First-generation Immigrants' Upward Mobility: Exploring the Effects of Language Spoken At Home using Blinder-Oaxaca Decomposition Method

Natania Wong[*]

Advised by Professor Nano Barahona

26 April, 2024

Abstract

This paper delves into the intricate dynamics of the wage gap between immigrants and native workers, with a focus on the often-overlooked factor of language spoken at home. Utilizing the Integrated Public Use Microdata Series (IPUMS) Census/ American Community Survey (ACS) 2000-2021 Data [1], I employ Ordinary Least Squares (OLS) regression and the two-fold Blinder-Oaxaca decomposition method to uncover the impact of language on both explained and unexplained components of the wage gap. The OLS regression analysis reveals that after controlling for fixed effects, native workers earn 16% more and English speakers at home 11% more. Race also plays an interactive role with nativity and language spoken at home. The Blinder-Oaxaca decomposition highlights that among non-natives, each additional year of education beyond college for English speakers explains a wage gap of 2.5%, while each additional year of immigration explains a gap of 22.5%. As post-graduate education and longer immigration periods are major contributors to the wage gap, this research offers a nuanced understanding of the economic challenges faced by immigrants, highlighting the imperative to factor linguistic considerations into strategies aimed at fostering economic equality. The study underscores that tailored interventions addressing language barriers and educational attainment can be instrumental in creating opportunities for more inclusive economic growth, benefiting both immigrant and native workers.

[*] University of California, Berkeley (email: natania.wong@berkeley.edu). I want to express my deep appreciation to Professor Julien Lafortune, Professor Ryan Edwards, Tim Cejka, and David Wu for their valuable guidance and support.

1 Introduction

Immigration has played an essential role in the history of the United States. According to Pew Research Center, immigrants in the United States have made up almost 15% of the country's population, and have accounted for over one-fifth of the migrant population worldwide. People immigrate for various reasons, whether they are seeking asylum, reunification with families, or are sponsored by companies. All immigrants from various backgrounds share a common motivation: they are seeking improved opportunities for themselves and future generations. However, they face a variety of barriers, usually arising from language barriers. These barriers might hinder their success and bring an impact on the following generations.

The current literature review reveals that most research conducted has been looking into how other factors like race, education level, sex, and age affect the wage gap, particularly in comparison between white, black, or Latino immigrants from lower-income backgrounds. However, this analysis lacks context regarding one of the main reasons contributing to the wage gap: language barrier. With different demographics of immigrants possessing various cultural norms, language is a common and crucial bridge to their heritage. A broader perspective should therefore be implemented by examining whether the language of heritage being carried over to the United States is an indicator of how cultural norms may play a role in one's social mobility, answering the critical questions of whether bilingual proficiency before immigration and increased years of residence in the United States positively impact social mobility. In this paper, I will delve into how much of the wage gap between natives and first-generation immigrants could be explained by language spoken at home in the United States. Specifically, I aim to discern to what extent these disparities can be attributed to the language spoken at home or potentially other factors through the Blinder-Oaxaca decomposition method.

In order to carry out this research, I will be utilizing the datasets from administrative data collected by the American Community Survey (ACS). Some of the variables that are deemed useful for the dummy variable indicator include birthplace, foreign birthplace or parentage, year naturalized, year of immigration, years in the United States, and language spoken, etc. I conducted data manipulation before computing the models, including adding dummy variables and subsetting the data into cohorts by education, race, and years in the USA.

In the first part of the paper, I computed Ordinary Least Squares (OLS) regression models with the treatment variable being one's nativity and one's language spoken at home, where the outcome of the model would be the log wage of the person. Fixed effects and interaction terms are also added to the models. The regression analysis reveals several significant findings. Initially, native workers earn 26% less than foreign-born workers, while those who speak English at home earn 23% more, contrary to the initial hypothesis. After controlling for year-fixed and state-fixed effects, native workers earn 16% more than foreign-born workers, and English speakers at home earn 11% more, supporting the hypothesis. Including interaction terms show that being native has a positive effect only for Whites, and speaking English at home has a positive effect only for Blacks and Whites, suggesting that race influences household norms and language choices, impacting wages.

In the second part of the paper, I utilize the Blinder-Oaxaca decomposition method to delve into the two-fold results, observing the explained and unexplained components that contribute to the wage gap between those who speak English at home and those who do not. Results show that differences in age and years immigrated contribute significantly to the wage gap, with English speakers tending to be older. Among all immigrants, the wage gap is mainly driven by the presence of post-graduate degree holders among English speakers and the longer period of immigration. However, the unexplained portion of the wage gap, particularly among non-college attendees, is influenced by age and years immigrated, suggesting contradictory results that older age and fewer years of immigration among English speakers who did not receive college education contribute positively to the wage gap, possibly indicating a higher familiarity with English among immigrants from English-speaking countries.

The rest of the paper is divided as follows. Section 2 reviews the existing literature. Section 3 describes the data used. Section 5 presents the main regression of the paper, where section 6 describes the decomposition results. Finally, section 7 and 8 concludes.

2 Literature Review

While immigrants are more likely to be lower-income, Papademetrious' paper delves into the underlying factors of immigrants facing downward mobility due to coming from "either low education levels or with ... experience that is not relevant to the host-country labor market" [2]. The downward mobility might also be a result of the lack of "sufficient language skills to perform jobs equivalent to their last occupation" [2]. This raises the question of whether social mobility would be improved if immigrants came from bilingual countries and possessed sufficient language skills before moving to the United States. The literature piece also mainly focuses on an individual's country of origin, and might not be able to bring a bigger picture of how the language barrier comes into place when immigrants strive to move up the social ladder.

While empirical studies shed light on the significant role of language barriers in contributing to the wage gap between immigrants and natives, they also highlight a nuanced reality. Research suggests that "immigrants who speak English poorly are more visibly foreign than others", and despite possessing adequate

English skills, immigrants might possess an accent as a second language speaker, potentially "facilitat[ing] discrimination on the part of natives, and contribute to social isolation and ghettoization" [6]. The language one chooses to speak at home often reflects ties to immigrants' heritage and cultural norms, as well as how much their first language resembles English. Therefore, while much attention has been given to the wage disparity between immigrants and native workers, with socioeconomic factors often cited as contributing factors to natives earning more on average, there remains a notable gap in the literature. Few studies have delved into the wage discrepancies among immigrants themselves, particularly those originating from similar regions and sharing similar socioeconomic backgrounds. Exploring how the language spoken at home influences subsequent wages among immigrants would provide valuable insights into these wage disparities.

The existing literature on the wage gap among immigrants often overlooks the internal disparities within immigrant communities and fails to acknowledge the impact of insufficient English proficiency. My research methodology aims to address these gaps by adopting a broader and more comprehensive perspective on the relationship between English proficiency and immigrants' social mobility in the United States. By examining the intricate interplay between language skills and economic outcomes, my study seeks to provide policymakers with valuable insights into the potential role of language education in shaping immigrant integration and success in the workforce.

3 Data

3.1 Institutional Background

The data in this paper comes from the American Community Survey (ACS). I was able to gain access to the data via the Integrated Public Use Microdata Series (IPUMS) [1].

ACS has an annual sample size of about 3.5 million Americans and chooses 1 in every 100 people in the population randomly. Since the default data extraction includes data dating back to the 1850s, there are a total of 34523507 observations in the dataset, with each observation referring to one person being surveyed.

