Return to English fluency among race and education

March 16, 2020

Quang Pham
qpham@ucsb.edu

Department of Economics
University of California, Santa Barbara

Advisor: Professor Olivier Deschenes

Abstract:

Language fluency is defined as an important factor of human capital for immigrant workers. The fluency premium (FP) is the incremental payment for English proficiency between similar workers. Using data from the American Community Survey, this paper investigates the overall English fluency premium for immigrant workers and the difference in the return to English fluency among 4 major race groups while controlling for counties, jobs, and year fixed effects It is determined with OLS that generally, the FP is about 14.5% for all immigrants and around 5% for immigrants who arrived before their 18th birthday. After applying IV, the FP is measured at 23.5%. There are also significant differences in the FP in both samples among the 4 major race groups and skill levels.

I. INTRODUCTION[1]

Effective communication and language skills are important aspects of human capital. An improvement in native language proficiency boosts productivity by allowing for more effective communications with co-workers and clients. Thus, higher English proficiency (EP) may lead to higher compensation. A fluency premium (FP) is the difference in wage compensation between a fluent and non-fluent English speaker who are otherwise identical. Existing economics literature on the effect of English proficiency (EP) on an immigrant's wage suggests that there is a strong and positive correlation between EP and compensation (McManus 1983; Berman, Lang, and Siniver 2003; Bleakley and Chin 2004).

In linguistics, language proficiency is measured by intelligibility and clarity in communication, but the same level of EP differs among ESL immigrant groups due to tonal and regional variations, or accents. For example, a French-Canadian immigrant speaking at a C2-level EP will have variations that will be different from a Vietnamese immigrant speaking at a C2-level EP. Accent prestige theory in sociolinguistics suggests that preferences change the way people perceive both the speaker and speech based on language variations (Labov 1966; Butler 2007; Derwing 2003). Recent quantitative analysis indicates that regional speech variations of English have tangible effects on wages (Grogger 2019). Preferences over accents and speech variations could alter the native language FP among immigrant groups who are from various linguistic backgrounds.

Although immigrant groups of the same race tend to settle in high density areas, some groups are more heterogeneous than others. Hispanic immigrant groups are more

[1] I would like to thank Professor Olivier Deschenes, my advisor, and Professor Shelly Lundberg, my instructor, for their invaluable guidance and overwhelming support in my first research endeavor. I could not have accomplished this without them.

homogenous culturally and linguistically with Spanish as their major language; therefore, the need for assimilation may not be pressing due to community effects. Meanwhile, Asian immigrants are more heterogeneous culturally and linguistically with Mandarin, Cantonese, Vietnamese, etc. as their primary language groups. As a result, the demand for EP and improvement may be higher for Asian immigrants and other heterogeneous immigrant groups. In addition, Asian and African language groups have larger linguistic distances from English which drives the marginal cost of native language acquisition (Chiswick 2004). These factors combined drive the margins for the FP of Asian and African immigrants.

The goal of this paper is to determine the magnitude and heterogeneity of the FP among 4 major immigrant race groups and education. The analysis is conducted using micro-data from the ACS 2013 5-year sample from IPUMS in states with high immigrant concentration. In the second phase, the sample is filtered to only childhood immigrants (those who arrived before their 18^{th} birthdays) to avoid endogeneity issues since adult migrants tend to move in response to labor opportunities and incentives. The FP is estimated using OLS with county, year, and job fixed effects, and the heterogeneity in the FP among race groups are measured using interaction terms between race and EP. The magnitude of the FP using the full sample of immigrants was measured at 14.5% with significant differences in the return to EP for all race groups. Meanwhile, using the restricted child migrant sample, the FP is measured at 5.4% with significant differences in Asian and African migrants.

The FP was allowed to differ between race groups among education levels by adding a second interaction term between EP and race, and a third interaction term between EP,

education, and race. The estimates of FP vary significantly among immigrant race groups and across education levels. Specifically, the FP is consistently higher for Asian immigrants in both the full and restricted sample whereas for Hispanic immigrants, it was lower in both samples.

The final part of the analysis uses age at arrival and the interaction term between age at arrival and non-English-speaking origin as instrumental variables for EP on the restricted child migrant sample to determine the magnitude and the heterogeneity of the FP among race groups. The magnitude of the FP using the IV approach was determined to be 23.5%; however, heterogeneity analysis of the FP with IV was inconclusive.

The rest of this paper will be organized as followed: Section II provides a brief literature review of the existing economics and sociolinguistics literature on the effect of EP on wage and speech preferences. Section III examines the survey data used in this paper with descriptive statistics and data visualization of key variables used in the analysis. Section IV discusses the empirical strategies used to investigate the magnitude and the differences of the FP. Section V presents the results and its implication. Section VI concludes.

II. LITERATURE REVIEW

There has been extensive research into the effect of native language fluency on wage. Early research into this effect is typically based around regressions of income on self-reported EP data while controlling for several relevant metrics such as education, occupation, etc. (Kassoudji 1988, Chiswick 1991). Chiswick (1991) designed a survey to measure the return to 2 different types of EP (reading and speaking) for low-skilled immigrants. The survey provided a special opportunity to collect data of variables like

fluency by type and fluency at arrival that are not collected by the Census. The study concluded that EP increases with longer duration in the United States, and the improvement in EP is greater and faster for those with higher education and young migrants. Chiswick (1991) also determined that the increase in EP with duration in the US was slower for Mexican immigrants; the study attributes this to the adverse effects of living in large ethnic enclaves on EP. The major implication of this study is that reading fluency has stronger implications on wages than speaking fluency because it can reflect other soft skills, for it is highly complementary to education and acquisition of other important labor skills. Kassoudji (1988) used Survey of Income and Education data from 1976 to determine the effect of EP on income for Asian and Hispanic immigrants by profession. Kassoudji analyzed East Asian and Hispanic immigrants separately across 6 major job categories. The study suggests that low EP abilities not only have a significant large and negative effect on wage (depending on occupations) but also impose restriction to certain job opportunities.

Recent studies by Grogger (2018) found that speech variations also have tangible effects on wages. To determine speech variation, Grogger collected data of one formal speech recording and one informal speech recording from the NSLY97 participants. In addition, anonymous listeners listened to these recordings and answered a series of questions to determine the participants' speech variation. Grogger discovered that the wage premiums for mainstream speech is robust to controls for a myriad of variables. His study implies there is a positive correlation between mainstream speech and wages after controlling for relevant factors. In addition, acquisition of mainstream speech is more likely in areas where the premiums to mainstream speech are higher as a response to the economic incentives. More recently, Grogger (2019) used instrumental variable strategy to isolate the

causal effect of regional dialects and accents on wages. The study estimates that the regional accent penalty is 20% of wages.

Butler (2007) used a randomized control trial to determine the effect of nonnative teachers' nonnative accent on students' English learning outcome and students' perception of the teachers' accents. The study split 312 students into 2 groups: one native English-speaking teacher and one with nonnative-speaking teacher. It determined that there were no significant effect of the teacher's accented English on student's outcomes. On the other hand, the students preferred a native English speaker accent over a nonnative speaker accent.

There are two primary issues when measuring EP's effect on wage. Firstly, the OLS estimator will be upwardly biased because EP is endogenous with other factors that determines wage for immigrants such as EP at arrival, education, location, etc. (Borjas (1994)). Secondly, there are some measurement errors in self-reporting of EP from immigrants that will have a downward bias on the OLS estimator for EP on wage. More recent researchers have developed their own methods to deal with these endogeneity issues (Dustmann & Van Soest (2002), Bleakley & Chin (2004 & 2010)). Using panel data, Dustman and Van Soest (2002) found that the downward bias from misreporting fluency is much larger than the upward bias from endogeneity of EP. To determine the inconsistencies in reporting fluency, they classified the errors into time consistent and time varying categories. The results imply that the potential misclassification of fluency levels can have strong and downward effect on the implications of fluency on wage. Overall, it can potentially outweigh the upward bias from endogeneity issues.

