CENTRE FOR ECONOMIC HISTORY
THE AUSTRALIAN NATIONAL UNIVERSITY
DISCUSSION PAPER SERIES

THE ROLE OF ENGLISH FLUENCY IN MIGRANT ASSIMILATION: EVIDENCE FROM UNITED STATES HISTORY

ZACHARY WARD
ANU

DISCUSSION PAPER NO. 2016-06

OCTOBER 2016

THE AUSTRALIAN NATIONAL UNIVERSITY
ACTON ACT 0200 AUSTRALIA
T 61 2 6125 3590
F 61 2 6125 5124
E enquiries.eco@anu.edu.au
https://www.rse.anu.edu.au/research/centres-projects/centre-for-economic-history/

The Role of English Fluency in Migrant Assimilation: Evidence from United States History*

Zachary Ward†
The Australian National University

October 2016

Abstract

I estimate the premium for speaking English and the rate of language acquisition in the early 20th century US using new linked data on over half a million migrants. Compared with today's migrants, early 20th century migrants arrived with much lower levels of proficiency, yet many acquired language skills rapidly after arrival. Learning to speak English was correlated with a small upgrade in occupational-based earnings (2 to 6%); the premium has at least doubled between 1900 and 2010, revealing that English fluency has become an increasingly large barrier to migration over time.

JEL Classification: F22, J24, J61, J62, N31, N32
Keywords: English fluency, language, migrant assimilation

*I would like to thank Brian Cadena, Ann Carlos, Dustin Frye, Tue Gorgens, Tim Hatton, Priti Kalsi, Ling-Yu Kong, Edward Kosack, Martine Mariotti, Xin Meng, Amber McKinney, Julie Moschion, John Tang and Jose Tessada for helpful pointers and discussions. I also thank the audience members at the Australian National University, La Trobe University, the University of Adelaide, the Australasian Cliometric Conference, the Natural Experiments in History Workshop, the EH-Clio conference at Pontificia Univerisidad Catolica, the University of Colorado, the University of Melbourne, and the University of Queensland. Many thanks go to Lee Alston who helped me to gain access to the full-count census data. All errors are my own.

†Email: Zach.A.Ward@gmail.com, Research School of Economics, HW Arndt Building 25A, College of Business and Economics, The Australian National University, Canberra, ACT 2600, Australia.

1 Introduction

Migrants' earnings for the last fifty years of American history have been characterized by two key facts. First, newly arrived migrants earn much less than natives. Second, in the decades after arrival migrants learn United States-specific human capital, which leads to higher earnings and a smaller wage gap (see Figure 1).[1] Today the most important skill to acquire, by far, is the ability to speak English; evidence suggests that the English premium is almost as large as the wage gap between natives and arriving migrants.[2] Indeed, a large English premium, a lack of English fluency at arrival, and the acquisition of English after arrival are consistent with the assimilation patterns observed in the late 20th century.

However, these patterns from the last fifty years are in stark contrast with the assimilation profile of the early 20th century, also shown in Figure 1.[3] Migrants who arrived one hundred years ago held *similarly skilled* jobs to natives at arrival, and they surprisingly had *no improvement* relative to natives after arrival. While similar skills near arrival could be explained by migrants coming from sources with levels of development comparable to the United States, the lack of increase after arrival is particularly surprising because it suggests little to no return for United States-specific human capital − or, in the context of language, little to no benefit for learning how to speak English. Alternatively, migrants could have arrived with already high levels of English fluency or migrants may not have invested in English language skills after arrival.

[1]For the major studies on this phenomenon, see Chiswick (1978), Borjas (1985, 1995; 2015) and Lubotsky (2007).These are not the only reasons for a wage gap at arrival; for example, the pre-migration human capital likely loses value across borders. Further, migrants may have less human capital on average relative to natives, perhaps because they come from relatively poor countries.

[2]Chiswick and Miller (2014, Table 5.4) use the 2005-2009 American Community Survey to estimate a 33 percent penalty for not speaking English relative to speaking English very well, without correcting for endogeneity or measurement error. Bleakley and Chin (2004) use an IV strategy and estimate that improving from speaking no English to speaking English very well increases income by over 100 percent. These results are compared to the migrant/native wage gap, where migrant arrivals earn approximately 50 percent of what natives earn in their first five years of arrival and 25-35 percent less after ten years (Lubotsky, 2007, Table 1).

[3]This assimilation profile is estimated for migrants who arrived between 1880 and 1900 in Abramitzky, Boustan and Eriksson (2014). Note that this assimilation profile is estimated based on occupational scores rather than wages; this is because wage data does not exist in the United States Census prior to 1940.

In this paper, for migrants who arrived at the end of the Age of Mass Migration (1890-1919), I estimate the association between speaking English and occupational upgrading, how many migrants were able to speak English at arrival, and the rate of acquisition after arrival; these values are then used to understand migrant assimilation in the early 20th century. More broadly, these values also help us to understand the relative importance of communication skills over time.

Due to recent advances in data, it is now possible to estimate English acquisition rates more precisely than ever before. Specifically, because of the mass digitization of historical census files, I am able to form new panel data by tracking migrants from census to census. Using full-count census files, I am able to track over 350,000 migrants from 1910 to 1920, and 300,000 migrants from 1920 to 1930.[4] Panel data is necessary to properly estimate the speed of language acquisition; otherwise, inference from following a synthetic cohort may be biased by the selective return of those with poor English skills (Borjas, 1985; Lubotsky, 2007). Using panel data is especially important in the context of the Age of Mass Migration since the return flow was large ($\geq$ 40% of inflows) and highly selective (Bandiera, Rasul and Viarengo, 2013; Ward, 2016); for example, Abramitzky, Boustan and Eriksson (2014) show that repeated cross sections overstate the occupational growth for migrants who arrived between 1880 and 1900. A further benefit of panel data is that when estimating the premium to English, it can be leveraged to reduce bias from unobservables, a key threat to validity.

I show that non-Anglophone migrants who arrived between 1890 and 1919 came with low English skills: only about 29 to 38 percent of permanent migrants were recorded as being able to speak English within one year of arrival. Yet while these migrants arrived with a low level of English proficiency, they had a high rate of language acquisition: within ten years of arrival, English fluency rates increased by 50 percentage points. Importantly, this is found in the panel data, which shows that the fast rate of acquisition is not biased by the selective return of those with lower English skills. However, the repeated cross-sectional data

[4]I also track about 10,000 between 1900 and 1910, which is much less because the digitized version of the 1900 census has not been released yet.

only slightly overestimates the speed of language acquisition in the panel, suggesting that the selection of return migrants on English-speaking ability was mildly negative.

The low level of English fluency at arrival and the sharp increase in the level after arrival contrast with the flat assimilation profile centered at zero in Figure 1.[5] This suggests that gaining language human capital had little to no effect on one's occupation. I use a variety of empirical strategies to verify if this is true. Primarily, I exploit the over half a million migrants in the panel data by estimating an individual fixed effect model; this method eliminates time-invariant individual-specific unobservables that may be correlated with English skills, such as ability. The analysis reveals that the relationship between speaking English and occupational outcomes was weak in the early 20th century, when those who gained English fluency had an associated increase in occupational-based earnings of 2 to 6%. There is evidence that the premium was higher in 1900 and declined throughout the next few decades, consistent with a declining return to skill between 1900 and 1950 (Goldin and Katz, 2008). Alternative empirical strategies, such as instrumenting for English fluency with age at arrival as in Bleakley and Chin (2004, 2008, 2010), lead to the same conclusion that language skills were relatively unimportant for upgrading jobs.

Therefore, a main reason for a flat assimilation profile in the early 20th century is that there was little penalty for not speaking English and little benefit to gaining fluency. Since then and over the past 100 years, migrants' wages relative to natives' have gradually worsened, which could be due to an increased English premium and worse levels of English proficiency. Using the 2000 Census and 2008 to 2012 ACS, I compare how English fluency levels have changed over the past century. While the comparison of English proficiency over time is only suggestive because the variable for English fluency is not recorded in the same manner, a reasonable merging across time shows that early 20th century migrants arrived with much lower levels of basic English skills relative to today's migrants (29-38 percent v. 74 percent).[6]

[5]Note that if one drops English-speaking sources from Abramitzky, Boustan and Eriksson's (2014) sample, the same flat assimilation profile centered at zero holds.

[6]The early 20th century data was recorded by an enumerator as a binary variable (0=cannot speak English, 1=can speak English). The late 20th century data was self-reported on a more qualitative scale (0=speaks no

Yet migrants in the early 20th century caught up to the levels of the late 20th century after 15 to 20 years of duration, when 94% of migrants were able to speak English.

Thus, the widening wage gap between natives and migrants over the last 100 years is not due to migrants are arriving with fewer English skills; rather, one reason is that the English premium has increased over the past century. The *wage* premium for English skills is estimated, using recent data, to be above 20 percent (Chiswick and Miller, 2014), with some estimates much higher (Bleakley and Chin, 2004). A similar OLS regression finds that the *occupational* premium is 12 percent, a number double the results from the early 20th century.[7]

A doubling of the English premium over time is reflective of the supply of and demand for those with English skills. The relative supply of English speakers appears to be increasing since recent migrants are arriving with more English skills than migrants from the past; therefore, a rising English premium is not reflective of a reduced supply of those with English skills. Rather, an increasing premium is likely due to an increasing demand for those with the ability to speak English. In particular, there is robust evidence that technological shifts in the economy have favored those with higher skill levels; these shifts have likely also benefited English speakers as well (Goldin and Katz, 1998, 2008; Lafortune, Tessada and Lewis, 2016). Thus, the technological setting is key for understanding the relative performance of migrants and natives in the early 20th century, just as it is for understanding the outcomes of migrants in more recent decades (Lalonde and Topel, 1992; Lubotsky, 2011; Perlmann, 2005).

English, 1=speaks English not well, 2=speaks English well, 3=speaks English very well). The results in this paragraph treat people who report any English skills (values 1-3) in the late 20th century as able to speak English.

[7]Further, there is evidence that an OLS estimate of the English premium underestimates the true return due to measurement error (Dustmann and Van Soest, 2001), suggesting that the increase between 1900 and 2010 is more than double.

2 Historical Context

Since 1850, there has been a secular decline in the percentage of migrants from English-speaking sources (see Figure 2).[8] For example, about 70 percent of the migrant stock in 1850 were from England or Ireland but soon non-English-speaking countries such as Germany and Norway became major senders to the United States; by 1880, the percentage from English-speaking sources had dropped to 50.[9] The 1880s also marked a turning point for the geographical composition of the flow toward lower-income countries from Southern and Eastern Europe. Hundreds of thousands began to arrive yearly from Italy, Greece and Russia and by 1910, the migrant stock had only 30 percent from an English-speaking source. While many of these migrants learned to speak English after arrival, the newer sources appeared to arrive with less proficiency and pick up English at lower rates than prior European arrivals: for example, the stock's ability to speak English decreased rapidly between 1900 (82%) and 1910 (70%).[10]

The worsening perceived quality of Southern and Eastern European migrants led to a severe nativist backlash against the country's open migration policies, with migrants' decreasing ability to speak English a particularly salient feature for natives to criticize.[11] This view was also held by some of the foremost migration scholars of the time: after analysing data on thousands of migrants, Jeremiah Jenks and W. Jett Lauck, concluded that "the greatest obstacle to a more rapid [assimilation] is that the recent immigrant cannot speak English," (Jenks and Lauck, 1926; pg 335). This statement was reflective of the "Americanization" fervor during the 1910s and 1920s, a movement focused on assimilating migrants through

[8]The major areas where English is the official language but not dominantly spoken include India, Philippines, Hong Kong and several African countries.