Given that both immigration policies and the labor market have evolved significantly over the past few decades, I decided that the observations and data back in the 1850s would be too distant and irrelevant for me to grasp recent developments in labor and immigration policies. According to IZA World of Labor, immigrants have a lower level of earnings compared to native workers due to various reasons including "education and experience acquired abroad are not perfectly transferable across borders" [5]. However, immigrants' wages grow over time and results show that on average, "the wages of first-generation immigrants ... are close to native wages after 20-30 years in the host country" [5], where there are also slight variations among immigrant groups. Therefore, I downloaded data dating back to two decades ago, including the years 2000, 2006, 2011, 2016, and 2021, allowing my dataset to align with my literature review results. The raw dataset I imported has 15309488 rows (observations), with each unit of observation representing one individual. The following sections will describe further data cleaning and manipulation carried out before regression models are built.

3.2 Data Description

The dataset has 29 columns (variables). Among those, I am interested in the following: Census year, Household Weight, Sex, Age, Birthplace, Citizenship status, Year naturalized, Year of immigration, Years in the United States, Language spoken, Speaks English, Educational attainment, Occupation, Industry, and Wage and salary income. While the treatment variables delve into the demographics of each individual, notably their age, sex, race, nativity, education attainment, and most importantly, their language spoken at home and English fluency, they will be explored and held constant in the models to minimize the biases induced. The outcome variable includes the salary income, in hopes of revealing the socioeconomic status of these individuals. We could then explore patterns among different immigrant groups and whether any of the treatment variables contribute to notable patterns between the outcome variables.

Since I have subset the ACS results to the most recent 20 years, the data is up to date and will give me a representative picture of the recent trend of foreign-born and native workers. All columns are in the format of numeric values, the codebook provided by IPUMS has described in detail what each column represents and what types of values they take in.

3.3 Data Cleaning

Data cleaning and data manipulation are necessary before carrying out further explanatory data analysis, namely because of the reasons below: Firstly, the dataset is filtered to include only entries from the past 20 years to ensure the most current and representative portrayal of trends among foreign-born and native workers.

Secondly, many variables consist of codes specific to the survey. For example, each entry in the birthplace column corresponds to a code, with codes 00100 to 12092 indicating locations within the United States territory. As a result, new dummy variables were created for nativity status, citizenship, education level, and English language usage at home, transforming the original columns to binary indicators of $TRUE$ and $FALSE$ accordingly.

Thirdly, the dataset is refined to include only individuals who identified themselves as part of the labor force upon further exploration of the income distribution among different demographic groups. Upon this filtration, approximately 7% of the observations within the labor force still reported zero wages. Given that my subsequent analyses will involve the use of logarithmic wages, these zero-wage entries pose a challenge as they would result in undefined values. Hence, I decided to exclude these zero-wage entries from further analysis. It is plausible that some respondents may have opted not to disclose their actual wages, leading to the presence of these zero-wage entries in the dataset.

3.4 Data Limitation

Despite the data has provided a broad view of insights on an individual level, there are rising problems with the inconsistency of price levels across different states and the span of 20 years. Price levels vary drastically in different parts of the United States, as well as across the period of 20 years as shown in time-series plots in the following section. Therefore, time-series data might be problematic in terms of providing insight on correlation due to spurious correlation. Additionally, since some of the columns are on a continuous scale, whereas some others are dummy variables, further data manipulation and understanding would be necessary to understand what each of the indices represents and how the regression model could be built.

Before building the models, it is important to note that we could not establish any causality in any of the regression models built below since the results are from observational instead of randomized control trials. the data should be explored to get a clearer picture of patterns among different groups and to identify confounding variables before building the regression models.

4 Descriptive Statistics

After data cleaning and manipulation, the updated dataset has a total observation count of 7401759, with 39 variables as stated. Summary statistics detailing the distribution of income salary among different groups are computed below:

Table 1: Summary Statistics of Income Wages

Characteristics		Income Wages ($)				
Native	Speaks English at Home	Min	1st Qu.	Median	3rd Qu.	Max
False	*False*	4	14,000	26,000	50,000	770,000
False	*True*	4	17,000	36,600	70,000	787,000
True	*False*	4	10,000	25,000	46,000	787,000
True	*True*	4	13,000	30,000	55,000	787,000

Note: American Community Survey (ACS) respondent's total pre-tax wage and salary income measurements (not adjusted for inflation), 5-year estimates 2000 to 2021.

In Table 1, the dependent variable of my regression model, i.e. the individual level of income wages is being explored. Table 1 compares the income wage distribution among 4 different groups, notably (1) Non-native & Does not Speak English at Home; (2) Non-native & Speaks English at Home; (3) Native & Does not Speak English at Home; (4) Native & Speaks English at Home. While the minimum and maximum wages in all 4 groups do not differ a lot, there is a huge median difference between the two non-native groups: For the non-natives, those who speak English at home on average, earn over 10,000 more than those who do not. The non-native group who speak English at home also exhibit the highest income wage for each benchmark out of all 4 groups. In general, those who choose to speak English at home exhibit higher income compared to those who do not both natives and non-natives. Interestingly, the natives who do not speak English at home exhibit the lowest income out of all 4 groups.

I further explored the trend of the wage by plotting a time series graph below:

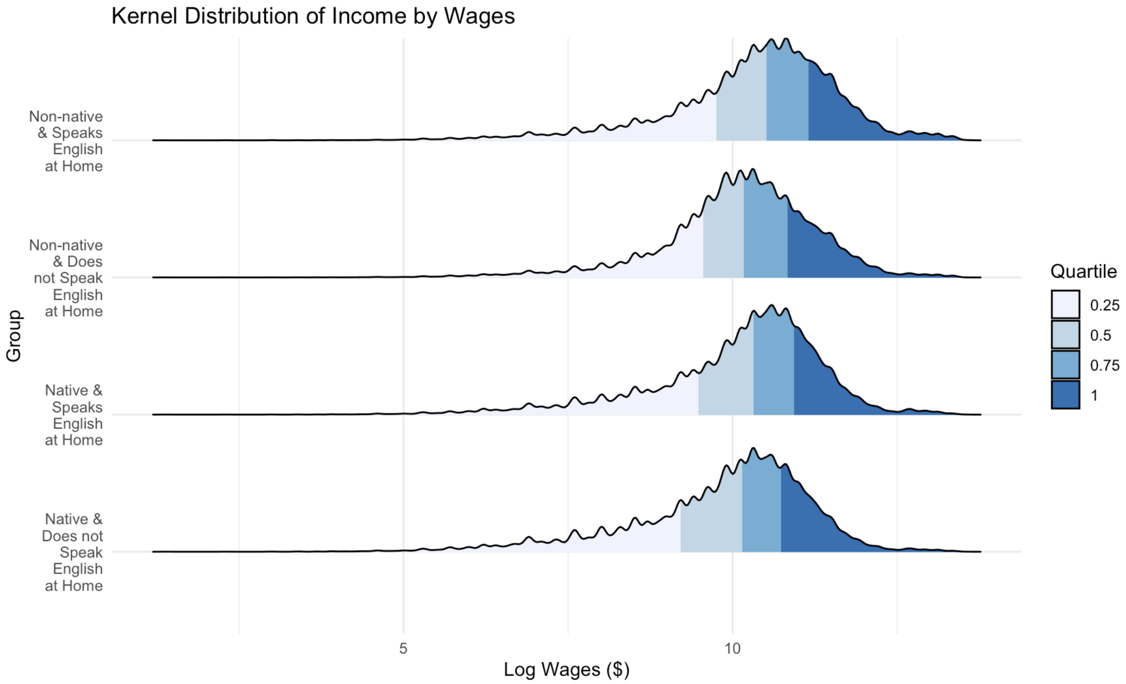

Figure 1: Kernel Distribution of Wages by Nativity and Language Spoken at Home

In Figure 1, the density plots for log income wages are plotted for each group. From the 4 density plots, it is evident that the income distribution is more left-skewed for both native groups when compared to the non-native groups. However, the group that is non-native and does not speak English at home seemingly has the highest median and the highest population at the right-tailed, meaning that it has a higher population of the highest income. Both groups that do not speak English at home have shown the lowest median log wages among all 4 groups.

