The Economic Returns to English Proficiency in India: Evidence from 2005 and 2012 Panel Data

Murari Ganesan

December 2024

Abstract

This study examines the economic returns to English proficiency in India, focusing on its potential role in shaping the country's development trajectory relative to other nations like China. Utilizing panel data from 2005 and 2012, a fixed effects model is employed to estimate the impact of possessing at least some English proficiency on individual income. The analysis reveals that basic English proficiency is associated with an income increase of approximately 600 INR, controlling for a wide range of variables. While the comprehensive use of controls supports a causal interpretation, the lack of robustness checks and the age of the data temper the findings' conclusiveness. The results, drawn from older data, may underestimate the current returns to English proficiency, given India's significant economic growth since 2012. Future research should revisit this question using more recent and granular data, incorporating additional controls such as industry-specific factors. Investigating the returns for college graduates relative to the general population is another promising avenue for deeper insight.

1 Introduction

India has surpased China as the most populus country. And like China some 20 - 30 years ago, India is about to expirience the gift of a demographic dividend. As defined by the United Nations Population Fund, a demographic dividend is "the economic growth potential that can result from shifts in a population's age structure, mainly when the share of the working-age population (15 to 64) is larger than the non-working-age share of the population (14 and younger, and 65 and older)" In other words, it is a boost to productivity for a country when due to having a large ammount of working age people relative to the dependent population leads to a country able to rapidly increase its productivity by not having to expend a lot of relative recourses on the dependents thanks to the rapid increase in those working boosting productivity.

The question this paper attempts to understand is how english literacy in India may augment the country's ability to develop during this time period where it can harness the demographic dividend.

As a result of British colonization of India, unlike other countries that developed rapidly to surpas the middle income trap, India already has a higher english fluency relative to those countries. The core hypothesis is that due to the dominance of the service sector in developed countries, India may be able avoid at least some of the pain that has come with rapid development, including ponentialy the middle income trap. This advantange has already shown its head in India, through things like the growth of India's IT sector and Indian call centers. Essentially, english fluency may enable India to avoid the fate of China, where it is stuck things like the middle income trap. The core question at hand is to explore how English fluency might aid India in achieving and sustaining economic growth.

English fluency may contribute to economic growth in multiple ways. First, it opens up opportunities in the global services sector, particularly in information technology (IT), finance, consulting, and other high-value, knowledge-intensive industries that are driven by international markets and English-language standards. For instance, India's already thriving IT and outsourcing industries are in part a product of English fluency, allowing companies to serve global clients and participate in a supply chain that requires proficiency in English-based software, contracts, and

technical documentation.

Furthermore, as countries reach middle-income status, they often face a "middle-income trap," where growth slows due to rising wages and a need to transition from labor-intensive manufacturing to higher value-added industries. China has encountered challenges in this area, particularly in shifting its economy toward innovation and high-value services. India, however, might leverage its English proficiency to better navigate this transition by moving more effectively into service-based industries, allowing it to avoid stagnation in middle-income status.

Additionally, English proficiency could bolster foreign direct investment (FDI) in India by lowering communication barriers with multinational corporations, making it easier for foreign firms to establish operations, hire local talent, and integrate Indian workers into global teams. This advantage may also enhance India's appeal as a regional hub for global businesses seeking a foothold in Asia, particularly in sectors such as business process outsourcing (BPO), customer service, and technology.

Moreover, English proficiency can empower India's workforce by enhancing access to global knowledge resources, educational materials, and technical expertise, which are predominantly available in English. This access not only enriches human capital but also facilitates the transfer of advanced skills and technologies into the Indian economy, fostering innovation and productivity.

The question of English fluency's impact on economic growth is thus essential to understanding whether India can leverage this linguistic asset to boost its competitiveness, diversify its economy, and sustain long-term growth. By exploring the link between English proficiency and income at the individual level, this study aims to shed light on how English fluency could drive broader economic outcomes in India.

2 Litrature Review

What I am exploring is extremely similar to Azam et al., 2013. This paper looks at returns to english-language skills in India. Their focus was on how english skills affect an induvidial's earn-

ings. This meant the limited the sample to those aged 18 - 65 who reported earning a wage, leading to a concern of selection bias due to the prevalence of family farms and businesses, so They also look at english skills with respect to household income and consumption which wouldnt have that selection bias. They find that english skills raise earnings, but find that age (more clearly, work expirience) is associated with even higher earnings than the increase in earnings for english speakers in India. More clearly, Indian firms seem to prefer more expirience over just english skills. They utilize the IDHS survey data to run a regression where they control for location fixed affects. They utilize the following regression equation:

$$y_i = \alpha_r + \beta English_i + \delta Schooling_i + \gamma Ability_i + \pi X_i + e_i$$

where $Urban_i$ is a dummy for an induvidial living in an urban part of India, α_i is location fixed affects, $Schooling_i$ is the years of education completed, and $Ability_i$ are proxies for ability like father's education, SSLC exam performance, and failing or repeating a grade. They only utilize data from the 2005 version of the IDHS survey.

This paper is the closest to what I am attempting to do, yet the differences are quite clear. the key difference in what I am attempting relative to what they have done is that I intend to utilize a lot more controls, to help further reduce the concerns of omitted variable bias amd I also have acess to more data from the same induvidials in 2012.