Bleakley and Chin (2004) created an instrumental variable (IV) by interacting age of arrival with non-English speaking origin dummy variable after restricting the samples to immigrants who arrived before their 18th birthday. The IV can be interpreted by the potential difficulty in adapting to the new country and language for immigrants who came to the US from a non-English speaking origin, so the coefficient of the IV on fluency was negative. The study determined that younger migrants from non-English speaking countries, on average, will have comparable EP levels to immigrants from English speaking countries. Meanwhile, older children immigrants from non-English speaking countries, on average, had lower EP. Furthermore, child immigrants from non-English speaking countries had overall lower educational levels than their counterparts. Using this method to control for endogeneity, Bleakley and Chin estimate the magnitude of FP to be roughly 22.5% in OLS and around 34% in 2SLS. The results also reinforce the findings of Dustmann and Van Soest that the downward bias from measurement errors is greater than the upward bias of endogeneity issues. They also found that EP has a positive effect on intermarriage and divorce rates while decreasing fertility and ethnic enclave residence (Bleakley & Chin (2010)).

III. DATA

1. Background on data

The data for the analysis are drawn from the American Community Survey (ACS) of 2013 5-year sample[2] from the Integrated Public Use Microdata Series (IPUMS) database. The self-reported ACS data from IPUMS is reliable and well-organized. Some key

[2] The ACS 2013 5-year weighted sample contains all households and persons from the 1% PRCS sample for 2009, 2010, 2011, 2012, and 2013 identifiable by year.

variables include EP, income, years in the US, birthplace, weekly hours worked, food stamped recipiency, county of residence, immigration status, occupation, educational level, etc. The occupation variable has over 900 identifying codes; fortunately, IPUMS organized these occupation codes into 6 major and relevant categories: management, service, sales, farm, construction, and production. The ACS data allows for flexibility in Hispanic ethnicity identification such that one can identify as ethnically Hispanics and racially white. Based on convention and for the ease of analysis, any immigrant who identifies as Hispanics have been assigned to the Hispanic group exclusive of Caucasian, African descent, and Asian.

A binary variable for English fluency was created. Those who reported their English fluency level as "Does not speak English" or "Yes, but not well" were assigned 0 for their fluency binary while those who reported "Yes, speaks only English", "Yes, speaks very well", "Yes, speaks well" were assigned 1. The assignment of "Yes, but not well" into the not fluent category is due to the overall tendency to overreport EP levels.[3]

The age at arrival variable was created by subtracting years in the US from current age. From age at arrival, a categorical variable age arrived was created. Age arrived divides immigrants into categories based on their age at arrival in increasing order. The categories are childhood migrants who arrived before their 12^{th} birthday, young migrants who arrived between their 12^{th} and 18^{th} birthday, young adult migrants who arrived between their 18^{th} and 30^{th} birthday, adult migrants who arrived between their 30^{th} and 40^{th} birthday, and middle age migrants who arrived after their 40^{th} birthday. After subseting the variable, age arrived becomes a binary variable.

[3] Dustmann & Van Soest (2002).

An indicator variable for the immigrant's birthplace linguistic background was created to identify whether immigrants came from a non-English-speaking country or an English-speaking country. The classification of origin is based on The World Factbook by the Central Intelligence Agency of 2019, linguistic texts, and country's population census. Countries where over 60% of its population speaks English are identified as English speaking while those with under 60% are identified as non-English speaking. For example, the Philippines, Sweden, and Jamaica are classified as English speaking countries due to their high percentage of English speakers.

The sample is restricted to California, Florida, New Jersey, New York, and Texas where there are high numbers of existing and recent immigrants from diverse backgrounds. Further, age is restricted to immigrants between 20 and 62 years old because labor and socioeconomic measures will be well-observed while eliminating young students and retirees. Citizenship status is restricted to "Not citizen", "Naturalized citizen", or "Born abroad of American parents". Given these restrictions, the sample size of all immigrants across 5 states is 815,125 observations for the initial analysis. After restricting the initial sample to only child migrants, the sample size of immigrants across 5 states is 304,628.

2. *Descriptive statistics and data visualization*

Table 1: Descriptive statistics for full sample of immigrants in 2015

	Race groups			
Variables	Caucasian	Hispanics	Asian	African descent
Income	53900	22900	46000	32200
	(85400)	*(35200)*	*(66500)*	*(42900)*
English profienciency	0.93	0.59	0.82	0.94
	(0.253)	*(0.411)*	*(0.257)*	*(0.383)*
Education	8.32	5.26	8.23	7.62
	(2.41)	*(2.94)*	*(2.8)*	*(2.42)*
Weekly hours worked	31.8	29.6	30.3	30.9
	(20.3)	*(19.2)*	*(19.5)*	*(18.3)*
Food stamp recipiency	0.076	0.215	0.071	0.069
	(0.265)	*(0.411)*	*(0.257)*	*(0.383)*
Age at arrival	21.5	20.6	24	23.3
	(13.3)	*(11)*	*(11.6)*	*(11.6)*
Observations	27530	82413	46744	11869

Note: Standard deviations are in parentheses.

Table 1 provides descriptive statistics for all immigrants in the selected sample by 4 major race groups (Caucasian, Asian, African descent, and Hispanics) for 2015. There are some significant gaps in the mean of key socioeconomic variables that are pertinent to this analysis between immigrant groups. In 2015, the fluency gap between Hispanic immigrants and their Caucasian and Asian counterparts is roughly 34% and 23%, respectively. The ACS education variable is scaled such that 3-6 each indicates the high school class level completed with 6 being the completion of 12^{th} grade or the attainment of a high school degree. Subsequently, an increase of 1 in the education variable after 6 is equivalent to completing 1 more year of college. In the full sample, Hispanic immigrant workers are lower skilled than their Caucasian and Asian counterparts based on the self-

reported education level variable. On average, Hispanic immigrants are .74 year short of finishing high school; meanwhile, Caucasian, Asian, and African descent immigrants have received at least 1 year of college education. Hispanic immigrants earned about 43% and 49% of what their Caucasian and Asian counterparts earned, respectively. Consequently, Hispanics immigrants are 13% more likely than their Caucasian and Asian counterparts to be food stamp recipients.

Table 2: Descriptive statistics for recent immigrants in 2015

	Race groups			
Variables	Caucasian	Hispanics	Asian	African descent
Income	39850	16600	26950	15660
	(75160)	*(34100)*	*(48730)*	*(34640)*
English profienciency	0.865	0.405	0.754	0.816
	(0.342)	*(0.491)*	*(0.431)*	*(0.388)*
Education	8.65	6.07	8.6	6.79
	(2.37)	*(3.13)*	*(2.69)*	*(2.6)*
Weekly hours worked	26.4	26.3	22.3	22.4
	(21.6)	*(20.2)*	*(20.8)*	*(20.2)*
Food stamp recipiency	0.092	0.219	0.051	0.195
	(0.29)	*(0.413)*	*(0.221)*	*(0.397)*
Age at arrival	33.3	33.3	32.6	33
	(10.8)	*(11)*	*(11)*	*(11)*
Observations	3071	4904	5450	1110

Notes: The sample has been restricted to immigrants who arrived in America between 2011-2015. Standard deviations are in parentheses.

[4]

[4]The means of age of arrival is likely to be overstated for recent immigrants due to the exclusion of those due to the exclusion of those under 20 years old in sample selection process.

Table 2 examines descriptive statistics for immigrants who have been in the US for less than 4 years by race groups for in 2015. The gaps in these characteristics widen among race groups of recent immigrants. The difference in income persists among recent immigrants such that Hispanic immigrants earn about 42% and 61% of their Caucasian and Asian counterparts, respectively. The flow of recent Hispanic immigrants are, on average, low-skill workers who tend to have lower EP and education levels; in contrast, recent Caucasian and Asian immigrants are higher skill with higher education and EP levels. From table 1 and 2, there is high heterogeneity in socioeconomic status among race groups for both existing and recent immigrants.

Table 3: Descriptive statistics for early immigrants in 2015

	Race groups			
Variables	Caucasian	Hispanics	Asian	African descent
Income	50490	26200	49220	33060
	(76600)	*(36800)*	*(68730)*	*(42960)*
English profienciency	0.985	0.78	0.942	0.988
	(0.128)	*(0.431)*	*(0.247)*	*(0.119)*
Education	8.16	5.73	8.41	7.62
	(2.26)	*(2.71)*	*(2.32)*	*(2.11)*
Weekly hours worked	32.3	30.9	31.7	30.6
	(20)	*(18.9)*	*(18.8)*	*(18.4)*
Food stamp recipiency	0.0733	0.201	0.0732	0.183
	(0.262)	*(0.403)*	*(0.263)*	*(0.385)*
Age at arrival	6.8	9.97	9.6	10.2
	(6.12)	*(5.86)*	*(5.66)*	*(5.41)*
Observations	10437	35145	13443	4106

Notes: The sample has been restricted to immigrants who arrived in America before their 18th birthday. Standard deviations are in parentheses.