[9]The migrants in this article arrived during the Age of Mass Migration (1850-1913), which has been studied extensively, most notably by Hatton and Williamson (1998). See Abramitzky and Boustan (2016) for a recent overview of the literature on historical migration to the United States.

[10]These results are based on a sample of foreign-born individuals over the age of 16, which was drawn from IPUMS. See Figure A1 for the rate of English fluency across the 20th century.

[11]Many began to lobby the government to maintain the "national origins" of the American population – with some success. For example, in 1906, English fluency was added as a requirement for citizenship. However, millions still came.

language instruction for children and adults (Bloch, 1920; Lleras-Muney and Shertzer, 2015).

Despite the heightened focus on English during this time period, the importance of English fluency for occupational upgrading still remains unknown. Jenks and Lauck (1926) argued the importance of English by showing the positive associations of staying longer in the United States, English proficiency, and skill, suggesting that those who stayed longer were able to learn English and upgrade their occupation. However, as is well known today, these results could simply reflect selective return migration of low-skilled non-English-speaking migrants.[12] I improve on Jenks and Lauck's original study and more recent research on English acquisition during this time period (e.g. Vigdor, 2010; Kuziemko and Ferrie, 2014; Jasso and Rosenzweig, 1989) by creating panel data, reducing the influence of bias from selective return migration.

Further, the benefit for learning how to speak English in the context of a developing economy is unclear.[13] The early 20th century economy still had a large focus on brawn rather than brain: instead of being dominated by white-collar jobs, the economy was largely agricultural in midst of a structural shift toward manufacturing and urban centers (Katz and Margo, 2014).[14] The shift away from agriculture coincided with increasing urbanization rates; as population density increased, the tasks performed by workers also changed toward more interaction (Boustan et al., 2013; Michaels et al., 2012). Other technological changes, such as the electrification of the manufacturing sector, led to more positions where managers oversaw operatives in factors (Gray, 2013; Katz and Margo, 2014). Given increased urbanization and

[12]See Abramitzky, Boustan and Eriksson (2014) and Borjas (1985) for a discussion of this problem of estimating the rate of occupational upgrading with repeated cross sections. Kuziemko and Ferrie (2014) show the same set of correlations between length of stay, English fluency and skill using micro-data from IPUMS. See Vigdor (2010) for a nice discussion of the English acquisition of migrants across the early and late 20th century; however, as is noted by the author, the results also may be biased by selective emigration.

[13]There are a few studies who estimate the association between English proficiency and occupational outcomes (Blau, 1980; Lleras-Muney and Shertzer, 2015). Most notably, Jasso and Rosenzweig (1989, 1980) estimate the premium of English and acquisition rates in 1900 for Germans and 1980 for Mexicans. They find a larger "return" for speaking English in the early 20th century, but the interpretation is unclear because the estimated return in 1900 is on occupational prestige while the return in 1980 is estimated for hourly wages.

[14]Despite the large agricultural sector for the entire population, migrants did not most often work in agriculture, though these were jobs that many of the foreign-born did in the source country. Rather, migrants most often worked at manufacturing jobs in urban centers (Wyman, 1993). Besides manufacturing (30%) and agriculture (13%), the foreign-born also worked in retail (16%), construction (8%), and mining (5%). Four of these five industries (excluding retail) did not necessarily require a strong command of English, but rather needed workers who were strong and able to endure long work hours.

the technological advance that favored high-skilled workers, it is likely that the demand for those with English skills was relatively low but increasing rapidly between 1900 and 1930 (Goldin and Katz, 1998; 2008; Lafortune, Tessada and Lewis, 2015). This is consistent with evidence from Canada that the association between English proficiency and wages increased between 1911 and 1931 (Inwood, Minns and Summerfield, 2016).

3 Data

3.1 Measuring English Skills in the Early 20th Century

The main data source on a migrant's English proficiency comes from the 1890 to 1930 United States Censuses. Unfortunately, the government stopped asking about English fluency after 1930, leaving a gap of 50 years until the 1980 Census regathered data on migrants' language skills.[15] Further, the 1890 Census micro-data was lost in a fire in the 1920s, leaving the analysis to the 1900 to 1930 Census.

Between 1900 and 1930, the census was taken by enumerators from door to door, and thus the ability to speak English ("Yes" or "No") was a judgment by the enumerator rather than self-reported as in recent Census data. Enumerators did not have an explicit cut-off point for whether a respondent was able to speak English between 1900 and 1930. The 1890 Census gave instructions to record English fluency based on whether a migrant was "able to speak English so as to be understood in ordinary conversation" − a higher bar than simply knowing a few words. However, this guidance was not in the instructions for the 1900 to 1930 Censuses, leading to a familiar problem of measurement error in language studies since there

[15]One potentially problematic issue regards how the census takers recorded a migrant's fluency (Stevens, 1999). For example, in 1900, three census questions under the broad heading of 'Education' were asked in a row: whether an individual could read, could write, and could speak English. The Census Bureau noted that some census takers simply recorded 'yes' or 'no' three times in a row - this problem was discovered as it appeared that black individuals had low rates of English proficiency when they likely only had low literacy rates (Census Bureau, 1913, page 1265). It is unclear how extensive this problem was for the foreign-born population. By 1910, the census sheets were corrected so as to not have the questions in order, but as in other studies, it is likely that the English variable was measured with error. However, the bulk of the data in this paper are from linked samples between 1910-1920 and 1920-1930, limiting the bias from this problem.

was no objective metric of English proficiency (Bleakley and Chin, 2004; Dustmann and Van Soest, 2001).

To provide a clearer idea of this early 20th century measure of English ability, I compare it to the more studied post-1980 US census data when migrants self-reported whether they spoke English either "not at all", "not well", "well", or "very well."[16] While an objective measure to compare the variables over time is unavailable, one can compare measures by relying on the linguistic hypothesis that second-language acquisition is more difficult at older ages. Importantly, this is thought to be be related to neurobiological changes in the brain prior and during puberty (Singleton, 1999). Evidence for this hypothesis has been well-documented by Bleakley and Chin (2004, 2008, 2010) who show in the 1990 and 2000 US census that migrants who arrived at older ages (≥ 8) had lower levels of English proficiency as adults.

Given that the relationship between age and second language acquisition is thought to be neurobiological, it is likely that it is roughly constant in the one hundred years apart; in other words, the age-at-arrival and English proficiency profile should be similar in early 20th and 21st century. Therefore, I estimate the age-at-arrival profile by regressing English ability on age at arrival and other controls such as country of birth, age, sex and fraction of migrants from their own country of birth in the county; I do this separately for the pooled 1900-1930 Census and the pooled 2000 Census and 2008-2012 ACS.[17]

I plot the estimated age-at-arrival fixed effects in Figure 3 using various codings of the dependent variable in the post-2000 data. For reference, in the early 20th century data a migrant who arrived at age 17 was approximately 8.6 percentage points less likely to speak English as an adult compared with a migrant who arrived at age zero. This profile matches

[16]This was only reported by migrants who spoke a language other than English at home. Since I will group migrants who spoke English "not well" to "very well" in one category, I place migrants who spoke English at home into this category.

[17]The sample for this regression are all migrants from non-English-speaking countries aged 25 to 55 and those who arrived under the age of 17. The regression controls for age, country of birth, sex, cohort of arrival, year and fraction of county from same country of birth. I do not use the 1980 and 1990 Census because they do not record a specific year of arrival, making it impossible to back out a precise age at arrival.

well with the profile for early 21st century data if one codes those who spoke *any* English, whether not well, well or very well as able to speak English. With this coding, a 17-year-old arrival was 9.1 percentage points less likely to speak English compared with a zero-year-old arrival. On the other hand, if one follows the more common method where those who speak English "not well" are instead placed in the unable to speak English group (Chiswick and Miller, 2014), then a 17-year-old arrival would be 30 percentage points less likely to speak English − much too steep of a decline relative to the early 20th century data. Therefore, I interpret the English variable in the early 20th century as reflective of basic English skills, where there was a low bar to clear for being recorded as able to speak English. Further, whenever I compare English proficiency across the early and late 20th century, I will make the assumption that those who self-reported any English ability in the late 20th century would have been recorded as able to speak English in the early 20th century.

3.2 Building New Linked Data

With this measure of English ability, I aim to estimate how many migrants arrived with English skills, the rate at which migrants learned to speak English after arrival, and the return to speaking English. To answer the research question on the speed of language acquisition, I need to create a panel that tracks individual migrant's ability to speak English over time; this is because tracking the same person is important for estimating the rate of language acquisition without bias. The alternative to using a panel is to pool multiple cross-sections together and follow a synthetic cohort's English proficiency over time. However, this leads to the well-documented problem of selective return migration. For example, if one observes an increase in a synthetic cohort's English skills, this could either reflect a true increase or that those unable to speak English returned home (see Abramitzky, Boustan and Eriksson (2014) for a discussion of this bias).[18] This issue is especially important for the Age of Mass

[18]Lubotsky (2007) also discusses another problem of repeated entries, which could influence the assignment of arrival cohort. However, this is only a problem if migrants report differing years of arrival across observations (e.g. whether first or most recent arrival). The 1910 and 1930 Censuses required respondents to

Migration: return migration rates were high (Bandiera, Rasul and Viarengo, 2013), and Abramitzky et al. (2014) have already shown that selective out-migration can bias estimated rates of occupational upgrading.

Therefore, I build a new panel to fix any bias that arises from selective out-migration. To do this, I take individuals first observed in the 1900 5% IPUMS sample, the 1910 full-count Census and the 1920 full-count Census, and then link them ten years later to the 1910, 1920 and 1930 censuses, respectively. Note that I do not use the full-count 1900 Census because the English variable has yet to be digitized. To link the data across years, I find similar matches based on first name, last name, year of birth, country of birth and year of arrival. To find the best match, I use probabilistic methods which have used by Massey (2016), Mill and Stein (2016) and Pérez (2016); this approach aims to improve on Abramitzky et al.'s (2014) matching algorithm in that I keep the best link based on year of birth, year of arrival and string similarity for first and last name, as opposed to only relying on differences in year of birth. These methods are described further in the Appendix B.

For each of the base samples, I do not link forward the entire set of migrants. Primarily, I drop migrants from English-speaking countries such as England, Ireland and (English) Canada because I am interested in how non-native speakers acquired human capital after arrival. Second, I only keep migrants who arrived within the past ten years in order to track specific migrant cohorts over time; I drop those who arrived in the same year as the Census (e.g. 1900 arrivals in 1900) since the Census does not cover the entire year of arrivals. Third, I drop migrants who were under the age of 10 at first observation because they were not asked about their ability to speak English; I also drop those who are older than 40 to ensure that no one would be older than 50 ten years later − this is to reduce bias from death. After these modifications, I am able to track 10,422 males between 1900 and 1910, 367,036 males from 1910 to 1920, and 307,529 males from 1920 to 1930. The linking rates for 1900 to 1910

report their *first* year of arrival, limiting biases from this problem. However, the 1900 Census did not specify the first arrival, which could bias estimates since re-entries were relatively common in the early 20th century (Keeling, 2010).