The paper Munshi and Rosenzweig, 2006 was a paper cited by Azam et al., 2013. This paper intended to explore how traditional institutions interact with globablization, specifically, how the Hindu Caste system interacts with the increased career choices, using Mumbai school enrolment survey data. They find that men tend to remain limited by the caste system in their employment opitons and school choice, wheras women seem to be embracing the increased oppourtunites offered by globablization. For my own work, this seems to indicate value in exploring the gender dynamic of english literacy.

A relevant paper was Grin, 2001. This paper explores the economic value of english. This

paper uses a simple OLS regression in the Swiss context to estimate the value of English skills in the Swiss labor market (by regressing earnings on self reported english skills). This methedology is a bit simple for a modern paper (it was writen 23 years ago), and so a key value of my paper would be in determining a more modern strategy for English skill

Chaudhary and Fenske, 2020 is another paper, but is considerbaly more recent. They attemtpt to see how the development of the railroad in the British Raj affected english literacy. They use two identification strategies. They synthetic panel variation from cohort-specific literacy rates due to differences in the timing of railroad exposure, and they use distance from an early railway plan as an instrument for district railway exposure. They use the decennial censuses of 1881 to 1921, which they use to measure literacy I have looked into this data source, but it seems to be in old reports, rather than an easily accessable data format, meaning, give the time constraints of my paper, I don't think I'll be able to utilize the data.

If we start to explore the most comparable context, that being China, Guo and Sun, 2014 is a paper that comes up. This paper explores the economic returns to English proficiency for college graduates in China. They utilize data from the China Data Center of Tsinghua University to get data from a survey that randomly selected 19 colleges from 2305 colleges in Mainland China they use CET-4 as their metric for english proficiency. They utilize a an OLS regression to get to their estimates. They find that for starting salaries, for every 10 point increase in CET-4 score, there is an associated 2.4% increase in a graduate's starting salary. With all of their utilized controls, the estimate of their model droped to a 0.6% increase in starting salary for every 10 point increase in CET-4 score. While they didn't explore through a regression model how the industry a graduate works in affects their starting income, they were able to see that "CET-4 scores statistically significantly improved college graduates' probability of working in foreign companies, where the average starting salary was the highest."

For this paper, Guo and Sun, 2014 allows us to see a scenario where english proficiency does help, but in a context with more develpment than India, but still not enough development to be comparable to returns to english proficiency in a developed reigon.

3 Improvements

In a sense, this paper attempts to explore a similar question to what Azam et al., 2013 explored. But the key difference exists with the scope and methedology they have used. Azam et al., 2013 only utilized the results of the first IDHS survey carried out in 2005. This paper has access to both IDHS surveys, for the 2005 version and the 2012 version, across the same induvidials. This allows us to control for timed fixed affects, as well as the more variables than Azam et al., 2013 had chosen to control for. This will allow this paper to come up with a more accurate estimate for the affect of english proficiency.

India also presents a much more unique scenario to explore, when considering the present literature other than Azam et al., 2013. When you consider Guo and Sun, 2014, you see an example of a more developed context, and also, a much more limited context, with them primarily exploring the question in terms of college graduates.

This question seems like a simple, yet unexplored area of the Litrature The closest papers to what I am attempting, Azam et al., 2013 but they only explore the affect it has on induvidial earnings, and don't attempt to look at it from a macroeconomic point of view. If I am successful, then we would be able to see how english literacy could act as an advantage for India in it's long term development. The other similar paper, Chaudhary and Fenske, 2020 is also limited in it's exloration at the induvidial level, and in how it explores prior to India's independence.

4 Reserach Design

We seek to understand how English proficiency affects an induvidal's income within the context of India. to do this, it is natural to use use a Mincer earnings model to estimate earnings under the follwing model:

$$ln\ w = ln\ w_0 + \rho s + \beta_1 x + \beta_2 x^2$$

where w_0 is your inital wages, s is your years of schooling, and x is an individual's years on the labor market. The problem with this model in our context is that for our individuals, we do not have a variable for years worked. Furthermore, in the context of India, this variable may not always serve as a strong explanatory factor for one's income. Due to the diversity of India's economy, including its substantial agricultural and informal service sectors, the relationship between years of labor market experience and income is less consistent than in other contexts. For those working in agriculture, incomes tend to be highly influenced by external factors such as weather patterns, crop yields, and market prices for specific agricultural products. These external dependencies weaken the direct link between years worked and income.

Similarly, in the service sector, particularly the informal and gig economy, income levels often depend more on the nature of tasks performed, clientele, and local economic conditions than on cumulative work experience. For these reasons, while years of labor market experience might capture productivity gains through experience in some contexts, it is less relevant in India's case, necessitating alternative modeling approaches. In this paper, we aim to adjust the standard Mincer framework to better suit the unique economic structure and data constraints of the Indian context.

We will still utilize an induvidial's age as a control, so by proxy, we will still be able to control for years worked, if you were to assume that most induviduals start employment in the age range of 13 to 22. But this paper will not make this assumption due to the uncertainty arround one's starting age.

The idea is to explore income with repsect to English proficiency. To investigate this relationship, this paper utilizes a fixed effects regression model. This approach allows us to estimate the effect of English fluency on income while controlling for unobserved, time-invariant factors that may affect individual's earnings. By focusing on within-individual variations over time, we aim to isolate the impact of English proficiency on income, addressing any potential omitted variable bias associated with individual-specific characteristics that do not change over time. We utilize this inital regression model:

$$Inc_{it} = \alpha_i + \beta Eng_{it} + C_{it} + \epsilon_{it}$$

In this equation, Inc_{it} denotes the esitmated income of an induvidial at time t. α_i is the time invariant fixed affect for induvidial i. This fixed effect captures unobserved characteristics unique to each individual that may influence their income, such as innate abilities and family background. C_{it} contains all relevant time-varying control variables that may influence income, such as age, work experience, education level, geographic location, and other socioeconomic factors that change over time. These controls will help ensure that our estimates of the effect of English proficiency on income are robust and account for other influential factors.