Table 3 presents the descriptive statistics for the childhood immigrant restricted sample for 2015. The differences in socioeconomic characteristics among races groups decreases moderately. Specifically, early Hispanic immigrants earn 52% and 53% of what Caucasian and Asian immigrants earned. Further, differences in English proficiency decreased significantly in this sample; meanwhile, the gap in educational attainment between Hispanics and the other 3 immigrant groups remains consistent in this sample. These results suggest strong implications of age at arrival on socioeconomic outcomes of immigrants. Therefore, in the empirical analysis, the effect of EP on wage is analyzed separately for early immigrants.

Table 4: Descriptive statistics by fluency

	NOT PROFICIENT	PROFICIENT
Age at arrival	26.58	20.15
	(10.64)	*(11.63)*
Sex	0.45	0.49
	(0.5)	*(0.5)*
Age	44.38	42.25
	(10.87)	*(11.48)*
Education	4.29	7.62
	(2.89)	*(2.73)*
Weekly hours worked	25.97	31.77
	(20.15)	*(18.94)*
Wage income	15235.43	42229.78
	(24640.25)	*(64052.62)*
Food stamp recipiency	0.25	0.11
	(0.43)	*(0.32)*
Caucasian	0.04	0.2
	(0.2)	*(0.4)*
Hispanics	0.75	0.38
	(0.43)	*(0.49)*
African Descent	0.02	0.09
	(0.13)	*(0.28)*
Asian	0.19	0.3
	(0.39)	*(0.46)*

Note: Standard deviations are in parentheses.

Table 4 displays the difference in immigrant characteristics by EP levels. There is a drastic difference in Age at Arrival between the Not Proficient and Proficient groups. The result reinforces findings in previous studies that younger immigrants tend to improve on EP much quicker and better than older immigrants[5]. The difference in weekly hours between the two groups suggests that EP has strong implication on immigrant labor supply. Table 4 implies that these characteristics should be in the controls when examining the effect of fluency on wages.

Table 5: Fluency for immigrants by race among education group for 2013 (full sample)

	Caucasian	Hispanics	Asian	African descent
College grad	.971 (.1672)	.846 (.3606)	.94 (.2358)	.982 (.1331)
No college	.889 (.3142)	.534 (.4987)	.676 (.4679)	.917 (.2761)

Note: Standard deviations are in parentheses

Table 5 examines heterogeneity in English fluency by education groups for the full sample of immigrants. Among non-college immigrants, there are drastic differences in fluency for Asian and Hispanic immigrants. With college graduates, English fluency is similar with smaller variations across race groups. The low fluency trend in Hispanic immigrants continue for high skill workers as well.

Figure 1 and table 6 in the appendix point out the heterogeneity in English fluency by education groups for the childhood restricted immigrants. Among Caucasian and African immigrants, English fluency is similar in both non-college and college graduate at roughly

[5] Chiswick (1991).

95% or above. On the other hand, for Asian and Hispanic immigrants, there is a large difference in fluency among education groups. The gaps are 24.2% and 9% for Hispanics and Asian immigrants, respectively.

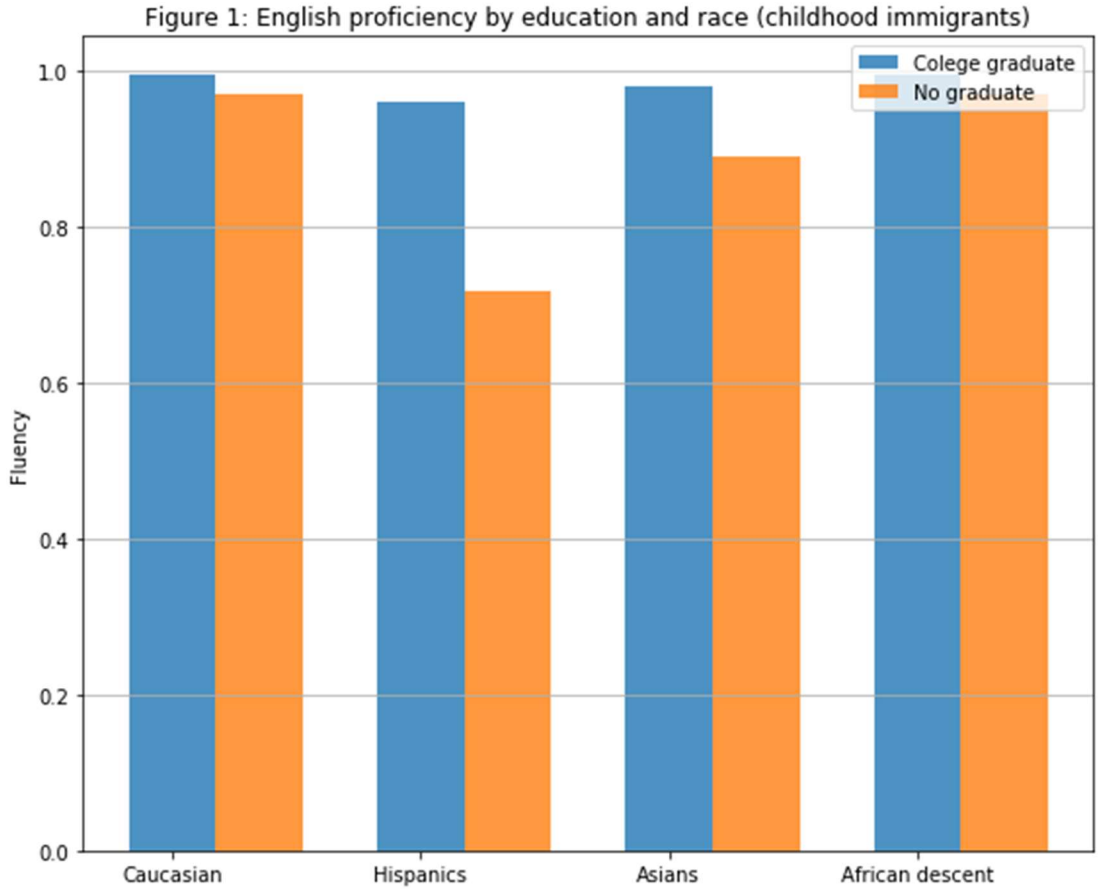
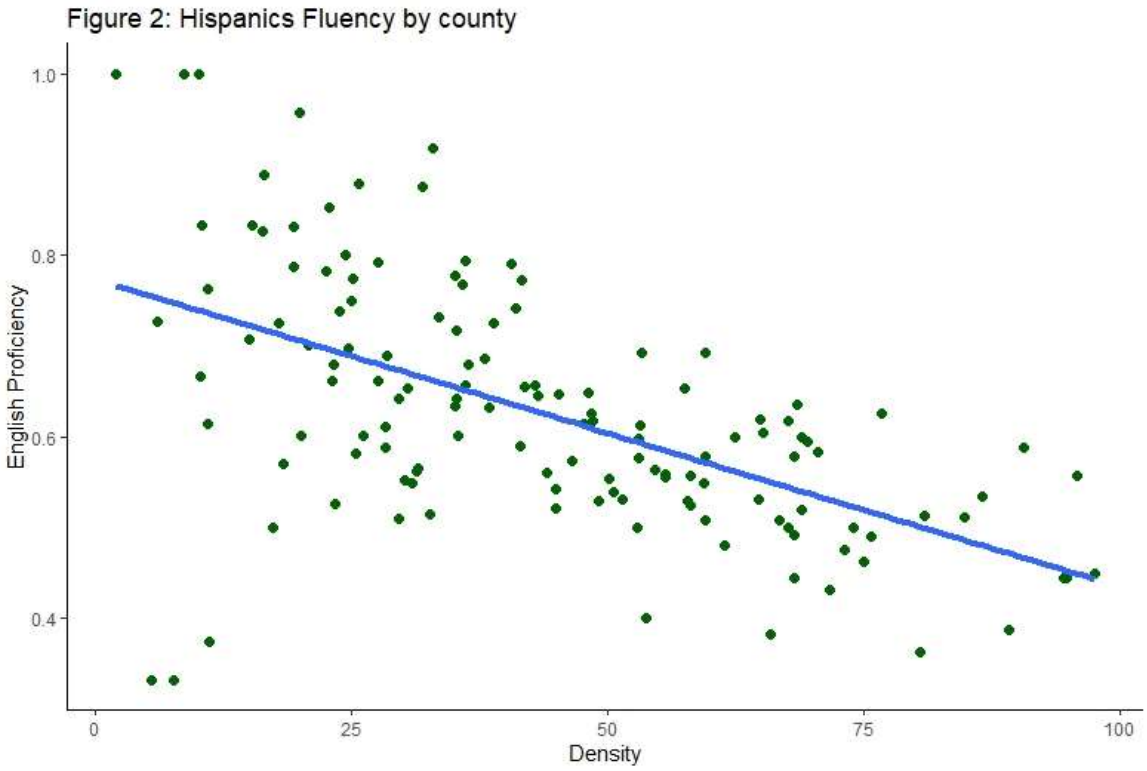