(22.1%), 1910 and 1920 (16.7%), and 1920 to 1930 (20.0%) are comparable to other historical studies that match migrants who may potentially return home (Abramitzky, Boustan and Eriksson, 2014, Ward, 2016).

While the linked datasets solve the problem of selective return migration because it is certain that these migrants stayed, there are also a few limitations. Primarily, linked datasets are non-random, as individuals with very common names, those who died or those who changed their name (e.g. females after marriage or anglicization of first name) cannot be linked forward (Biavaschi et al, 2013). This tends to make linked datasets more skilled than their underlying population. For this paper, I am particularly concerned that a successful link is related to better English proficiency; if so, then I would mistakenly infer that permanent migrants had better English skills at arrival when it would actually just reflect a bias from the linking process.

To gauge the representativeness of the sample, I compare an arrival cohort in the cross-sectional sample from IPUMS to the same cohort in the linked sample in the second year of observation (i.e. 1910 for the 1900 to 1910 linked sample). The linked and cross-sectional sample should contain the same information since each migrant has stayed in the United States for 11 to 20 years; however, while the IPUMS cross section is random, the linked sample may not be. The results for each sample are shown in Table 1. The linked samples are indeed biased; they contain migrants with slightly higher English-speaking ability, by 1 percentage point for the 1910 to 1920 and 1920 to 1930 samples. The discrepancy may be because New source countries are less likely to be linked, perhaps due to misspelled names. Despite differences in representativeness on English ability, there are small differences in occupational categories, with low-skilled laborers and operatives less likely to be linked. Overall, just as in other historical linked sample, it is slightly higher skilled than the cross section. To make the linked sample more representative, I reweight it to match the cross-sectional sample on English ability, literacy status, age and birth place. While some differences between the weighted sample and the cross section are statistically significant due to the sample size, they

are economically small; I will use the weighted sample throughout the rest of the analysis.

4 English Fluency Rates

4.1 Speed of English Acquisition

I estimate the rate of English acquisition for arrivals between 1890 and 1919 using the following flexible form:

$$SpeakEnglish_{ict} = \phi_c + \mu_{t-c} + \Pi' X_{it} + \varepsilon_{it} \tag{1}$$

Individual i from arrival cohort c's ability to speak English in census year t is modeled as a non-linear function of years in the United States (μ_{t-c}), incorporated as fixed effects for every two years (e.g. 0 to 1 years, 2 to 3 years, etc.). This parameterization captures the quick acquisition of English within the first ten years of stay and a leveling off in the second ten years. I also estimate arrival cohort fixed effects (ϕ_c) for every five year entry cohort to capture changes in the cohort quality in terms of English speaking ability (e.g. 1890-1894 arrivals, 1895-1899 arrivals, etc.). In various regressions I include control variables in X_{it} such as the country of birth and age at arrival. To capture potential biases from selective return migration, I run the regression twice, once with the panel data and once with repeated cross sections. Note that this regression assumes the same rate of acquisition across cohorts, a reasonable assumption when examining the underlying data.[19]

The estimated rate of acquisition is shown in Figure 4 for the 1890-94 cohort; I will later show how estimates vary across entry cohorts. About 38 percent of migrants in the panel data, or those who stayed at least ten years, knew how to speak English within the first year of arrival. After this low start, the rate of English proficiency increased rapidly within ten years of arrival: for those who had stayed ten to eleven years, about 87 percent of migrants were able to speak English, almost 50 percentage points higher than arrivals in their first

[19] See Figure A2.

year. This estimate is reasonable as second language acquisition can take only a couple of years (Krashen et al., 1979). After ten years of stay, the rate of acquisition levels off as most migrants knew how to speak some English; however, even after 20 years of stay, about 6% of migrants were still unable to speak English.

The estimated rate of acquisition from the repeated cross sections are also shown in Figure 4. The repeated cross section would estimate a higher rate of English acquisition since arrivals start at a lower percentage of 30%, but end at the same 94% fluency after twenty years. This is consistent with the arguments by Lubotsky (2007) and Abramitzky, Boustan and Eriksson (2014) that repeated cross-sections tend to overestimate improvements in migrants' attributes because of negatively selected return migration. In this case, migrants with worse English proficiency at arrival tended to return at higher rates. However, the degree of negative selection does not appears to be very strong as there is only a 8 percentage point gap between recent arrivals in the panel and recent arrivals in the repeated cross section, suggesting that other research estimating acquisition rates with repeated cross sections were not far from the truth (Kuziemko and Ferrie, 2014; Vigdor, 2010).

How does this speed of acquisition compare across the earlier and more recent migrant cohorts? In Figure 4, I also plot the mean English fluency over time of the 1990-94 arrival cohort, estimated in the same way as the early 20th century using repeated cross sections, but this time pooling the 2000 Census and 2008-2012 ACS. Note that with the 1990s cohort, I code an observation as able to speak English if they spoke English not well, well or very well. The figure shows two main conclusions. First, arrivals from non-English-speaking countries in 1990 had much higher levels of English fluency at arrival than migrants from the early 20th century cross section (74% v 30%). This could be due to the spread of English internationally compared with the beginning of the 20th century. It also could be because today's migration restrictions lead to longer delays for entry into the United States: lags associated with visa applications and entry requirements may cause potential migrants to invest in English skills prior to arrival. In contrast, for the pre-World War I arrivals, there was nothing stopping

individuals from freely arriving within a 2-week trip from Europe.

Another conclusion from Figure 4 is that migrants during the Age of Mass Migration acquired English fluency at much faster rates after arrival than migrants from the late 20th century. Of course, this is due to a lower starting point for the migrants who arrived between 1890 and 1894, but they do catch up to the 1990s arrivals within 15 to 20 years of stay when 94% of migrants could speak some English. Unfortunately, selective out-migration cannot be corrected for in the late 20th century data, but if return migrants were negatively selected on English ability as they were on income (Lubotsky, 2007), then return migration would not overturn these results that late 20th century migrants arrived with higher levels of English fluency but had lower rates of acquisition.[20]

4.2 Cohort Effects

The other main benefit from using panel or repeated cross sections is that one can estimate how cohorts changed in their ability to speak English near arrival. These estimates for arrival cohorts from 1890 to 1919 are shown in Figure 5, which display a U-shaped decline and increase in arrival cohort's English skills over time. While the 1890 to 1894 cohort had approximately 38% of migrants able to speak English near arrival, this dropped to 34% for the 1895 to 1899 cohort, and ultimately to 29% for the 1905-1909 cohort. By the 1910s, the English ability increased back to 34 and 38%. This general U-shaped trend is also found in the repeated cross-sectional data, but with a distinct level shift below those in the panel. Once again, this reflects that migrants were negatively selected on English ability between 1890 and 1919, where the negative selection − or gap at arrival between permanent and cross-sectional migrants − grew slightly during the 1900s decade compared with the 1890s

[20]This evidence is only suggestive because the variables do not match precisely. For example, as opposed to coding the post-1980 self-reported English proficiency of "not well" as "able to speak English," I recode it as "unable to speak English." When one does this, the results that late 20th century migrants have a slower *rate* of acquisition remain, but the results on starting *levels* differ depending on the matching of English variables. In fact, this way of merging variables suggests that the early 20th century migrants assimilated much more quickly than recent migrants in terms of English skills. However, based on the age at arrival/ English fluency profiles across time, I do not believe this is the best way to match the data.

decade.

The decline in arrival cohort's English ability between 1890 and 1909 could reflect either a shift in the composition toward poorer sources in Southern and Eastern Europe, similar to Borjas' (2015) argument for the decline in cohort quality following the Immigration and Nationality Act of 1965, or it could reflect a decline of quality *within* country, as suggested by Abramitzky, Boustan and Eriksson (2014) for late 19th century arrivals. I re-estimate the cohort fixed effects after controlling for country of birth fixed effects and plot them in Figure 6. The figure shows that the decline between 1890 and 1909 is significantly less when controlling for birth place, suggesting that the reason for a decline in English skills is because of a shift toward poorer countries. Interestingly, following 1910, arrival cohorts within a given country appear to rapidly increase their English proficiency at arrival, perhaps because migrants were investing more in pre-migration skills between 1910 and 1914, or because interruptions to migration during World War I limited non-English-speakers between 1915 and 1919.

The compositional shift toward poorer countries between 1890 and 1909 led to arrival cohorts with less English skills at arrival; this can be clearly seen in Figure 7, which plots the English fluency levels by language of origin, proxied by mother's tongue.[21] The figure is sorted by English fluency at arrival, where the leftmost languages had the lowest levels of English skills. Southern and Eastern European ethnicities such as Slovaks, Hungarians, Ukrainians, Poles, and Italians all dominate the left-hand side, where 10% to 30% of migrants were able to speak English within one year of arrival. Note that Jewish migrants from Eastern Europe had higher English fluency rates at arrival than others from the same region, near 40 percent.

These fluency rates for Eastern and Southern Europeans are mostly lower when compared with migrants from Northern and Western Europe; German-speaking arrivals had the lowest fluency levels for this group, where 30% of migrants were able to speak English at arrival. Other ethnicities such as the Dutch, Norwegians, Swedes and Danish all had higher arrival

[21]This figure is for migrants who arrived between 1900 and 1909. I do not separately show the 1890 to 1899 arrival cohort by country of birth due to smaller sample sizes. See Figure A3 for the figure by country of birth.

rates, from 45 to 65%. Interestingly, Chinese (45%) and Japanese (57%) had relatively high rates of English proficiency near arrival. After fifteen plus years in the United States, most ethnicities had over 80 percent of their group as able to speak English, with Northern and Western Europeans at higher rates above 90 percent.

5 The English Premium

5.1 The English Premium in the Early 20th Century

Migrants acquired English skills at relatively fast rates in the early 20th century; this could reflect that English was highly valuable for improving outcomes. In this section, I estimate the English premium between 1900 and 1930. Estimating the premium for English has straightforward econometric issues: primarily, the ability to speak English could be correlated with an unobserved omitted variable that could positively bias the estimate. Instead, I leverage the panel features of the linked dataset to estimate the association between upgrading one's occupation and learning to speak English. Note that here I only aim to estimate the association here while reducing the threat of unobservables; other methods to provide exogenous variation in English ability will be discussed later.

I estimate the association between speaking English and a migrant's occupation instead of wage because wage is unavailable in the Census until 1940. First, I group migrants into six occupational categories: high-skilled white collar (e.g. managers and doctors), medium-skilled white collar (e.g., salesmen and clerks), semi-skilled workers (e.g. craftsmen), farmers, low-skilled service/manual workers (e.g. waiters and operatives), and laborers. Later I will assign alternative occupational scores for occupations at a much finer level of detail in order to understand how learning to speak English is correlated with an increase in earnings, but here I provide descriptive evidence based on occupational categories.

I estimate the rate at which one changes occupations in the following linear probability

model:

$$OccupationGroup_{it} = \gamma_0 + \gamma_1 SpeaksEnglish_{it} + \varphi_i + \Pi' X_{it} + \varepsilon_{it} \quad (2)$$

The dependent variable is a zero / one variable for whether one belongs to one of six occupational groups. I run the regression six times − once for each of the groups in order to estimate how learning to speak English affects the net flow into or out of an occupational group. After controlling for individual fixed effects φ_i, the coefficient γ_1 will produce an estimate of the effect of English while accounting for numerous unobservable factors that are constant within an individual i such an unobserved ability. Note that the coefficient on the ability to speak English can only be estimated with individuals who change their English-speaking ability; given the low level of English ability at arrival and the high rate of acquisition afterwards, this provides me with a large amount of variation to estimate the coefficient.