The key asumptions of a fixed affects model is that the error term $\epsilon_i t$ has a conditional mean of 0, large outliers are unlikley, and that there is no perfect multicollinearity. These asumptions can be verified by robustness checks.

Even with the isolation offered by controling for the time variant factors, we still need to worry about the influence of time variant controls. This is where this paper is able to improve on the work of Azam et al., 2013.

With our data we are able to add considerbaly more controls than Azam et al., 2013 utilized, enabling us to get a more accurate estimate.

5 Data Sources

My primary data source will be the India Human Development Survey (IHDS) which refers to two surverys carried out in 2005 and in 2011-2012 by the University of Maryland. This survey has hundreds of variables for 200,000 people across the two listed time periods.

The primary varaibles I intend to use are English literacy, Completed years of education, Location (urban or rural), income, martial status, Computer ownership, sex, farm ownership, business ownership, age, and dificulty walking 1 kilometer. The key point of interest is English literacy on income, but we would use the rest of the variables as controls.

Completed years of education can vary over time and effect english literacy as I would expect that the more years of education completed would increase the chance that an induvidial is fluent in english due to learing enlgish in school. Location could affect enlgish literacy as I would expect that english would be used in cities in India as an intermediary language, making it a valuble skill in cities that could incentivise or simply make it easier to learn the language. Marital status could be corelated and vary over time due to the different reasons why one may or may not be maried. Age is a natural control for this study as one would expect that as someone ages, they develop more skills including english literacy and increase their income. The reason I include ability to walk 1 kilometer as a control, as it is the best measure of total health included in the survey that does not contain missing values. I would expect health to be corelated with both english literacy and income as those with a high volume of health conditions would likley struggle to find high paying work, and also have less need to learn english. The reason to include computer ownership is because it is a varaible that would vary over time, and it also could impact income through potential sumplementary income from internet activities.

There is an industry variable to denote the industry and induvidial works in, but this data was only collected in the 2011 - 2012 survey. Controling for the industry would be quite helpful, but due to this limitation, it would require assuming that the induviduals do not change industry over time. If this assumption is made, it would anyways be controled for in the fixed affects, making it no longer a usefull control.

I have two variables related to businesses ownership. One is a dummy for ownership of any non-farm businesses, and one for ownership of farm businesses. The issue is that it isnt likely a very time variant varaible, as if you own a businesses right now, you will probably still own a business seven years later. But since it isnt necicarily the case, as someone could have started a business that augmented their income and required them to become more fluent in english, it still makes sense to include it as a control.

6 Potential Problems

The main potential problems comes from the nature of fixed affects regressions. The ones I see are Limited Within-Individual Variation in Key Variables, Endogeneity and Reverse Causality, Time-Varying Omitted Variable Bias, Measurement Error in Key Variables, Potential Multicollinearity among Control Variables, and Lack of Generalizability and External Validity.

It is possible that the variation across induvidials less than what is needed for a fixed affects regression. The fixed effects regression model relies on within-individual variation over time to identify effects. However, English literacy may not change significantly for most individuals, especially if they have completed their education before the survey period. If English literacy is mostly constant, it will be difficult to estimate its impact on income using a fixed effects model. If this ends up being the case, the best option I can see is to split the survey sample into subpopulations and see if there exists variation across the subpopulations.

It is possible that English literacy might be endogenous to income. That is, higher income may allow individuals to invest in English education, resulting in reverse causality. If this ends up being the case, it will be hard to claim causality.

Even though I am controling for a lot of potential sources of ommited variable bias, its still possible that I do not account for unobserved factors that vary over time. For example, economic shocks, industry demand shifts, or local language policies could affect both income and English literacy.

As the reverse, it is possible that the asumptions behind the variables I do control for are inaccurate. For instance, the assumption that completed education years would directly increase English literacy could be inaccurate if English is only a minor part of the curriculum in some regions. To handle this concern, I could validate the assumptions using descriptive statistics or robustness checks. For industry changes, conduct sensitivity analyses by removing or including the industry variable and observing the stability of results.

Some control variables, such as education, urban location, and computer ownership, may be highly correlated with each other. Multicollinearity can inflate standard errors, making it harder

to identify significant relationships. I will have to calculate variance inflation factors to check for multicollinearity.

While I do belive using ability to walk 1 kilometer will be a a suitable proxy for physical health, it will not work for any mental health or coginitive troubles across induvidals. This means that these other health factors may lead to ommited variable bias.

With only two survey waves (2005 and 2011-2012), the study has limited time points to track income and English literacy changes. This makes it challenging to observe gradual transitions or short-term shocks, and the longer period between surveys may dilute the impact of time-varying variables.

This research may not be generalizable to contexts outside of India. British colonial influence in India may have been fundementally different than other colonized nations like Myanmar or South Africa.

Once the data is clean, I will be able to examine patterns in English literacy and other relevant variables across each state, aiming to identify any state groupings where parallel trends may hold. If there is no subset of states exhibiting parallel trends, I may need to reconsider the identification strategy, although determining an alternative approach could be challenging.