I further explored the mean log wages over time by plotting time-series plots for the mean log wages by nativity and by language spoken at home, as shown below:

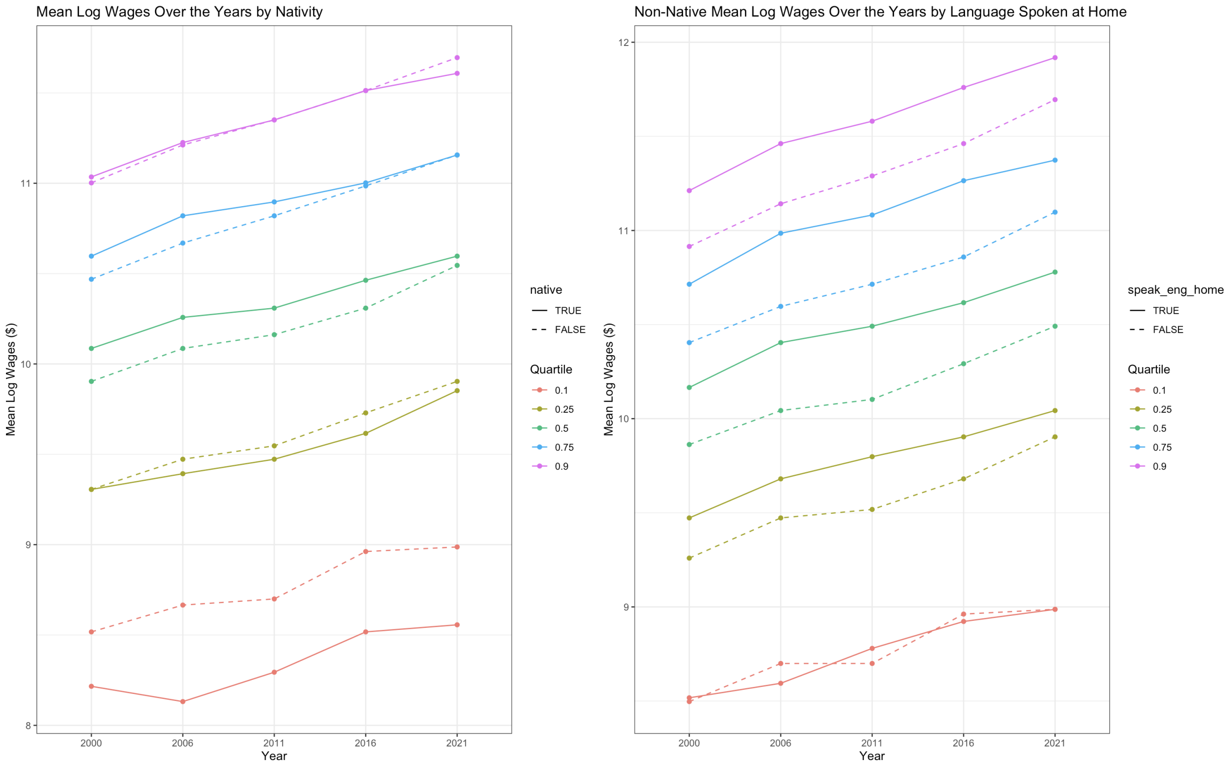

Figure 2: Time-series Plot of Mean Log Wages by Nativity and Language Spoken at Home

In the left plot of Figure 2, mean log wages over the years are plotted and grouped by the nativity of the individual. Natives are represented by a solid line, whereas non-natives are represented by a dashed line. As seen in the plot, the disparity of wages between natives and non-natives decreases as income increases: non-natives earn more than natives at the 10th percentile, and the gap decreases at the 25th percentile. The earning ratio is around the median of the distribution. At the median, natives earn more than non-natives, and the gap further decreases at the 75th percentile. At the 90th percentile, the gap almost entirely closes up. It is also interesting to note that while the gap closed up among groups across the overall distribution, the gap between natives and non-natives also closed up over 20 years except for those at the 10th percentile. In around 2016, the 90th percentile non-native earners were found to be earning the same as those who are native, and that trend persists, where the gap further widens and non-natives within the top-earner category end up earning significantly more than those who are native by the end of 2021. I hypothesize that the top earners from the technology sector might have emerged in recent years, contributing to the change of trend.

In the right graph of Figure 2, mean log wages over the years for non-natives are plotted and grouped by the indicator of whether they speak English at home. The group that speaks English at home is represented by the solid line, whereas the group that does not is represented by the dashed line. As seen in the plot, except for the group earning at the 10th percentile, those who speak English at home consistently earn much more than those who do not. At the 10th percentile, speaking English at home seems not to have a consistent impact on one's earnings. However, glancing at the rest of the distribution, those who speak English at home have a significantly higher income compared to those who do not speak English at home over the last 20 years. The disparity is the largest at 50th percentile and 75th percentile. This explains how the language one chooses to speak at home may explain some part of the wage gap among immigrants. Therefore, I will further delve into different subsets of the data, notably, whether years in the USA and race group affect the trend of results.

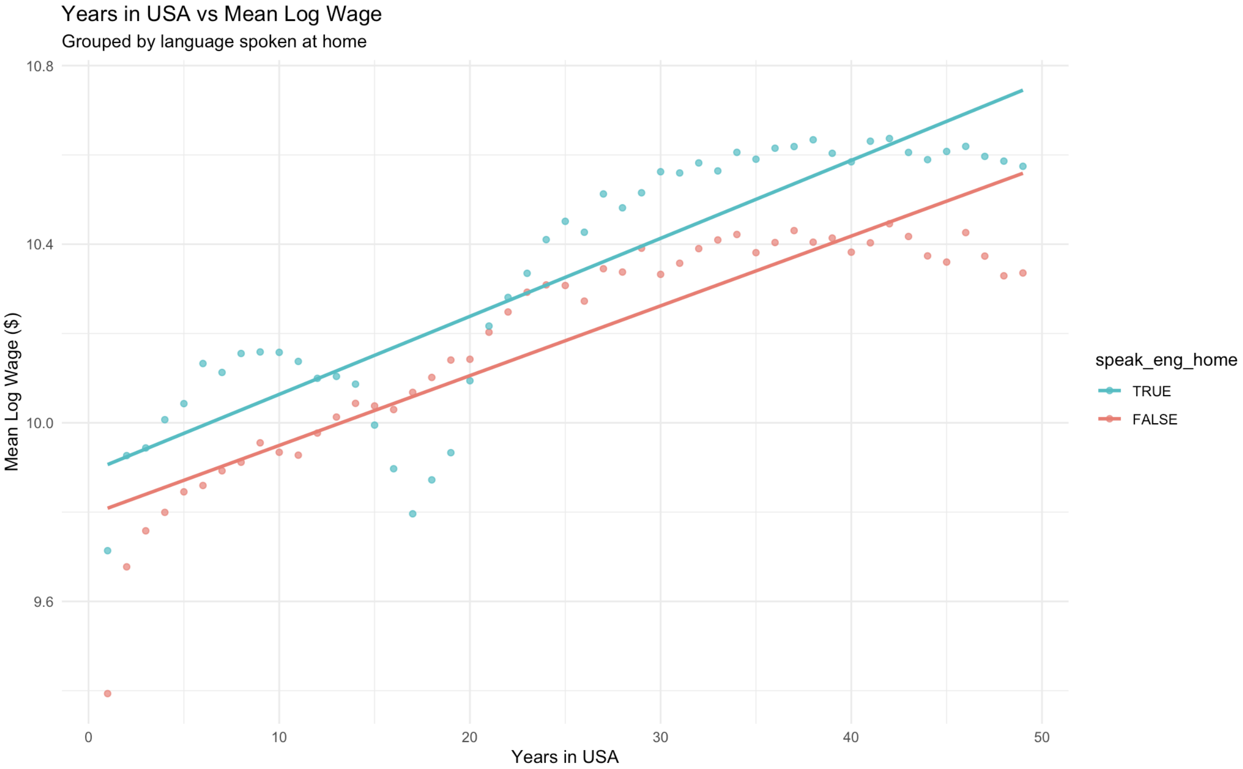

Figure 3: Scatter Plot of Years in USA vs Mean Log Wage (by language spoken at home)