For controls, I include the year of observation, which accounts for the average shift in occupational group over time as identified by those who either do not acquire English skills or those who had already acquired English skills. I further interact year with age at first observation (grouped into five-year intervals) to allow for job switching to vary by points in the life cycle. I also interact years in the United States at initial observation, grouped into two-year intervals, with year for the same reason. Finally, I include controls for literacy, logged population in a county and fraction of migrants from the same birthplace in county to account for changes in general human capital, size of network, and population density. I drop individuals for whom I do not observe jobs in both censuses; this mostly leads to dropping children. In all regressions I calculate the standard errors by clustering on country of birth.

The results are shown in Table 2.[22] The table is split into three panels, one for the years 1900 to 1910 (or migrants who arrived between 1890 and 1899), one for years 1910 to 1920,

[22]For the raw data on transitions across occupational groups for those between 1910 and 1920, see Table A1 for those who switched English status, Table A2 for those who knew how to speak English at first observation, and Table A3 for those who never learned how to speak English. The tables show that there were both transitions into and out of each occupational group.

and one for years 1920 to 1930. The results show that between censuses those who learned to speak English left laborer jobs to enter into other categories. For example, learning to speak English led to almost a third fewer laborers between 1910 and 1920; it appears that laborers most commonly transitioned into either unskilled service or operative jobs, or semi-skilled blue-collar jobs. Note that learning to speak English was indeed correlated with positive flow into holding a white-collar job, either high-skilled professional ones or medium-skilled sales jobs, however this correlation was relatively weak; rather, speaking English led migrants to move up the occupational distribution only slightly. The net flow into farming occupations was zero; rather, migrants mostly worked in other types of jobs, perhaps because migrants often settled in urban areas.

The occupational categories provide a broad picture of the net flows into and out of occupation groups for English speakers, but do not give a simple estimate of the English premium. To estimate this, one needs to assign each of the nearly 250 occupational codes an occupational score.[23] Unfortunately, there is no representative occupational score at this level of detail for each decade between 1900 and 1930; therefore, I resort to other occupational scores used in the literature. One occupational score is based on wages by occupation as observed in the 1901 Cost of Living Survey (CLS), a sample of families in 99 cities; note that this is a sample of families in urban centers, which is perhaps unrepresentative of rural areas or single individuals.[24] Second, I use an occupational score based on average earnings by occupation, estimated separately for "New" and "Old" Source countries. This score, originally created in Ward (2016), improves on the prior score by being representative of migrant earnings, but does not include business or farm income. Finally, I also use the 1950 occupational score provided by IPUMS; this improves on the prior 1940 score by having self-employed income, but is measured many decades after the sample period. While these scores

[23]This is based on the standardized occupational codes variable occ1950 in IPUMS.

[24]This score is used by Abramitzky, Boustan and Eriksson (2012) when estimating the return to migration in the United States. For less than 2% of observations there is no occupation score in the 1901 Cost of Living Survey. For the missing occupations, I calculate its position in the 1950s occupational score distribution, which has scores for all occupations. I assume the missing occupation's point in the 1901 distribution is the same as in the 1950 distribution, and then fill in its score based on its predicted wage.

represent wages for various populations, the more important difference between these scores reflects the compression of wages across the early 20th century (Goldin and Katz, 2008). Therefore, it is expected that the highest return to speaking English will be from the 1901 score when wages are least compressed, and the lowest return to speaking English will be from the 1950 score, when wages are most compressed.

The results from running the Equation (2) with logged occupational score as the dependent variable are shown in Table 3. The first column uses the 1901 CLS score and finds that speaking English led to a 5.7% upgrade between 1900 and 1910, a 5.4% upgrade between 1910 and 1920, and a 6.9% upgrade between 1920 and 1930. Holding skill prices constant, it appears that English became slightly more valuable over time for improving one's occupation, but was still relatively unimportant when compared with today's estimates of the return to English. However, the magnitude of the English premium changes when one uses either the 1940 and 1950 occupational score; for example, the score based on 1940 income finds that the associated increase for speaking English and occupational outcomes is only 1.7 to 3.1 percent, less than half the return compared with the results when using 1901 skill prices. This suggests that the English premium was actually *decreasing* between 1900 and 1930 as the wage distribution was becoming more compressed and the return to education fell (Goldin and Katz, 2008).

Finally, the third column shows when using the occupational scored based on 1950 wages, the favored metric by Abramitzky, Boustan and Eriksson (2014) in their assimilation study, the premium for learning to speak English is 0 percent between 1900 and 1920, and 2.8 percent between 1920 and 1930. This finding of essentially no return to speaking English is consistent with their result of migrants' having little to no increase in economic position relative to natives after arrival. However, note that the variable for speaking English is not exogenous, but may be correlated with other factors - such as other types of United States specific human capital - that change over time; however, these other factors likely positively affect labor market outcomes, suggesting that the individual fixed effect estimate is an upper

bound of the true return to English skills.

5.2 Robustness of the Estimate

The estimate from the linked sample shows a relatively low return to English. However, the estimate has a few limitations: mainly, learning to speak English is not exogenous. I use an additional empirical strategy to estimate the English premium in the appendix, where one could exploit the well-defined relationship between age at arrival and the ability to speak English (Bleakley and Chin 2004, 2010).[25] Bleakley and Chin (2004) use this relationship to instrument for the ability to speak English based on whether one arrived at an older or younger age, and whether the migrant was born in an English-speaking country.[26] For this empirical strategy to be successful, the non-language effects of age at arrival on adult occupation should be similar across English and non-English-speaking source countries; however, this does not appear to hold in the early 20th century, which is why I do not include it in the main analysis. Yet, when using the strategy one is able to sign the bias such that language human capital was much less important than non-language human capital, reaffirming that English fluency was relatively unimportant in the early 20th century (see Appendix D for details). These results flip in the late 20th century when English was highly important for adult outcomes. Therefore, the age-at-arrival analysis, the individual fixed effects regressions, and a simple comparison of the quick rate of language acquisition in this study with the lack of occupational upgrading found in Abramitzky, Boustan and Eriksson (2014) all make the same point that English fluency was relatively unimportant for occupational outcomes one hundred years ago compared with today.

A separate issue raised by Dustmann and Van Soest (2001) is that measurement error for language skills biases OLS coefficients downward. Thus it may be argued that the reason

[25]The older a migrant arrives − especially after the ages of 8 to 10 − the less likely he is able to fully speak English as an adult. This is because of neurobiological changes during and following puberty that makes the costs of acquiring a second language higher (Singleton, 2001).

[26]In essence, this strategy is a ratio of the reduced form shown in Figure A5 to the first stage shown in Figure A4.

why the early 20th century premium is much lower than estimated English premia post-1980 is because there is simply more measurement error. This likely is not the case because of the method of recording the English variable: the late 20th century measures of English proficiency were self-reported on a wide, subjective scale from zero to three. A binary scale, as used in the early 20th century, may reduce measurement error (Chiswick and Miller, 2014); further, the measurement was externally done by an enumerator which may lead to less measurement error. Nevertheless, the estimates in Table 3 may be downwardly biased.

How does this English premium compare to the premium from recent years? There are various estimates of the English premium in recent decades, with most OLS estimates using pre-2000 United States data ranging from 10 to 20 percent (Chiswick and Miller, 2014, Table 5.5), and studies using post-2000 data finding correlations above 20 percent. For example, the 2008 to 2012 ACS shows that the association between speaking English and log *wages* is 25%; however, when one uses a logged *occupational score* instead, then the correlation falls to 12%, showing that using occupations only captures about half of the wage premium.[27] Therefore, an occupational premium in the early 20th century of between 2 to 6 percent suggests that the occupational return of English skills have at least doubled to 12 percent in recent years; further, the increase in premium has likely more than doubled between 1930 and 2010 given that the estimated premium when using the occupational score closest to 1930 (i.e., the score based on 1940 skill prices) yields a premium of 2.9%.[28]

5.3 Heterogeneity Across Subgroups

The association between English skills and occupational upgrading may have varied across different subgroups; for example, it could be that New source countries faced substantial

[27] This is based on a regression from the 2008 to 2012 ACS, where the association between speaking English and wages is 25 percent, and the association between speaking English and logged occupational score, based on wages by occupation, is 12 percent. See Table A4.

[28] Additionally, note that Bleakley and Chin (2004) aim to improve on these OLS estimates by instrumenting for English ability with the interaction between age at arrival and being from an English-speaking country, and find that an IV estimate for speaking English is approximately 50 percent higher than the OLS estimate, suggesting that the occupational return to speaking English may be closer to 21 percent for recent years.

discrimination from natives and were not promoted despite learning how to speak English, while Old source country migrants faced less discrimination (at least prior to World War I for Germans). Alternatively, speaking English and occupational upgrading may be different for child arrivals compared to adult arrivals (Hatton, 1997). Given that I have panel data on over 500,000 migrants, it is straightforward to analyze subgroups of the original sample.

I re-estimate the English premium with various subsamples in Table 4. The first column provides the reference when using the entire sample and the 1901 Cost of Living Survey score; the second and third column splits the sample into Old source countries (i.e. Northern and Western European) and New source countries (i.e. mostly Southern and Eastern European). Despite the different levels of English ability for arrivals, the association between speaking English and occupational outcomes is very similar between the two source groups. This is suggestive evidence that discrimination against New source countries is not driving a relatively low English premium; rather, a relatively low English premium may be a consequence of a lack of white-collar jobs in the economy.

The forth and fifth columns separate migrants into adult arrivals, or those who arrived over the age of sixteen, and child arrivals. Results are mixed, but generally there appears to be a relatively constant return for adult arrivals over time (6 to 7%), while the return for child arrivals changed from zero percent between 1900 and 1910 to 7.5% between 1920 and 1930. It is unclear what is driving this increase for child arrivals, and perhaps relates to new laws enforcing child migrants to attend school (Lleras-Muney and Shertzer, 2015). Finally, the sixth column considers those who arrived within the last year (e.g. 1899 arrivals in 1900), and finds a similar return to these arrivals and the entire sample.

6 Concluding Remarks

Surprisingly little is known about the importance of English for migrant outcomes one hundred years ago. Using new linked data, I estimate both the rate of English acquisition and the

return to acquiring English skills in the early 20th century. I show that the English premium was relatively low; in other words, the inability to speak English was not a large barrier to migration. Interestingly, even though the English premium was small in the early 20th century, migrants acquired English proficiency at relatively quick rates. The rapid acquisition of English aligns with other research suggesting that migrants during the Age of Mass Migration assimilated quickly in cultural (rather than economic) terms (Abramitzky, Boustan and Eriksson, 2016). This contrasts with recent arrivals who came to the United States with relatively high rates of English proficiency. After fifteen to twenty years of duration, early 20th century migrants caught up to the English levels of late 20th century migrants. If one considers the English premium in a supply and demand framework, then the results suggest that the relative supply of English speakers has been constant or has increased over time.