For the English literacy data, my initial inspection indicates that significant data cleaning will be necessary to make it suitable for analysis. After cleaning, I will evaluate the sample size and check for any remaining biases that could impact the analysis.

7 Preliminary Statistics

Here is a preliminary summary statistics table. The columns labeled (D) are itended to be dummy variables. This inital summary table only includes the columns of data that do not have any Nan values. With this table we can see the total size of the raw data contains 420,311 observations. The rows with a N less than 420,311 indicates that the row contains empty values. Though we need to divide this by two as this includes both time periods, meaning our dataset utilizes data from about

Table 1: summary table before cleaning

Statistic	N	Mean	St. Dev.	Min	Max
Sex..D.	420,311	1.496	0.500	1	2
Age	420,311	28.552	19.887	0	116
Martial.status	420,308	1.583	0.651	0	5
Nonfarm.Business.ownership..D.	420,311	0.232	0.422	0	1
Farm.Ownership..D.	420,311	0.469	0.499	0	1
Literate.Any.Language..D.	419,169	0.660	0.474	0	1
English.Ability	414,338	0.261	0.526	0	2
Attended.School..D.	418,772	0.688	0.463	0	1
School.Years.Completed	418,731	4.978	4.771	0	15
Uses.Computer..D.	420,311	0.059	0.236	0	1
Dificulty.walking.1.km..D.	420,311	0.032	0.215	0	2
INCOME	420,311	126,846.000	211,451.200	−1,037,040.000	12,032,845.0

210,155 unique induvidals. We can see the data has induvidials with a mean age of approximately 28 years of age, with a standard deviation of 19 years of age. This is a good sign in terms of the appliciability of this sample, as we know that the middle of the dataset is fairly young, but still has school done. Our main outcome variable, income is in Indian Rupees, which is why it has such high values. We see a mean of 126,846 INR, with a standard deviation of approximately 211,451.2 INR. We can also see that our minimum and maximum values are extreme. this extreme minimum and maximum relative to the mean, when taken with the standard deviation indicates that it would be wise to either filter down the data, or consider imputing the extreme values with the mean or median $\pm$ some decided noise.

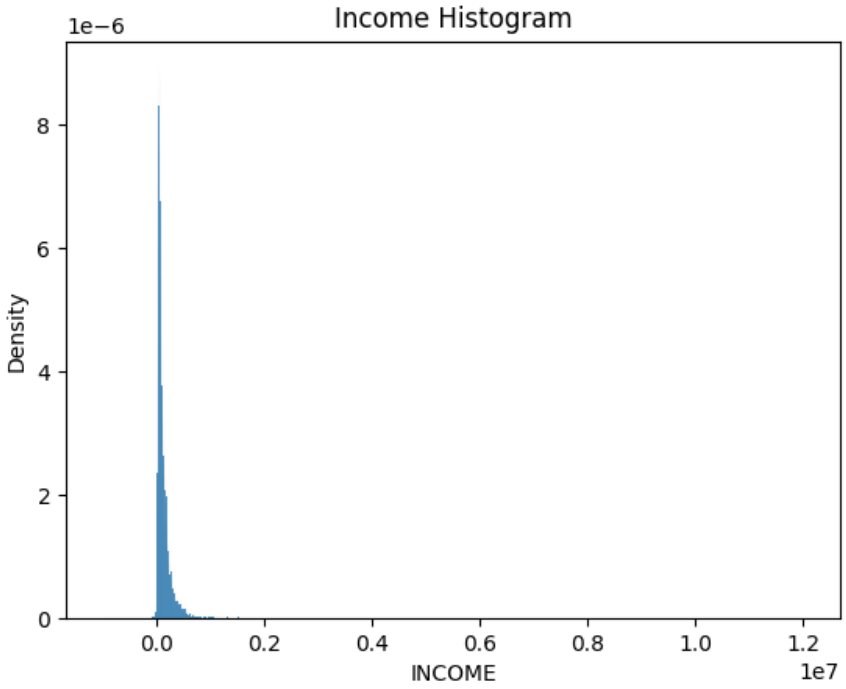

Figure 1: Income Histogram before fitering

Figure 1 shows the skew of the data. It shows that almost 80% of the data is close to 0 in terms of income. Also, due to the length of the x-axis, it seems like this data might have a lot outliers. If we Follow the standard statistical method of filtering out $\pm 1.5 * IQR$, you get a more clear understanding of the data. Simply removing this data is justified due to the size of the dataset. This filter is performed leading to a more refined dataset.

We see that there isn't much data lost after the fitering, with there still 381,167 observations to work with across the two surveys. Some odd things to note is the negative Income listed as the minimum value. You can also see in Figure 2, which had an updated histogram of income data after removing the outliers, and is grouped by the survey, that there is some data where they have negative incomes. This is odd, and not mentioned as holding some special meainng in the codebook. Therefore, I am concluding it is some sort of net income, or a negative income result for business owners.