In Figure 3, the relationship between years in the United States and mean log wages is being explored. It is evident that from years 0 to 50, there is a positive relationship between years in the USA and mean log wages for both those who speak English at home and those who do not. Those who speak English at home start at a higher mean log wage than those who do not, and the gap between the two groups further slightly increases with the increase of years in the USA. The increase also exhibits diminishing returns, where the mean log wages increase at a decreasing rate. In the following plots, data will be further subset and examined with additional conditions.

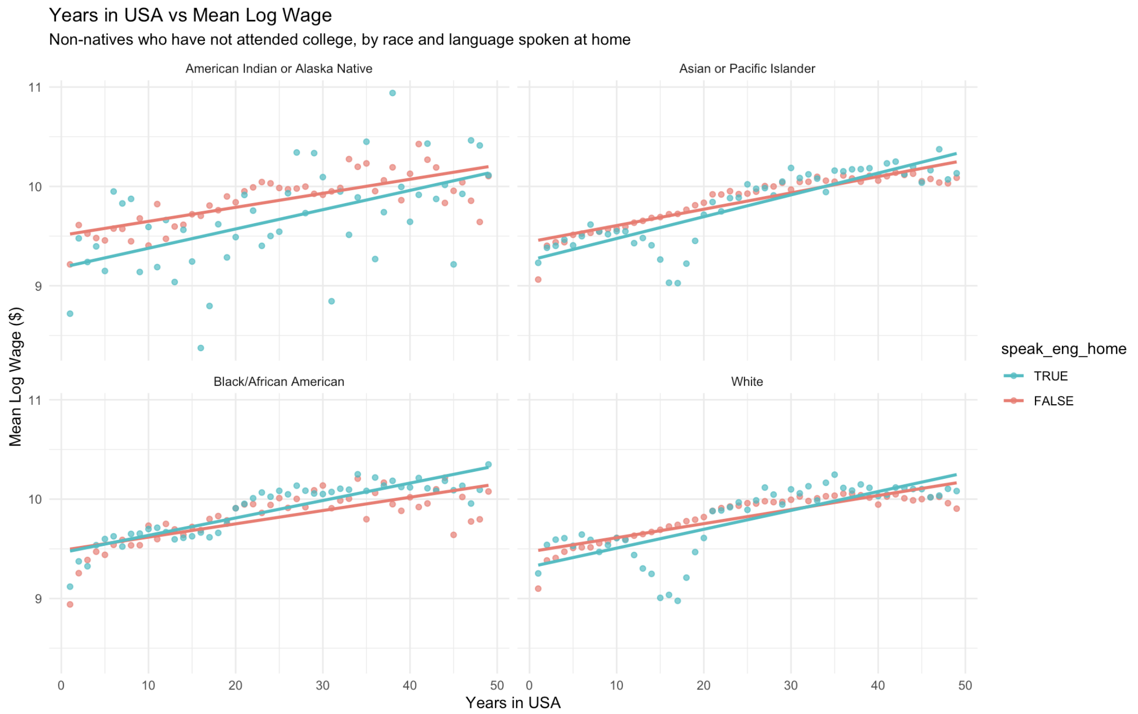

Figure 4: Scatter Plot of Years in USA vs Mean Log Wage (by race and language spoken at home)

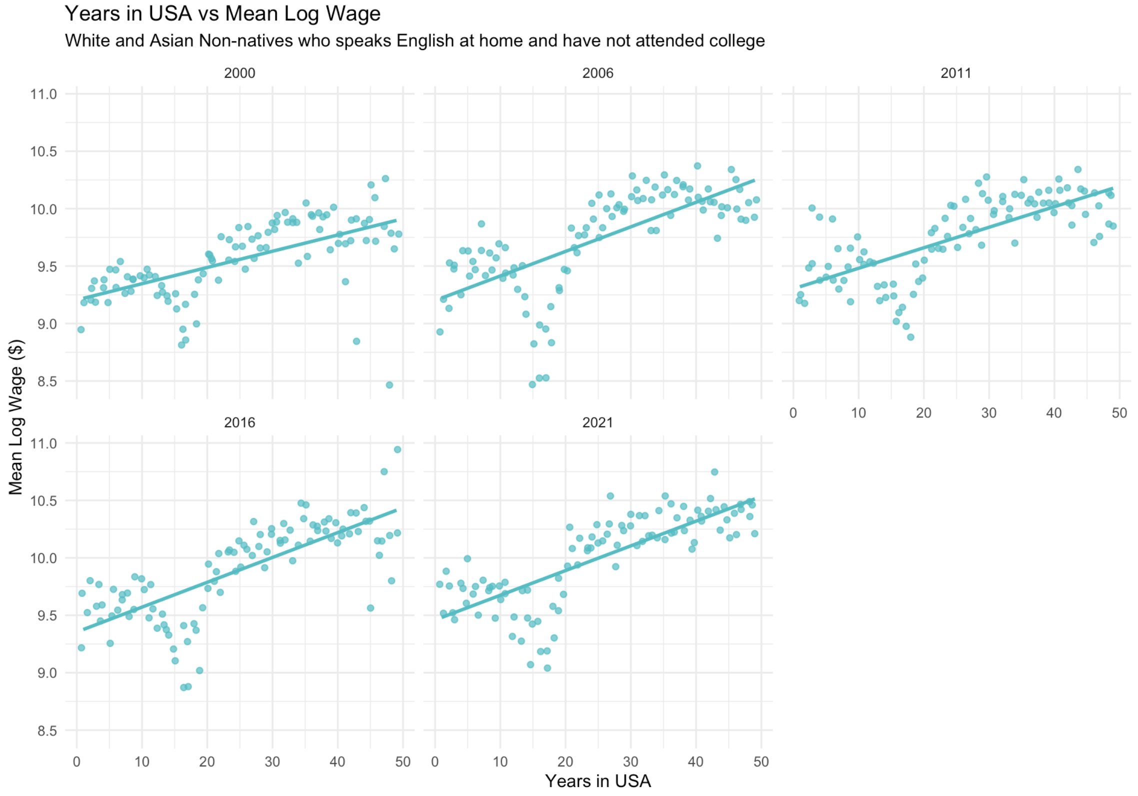

Figure 5: Scatter Plot of Years in US vs Log Income Wages (White and Asian and Pacific Islander non-natives who speak English at home, by Census Year)

Figure 3 exhibits an increasing relationship between years in the USA and mean log wages for both language groups. In Figure 4, I am hoping to further delve into the relationship of these two variables, but with different subsets of the data, notably, different race groups, and with the condition of education level added.