Therefore, since the relative supply of English speakers has not decreased, it is likely that the main force driving an increased English premium over the last 100 years is related to demand for English skills and the premium for general human capital. This is easily seen when examining the structure of the labor force over the past century in Figure 8, in which the fraction of white-collar jobs has tripled, agricultural jobs have been all but eliminated, and the proportion of blue collar jobs has decreased. The importance of interaction and social skills appear to have grown not only in the past few decades, but also over the last century (Deming, 2015; Michaels et al, 2013). These demand shifts likely increased the premium for English, causing English fluency to be of primary importance for migrants to succeed in the United States in the 21st century. However, while tasks have shifted to be more interactive, an increase in the English premium in recent decades is also related to the increasing return to human capital. A correlation between the English premium and general skill premium is observed in the early 20th century, as there is evidence that the English premium was higher in the early 20th century, but then decreased over the next few decades, just as the skill premium was decreasing in the overall economy (Goldin and Katz, 2008).

While a low English premium in the early 20th century is likely the consequence of the

technological setting, it could be that some foreign born were not promoted because of discrimination. Indeed, discrimination appears to have been rampant; for example, those who changed their name to be more "American" received a premium in the labor market (Biavaschi, Giulietti and Siddique, 2013), and brothers who had more American-sounding names earned more than brothers who had more foreign-sounding names (Abramitzky, Boustan and Eriksson, 2016). Yet these two results suggests that there is a positive return to becoming more "American", and a similar positive return would likely hold for becoming more American by learning to speak English. Further, I show that the English premium is similar for migrants who experienced more discrimination (Southern and Eastern Europeans) and migrants who experienced less discrimination (Northern and Western Europeans), suggesting that discrimination was not the main driver of the low English premium.

While I stress the importance of English fluency for understanding the variation in migrant assimilation profiles over time, it is not the only determinant of the profile. In particular, migrants' earnings relative to natives' also depend on their pre-migration human capital; indeed, this point has been stressed by Borjas (1985, 1995, 2015) as migrant sources have shifted to poorer countries following the Immigration and Nationality Act of 1965. Therefore, another reason for the difference in assimilation profiles across time may be that migrants in the past had pre-migration human capital levels similar to natives, compared with today's difference in human capital between natives and migrants (Abramitzky and Boustan, 2016).

If current trends continue, then the English premium will increase even further in future decades. If technological shifts continue to favor those with skill, especially social skills (Goldin and Katz, 2008; Deming, 2015), and if migrants do not have higher rates of investment in English skills either pre-arrival or post-arrival (Borjas, 2015), then the value of English skills will increase and the average migrant's economic position will gradually worsen over time.

References

Abramitzky, Ran and Leah Platt Boustan, "Immigration in American History," *Journal of Economic Literature*, 2016.

_ , _ , and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *The American Economic Review*, 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , and _ , "Cultural Assimilation during the Age of Mass Migration," Working Paper 22381, National Bureau of Economic Research July 2016.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique, "The Economic Payoff of Name Americanization," Technical Report, Institute for the Study of Labor (IZA) 2013.

Blau, Francine D, "Immigration and labor earnings in early twentieth century America.," 1980.

Bleakley, Hoyt and Aimee Chin, "Language Skills and Earnings: Evidence from Childhood Immigrants," *Review of Economics and Statistics*, 2004, *86* (2), 481–496.

_ and _ , "What holds back the second generation? The intergenerational transmission of language human capital among immigrants," *Journal of Human Resources*, 2008, *43* (2), 267–298.

_ and _ , "Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants," *American Economic Journal: Applied Economics*, 2010, pp. 165–192.

Bloch, Louis, "The Ability of European Immigrants to Speak English," *Quarterly publications of the American Statistical Association*, 1920, *17* (132), 402–416.

Borjas, George J, "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics*, 1985, *3* (4), 463–489.

_ , "The economics of immigration," *Journal of economic literature*, 1994, pp. 1667–1717.

_ , "Assimilation in Cohort Quality Revisited: What Happened to Immigrant Earnings in the 1980s?," *Journal of Labor Economics*, 1995, *13* (2), 211–245.

_ , "The Slowdown in the Economic Assimilation of Immigrants: Aging and Cohort Effects Revisited Again," *Journal of Human Capital*, 2015, *9* (4), 483–517.

Boustan, Leah Platt, Devin Bunten, and Owen Hearey, "Urbanization in the United States, 1800-2000," Technical Report, National Bureau of Economic Research 2013.

Chiswick, Barry R, "The Effect of Americanization on the Earnings of Foreign-born Men," *Journal of Political Economy*, 1978, *86* (5), 897–921.

_ and Paul W Miller, "Do Enclaves Matter in Immigrant Adjustment?," *City & Community*, 2005a, *4* (1), 5–35.

_ and _ , "Linguistic distance: A quantitative measure of the distance between English and other languages," *Journal of Multilingual and Multicultural Development*, 2005b, *26* (1), 1–11.

_ and _ , "International migration and the economics of language," *Handbook of the Economics of International Migration, 1A: The Immigrants*, 2014, *1*, 211.

Cutler, David M, Edward L Glaeser, and Jacob L Vigdor, "Is the melting pot still hot? Explaining the resurgence of immigrant segregation," *The Review of Economics and Statistics*, 2008, *90* (3), 478–497.

Deming, David J, "The growing importance of social skills in the labor market," Technical Report, National Bureau of Economic Research 2015.

Dustmann, Christian and Arthur Van Soest, "Language fluency and earnings: Estimation with misclassified language indicators," *Review of Economics and Statistics*, 2001, *83* (4), 663–674.

_ **and Francesca Fabbri**, "Language proficiency and labour market performance of immigrants in the UK," *The Economic Journal*, 2003, *113* (489), 695–717.

Feigenbaum, James J, "A Machine Learning Approach to Census Record Linking," 2016.

Ferrie, Joseph P, "A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 1996, *29* (4), 141–156.

Goldin, Claudia and Lawrence F Katz, "The origins of technology-skill complementarity," *The Quarterly Journal of Economics*, 1998, *113* (3), 693–732.

_ **and** _ , *The race between education and technology*, Harvard University Press, 2008.

Gray, Rowena, "Taking technology to task: The skill content of technological change in early twentieth century united states," *Explorations in Economic History*, 2013, *50* (3), 351–367.

Greenwood, Michael J and Zachary Ward, "Immigration quotas, World War I, and emigrant flows from the United States in the early 20th century," *Explorations in Economic History*, 2015, *55*, 76–96.

Guven, C and A Islam, "Age at migration, language proficiency, and socioeconomic outcomes: evidence from australia.," *Demography*, 2015, *52* (2), 513.

Hatton, Timothy J, "The Immigrant Assimilation Puzzle in Late Nineteenth-Centuty America," *The journal of economic history*, 1997, *57* (01), 34–62.

_ **, Jeffrey G Williamson et al.**, "The age of mass migration: Causes and economic impact," *OUP Catalogue*, 1998.

Inwood, Kris, Chris Minns, and Fraser Summerfield, "Reverse assimilation? Immigrants in the Canadian labour market during the Great Depression," *European Review of Economic History*, 2016, *20* (3), 299–321.

Jasso, Guillermina and Mark R Rosenzweig, "Language Skill Acquisition, Labor Markets and Locational Choice: The Foreign-Born in the United States, 1900 and 1980," in "Migration and Labor Market Adjustment," Springer, 1989, pp. 217–239.

_ **and** _ , *The new chosen people: Immigrants in the United States*, Russell Sage Foundation, 1990.

Jenks, Jeremiah Whipple and William Jett Lauck, *The immigration problem*, Funk & Wagnalls Company, 1926.

Katz, Lawrence F and Robert A Margo, "Technical change and the relative demand for skilled labor: The united states in historical perspective," Technical Report, National Bureau of Economic Research 2014.

Keeling, Drew, "Repeat migration between Europe and the United States, 1870–1914," in "The Birth of Modern Europe," Brill, 2010, pp. 157–186.

Krashen, Stephen D, Michael A Long, and Robin C Scarcella, "Age, rate and eventual attainment in second language acquisition," *Tesol Quarterly*, 1979, pp. 573–582.

Lafortune, Jeanne, José Tessada, and Ethan Lewis, "People and Machines: A Look

at the Evolving Relationship Between Capital and Skill In Manufacturing 1860-1930 Using Immigration Shocks," Working Paper 21435, National Bureau of Economic Research July 2016.

LaLonde, Robert J and Robert H Topel, "The assimilation of immigrants in the US labor market," in "Immigration and the workforce: Economic consequences for the United States and source areas," University of Chicago Press, 1992, pp. 67–92.

Lleras-Muney, Adriana and Allison Shertzer, "Did the Americanization Movement Succeed? An Evaluation of the Effect of English-Only and Compulsory Schooling Laws on Immigrants," *American Economic Journal: Economic Policy*, 2015, *7* (3), 258–90.

Lubotsky, Darren, "Chutes or ladders? A longitudinal analysis of immigrant earnings," *Journal of Political Economy*, 2007, *115* (5), 820–867.

_ , "The effect of changes in the US wage structure on recent immigrants' earnings," *The Review of Economics and Statistics*, 2011, *93* (1), 59–71.

Massey, Catherine G, "Playing with Matches: An Assessment of Accuracy in Linked Historical Data."

Michaels, Guy, Ferdinand Rauch, and Stephen J Redding, "Urbanization and Structural Transformation," *The Quarterly Journal of Economics*, 2012, *127* (2), 535–586.

_ , _ , and _ , "Task specialization in US cities from 1880-2000," Technical Report, National Bureau of Economic Research 2013.

Mill, Roy and Luke CD Stein, "Race, skin color, and economic outcomes in early twentieth-century America," *Available at SSRN*, 2016.

Parman, John, "Childhood health and sibling outcomes: Nurture Reinforcing nature during the 1918 influenza pandemic," *Explorations in Economic History*, 2015, *58*, 22 – 43.

Pérez, Santiago, "The (South) American Dream: Occupational Mobility of Immigrants in 19th Century Argentina," 2016.

Perlmann, Joel, *Italians Then, Mexicans Now: Immigrant Origins and the Second-Generation Progress, 1890-2000*, Russell Sage Foundation, 2005.

Singleton, David, "Age and second language acquisition," *Annual review of applied linguistics*, 2001, *21*, 77–89.

Stevens, Gillian, "A century of US censuses and the language characteristics of immigrants," *Demography*, 1999, *36* (3), 387–397.

Vigdor, Jacob L, *From immigrants to Americans: The rise and fall of fitting in*, Rowman & Littlefield, 2010.

Ward, Zachary, "Birds of Passage: Return Migration, Self-Selection and Immigration Quotas," *Explorations in Economic History*, 2016.

Wyman, Mark, *Round-trip to America: the immigrants return to Europe, 1880-1930*, Cornell University Press, 1993.