Figure 2 – 3 examine the relationship between county density and overall FP for Hispanic and East Asian immigrant groups. County immigrant density is created by dividing the number of immigrants from a race group by the total number of immigrants in a specific county, and EP is the mean of the binary fluency variable of the individuals

within an immigrant group for that county. As expected from the existing literature, there is a negative correlation between density and EP for Hispanics immigrants in figure 2.[6]



The same analysis of the effect of density on fluency for Asian immigrants is more complex. Asian immigrants should not be treated as a homogenous group because within this group, there is a large level of cultural and linguistic heterogeneity. The 4 largest Asian immigrant groups in the United States (Chinese, Indian, Filipinos, and Vietnamese) originate from 4 different linguistic backgrounds with varying home language linguistic differences from English. Also, Filipinos and Indian immigrants should be excluded from this analysis due to their high level of EP in their native countries.

Figure 3 shows the negative correlation between EP and density among East Asian immigrants. The negative correlation was not apparent before omitting Filipinos and Indian

[6] McManus (1990).

immigrants from the sample. It suggests that there could be complications in the analysis if English ability if birthplace is not in the controls. The visualization of the relationship between density and EP points out the potential issues in the analysis of fluency for Asian without effective control variables for English speaking origin.

.

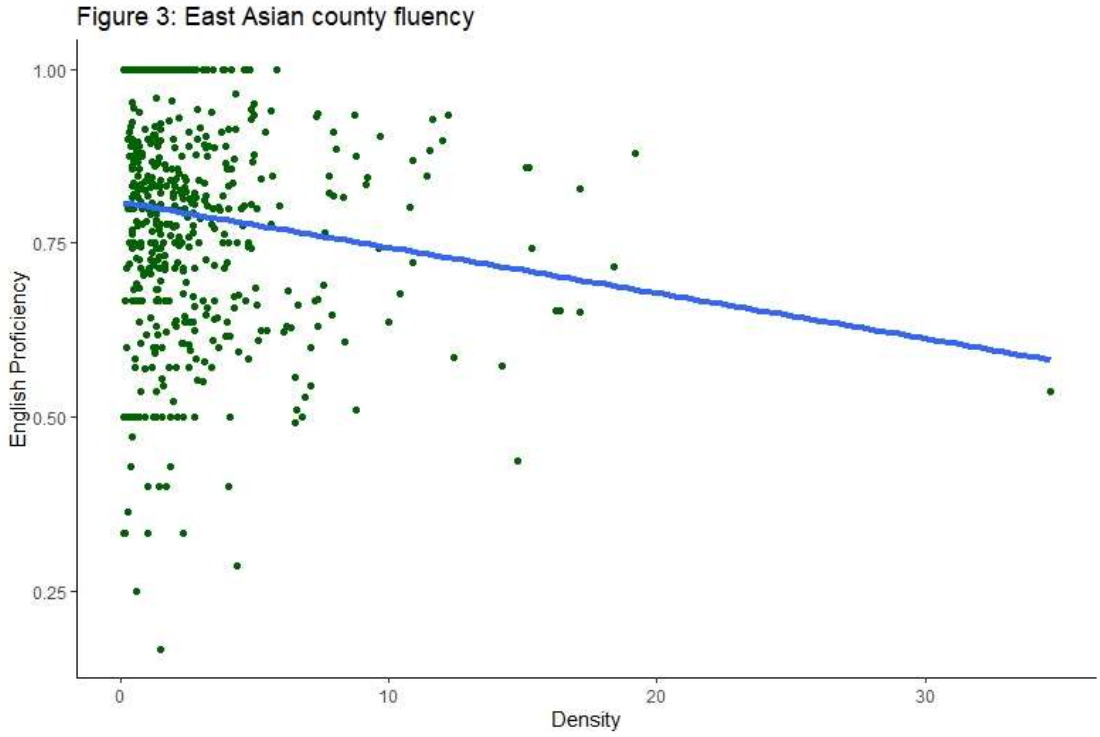


Figure 4 – 6 presents the heterogeneity in fluency within county lines in California among immigrant race groups. In California, EP for Hispanics and Asian immigrants is lower for the Central Valley possibly due to its farming industry and for Southern Coastal areas possibly due to higher levels of immigrant density. Overall, Asian and Caucasian immigrants have higher EP levels than Hispanic immigrants for all counties within California. From this map, there are some county or regional variations possibly due to industrial, infrastructural, and geographical differences or varying levels of immigrant

density that could have a within county effect on fluency. Therefore, a county fixed effect is included to control these unobserved factors that vary across counties.