Data is being manipulated into aggregated race groups, namely American Indian, Asian and Pacific Islander, Black or African American, and White. According to the United States Census Bureau [7], the definition of the race groups is as follows:

- American Indian: those who "have origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment"
- Asian: those who "have origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent"; Pacific Islander: those who "have origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands"
- African American: those who "have origins in any of the Black racial groups of Africa"
- White: those who "have origins in any of the original peoples of Europe, the Middle East, or North Africa"

In Figure 4, years in the USA are plotted against mean log wage for a subset of the data who have not attended any college. The two colors of points and smooth lines represent the two language groups respectively. Each plot of Figure 4 represents one race aggregated above.

All four plots have shown a positive trend for mean log wage over the years. Within the subset of American Indians, those who do not speak English at home start with a much higher wage than those who do, but increase at a much slower rate over the years, hence, the mean log wage has later converged with those who speak English at home at around year 50.

Both language groups for the Black and African Americans start at a similar mean log wage at year 0, with the higher increase rate of earnings of those who speak English at home, hence, increasing the mean log wage gap between the language groups.

Lastly, the trend of the Asian and Pacific Islanders and that of the Whites resemble each other. Despite the overall trend of the two language groups for these two racial groups do not differ a lot over time, it is interesting to note that those who not speak English at home start with a higher mean log wage and has later converged and become overtaken with those who speak English at home at around year 30. It is also important to note that there is a sharp drop in the mean log wage for both Asians and Whites at around Year 15, further exploration is carried out in Figure 5.

Figure 5 shows the relationship between years in the USA and mean log wage for only Whites and Asian and Pacific Islanders, arranged by Census year. The objective of plotting figure 5 is to explore whether the sharp drop happens exclusively during a census year. Results show that there is a consistent pattern of a sharp drop in mean log wages for White and Asian immigrants after they have immigrated for around 16-18 years throughout all the census years being included in this dataset. Since outliers that exhibit a wage lower than 1.5 times that of the interquartile range (IQR) have been dropped, further study beyond that of this paper may be needed to explain this phenomenon.

Despite exhibiting an overall similar trend, the above Figure 4 shows that there may be varying trends and results within each racial group due to different cultural norms. These different cultural norms may subsequently play a role in the language they choose when interacting with family members, and their income wages. Therefore, race should also be considered when computing the models in the next section.

5 Relationship Between Income and Language Spoken at Home

Two methodologies will be employed to explore the relationship between income and language spoken at home in this research, namely the Ordinary Least Squares (OLS) regression and Blinder-Oaxaca Decomposition. The goal of using different methodologies is to gain a more nuanced understanding of how language factors contribute to variations in wage outcomes. Ultimately, the goal is to decompose the effects of the language spoken at home on one's earning potential in the labor market and to provide a comprehensive analysis that unveils the intricate mechanisms through this topic.

The following section will be a thorough walk-through of the OLS regression. I am hoping to explore the different facets of the relationship between language and wages, such as potential nonlinear effects, interaction effects with other variables, or the presence of mediating factors.

5.1 Empirical Strategy

The first Ordinary Least Squares (OLS) regression is widely regarded as the predominant method for fitting linear statistical models. The base model of the OLS regression model can be represented by

$$\log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{1}$$

where Y_i is i^{th} observation's value on the continuous dependent variable of income wages; β_0 is the regression intercept, X_{i1} is i^{th} observation's value on the independent (predictor) dummy variable showing if one is born outside of United States (*native*), β_1 is the corresponding regression coefficient; X_{i2} is i^{th} observation's value on the second predictor dummy variable showing if one speaks English at home (*speak_eng_home*), and β_2 is the corresponding regression coefficient. Both *native* and *speak_eng_home* are indicator variables, whereas income wages is a continuous variable. The error term ε_i accounts for the unexplained variation by independent variables in the dependent variable.

The main assumptions being held when building the following OLS models include (1) Linearity: The dependent variable Y_i is the sum of each product of the independent variable X_ik and the model is linear in parameters; (2) No multicollinearity: no independent variable should be perfectly linear to another independent variable; (3) The error term exhibits homoscedasticity, meaning that it has a constant variance for all independent variables, denoted as $Var(\epsilon \mid X) = \sigma^2$; (4) No autocorrelation: all error terms are not related to each other, which can be denoted as $\text{Cov}(\epsilon_i, \epsilon_j \mid X) = 0 \quad \text{for} \quad i \neq j$; (5) All error terms should be normally distributed, denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$.

It is important to note that while the assumption of the error term not being correlated with any of the independent variables and having a population mean of 0 (denoted as $E(\epsilon \mid X) = 0$) is usually being held in OLS models, the assumption is only necessary for establishing causal inference in the models. This paper emphasizes exploring the correlation between variables instead of establishing causality, and no causality can be established since there are factors not being taken into account when the Census is being conducted, for example, workers' family background or motivation to work. The model I built is the best linear predictor where β_k is minimized. Therefore, the error term is not equal to zero, denoted as $E(\epsilon \mid X) \neq 0$.

As mentioned in the previous section, there are confounding variables that are correlated to one's English proficiency while also affecting one's wages. Therefore, it is crucial to consider these factors and to add them into the model as control variables. The refined OLS model can be represented by

$$\log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i \tag{2}$$

where the additional X_{ip} added represents p^{th} observation's value on multiple control variables, including age, sex, occupation, and several dummy variables manipulated during data cleaning that represents one's level of educational attainment (attended some college, college graduate, post-graduate), years in the USA, one's residing state and year when the ACS was conducted; while β_k represents the corresponding regression coefficients.

As mentioned in the previous section, one's residing state and year are added as fixed effects since they affect the price level and would be confounding if they were not held constant across the span of the entire country and over 20 years. Furthermore, these additional control variables are closely related to one's socioeconomic status, hence, one's income level. For example, educational attainment often serves as the minimum qualification for many skill-based occupations, which tend to yield higher incomes. Leaving these variables out would lead to potential omitted variable bias (OVB), and affect the accuracy of the regression results.

As shown in differing mean log wage patterns among racial groups in Figure 4, race directly impacts cultural norms and hence, one's language-speaking habit at home. Therefore, the final OLS model can be represented by

$$\log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} * X_{i3} + \beta_5 X_{i2} * X_{i3} + ... + \beta_k X_{ik} + \varepsilon_i \tag{3}$$

where X_{i3} is a categorical variable representing one's race, and β_3 represents its corresponding coefficient. Additional terms including $\beta_4 X_{i1} * X_{i3}$ and $\beta_5 X_{i2} * X_{i3}$ represent two interaction terms respectively: (1) one's nativity status and race, (2) one's language spoken at home and race. I am hoping to explore how different racial groups interact with one's nativity and language spoken at home, as well as how that will impact one's language spoken at home, and ultimately mean log wages.

5.2 Results

The following regression results show the result of all three regression equations: 1, 2, 3. The coefficients and standard error (in parentheses) for each variable have been added to each column accordingly. No control variables have been added to the first model, while control variables are being added to both models 2 and 3.