Table 1: Representativeness of the Linked Samples

	1890-1899 Cohort in 1910			1900-1909 Cohort in 1920			1910-1919 Cohort in 1930		
	Cross	Panel		Cross	Panel		Cross	Panel	
		Difference from Cross			Difference from Cross			Difference from Cross	
		Unweighted	Weighted		Unweighted	Weighted		Unweighted	Weighted
Speak English	0.867	0.0391***	0.00193	0.879	0.0113***	0.00163	0.923	0.0128***	0.000735
	(0.340)	(0.00466)	(0.00575)	(0.326)	(0.00251)	(0.00255)	(0.267)	(0.00112)	(0.00116)
Literate	0.889	0.0467***	0.00131	0.847	0.0102***	0.00157	0.888	0.00708***	0.000352
	(0.314)	(0.00416)	(0.00552)	(0.360)	(0.00278)	(0.00281)	(0.315)	(0.00133)	(0.00136)
New Source	0.457	-0.0795***	-0.00306	0.765	-0.0676***	-0.00503	0.830	-0.0194***	-0.00106
	(0.498)	(0.00719)	(0.00788)	(0.424)	(0.00329)	(0.00331)	(0.376)	(0.00161)	(0.00160)
Age	35.53	0.364***	0.0177	35.43	0.415***	-0.0350	36.52	-0.0194	0.0281
	(7.721)	(0.108)	(0.124)	(7.129)	(0.0550)	(0.0556)	(7.070)	(0.0300)	(0.0302)
Age at Arrival	19.16	-0.0887	-0.216*	20.06	0.609***	0.216***	19.30	-0.259***	-0.173***
	(7.704)	(0.108)	(0.122)	(7.124)	(0.0549)	(0.0556)	(7.043)	(0.0298)	(0.0300)
Professional	0.137	0.000189	0.00220	0.117	0.00584**	0.0119***	0.119	0.00535***	0.00855***
	(0.344)	(0.00502)	(0.00543)	(0.321)	(0.00249)	(0.00297)	(0.323)	(0.00138)	(0.00141)
Sales/Clerical	0.0652	-0.00227	0.00199	0.0507	0.00248	0.00771***	0.0634	0.00239**	0.00332***
	(0.247)	(0.00358)	(0.00396)	(0.219)	(0.00170)	(0.00172)	(0.244)	(0.00104)	(0.00106)
Semi-Skilled	0.202	0.0152**	-0.00662	0.216	0.00701**	-0.00380	0.202	0.00852***	0.00157
	(0.401)	(0.00593)	(0.00602)	(0.411)	(0.00318)	(0.00321)	(0.401)	(0.00171)	(0.00171)
Unskilled Service/	0.269	-0.0409***	-0.0257***	0.289	-0.0161***	-0.0165***	0.288	-0.0144***	-0.0141***
Operative	(0.443)	(0.00632)	(0.00690)	(0.454)	(0.00350)	(0.00357)	(0.453)	(0.00192)	(0.00194)
Farmer	0.0982	0.0287***	0.0102**	0.0579	0.00608***	0.00483***	0.0412	0.00519***	0.00428***
	(0.298)	(0.00458)	(0.00475)	(0.234)	(0.00181)	(0.00184)	(0.199)	(0.000854)	(0.000862)
Laborer	0.211	-0.0188***	-0.00318	0.242	-0.0222***	-0.0224***	0.264	-0.0272***	-0.0238***
	(0.408)	(0.00587)	(0.00644)	(0.428)	(0.00331)	(0.00336)	(0.441)	(0.00186)	(0.00188)
Observations	8,525	Panel: 10,422		17,568	Panel: 367,036		67,434	Panel: 307,529	

Notes: Data is from linked samples between 1900 to 1910, 1910 to 1920, and 1920 to 1930; cross-sectional data is from IPUMS 1% samples in 1910, 1920 and a 5% sample in 1930 (Ruggles et al., 2015). *p<0.10, **p<0.05, ***p<0.01

Table 2: Speaking English and Occupational Transitions, Individual Fixed Effects

	I Professional/ Manager	II Sales/ Clerical	III Semi-Skilled	IV Unskilled Service/Oper.	V Farmer	VI Laborer
Panel A: 1900 to 1910 Census						
Speak English	0.0329**	0.0165**	0.0248	-0.0152	-0.000643	-0.0583
	(0.0134)	(0.00754)	(0.0164)	(0.0211)	(0.00609)	(0.0418)
Mean of Dep. Var. in 1900	0.0689	0.0587	0.171	0.300	0.0564	0.345
Number of ind	9,071	9,071	9,071	9,071	9,071	9,071
Panel B: 1910 to 1920 Census						
Speak English	0.00452	0.0220***	0.0424***	0.0454***	0.000992	-0.115***
	(0.00592)	(0.00406)	(0.00568)	(0.0123)	(0.00422)	(0.00755)
Mean of Dep. Var. in 1910	0.0582	0.0588	0.189	0.293	0.0293	0.372
Number of ind	319,817	319,817	319,817	319,817	319,817	319,817
Panel C: 1920 to 1930 Census						
Speak English	0.0134***	0.0163***	0.0398***	0.0278***	0.00616**	-0.103***
	(0.00367)	(0.00396)	(0.00635)	(0.00613)	(0.00248)	(0.00629)
Mean of Dep. Var. in 1920	0.0734	0.0535	0.204	0.312	0.035	0.322
Number of ind	254,130	254,130	254,130	254,130	254,130	254,130

Notes: Data is from linked samples between 1900 to 1910, 1910 to 1920, and 1920 to 1930. The results are from a regression of occupational category on the ability to speak English individual fixed effects and controls described in text. *p<0.10, **p<0.05, ***p<0.01

Table 3: Speaking English and Occupational Score, Individual Fixed Effects

	I CLS 1901 Wage	II 1940 Census Wage	III 1950 Census Wage + Bus/Farm Inc
Panel A: 1900 to 1910 Census			
Speak English	0.0565***	0.0227**	0.00284
	(0.0142)	(0.00907)	(0.0123)
Number of ind	9,071	9,071	9,071
Panel B: 1910 to 1920 Census			
Speak English	0.0541***	0.0167***	-0.00138
	(0.00529)	(0.00337)	(0.00660)
Number of ind	319,817	319,817	319,817
Panel C: 1920 to 1930 Census			
Speak English	0.0694***	0.0311***	0.0278***
	(0.00653)	(0.00628)	(0.00793)
Number of ind	254,130	254,130	254,130

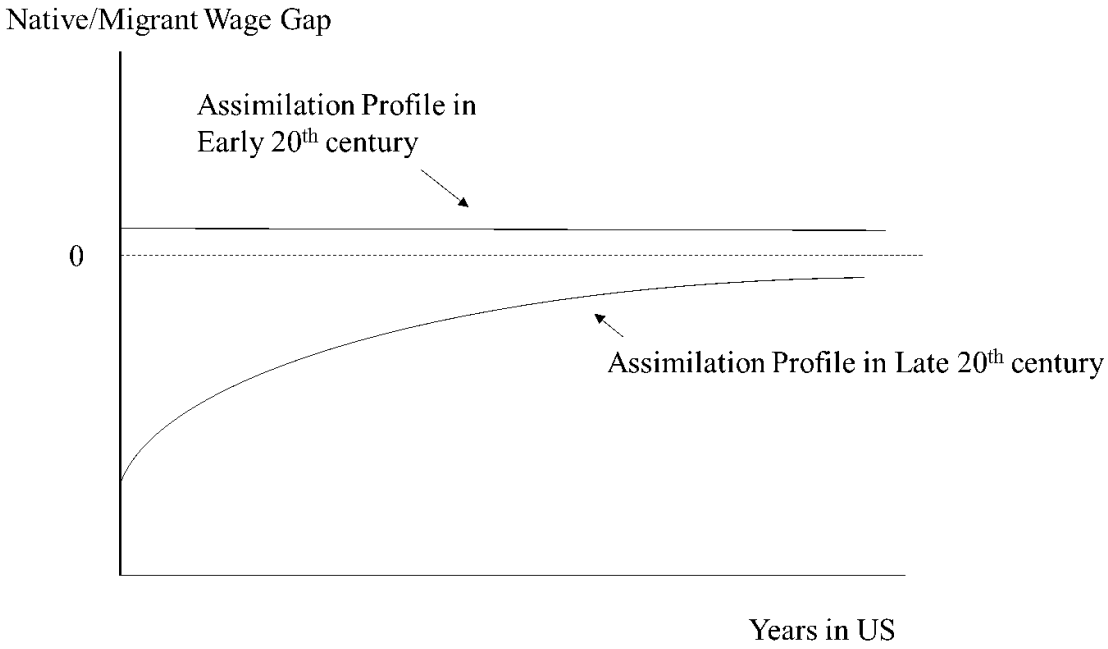
Notes: Data is from the 1900 to 1910, 1910 to 1920, and 1920 to 1930 linked samples. The results are from a regression of log occupational score on the ability to speak English with individual fixed effects and controls described in text.

Table 4: Speaking English and Occupational Upgrading, Alternative Samples

Sample:	I Base	II Old Sources	III New Sources	IV Child Arrivals	V Adult Arrivals	VI ≤ 1 Year Stay
Panel A: 1900 to 1910 Census						
Speak English	0.0565***	0.0614**	0.0558**	0.0170	0.0656***	0.0585
	(0.0142)	(0.0213)	(0.0212)	(0.0358)	(0.0136)	(0.0618)
Number of ind	9,071	5,678	3,393	2,163	6,908	374
Panel B: 1910 to 1920 Census						
Speak English	0.0541***	0.0438***	0.0557***	0.0460***	0.0555***	0.0654***
	(0.00529)	(0.00726)	(0.00608)	(0.00931)	(0.00505)	(0.0136)
Number of ind	319,817	97,766	222,051	54,531	265,286	36,168
Panel C: 1920 to 1930 Census						
Speak English	0.0694***	0.0699***	0.0708***	0.0740***	0.0690***	0.0981***
	(0.00653)	(0.00972)	(0.00734)	(0.0120)	(0.00610)	(0.0116)
Number of ind	254,130	49,165	204,965	52,506	201,624	7,616

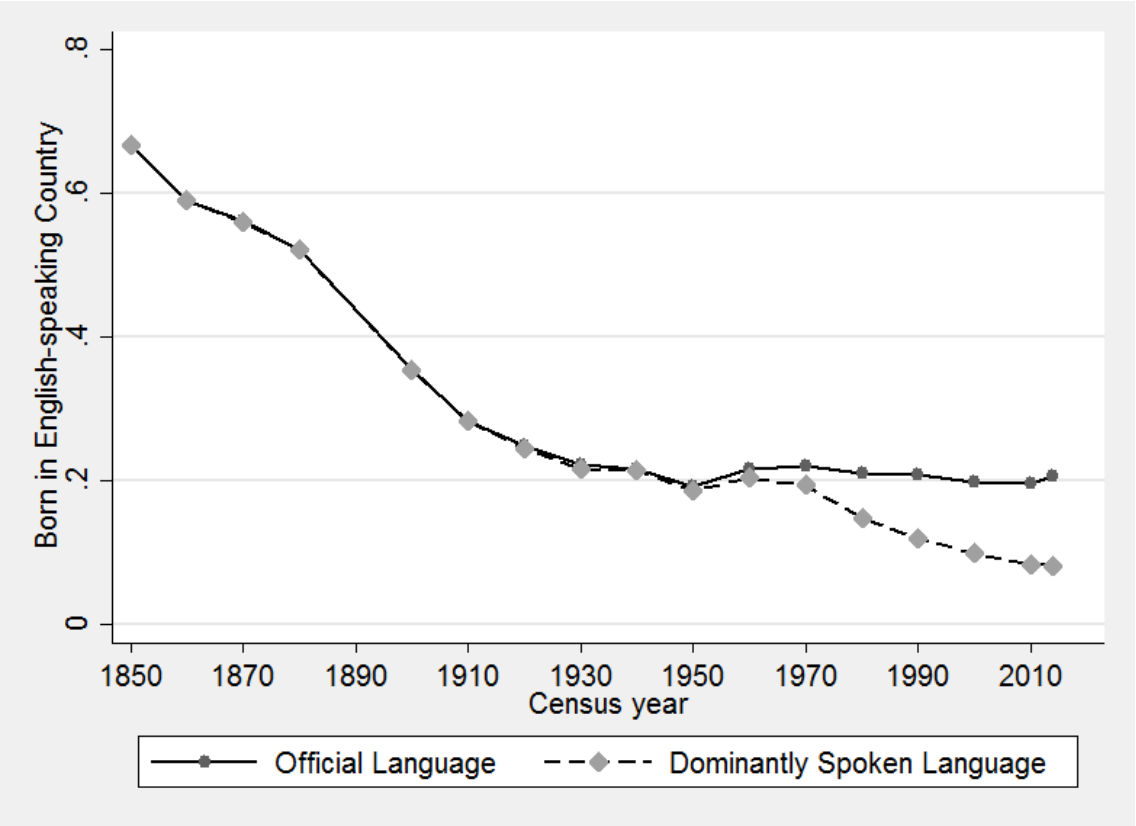
Notes: Data is from the 1910 to 1920, and 1920 to 1930 linked samples. The results are from a regression of log occupational score on the ability to speak English with individual fixed effects and controls described in text.