Figure 4: California Hispanic fluency

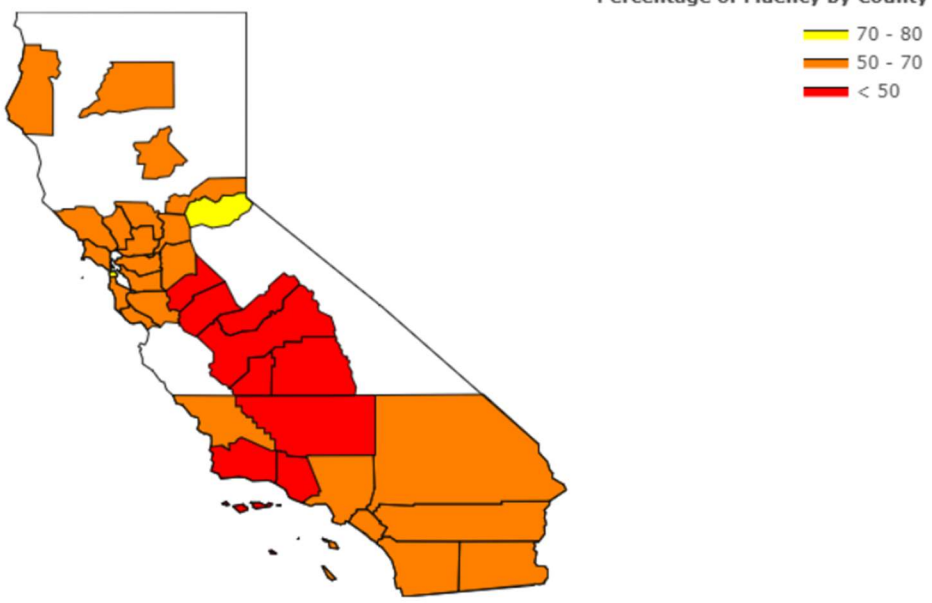


Figure 5: California Asian fluency

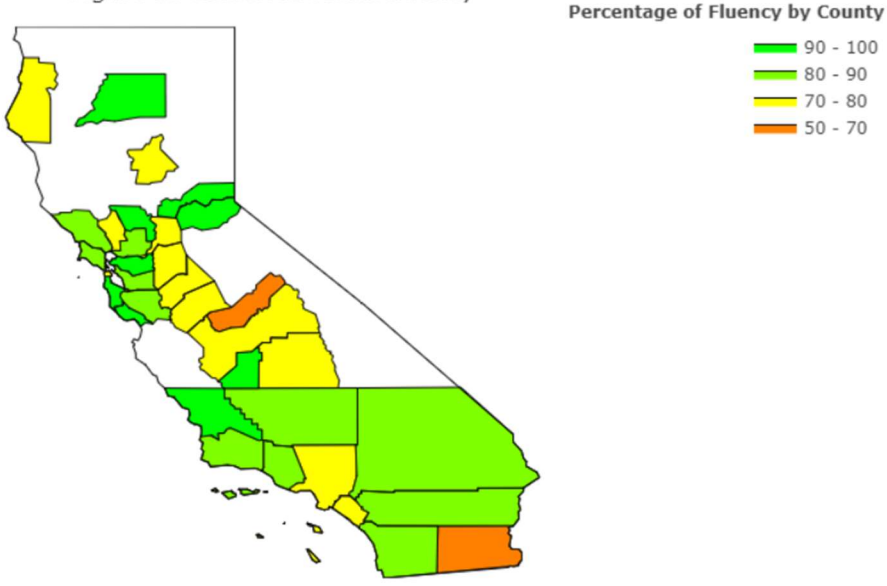
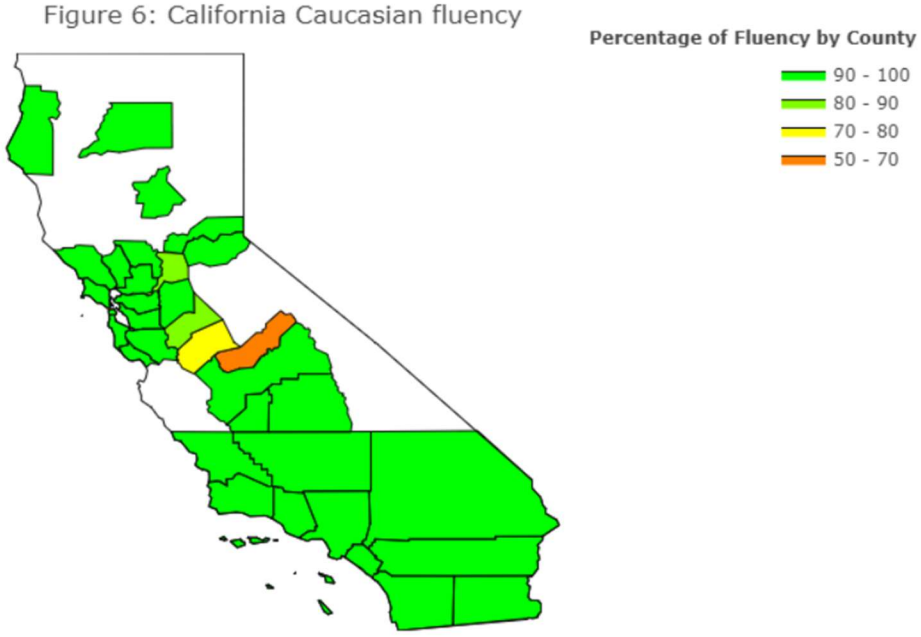


Figure 6: California Caucasian fluency



IV. EMPIRICAL STRATEGY

The first phase of the analysis is an OLS regression of log of wage on fluency, race indicator, and other control variables to obtain a baseline result.

Model 1:

$$log\ INCWAGE_{ict} = \beta_0 + \beta_1 FLUENCY_i + \beta_2 RACE_i + \beta_j OTHER_{ji} + u_i$$

In this model, *i*, *c,* and *t* index individual immigrants from a specific county during a specific year. *Log WAGE* is the natural log of wage income. $FLUENCY_i$ is the binary fluency variable, and R_i is the race identifying variable for 4 major immigrant groups. $OTHER_j$ are the various control variables such as age, gender, hours worked, education, age at arrival, origin, and years in the US. From the above data analysis, the model includes a county fixed effect to eliminate differences across county lines because there are

endogeneity issues with immigrants' tendency to enclave[7]. Furthermore, the model includes indicator variables for occupation to control for occupational differences in wages. The 2013 ACS 5-year is a 5-year linked data sample from 2009-2013, so the model includes year fixed effects.

To analyze heterogeneity in the FP, interaction terms between race groups are incorporated into the model. In this case, C_i, A_i, H_i, and AD_i indicates whether the immigrant is Caucasian, Asian, Hispanics, or African descent, respectively. The analysis of the heterogeneity of the FP incorporates year, occupational, county fixed effects, and similar controls as mentioned in the previous model.

Model 2:

$$log\ INCWAGE_{ict} = \beta_0 + \beta_1 F_i + \beta_2 C_i + \beta_3 C_i{*}F_i + \beta_4 A_i + \beta_5 A_i{*}F_i + \beta_6 H_i + \beta_7 H_i{*}F_i +$$

$$\beta_8 AD_i + \beta_9 AD_i{*}F_i + \beta_j OTHER_{ji} + u_i$$

Holding race and all else constant, the FP for an immigrant from a certain race group is the incremental payment of being fluent and marginal effect of being fluent interacted with his/her race. Therefore, the return of FP in a subgroup and its significance is determined by the linear combination between fluency and the interaction term between race of a group and fluency. For example, the FP for an Asian immigrant would be $\beta_1 + \beta_5$.

Education can mediate the return to fluency, so the third model in each phase of the analysis allows the difference in FP among race to be different among education groups by introducing a third interaction term between fluency, race, and education. The education

[7]McManus (1990) & Chiswick (1991)

level is determined by an indicator variable of whether the immigrant holds a college degree. The third model also includes year, job, and county fixed effect.

Model 3:

$$log\ INCWAGE_{ict} = \beta_0 + \beta_1 FLUENCY_i + \beta_2 COLLEGE_i + \beta_3 RACE_i +$$

$$\beta_4 COLLEGE_i * RACE_i + \beta_5 RACE_i * FLUENCY_i + \beta_6 RACE_i * FLUENCY_i * COLLEGE_i$$

$$+ \beta_7 FLUENCY_i * COLLEGE_i + \beta_j OTHER_{ji} + u_i$$

Holding race and all else constant, the FP of an immigrant without a college degree from a specific race group is determined by the linear combination of $\beta_1 + \beta_5$. The result reflects the difference in earnings of an immigrant without a college degree within a specific race group with respect to fluency. Meanwhile, the FP of an immigrant with a college degree from a specific race group is determined by the linear combination of $\beta_1 + \beta_5 + \beta_6 + \beta_7$. The result reflects the difference in earnings of immigrants with a college education within a specific race group with respect to fluency.

For stage 2 of the analysis, the sample will be restricted to only immigrants who arrived in the US before their 18th birthday. Subsetting the sample potentially eliminates some endogeneity issues in adult migration for employment, fluency at arrival, etc. because children often do not make the choice to migrate to the United States. The three previously established models in stage one will be used to analyze the FP for comparability and consistency in the childhood migrant sample.

The final model to estimate the magnitude and heterogeneity of EP on wage is an IV regression similar to Bleakly and Chin's approach to deal with endogeneity issues. Since $FLUENCY_i$ is endogenous,

$$FLUENCY_{ict} = \gamma_0 + \gamma_1 LATE_i + \gamma_3 LATEORG_i + \gamma_j OTHER_{ji}$$

$LATE_i$ is the indicator for whether an immigrant arrives before or after their 12th birthday, and $LATEORG_i$ is the interaction variable between $ORIGIN_i$ and $LATE_i$. Late arrival is excluded from the final stage because years in US is included in the final regression. The IV strategy also includes county, year, occupational fixed effects, and previous controls to remain consistent.

$$log\ INCWAGE_{ict} = \beta_0 + \beta_1 FLUENCY_i + \beta_2 RACE_i + \beta_j OTHER_{ji} + u_i$$

The interaction terms between fluency and race will be introduced to allow the difference among immigrant race groups. The final stage of the analysis will also include year, occupation and county fixed effect. Among other things, gender, hours worked, and education will also be included in the analysis as control variables.

$$log\ INCWAGE_{ict} = \beta_0 + \beta_1 FLUENCY_i + \beta_2 RACE_i + \beta_3 RACE_i * FLUENCY_i + \beta_j OTHER_{ji} + u_i$$

V. RESULTS

1. Phase 1 with full sample

Table 7 displays the results of the OLS regression with the full sample of log wage on fluency while controlling for race and other key variables. Hours worked was omitted from the first model to determine the relationship between the hours worked and fluency. Origin is a binary variable that determines whether an immigrant comes from a non-English

or English-speaking country (non-English = 1). In the second model, it is clear that there is a positive correlation between hours worked and fluency which suggests that fluency has a strong effect and positive effect on labor supply.