Table 2: The effect of nativity and speaking English at home on wages

Dependent Variable: Log Income Wages			
	(1)	(2)	(3)
(Intercept)	10.09***	8.62***	8.56***
	(0.00)	(0.01)	(0.02)
Native	−0.26***	0.17***	0.17***
	(0.00)	(0.01)	(0.02)
Speaks English at Home	0.23***	0.11	0.07***
	(0.00)	(0.00)	(0.01)
Race (Asian or Pacific Islander)			0.07***
			(0.01)
Race (Black or African American)			−0.03
			(0.02)
Race (White)			0.09***
			(0.01)
Native*Asian or Pacific Islander			−0.01
			(0.02)
Native*Black or African American			−0.06***
			(0.02)
Native*White			0.01***
			(0.02)
Speak English at Home*Asian or Pacific Islander			−0.02*
			(0.01)
Speak English at Home*Black or African American			0.06***
			(0.01)
Speak English at Home*White			0.04***
			(0.01)
Controls	No	Yes	Yes
R^2	0.00	0.23	0.24
Number of observation	7401759	7401759	7401759

Note: * 10%, ** 5%, *** 1% significance levels.

According to the first model (column) of the regression table 2, it is evident that native workers on average earn 26% less than foreign-born workers, holding all else equal. Those who speak English at home earn 23% more than those who do not, holding all else equal. All results are statistically significant since the t-value is greater than 2 and the p-value is lower than 0.05. The result is surprising since it contradicts my initial hypothesis. I am contemplating the effects of confounding variables and the lack of year-fixed and state-fixed effects.

The second model (column) of the regression table 2 shows that after adding control variables including year-fixed and state-fixed effects, native workers on average earn 16% more than foreign-born workers, holding all else equal. Those who speak English at home earn 11% more than those who do not, holding all else equal. The coefficient decreases for half compared to the first model but remains positive and significant. All results are still statistically significant since the t-value is greater than 2 and the p-value is lower than 0.05. It is also evident that R^2 increases to 23%. The result aligns with my hypothesis, that native workers on average earn more than non-natives.

All variables remain the same in the final, and the third model (column) of the regression table 2 compared to the second model except for two additional interaction terms. These interaction terms are between one's nativity and one's race and between whether one speaks English at home and one's race. The updated regression results show that native workers on average earn 17% more than foreign-born workers, holding all else equal. Those who speak English at home earn 7% more than those who do not, holding all else equal. The coefficient slightly decreases compared to the second model but remains positive and significant. Asians and Whites both have a positive coefficient, where Asian and White workers on average earn 7% and 9% more than those who are neither Asian, White, or Black. Being Black has shown a negative effect on log wages, where they on average earn 3% less than those who are not Asian or White. The interaction term shows that being native only has a positive effect on those who are also White, where they earn 1% more than those who are not native or White. Surprisingly, being native and being Asian or Black does not have a positive effect on their log wages. The other interaction term between whether one speaks English at home and one's race shows that being able to speak English only has a positive effect on those who are also Black and White, where they on average earn 6% and 4% more than those who are not. This proves that race plays an important role in the norms within a household, and what kind of language one chooses to speak at home, subsequently, one's wages.

R^2 remains at 23%. The result still aligns with my hypothesis and slightly increases compared to the result from the second regression.

6 Decomposing Effects of Language Spoken at Home using Blinder-Oaxaca Decomposition

6.1 Empirical Strategy

The Blinder-Oaxaca decomposition technique is a valuable tool for discerning and quantifying the distinct influences of group variations in measurable attributes, such as education, experience, and geographical location, on disparities in outcomes. This technique is widely used in the context of social science studies. In the context of our research, we are going to apply the Blinder-Oaxaca decomposition method to decompose group variations of English language usage, and how that contributes to disparities in wages. Using the results, I will compute bootstrapped standard errors for the estimates, and visualize the outcomes of the decomposition, hoping to disentangle the research question of how much of the wage gap could be attributed to English language usage at home.

As mentioned in Section 2, literature review has shown that past research has mainly focused on the wage gap between native and foreign-born workers. Therefore, I am subsetting the data to only include immigrants who have not attended any college, hoping to delve into a distinct direction in this research by decomposing the wage gap among only immigrants.

The subsetted dataset includes two distinct groups: those who speak only English at home (Group A) and those who do not speak English at home (Group B). The group assignment is arbitrary and the results will be symmetric regardless. However, most literature reviews have assigned the group that presumably has a higher wage as Group A, such that the mean difference later calculated, denoted by $Delta\overline{Y}$ will be positive. The outcome variable $(\overline{Y})$ can be estimated by using a multivariate linear model incorporating a set of measured predictors $(\overline{X})$. The mean outcome $(\overline{Y})$ of each group can be denoted as:

$$\overline{Y}_A = \beta_{0A} + \sum_{j=1}^{J} \beta_{jA}\overline{X}_{jA} \tag{4}$$

$$\overline{Y}_B = \beta_{0B} + \sum_{j=1}^{J} \beta_{jB}\overline{X}_{jB} \tag{5}$$

where β_0 represents the intercept coefficient; and $\overline{X}_j$ represents a collection of J measured predictor variables, where the vector β_{jB} denotes the slope coefficients that signify the relationship between the predictors $(\overline{X}_j)$ and $\overline{Y}$.

The mean difference in outcome ($\bar{Y}$) between immigrants who speak only English at home (Group A) and those who do not speak English at home (Group B) can then be denoted as:

$$\Delta\overline{Y} = \overline{Y}_A - \overline{Y}_B \tag{6}$$

$$\Delta\overline{Y} = (\beta_{0A} - \beta_{0B}) + \sum_{j=1}^{J}(\beta_{jA}\overline{X}_{jA} - \beta_{jB}\overline{X}_{jB}) \tag{7}$$

The Blinder-Oaxaca decomposition method decomposes the overall wage disparity into 3 components: (1) overall mean difference of X; (2) differences in values of intercepts and (3) slope coefficients. To transform the equation into 3 components, two hypothetical terms (i.e., β_{jA} and $\overline{X}_{jB}$) are added in Equation 6 above, where β_{jA} represents a vector of reference coefficients that has "typically been interpreted to be non-discriminatory, ..., [and that] the set of regression coefficients would emerge in a world of no labor market discrimination" (Hlavac, 2022) according to a literature review on labor market discrimination. Then, the difference in mean between Group A and B ($\Delta\overline{Y}$ in the Blinder-Oaxaca decomposition will be as follows:

$$\Delta\overline{Y} = (\underbrace{[\sum_{j=1}^{J}(\overline{X}_{jA} - \overline{X}_{jB})\beta_{jA}]}_{\text{explained}} + \underbrace{[(\beta_{0A} - \beta_{0B}) + \sum_{j=1}^{J}(\beta_{jA} - \beta_{jB})\overline{X}_{jB}]}_{\text{unexplained}} \tag{8}$$

As shown in 8, the two-fold Oaxaca-Blinder decomposition further decomposes the difference in mean outcomes into two components: "explained" and "unexplained". The "explained" component of the gap encompasses the collective group difference in Y, derived from disparities in the mean values of predictor variables; whereas the remainder of the equation is the "unexplained" portion, attributed to variations in intercepts and slope coefficient estimates. These differences signify variations in how the predictor variables correlate with outcomes across the two groups. Essentially, even if Group A were to possess equivalent mean levels of predictor variables as Group B, the differences in outcome between the two groups might remain. Consequently, this method would enable us to identify the factors that contribute to this disparity in the outcome variable, providing "pieces" of insight into how much of the differences could be accounted for in each explanatory variable.