Figure 1: Assimilation Profiles Across Time for Permanent Migrants



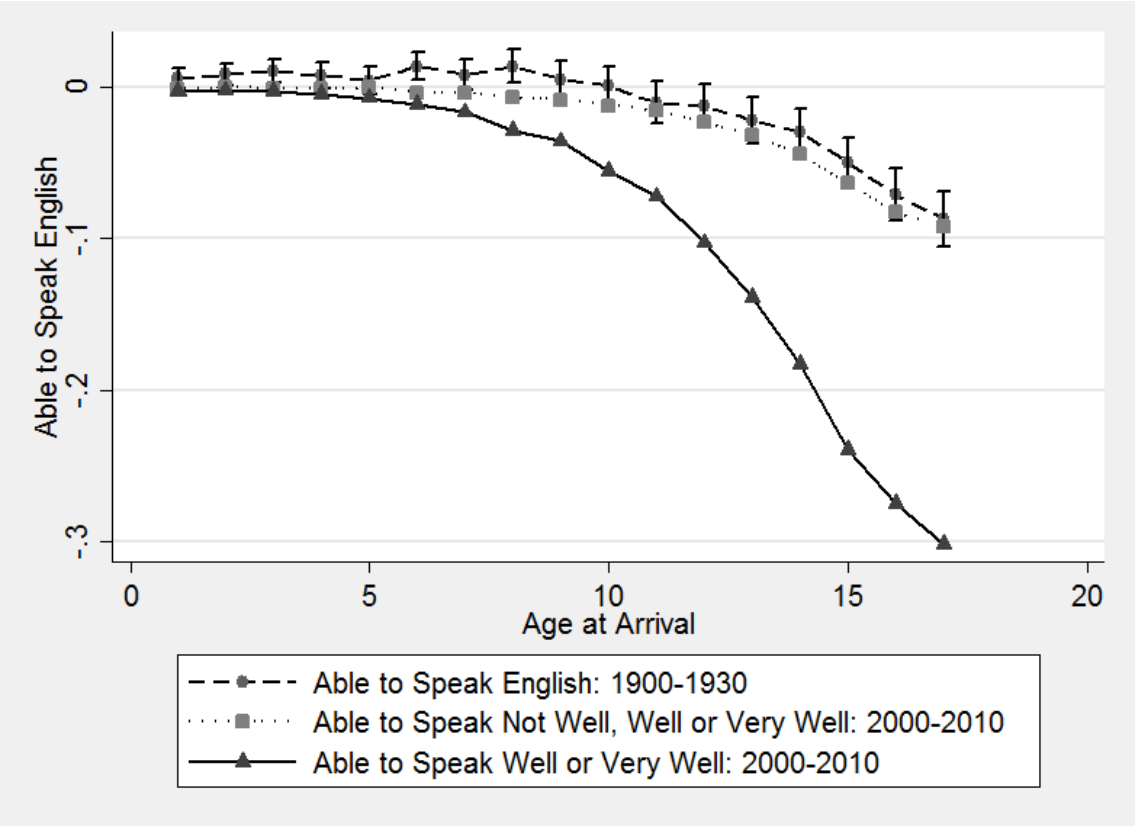
Notes: The typical assimilation profile in the early 20th century is found by Abramitzky, Boustan and Eriksson (2014); late 20th century by Lubotsky (2007). The findings only represent the assimilation of permanent migrants who stay throughout a panel.

Figure 2: Fraction of Migrant Stock Born in an English-Speaking Country



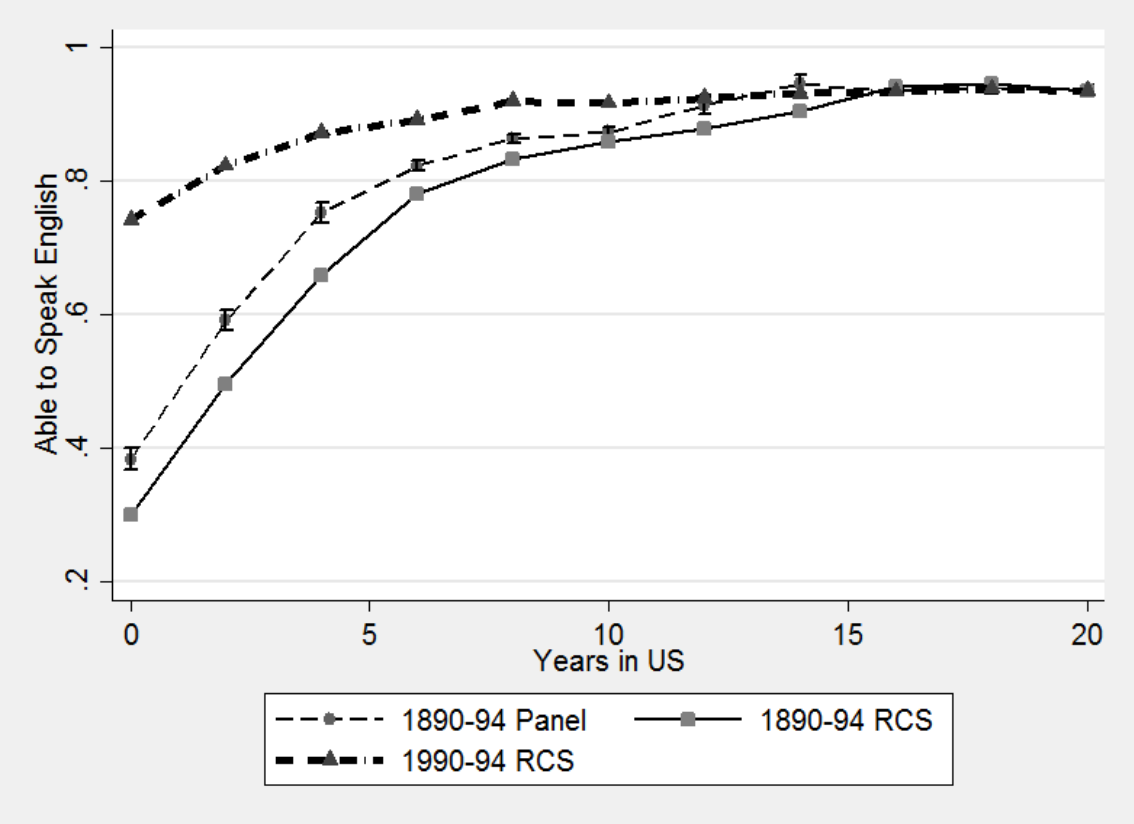
Notes: Data is from 1850-2014 IPUMS. The graph separates countries by whether English if an official language or dominantly spoken; for example, India and Philippines have English as an official language, but it is not predominantly spoken by the populace.

Figure 3: Age at Arrival and English Proficiency Profile, Early 20th and 21st Century



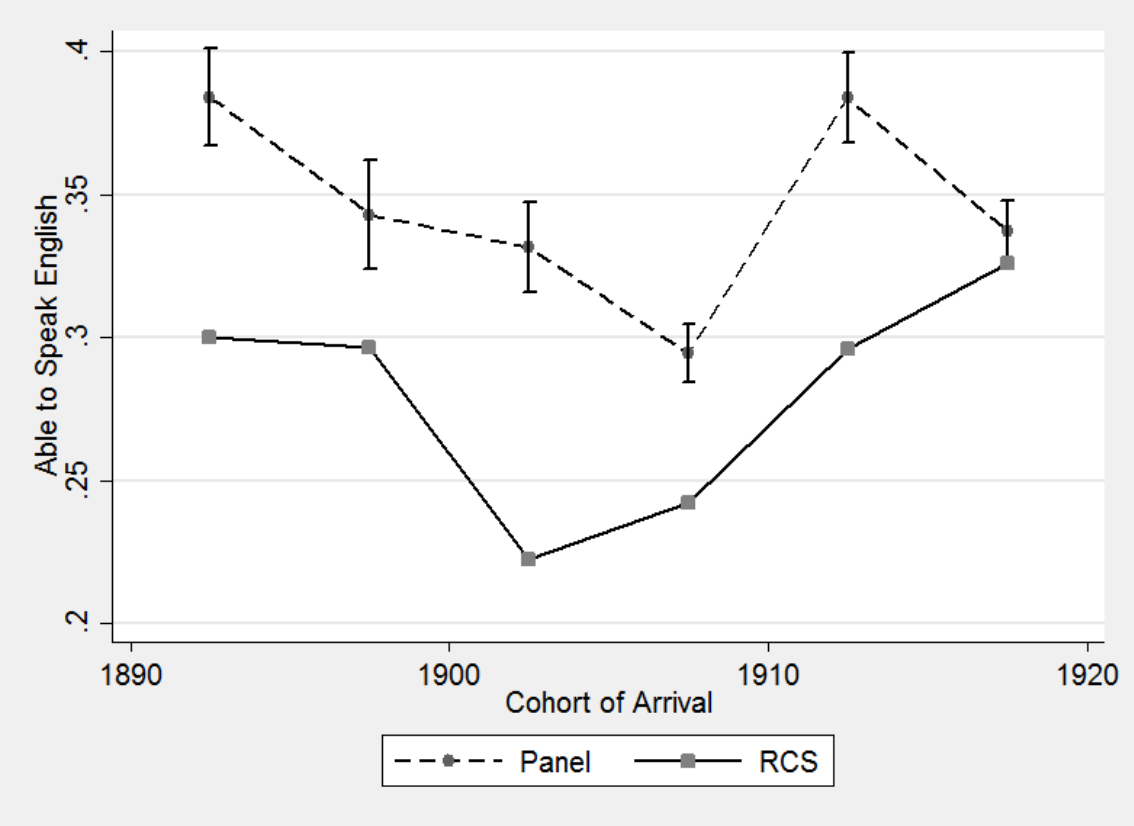
Notes: Data is from 1900-1930, 2000 Censuses and 2008-2012 ACS. The figure plots age-at-arrival fixed effects from a regression of ability to speak English on age at arrival, age, year, cohort of arrival, country of birth, sex, and fraction of migrants from same birthplace in county.