Table 2: cleaned summary table

Statistic	N	Mean	St. Dev.	Min	Max
URBAN	381,167	0.310	0.463	0	1
Sex..D.	381,167	0.503	0.500	0	1
Age	381,167	28.283	19.833	0	116
Martial.status	381,167	1.588	0.652	0	5
Nonfarm.Business.ownership..D.	381,167	0.221	0.415	0	1
Farm.Ownership..D.	381,167	0.469	0.499	0	1
Literate.Any.Language..D.	381,156	0.643	0.479	0	1
Attended.School..D.	381,046	0.671	0.470	0	1
School.Years.Completed	381,008	4.702	4.627	0	15
Uses.Computer..D.	381,167	0.047	0.211	0	1
Dificulty.walking.1.km..D.	381,167	0.032	0.216	0	2
INCOME	381,167	87,987.970	68,890.960	−120,410.000	306,868.400
English.Ability.D	381,167	0.196	0.397	0	1

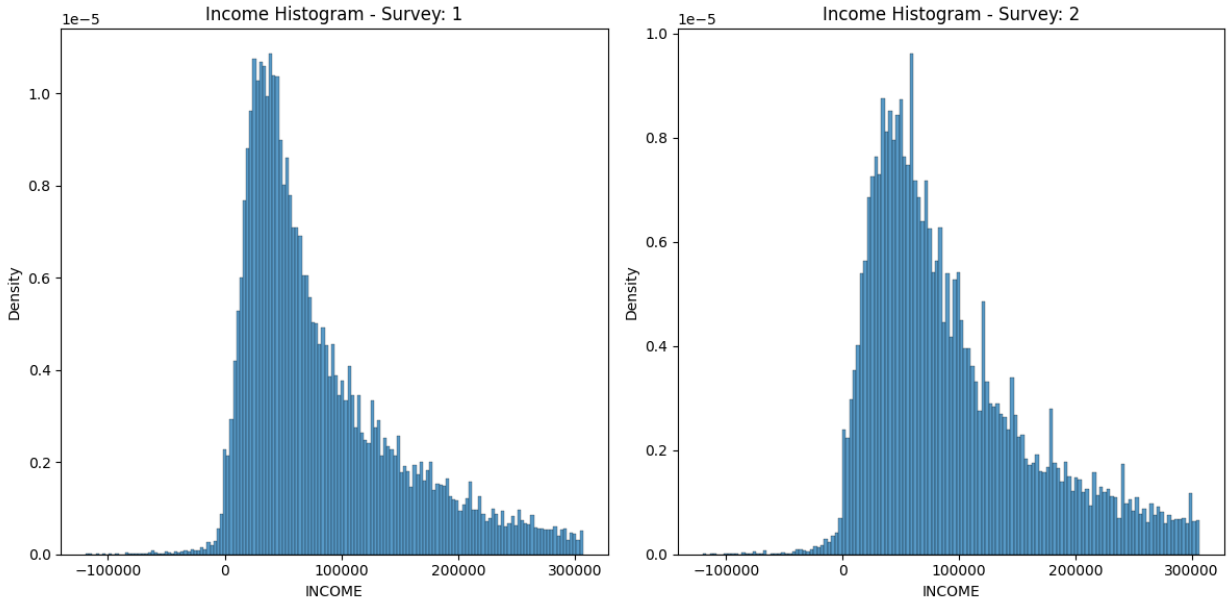

Figure 2: Income Counts by Survey

Figure 3 shows the size of the values for each category on english ability, where the category 0 is none, 1 is some, and 2 is fluent. We can see that over the 6 year gap between the surveys that english fluency increases. We can also see that the group with some english fluency is considerably

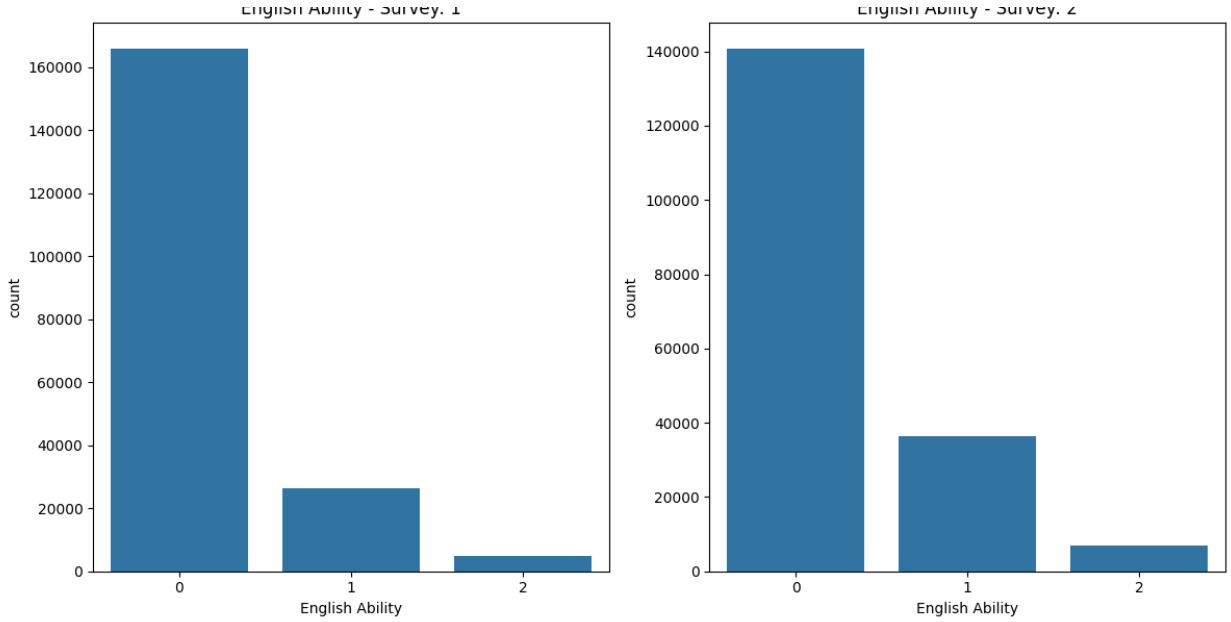

Figure 3: English Ability size Bar graph by survey; Higher is more English Ability

smaller than the group with no english fluency for both time periods.