6.2 Results

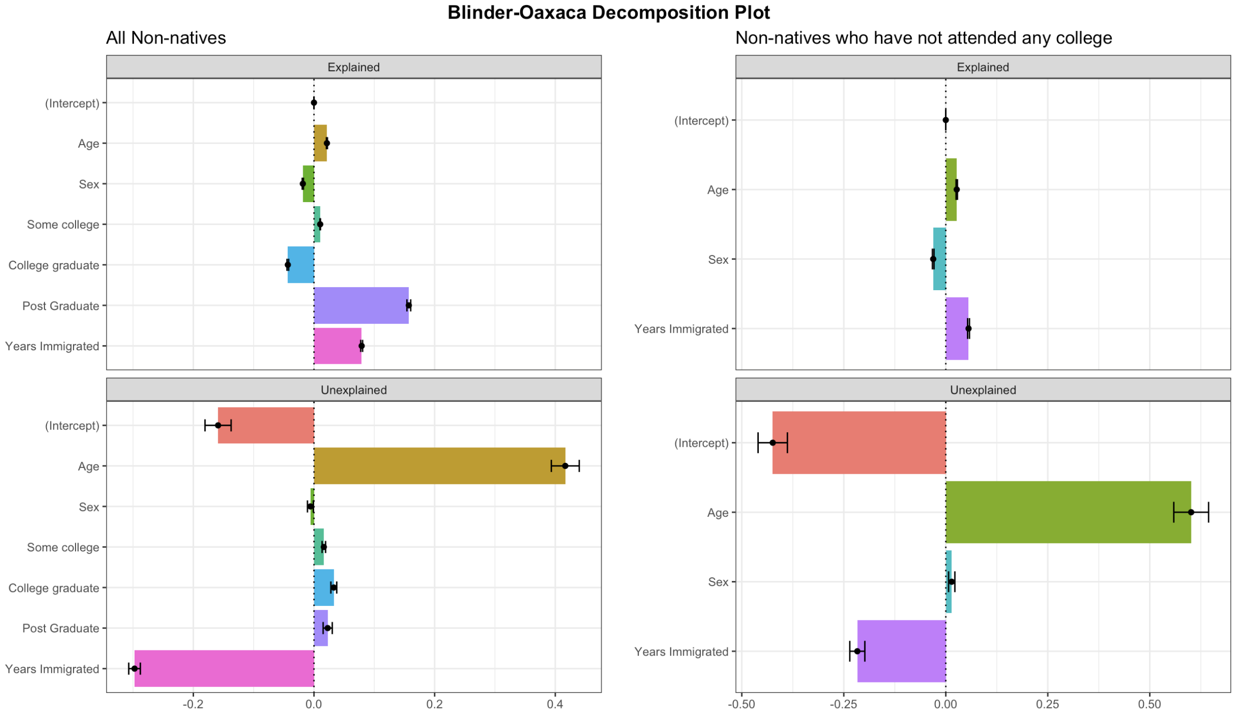

Figure 6: Blinder-Oaxaca Decomposition Plots (All non-natives vs non-natives who have not attended any college)

Using the R package Oaxaca, the wage gap between those who speak English and those who do not is decomposed into explained and unexplained parts for each variable. Figure 6 shows the visualization of the decomposition results, where the top part of each plot represents the explained part of the coefficient column of the wage gap, and the lower part represents the unexplained part of the coefficient column of the wage gap.

According to the documentation on "Blinder-Oaxaca Decomposition in R", the two negative weights in decomposition results indicate whether the "reference coefficients come from pooled regressions without (-1) or with (-2) the group indicator variable included as a covariate" [4], i.e. whether the variable *speak_eng_home* is included as a covariate in the decomposition. In the following results, I will mainly delve into the results that rely on pooled regression coefficients derived from a regression model that excludes the group indicator variable (whether one speaks English). The decomposition results are indicated by a weight of -1 in the weights column.

In the decomposition results where group weight equals -1, the coefficients in the explained column represent the contribution of each explanatory variable to the explained portion of the wage gap. For instance, the Age variable has a mean of 43.56 in group A and a mean of 41.84 in group B. It also has a coefficient of 0.021 in the explained column, denoted as β_{jA}, meaning that the differences in age between the two groups explain 2.5% of the wage gap. Applying these values to the explained part of Formula 8:

$$explained = \sum_{j=1}^{J} (\overline{X}_{jA} - \overline{X}_{jB})\beta_{jA} \tag{9}$$

$$Age\_explained = (43.56 - 41.84) * 0.021 \tag{10}$$

The coefficients in the unexplained column represent the total contribution of each explanatory variable (of groups A and B) to the unexplained portion of the wage gap. Similarly, the total unexplained portion of the Age variable is 42%. It is the sum of the unexplained portion of Group A of 32%, denoted as β_{jA}, and that of Group B: 9.5%, denoted as β_{jB}. The mean age of Group B denoted by $\overline{X}_{jB}$ is 41.84, making the unexplained portion of Formula 8:

$$slope\_coef\_unexplained = \sum_{j=1}^{J} (\beta_{jA} - \beta_{jB})\overline{X}_{jB} \tag{11}$$

$$Age\_unexplained = (0.32 - 0.095) * 41.84 \tag{12}$$

While the mean and coefficients for each component (variable) can be plugged into both the unexplained and explained portion of the formula above, it is difficult to interpret the size of each effect with only the numeric values. Therefore, the visualization in Figure 6 presents clearly the size of the coefficients for each variable in both explained and explained portions using bar-plot with error bars, where the x-axis represents the size of the coefficient and standard error.

In Figure 6, the left side shows the results of all non-natives, while the right side shows the results of the non-natives who have not attended any college. On the left side of Figure 6, it appears that the wage gap between the two language groups is mostly driven largely by the presence of workers who have earned a post-graduate degree, and the higher number of years immigrated among those who speak English at home (in the explained component). On the right side of Figure 6, it appears that among the non-natives who have not attended any college, the wage gap is driven mostly by a higher number of years immigrated to the United States, and older years of age.

I can further explore the unexplained component by examining the following variables for all non-natives: age and years immigrated. Through further visualization, I am hoping to anticipate how much of the unexplained portion of the wage gap can be attributed to the language one chooses to speak at home.

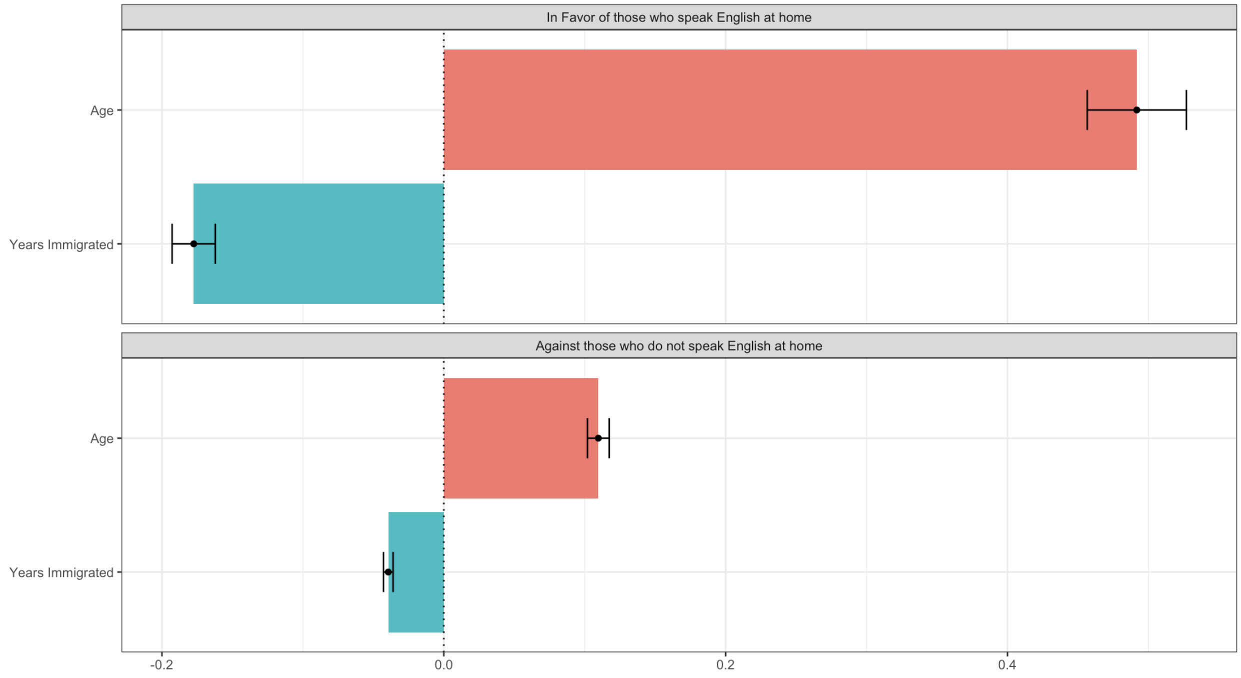

Figure 7: The unexplained portion of sub-components in twofold Blinder-Oaxaca Decomposition among non-natives who have not attended any college

