0.342** (0.069)