Figure 4 shows us the relationship between wealth and income. As expected, those with higher English ability have higher averaege incomes. This graph could either indicate systemic bias in this sample of about 190,000 people, where those who do have higher english fluency just happen to have higher incomes, or it can tell us what is expected, that english ability is a factor in explaining one's income, but could be corelated with factors like the years of education that someone has had.

8 Results

8.1 Inital Model

We now explore some inital regresison results. The inital regression we explore follows the following econometric model:

$$Income_{it} = \beta_1 EnglishAbility\ 1_{it} + \beta_2 English\ Ability\ 2_{it} + \alpha_i + \epsilon_{it}$$

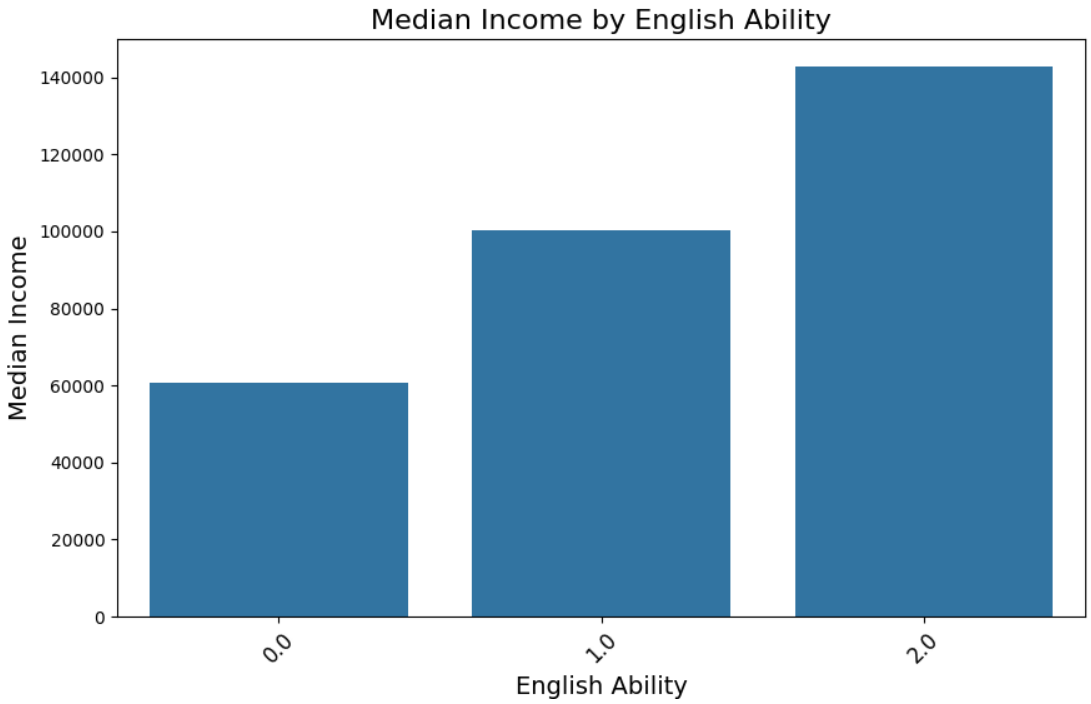

Figure 4: English Ability income Bar graph by survey; Higher is more English Ability

With α_i being induvidial fixed affects which are removed by running the fixed affects regression. This gives the results seen in table 3, row 1. This intial regression indicates that there is likely little relationship between the level of English ability an induvidial has and their income. This is due to there being no statistical significance for the estimated increase in income for those who are english fluency and only statistical significance at the 10% level for those with partial english fluency. Under this inital model, having partial english fluency is associated with a decrease in income of approximately 341 INR, and being fluent in english leads to a increase in an induvidial's income of aproximatley 218 INR.

These intial results are extremely unexpected. It would be expected that those who are able to at least speak some english would work in either an industry or region where there are more need for english, which tends to mean either the induvidial works in a higher paying industries or works near higher paying industries such that english fluency would be a more valuble skill for a worker to have.

One theory as to why there isnt statistical significance in this inital model is because of the sample size for each group of english ability being to small. If we refer back to Figure 3, we can see that relative to the group with no english ability, the group with any ammount of english ability is quite small, especially in the 1st time period. Even in the seccond time period, the group with no english ability is still at least 3 times larger than the entire group with at least some english ability. This leads us to consider whether simply grouping all of the induvidials with at least some english ability together to explore the relationship between any ammount of english fluency and an induvidial's income.

This leads us to the this updated model:

$$Income_{it} = \beta_1 EnglishAbility_{it} + \alpha_i + \epsilon_{it}$$

When you change the English Fluency to just a dummy, still running a fixed affects regression, you end up with the results seen in table 3, row 2. These results look much more promising, with statistical significance here prior to any controls. Here, we see that controlling only for time variant factors, over the 7 years between surveys that there was an averaege increase of 15,144.730 Rupees in income for in India given this sample. Having any ammount of english leads to an estimated increase in income of 1,892.77 rupees. Both of these are statistically significant at the 1% level. This result indicates that the explanation for the lack of statistical significance with the inital regression is likely what was happening.