Table 7: Pooled OLS regression with all immigrants (full sample)

	(1) log(wage)	(2) log(wage)	(3) log(wage)	(4) log(wage)	(5) log(wage)
Fluency	0.163***	0.198***	0.145***	0.146***	0.142***
	(0.0034)	(0.0030)	(0.0031)	(0.0031)	(0.0031)
Hispanics	-0.083	-0.066	-0.075	-0.074	-0.059
	(0.0039)	(0.0035)	(0.0035)	(0.0035)	(0.0036)
Asian	-0.028	-0.032	-0.011	-0.011	-0.008
	(0.0039)	(0.0035)	(0.0035)	(0.0035)	(0.0034)
African descent	-0.096	-0.063	-0.050	-0.050	-0.036
	(0.0057)	(0.005)	(0.0050)	(0.0050)	(0.0051)
Male	-0.084	0.222	0.226	0.226	0.227
	(0.0056)	(0.0024)	(0.0024)	(0.0023)	(0.0023)
Education	0.068	0.058	0.060	0.060	0.060
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Age	0.015	0.016	0.012	0.015	0.011
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)
Hours worked		0.038***	0.038***	0.038***	0.038***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Years in US			0 .007***	0.009***	0.005***
			(0.0001)	(0.0003)	(0.0001)
Age arrived				0.014***	-0.212***
				(0.0034)	(0.0034)
Origin					-0.062***
					(0.0039)
Constant	9.075	7.78	7.83	7.78	8.46
	(0.0117)	(0.0110)	(0.0110)	(0.0115)	(0.0184)
Observation	572,148	572,148	572,148	572,148	572,148
Number of groups	143	143	143	143	143
Overall R-squared	0.2896	0.4267	0.4300	0.4333	0.4320
R-squared	0.5889	0.7256	0.6913	0.6957	0.6860
F-test	9276.54	17401.05	16975.68	14227.48	15688.03

Note: The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001

Further, the third model suggests that there is a negative correlation between fluency and age arrived. The final estimates of the FP in stage 1 remains robust and statistically significant after the 3rd model at around 14.3%. Holding all else constant, an immigrant who is fluent in English makes 14.2% more than their non-fluent counterparts. The final

estimate of FP in column 5 of table 7 is 35.5 percentage or 8.3 percentage point lower than Bleakley and Chin (2004) 22.5% OLS estimate.

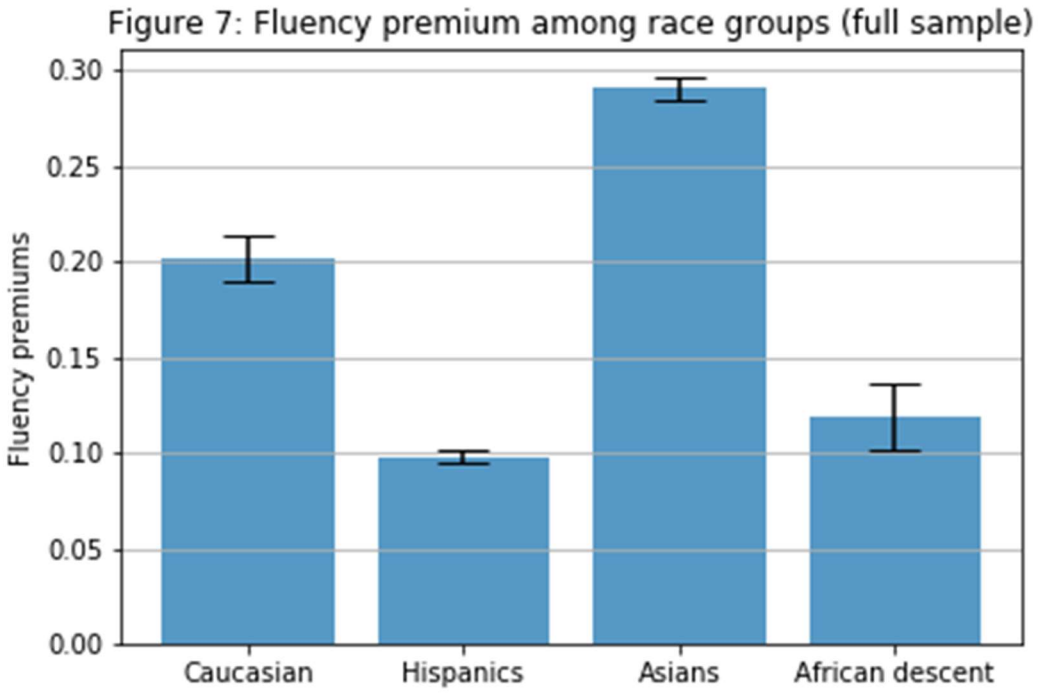


Table 8 in the appendix and figure 7 reports the FP among the 4 major immigrant race groups. The results in table 6 is derived by using the final model of table 5 to control for key variables and preserve consistencies. The FP for race groups and their significance are calculated using linear combinations of the coefficient of fluency and the interaction term between fluency and race. For example, holding all else constant, an Asian immigrant who is fluent will earn 29% more than a non-fluent Asian immigrant. The FP and the difference in FP for Caucasian, Hispanic, Asian, and African immigrants are statistically significant. The results in table 6 suggest that, for the general immigrant, the importance of fluency on wage depends on race.

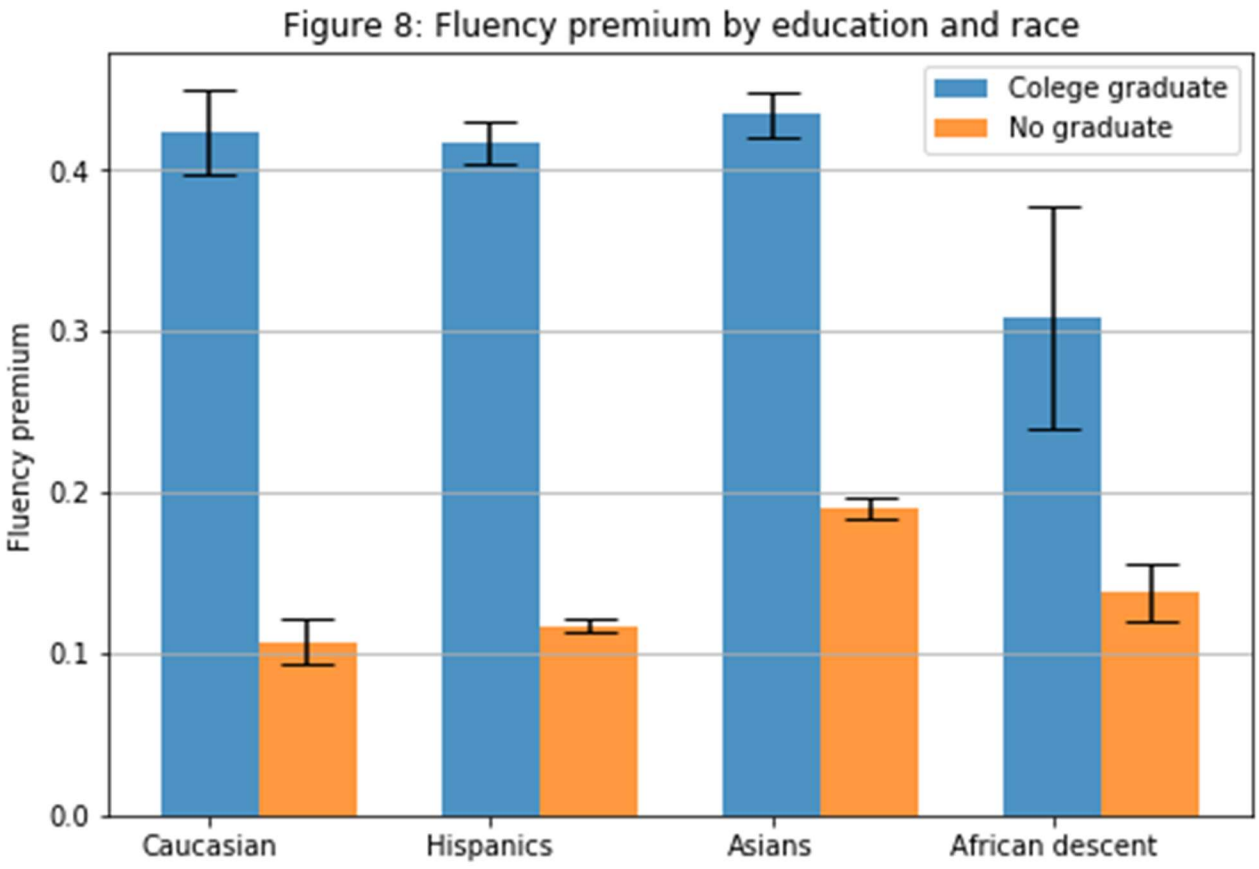


Table 9 of the appendix and figure 8 present the results of allowing the FP to be different among race and education groups. The FPs for college graduates are determined by the linear combination of the coefficient of fluency, interaction between fluency and race, interaction between fluency and university, and interaction between fluency, race, and university. For example, holding all else constant, a university educated English-fluent Hispanic immigrant will earn 41.7% more than a university educated non-fluent Hispanic immigrant. For college graduates, within the same race group, the FP is large and significant; however, across race groups, the FP is similar for Caucasian, Hispanics, and Asian immigrants but sharply lower for immigrants of African descent.