In Figure 7, it is evident that the group that speaks English at home has an older age, and that older age contributes drastically to the positive wage gap between those who speak English and those who do not, whereas the fewer amount of years those who speak English at home immigrated to the United States have also contributed positively to the wage gap. Although the general concept is given that younger immigrants can grasp language skills more effectively, and therefore, the group that speaks English at home should have a younger age, the phenomenon exhibited in the Blinder-Oaxaca decomposition method result above could also potentially be explained by the "older" group of immigrants being mostly from English-speaking countries, which make these immigrants naturally subject to a higher familiarity with English before moving to the United States. Hence, they will be speaking English at home.

7 Discussion

While the first OLS regression results suggest findings that contradict my hypothesis, results align with my hypothesis as additional control variables are being added, as well as the state-fixed and year-fixed effects. Results have shown that those who speak English at home earn a consistently higher income than those who do not, despite white and Asian non-natives experiencing a sharp drop in average log wage after moving to the United States for around 15 years. The reason behind the drop will require further studies into the patterns of immigration among these two racial groups, since the result has been consistent throughout the Census conducted over two decades, meaning that the phenomena do not happen in a specific Census year or a specific cohort group.

Results from the Blinder-Oaxaca Decomposition show that most immigrants who speak English at home have earned post-graduate degrees, explaining education level has close tights with one's language proficiency and subsequent ability to find jobs with higher wages. Subsequent results show that an older distribution of age within the group that speaks English at home also contributes to the positive wage gap. This could be explained by the demographic distribution of "older" immigrants and immigrant policies that change over time. Immigrants who speak English at home may belong to earlier waves of immigration to their host country, where these earlier waves of immigration occurred several decades ago when immigration policies were different. Immigrants from certain regions, such as Europe, were more likely to speak English as their first language. Therefore, older immigrants who speak English at home may represent earlier cohorts of immigrants who arrived when English proficiency was more common among newcomers. Furthermore, older immigrants may have had more opportunities to acquire English language skills through education opportunities or professional experiences in their home countries. For example, individuals who pursued higher education or worked in industries with international connections may have developed English proficiency over time, leading them to speak English at home even after they immigrate. Lastly, family circumstances and simulation may play a role in the language older immigrants choose to speak at home. Older immigrants who have stayed in the United States for an extended amount of time may have already built their families and roots here, hence, English becomes the primary language of communication within their household when their family members, such as children or grandchildren, are fluent English speakers or were born in the United States. They may prioritize using English at home to assimilate into the host country by adopting English as the primary language of communication.

The Blinder-Oaxaca Decomposition results also show that fewer years of immigration also contribute to the positive wage gap. Despite this contradicts the common belief that more years of immigration would enhance one's skill set and minimize one's language barrier, contributing to a higher wage, this can also be explained by several reasons. Firstly, immigrants on the left plot of Figure 6 may have a higher education level or specialized skills in the first place; these immigrants may possess credentials that are in high demand or transferable skills that are valued in the host country's labor market despite a shorter duration of residence. A higher education level also has a direct connection to their language proficiency and subsequent cultural fluency in the host country's business setting and their ability to adapt to the labor market, subsequently contributing to the positive wage gap compared to native-born workers who have less education. Secondly, some immigrants with specific skills may be placed in a geographical location that has higher mobility or receive job placement assistance from immigrant networks or agencies, allowing them to have better access to high-paying job opportunities despite having a shorter duration of residence. Finally, prior working experience and industry connections also affect the wages immigrants receive in their host country. These prior experience and knowledge would enable newcomers to capitalize on emerging opportunities or fill necessary gaps in the labor market.

By delving into the signs of each component, I can dissect the effect of each component on the wage gap. Understanding the sign of each effect and component is important because it gives a clearer picture of the nature of the wage gap, hence, providing context to the groups being analyzed and the factors being considered.

8 Conclusion

This study makes a significant contribution to the understanding of wage disparities between immigrants and native workers, offering valuable insights for policymakers. By incorporating language spoken at home as a crucial factor, often overlooked in previous studies, we shed light on a potential gateway to address the language barriers faced by immigrants. The utilization of the ACS panel data enables a nuanced examination of living habits, specifically the language spoken at home, and its subsequent impact on wages. The findings emphasize the importance of considering linguistic factors in policy discussions surrounding the wage gap. Policymakers can use this information to develop more targeted and effective strategies to promote economic equality, taking into account the influence of language on immigrants' earning potential. This study thus provides a fresh perspective that can inform policies aimed at fostering inclusive economic growth.

In conclusion, this research underscores the nuanced dynamics contributing to the wage gap between immigrants who speak English at home and those who do not. The results confirmed the impact of language on wages and highlighted the role of post-graduate education as a major contributor to the wage gap. The two-fold Blinder-Oaxaca decomposition method further unveiled that the time of immigration and age also play a substantial role in both the explained and unexplained components of the wage gap and that these factors may be closely related to one's language spoken at home. These findings offer a comprehensive perspective for policymakers aiming to address wage disparities. To create more equitable economic opportunities, it is crucial to recognize the influence of language and educational attainment, ensuring that policy interventions are tailored to the specific challenges faced by immigrants in the labor market.

References

[1] Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. *IPUMS USA: Version 15.0 [dataset].* Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D010.V15.0, Oct 22, 2023

[2] Papademetriou, D. G., Somerville, W., & Sumption, M. *The social mobility of immigrants and their children.* Washington: Migration Policy Institute, 2009.

[3] Kasinitz, P. *Becoming American, becoming minority, getting ahead: The role of racial and ethnic status in the upward mobility of the children of immigrants.* The Annals of the American Academy of Political and Social Science, 620(1), 253-269, 2008.

[4] Hlavac, Marek. *oaxaca: Blinder-Oaxaca Decomposition in R.* R package version 0.1.5., https://CRAN.R-project.org/package=oaxaca, 2022.

[5] Huang, Z., Anderson, K. *Can immigrants ever earn as much as native workers?* IZA World of Labor 2019: 159 doi: 10.15185/izawol.159.v2, 2019

[6] Bleakley, Hoyt, and Aimee Chin. *Language Skills and Earnings: Evidence from Childhood Immigrants*.* MIT Press, direct.mit.edu/rest/article-abstract/86/2/481/57467/Language-Skills-and-Earnings-Evidence-from?redirectedFrom=PDF, 1 May 2004

[7] The United States Census Bureau. *About the Topic of Race* U.S. Census Bureau, https://www.census.gov/topics/population/race/about.html, March 1, 2022