8.2 Education Controls

We now attempt to seek causality by controling for any omitted variabes that could vary over time. We will split up the regressions for the different types of control variables we have in the data.

The first regerssion we explore are any education related to education by utilizing the following model:

Table 3: Income vs English ability within controling for time

	Dependent variable:	
	INCOME	
	(1)	(2)
English.Ability_1	-341.075^{*}	
	(186.693)	
English.Ability_2	218.085	
	(398.238)	
English.Ability.D		$1{,}892.771^{***}$
		(188.929)
factor(SURVEY)2	$15{,}132.590^{***}$	$15{,}144.730^{***}$
	(131.765)	(130.896)
Observations	367,232	376,974
R^2	0.039	0.040
Adjusted R^2	-0.079	-0.080
F Statistic	$4{,}430.731^{***}$ (df = 3; 326991)	$7{,}066.613^{***}$ (df = 2; 334924)
Note:		*p<0.1; **p<0.05; ***p<0.01

$$Income_{it} = \beta_1 EngAbil_i + \beta_2 LitAnyLang_i + \beta_3 SchoolYrs_i + \beta_4 Computer_i + \alpha_i + \epsilon_{it}$$

Specifically, for this regresion, we control for school years completed, a dummy for literacy in any language, and a dummy for computer usege. the reason the dummy for computer usage is included in this regression controling for educational factors is due to potential computer usage in schools / universities, and the induvidial utilizing said computer to augment their income. For this model, $EngAbil$ is the dummy for an induvidial having at least some english ability, $LitAnyLang$ is a dummy for an induvidial being literate in any language, $SchoolYrs$ is a descrete continous variable for the ammount of year of school an induvidial has completed, and $Computer$ is a dummy variable for utilizing a computer.

In this model, we find that literacy in any language is associated with a decrease in income by about 1402.247 INR. For each aditional year of education, their income is estimated to increase by about 241.43 INR. Owining a computer is associated with an increase in an induvidial's income of about 3095.68 INR. Under this model, we find that knowing at least some english is associated with an increase in the induvidial's income of about 504 INR. Under this model, all of the controls are statistically significant at the 1% level while the dummy for some english proficiency is only statistically significant at the 5% level.

8.3 Business Controls

We now explore how business related control variabes affect the estimated effect of some english fluency under the follwing model:

$$Income_{it} = \beta_1 EngAbil_i + \beta_2 NonFarmOwner_i + \beta_3 FarmOwner_i + \alpha_i + \epsilon_{it}$$

The question this model adresses is how business ownership affects the returns from english fluency. The concern with $NonFarmOwner$ and $FarmOwner$ is that since business owners likely

have to deal with government documentation, they may be more likley to be eudcated or simply had to be literate in at least 1 language, and that may be corelated to learning english.

For this model, owning a non-farm business is associated with an increase in the induvidial's income of about 20,457.78 INR, while owning a farm raises an induvidial's income by approximately 8,676 INR. Having some English proficiency is associated with an increase in the induvidial's income of 1,802.31 INR. These results for our variable of interest indicates that busniess ownership is only ever so slightly related with whether an induvidual is at least slightly proficient in english due to how close this result is with our inital results.

8.4 Life Status Controls

We now move to the life status controls. These are all controls related to the status of an induvidial's life. It utilizes the following model:

$$Income_{it} = \beta_1 EngAbil_i \beta_{2-6} MS_{2-6} + \beta_7 Sex + \beta_8 1km + \alpha_i + \epsilon_{it}$$

In this model, we control for marital status, the sex of the induvidial and the induvidual's ability to walk 1 kilometer. We control for marital status through the use of 6 dummy variables for each possible marital status collected in the data. The 6 dummys are: 0 for the induvidial's spouse being absent, 1 for the induvidial being married, 2 for the induvidial being unmarried, 3 for the induvidial being widowed, 4 is for the induvidial being separated or divorced, and 5 refers to a North indian tradition called Guana, where the induvidial is bethrowed (but this bethrwodal is usually when the induvidal is a child).

For our life conditions model, we only see statistically siginficant results for if an induvidial is either widdowed or divorced/separated from their spouse, where we see that a widowed induvidial has an estiamted decrease in their income of about 2,740 INR, and being separated or divorced from your spouse is associated with a decrease in the induvidial's income of about 3,740 INR. another thing we can conclude is that all marital status controls are associated with a decrease in

the induvidial's income. The estimated effect of an induvidial's age by 1, which is statistically significant at the 1% level, reinforces the earlier hypothesis that the Mincer earnings model would not apply to this context, with an increase in age being only associated with an increase in income of 22 INR, which is about 0.26 USD. For our variable of interest, we see an estiamted increase in income of about 1,804 INR, statistically significant at the 1% level. The closness of this estiamte to the inital regression results indicate that these controls are also mostly independent of

8.5 Full Controls

The final model we explore is the results when utilizing all of the listed controls. Under this model, with all controls utilized, the model indicates that knowing at least some english is associated with an increase in an induvidial's income of about 605 INR. This is indicates that a significant part of the intial estimate came from the correlation of english proficiency and the education controls. This is because of the difference in the resulting estimate from the education controls model and the complete controls model is so small, it indicates that the education controls were significant ommited variables in the inital model.