Meanwhile, the FPs for non-college/low skilled immigrants are calculated with the linear combination of the coefficient on fluency, and the interaction term between fluency

and race. For example, an English-fluent Caucasian immigrant without a college degree will make 10.7% more than a non-fluent Caucasian immigrant without a college degree. Among immigrants without a college degree, the FPs are significantly lower than those of college graduates, but there is more heterogeneity across race groups. Namely, the FPs for both Asian and African descent immigrants are higher than those of Caucasian and Hispanic immigrants. The results suggest that the FPs are large and consistent across race groups for high-skill workers, but they are smaller in magnitude but have more variance across race groups for low-skill workers.

2. *Phase 2 with restricted child immigrants*

Table 10: Pooled OLS regression with childhood immigrants

	(1) log(wage)	(2) log(wage)	(3) log(wage)	(4) log(wage)	(5) log(wage)
Fluency	0.05***	0.054***	0.057***	0.057***	0.056***
	(0.0066)	(0.0057)	(0.0058)	(0.0058)	(0.0058)
Hispanics	0.012	-0.004	-0.005	-0.006	-0.003
	(0.0063)	(0.0055)	(0.0055)	(0.0055)	(0.0057)
Asian	0.043	0.054	0.051	0.051	0.051
	(0.0068)	(0.0059)	(0.0060)	(0.0060)	(0.0060)
African descent	-0.039	-0.013	-0.016	-0.016	-0.015
	(0.0096)	(0.0083)	(0.0084)	(0.0084)	(0.0085)
Male	0.363	0.187	0.187	0.187	0.187
	(0.0043)	(0.0038)	(0.0038)	(0.0038)	(0.0038)
Education	0.088	0.075	0.075	0.075	0.075
	(0.0010)	(0.0009)	(0.0009)	(0.0008)	(0.0009)
Age	0.032	0.024	0.025	0.026	0.026
	(0.0004)	(0.0002)	(0.0004)	(0.0004)	(0.0006)
Hours worked		0.041***	0.041***	0.041***	0.041***
		(0.0002)	(0.0002)	(0.0002)	(0.0002)
Years in US			-0.001**	-0.002***	-0.002***
			(0.0003)	(0.0006)	(0.0006)
Age arrived				-0.017**	-0.018**
				(0.0071)	(0.0053)
Origin					-0.009
					(0.0065)
Constant	9.08	7.32	7.83	7.31	8.5
	(0.0117)	(0.0169)	(0.0110)	(0.0173)	(0.0187)
Observation	223,036	223,036	223,036	223,036	223,036
Number of groups	143	143	143	143	143
Overall R-squared	0.2822	0.4438	0.4439	0.4439	0.4439
R-squared	0.4989	0.6029	0.6054	0.6054	0.6064
F-test	3610.01	7487.04	7175.52	7175.52	6888.85

Note: The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001

Table 10 displays the results of the pooled OLS regression of log wage on fluency in the childhood sample. After subsetting the sample, the robust estimate of the FP is around 5.7% for childhood immigrants; therefore, holding all else constant, an immigrant who is fluent in English makes 5.7% more than their non-fluent counterparts. Although the restricted sample reduced the estimate of the FP, the estimates are still statically significant. The return to education increases significantly which suggests that education has stronger implications on young immigrants than the previous full sample.

The restriction of the sample reflects a sharp decline in the FP. The change in the magnitude of the FP is potentially due to the varying effects of fluency at arrival on EP outcomes. Childhood migrants are more affected by fluency at arrival than adult migrants because it has strong implication on education decision and outcome among a myriad of other factors which in turn effect EP and the FP. Thus, fluency at arrival is crucial for childhood immigrants, but the OLS estimate of the FP in this sample does not account for how EP is affected by fluency at arrival. Therefore, an IV strategy is necessary to avoid endogeneity issues with fluency at arrival. Instrumenting fluency with age arrived and the interaction term between age arrived and origin considers the effect of arriving late and the marginal impact of arriving late from a non-English speaking country on fluency. This IV strategy accounts for factors at arrival that affect fluency.

Table 11 in the appendix and figure 9 examines heterogeneity in the FP among race groups. Again, the results in table 9 are derived by using the final OLS of table 8 to control for key variables and preserve consistencies. Similarly, the FP for race groups and their significance are calculated using linear combinations of the coefficient of fluency, race,

and the interaction term. The FP for Asian immigrants remains the highest among the 4 groups at 19.6% which is statistically significant. With the restricted sample, only the FP for Asian immigrants is statistically significant.

Figure 9: Fluency premium among race groups (childhood sample)

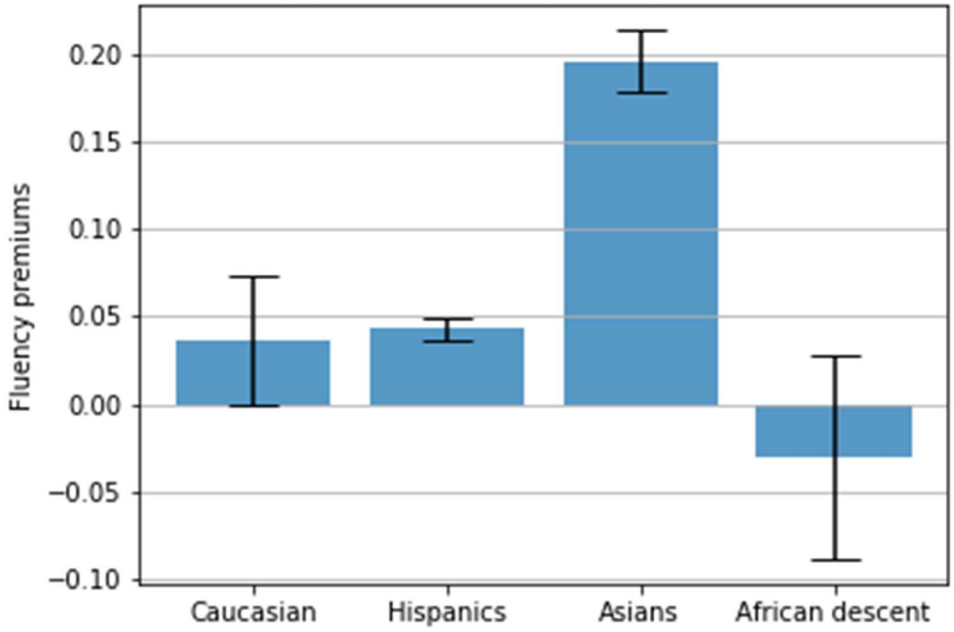


Table 12 displays the results of allowing heterogeneity in the FP in levels of education among race groups after restricting the sample. Among high skill workers, the FP for Hispanic immigrants is significantly larger than the FP for any other groups at 43.4%; meanwhile, for Asian immigrants, it is around 17%. The estimate for Caucasian and African descent immigrants are not significant because in both of these groups, immigrants have very high levels of FP across all education levels as seen in table 1.

For low skill workers, FPs for 8.7% and 15.1% for Hispanics and Asian immigrants, respectively. In this group, the FP among race for Asian and Hispanic immigrants are similar to the full sample estimate of table 9 and figure 8.