It is possible that this result is causal, as we have quite a lot of controls that may affect each induvidial's income. Furthermore, we have statistically significance in many of the controls, and statistical significance at the 1% level for the variable of interest. In this case, if this relationship is causal, we would have the ATT. More clearly, the averaege treatment affect on the treated is about 605 INR. This would mean that for the group that had at least some english proficiency relative to the group that had no english proficiency, the averaege increase in income from at least some english proficiency is about 605 INR. All of that is dependent on claiming causality, but without further robustness checks, it is hard to fully claim this causal channel.

8.6 Potential Issues and Biases

In terms of the model's utilized, the main concern remaining is that there may be further education related controls that by omiiting them, the final model is being inflated. But this isn't likely seeing

Table 4: Regression Results

	Dependent variable:		
	INCOME		
	(1)	(2)	(3)
AnyLang	−1,402.247***		
	(205.164)		
SchoolYrs	241.428***		
	(25.688)		
Computer	3,095.677***		
	(349.022)		
NoFarmBus		20,458.780***	
		(276.598)	
FarmBus		8,675.988***	
		(306.741)	
MS_1			−426.110
			(575.113)
MS_2			−769.966
			(588.344)
MS_3			−2,740.533***
			(655.146)
MS_4			−3,740.522***
			(1,160.341)
MS_5			−243.457
			(1,618.101)
Age			22.365***
			(6.398)
Sex			51.677
			(129.073)
Dif1km			−1,268.359***
			(311.371)
EngAbil	504.059**	1,802.311***	1,804.704***
	(226.118)	(189.845)	(194.076)
factor(SURVEY)2	14,974.460***	15,187.730***	15,074.470***
	(132.540)	(130.949)	(133.530)
Observations	380,997	381,167	381,167
R^2	0.039	0.057	0.039
Adjusted R^2	−0.080	−0.061	−0.080
F Statistic	2,784.159*** (df = 5; 338855)	5,079.946*** (df = 4; 339025)	1,384.385*** (df = 10; 339019)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Regression Results

	Dependent variable:
	INCOME
MS_1	−799.363
	(569.880)
MS_2	−1,267.221**
	(584.633)
MS_3	−2,672.474***
	(651.251)
MS_4	−4,103.705***
	(1,150.447)
MS_5	−1,086.462
	(1,603.094)
Age	13.241**
	(6.358)
Sex	−12.059
	(129.725)
Dif1km	−1,099.922***
	(308.763)
NoFarmBus	20,427.450***
	(276.596)
FarmBus	8,618.805***
	(307.041)
AnyLang	−1,228.394***
	(209.427)
SchoolYrs	189.991***
	(26.186)
Computer	3,178.839***
	(350.494)
EngAbil	605.837***
	(225.078)
factor(SURVEY)2	15,119.230***
	(132.996)
Observations	380,997
R^2	0.057
Adjusted R^2	−0.060
F Statistic	1,372.605*** (df = 15; 338845)
Note:	*p<0.1; **p<0.05; ***p<0.01

as we utilize years of education, which should be adequate at adressing how education for most education levels affects income.

A more concerning potential ommited variable is the Industry of the induvidial. Specifically, if we could control for the industry we could see how english returns may be related for those who may work in industries where english proficiency is more valuble like IT, healthcare, or finance, and how the estimate may be draged down by those in industries where it is less valuble like the service industry or agriculture.

Another area where these results are biased is in the sample size. While 200,000 induvidials is quite a lot, it is quite a small fraction relative to the size of India. This fact increases the likelihood that the data may be biased somehow, and that leads to results that either underestimate or overestimate the returns for english proficiency.

The final area of concern is the limitations that exist given the time period utilized. Specifically, the concern is that since there is only data from up to 2012, that the results found here are no longer the case. Even if you were to assume that these results have no issues, and the true increase in income from english proficiency is about 600 INR, it is still possible that this is no longer the case due to conditions in India changing since 2012, such that it is now no longer arround 600 INR. In fact, due to the trends of development in India, it is likely higher, but since we do not have any more recent data to work with, this is the best estiamte given the limitations of the data.

9 Conclusion

This paper seeks to explore the returns to English proficiency in India. The reason of interest for this question lies in how english prevalence in India may advantage it as it develops especially relative to how China has developed. By utilizing panel data from 2005 and 2012, this paper develops a fixed affects model to estimate how knowing at least some english affects an induvidial's income. The model with all included controls finds that at least some english proficiency is associated with an increase in the induvidial's income of about 600 INR. By utilizing a large ammount of controls,

the results indicate casuality, but without further robustness checks and controls, it isn't possible to say for certain. Even if it was extremely clear that the relationship was causal, these results are held back by the age of the data, with it being possible that since the most recent data we have for this panel data being from 2012, these results may no longer be true. But due to the development in India since 2012, it may be possible to argue these results are a floor for the returns to at least some english proficiency in India. The main place for further study indicated by this paper is to rexplore these results with more recent and more continois data, and by adding controls for things like industry. It also may be wise to explore the returns for college graudates, and see the results relative to these results for the general population.

References

Azam, M., Chin, A., & Prakash, N. (2013). The returns to English-language skills in India. *Economic Development and Cultural Change*, *61*(2).

Chaudhary, L., & Fenske, J. (2020). Did railways affect literacy? Evidence from India.

Grin, F. (2001). English as economic value: Facts and fallacies. *World Englishes*, *20*(1), 65–78.

Guo, Q., & Sun, W. (2014). Economic returns to english proficiency for college graduates in mainland china. *China Economic Review*, *30*, 290–300.

Munshi, K., & Rosenzweig, M. (2006). Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy. *American Economic Review*, *4*(96).