Table 12: Fluency premium difference among groups race groups for education childhood

	Caucasian	Hispanics	Asian	African descent
College grad	0.051	0.434***	0.170***	0.166
	(0.0980)	(0.0378)	(0.0435)	(0.1788)
No college	0.043	0.087***	0.151***	-0.027
	(0.0395)	(0.0063)	(0.0194)	(0.0612)

*Note: Education, Age, Years in USA, Origin, Age at arrival, Hours worked, and sex were controlled for in this model. The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001*

3. *Phase 3 with restricted child immigrant sample and IV strategies*

Table 13 presents the results of the IV regression. The age arrived variable becomes binary with the restricted sample. Some of the endogeneity issues in EP can be resolved by 2SLS. Since children tend to not make migration decision, restricting the sample to only child immigrants solves some of the endogeneity issues. Age at Arrival and Origin have strong implications on fluency because immigrants who arrived later in their lives from a non-English speaking country tend to have lower fluency. Furthermore, the interaction term between Age arrived and Origin allows for the extra effect of arriving later from a non-English speaking country. When Age Arrived, Origin, and the interaction term between Age Arrive and Origin were used to instrument fluency for the 2SLS regression of log wage on fluency, the FP is around 25.4% and statistically significant. In the second model, origin was included in the final stage of the 2SLS regression of log wage on fluency. Age arrived was excluded from the final stage because years in the US were included in the final stage and is a more precise control for log wage. The estimate of the FP in column 7 is around 23.5% and statistically significant. The lower result with the second set of instruments reinforces that there are implications of origin on wage, as seen previously in

tables 5 and 8. The results of 2SLS regression is 25.3 percentage and 10.5 percentage points lower than Bleakley and Chin's 2SLS estimate of 34% for the FP.

Table 13: 2SLS regression with childhood immigrant sample

Instruments	Fluency premium	Standard error
Age arrived, Origin & Age arrived * Origin	0.254***	(0.0758)
Age Arrived & Age *Origin	0.235**	(0.0781)

Note: The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001

The results of applying the heterogeneous return for fluency with IV were inconclusive. Once the interaction terms were applied to the model, the age arrived and the interaction term between age arrived and origin became weak instruments. Therefore, they lack predictive power on fluency and led to inconclusive results on the heterogeneity of the FP.

VI. CONCLUSION

1. OLS estimates of FP magnitude and heterogeneity

The OLS estimates of the FP suggests that the return to fluency for all immigrants is large and significant at roughly 14.5%. Furthermore, the model allowing for heterogeneity implies that there are significant differences in the return to fluency across the 4 major immigrant race groups. The lower FP for Hispanic immigrants is possibly explained by the high linguistic and cultural homogeneity and enclaving tendencies for Hispanic immigrants. These factors decrease the demand and premium for language improvement. Furthermore, Spanish is closer linguistically than the languages of other race groups within the sample, so the marginal cost of language acquisition is lower. Meanwhile, Asian immigrants observe a significantly higher FP than every other group within the sample.

Despite tendencies to enclave, the high estimation of the FP is possibly due to the high level of linguistic and cultural heterogeneity within Asian immigrant groups which increases the demand for EP. In addition, Asian languages are more distant from English linguistically which increase the cost of language acquisition. These factors perhaps combine to drive up the FP as observed by Asian immigrants.

2. Heterogeneity in race and education

In the full sample, the FP among high skill workers is large and significant; however, there is not much variation among race groups in the magnitude of the FP. Meanwhile, for low skill workers, the magnitude of the FP is significantly lower, but there are more variance in the return to fluency.

The differences between the FP in these two education groups can be explained by language fluency and skill complementarity. Labor skill is increasingly complementary with language fluency, as high-skill jobs typically demand high levels of communication with coworkers and clients. Non-fluent but high skilled workers are at a major disadvantage; in contrast, non-fluent low skilled workers are not as severely penalized. Among low-skill workers, heterogeneity in FP can possibly be explained in the similar manner as section 1 of the conclusion.

3. IV strategies

The 2SLS of FP is determined to be 23.5% after including appropriate fixed effects and controls. Despite multiple attempts to allow for heterogeneity, the results were inconclusive. The model failed possibly due to the lack of data classification for the origin variable, and there was not enough data to determine the FP within a race group with 2SLS. For example, African descent were grouped into 4 geographical classification in the ACS

survey, namely, East, West, North, and Central Africa, so it is not possible to correctly identify English speaking origin for these immigrants.

VII. APPENDIX

Table 6: Fluency for immigrants by race among education group for 2013 (childhood restricted sample)

	Caucasian	Hispanics	Asian	African descent
College grad	.994 (.0741)	.958 (.2001)	.979 (.1428)	.993 (.0853)
No college	.970 (.1699)	.716 (.4509)	.889 (.3145)	.968 (.1737)

Note: Standard deviations are in parentheses

Table 8: Fluency premium difference among groups for all immigrants (full sample)

	Race groups			
	Caucasian	Hispanics	Asian	African descent
Fluency premium	0.201*** (0.0119)	0.098*** (0.0035)	0.290*** (0.0061)	0.119*** (0 .0175)
Difference from Caucasian	0	-0.103*** (0.0122)	0.089*** (0.0131)	-0.082*** (0.0211)

*Note: Education, Age, Years in USA, Origin, Age at arrival, Hours worked, and sex were controlled for in this model. The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001*

Table 9: Fluency premium difference among groups race groups for education (full sample)

	Caucasian	Hispanics	Asian	African descent
College grad	0 .424*** (0.0263)	0.417*** (0.0127)	0.434*** (0.0138)	0 .308*** (0.0687)
No college	0.107*** (0.0134)	0.117*** (0.0037)	0.190*** (0.0070)	0.137*** (0.0181)

*Note: Education, Age, Years in USA, Origin, Age at arrival, Hours worked, and sex were controlled for in this model. The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * p < 0.05, ** p < 0.01, *** p < 0.001*

Table 11: Fluency premium difference among groups for childhood immigrants

	Race groups			
	Caucasian	Hispanics	Asian	African descent
Fluency premium	0.036 (0.0367)	0.043*** (0.0062)	0 .196** (.0175)	-0.030 (0.0580)
Difference from Caucasian	0	0.007 (0.0370)	0.161*** (0.0405)	-0.066 (0.0686)

*Note: Education, Age, Years in USA, Origin, Age at arrival, Hours worked, and sex were controlled for in this model. The models include year, occupational, and county fixed effects. Robust standard errors are in parenthesis. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Bibliography

Berman, Eli, Kevin Lang, and Erez Siniver. "Language-Skill Complementarity: Returns to Immigrant Language Acquisition." *Labour Economics* 10, no. 3 (06, 2003): 265-290.

Bleakley, Hoyt and Aimee Chin. "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants." *American Economic Journal: Applied Economics* 2, no. 1 (01, 2010): 165-192.

Bleakley, Hoyt and Aimee Chin. "Language Skills and Earnings: Evidence from Childhood Immigrants." *Review of Economics and Statistics* 86, no. 2 (05, 2004): 481-496.

Butler, Yuko Goto. "How are nonnative-English-speaking teachers perceived by young learners?" *Tesol Quarterly* 41, no. 4 (2007): 731-755.

Chiswick, Barry R. "Speaking, Reading, and Earnings among Low-Skilled Immigrants." *Journal of Labor Economics* 9, no. 2 (04, 1991): 149-170.

Chiswick, Barry R., and Paul W. Miller. "Why is the payoff to schooling smaller for immigrants?." *Labour Economics* 15, no. 6 (2008): 1317-1340.

Derwing, Tracey. "What do ESL students say about their accents?" *Canadian Modern Language Review* 59, no. 4 (2003): 547-567.

Dustmann, Christian and Arthur van Soest. "Language and the Earnings of Immigrants." *Industrial and Labor Relations Review* 55, no. 3 (04, 2002): 473-492.

Grogger, Jeffrey. “Speech and Wages” *Journal of Human Resources* 54, no. 4 (11, 2019): 926-952.

Kossoudji, Sherrie A. "English Language Ability and the Labor Market Opportunities of Hispanic and East Asian Immigrant Men." *Journal of Labor Economics* 6, no. 2 (04, 1988): 205-228.

Labov, William. "The social stratification of English in New York City." (1966).

McManus, Walter, William Gould, and Finis Welch. "Earnings of Hispanic Men: The Role of English Language Proficiency." *Journal of Labor Economics* 1, no. 2 (04, 1983): 101-130.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. *IPUMS USA: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D010.V9.0*