Figure 4: Speed of Language Acquisition Across the 20th century



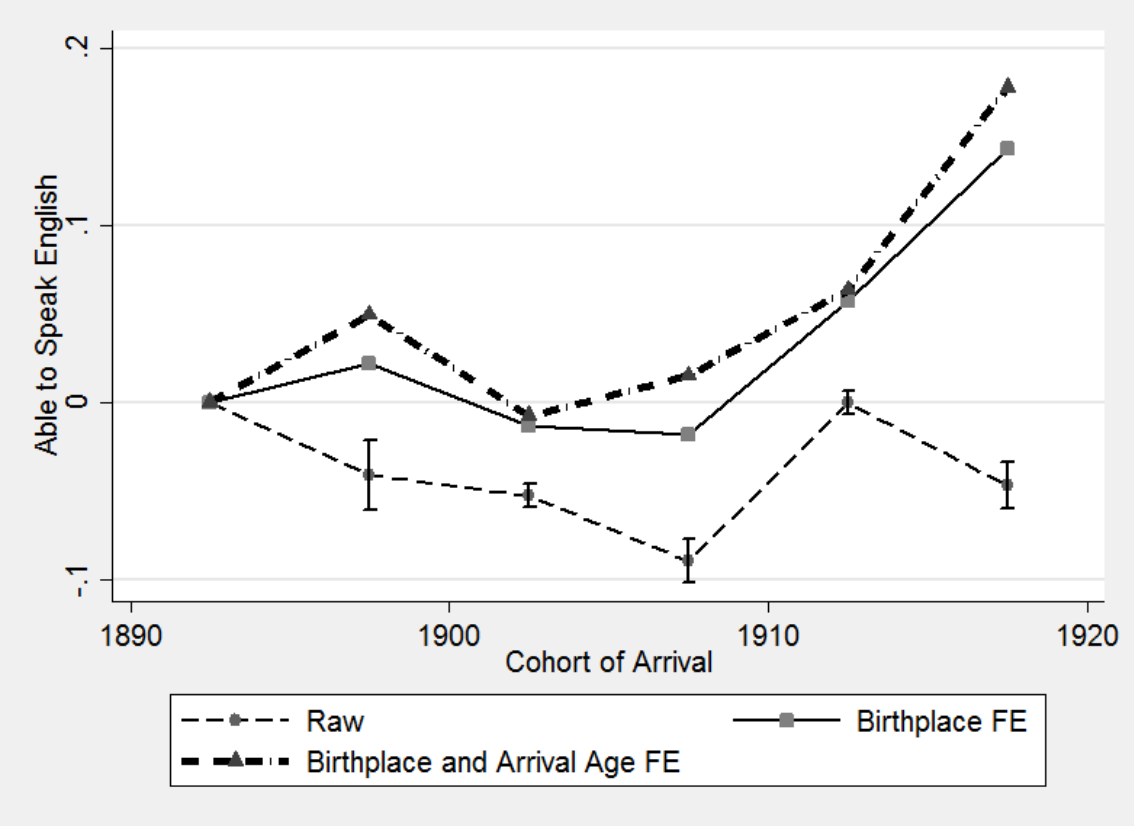
Notes: Data is from linked panel data 1900-1910, 1910-1920 and 1920-1930; the 1900-1930, 2000 IPUMS samples and 2008-2012 ACS. The figure shows the mean ability to speak English in the years after arrival. RCS stands for repeated cross section.

Figure 5: Cohort Effects



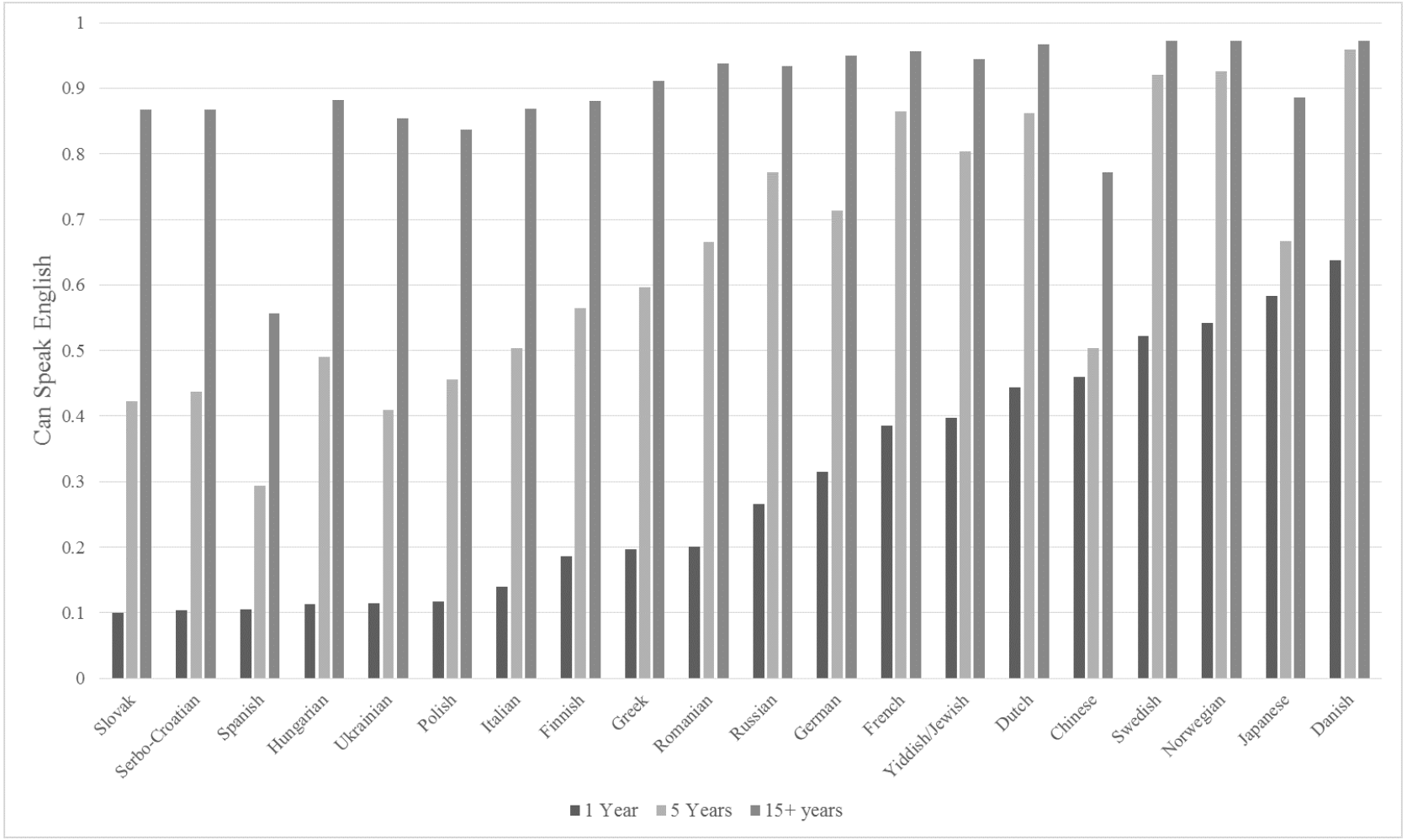
Notes: Data is from IPUMS (1900-1930) and linked samples (1900-1910; 1910-1920; 1920-1930). The figure shows the mean ability to speak English for arrivals by arrival cohort. RCS stands for repeated cross section.

Figure 6: Cohort decline largely due to shifting birthplace composition



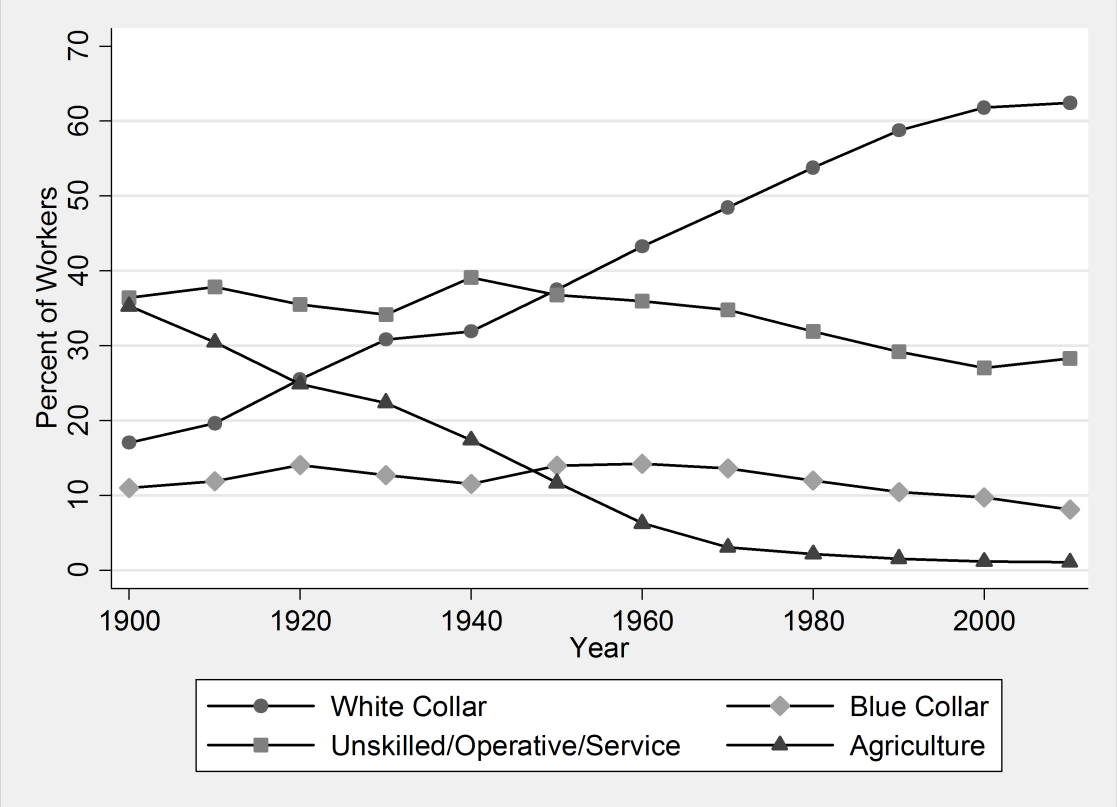
Notes: Data is from IPUMS (1900-1930) and linked samples (1900-1910; 1910-1920; 1920-1930). The figure shows the mean ability to speak English for arrivals by arrival cohort. RCS stands for repeated cross section.

Figure 7: Speed of Language Acquisition by Ethnicity, 1910 to 1920



Notes: Data is from linked sample between 1910 and 1920. The figure shows the mean ability to speak English by ethnicity, as proxied by mother's tongue in the 1920 census.

Figure 8: Share of Labor Force by Sector, 1900-2010



Notes: Data is from Katz and Margo (2014).

Online appendix, not meant for publication

Table A1: Occupational group transitions for migrants who learned to speak English (1910-1920)

Occupational Group in 1910		Occupational Group in 1920						
		Professional	Sales/ Clerical	Semi-skilled	Unskilled Manual/Oper	Farmer	Laborer	Total
	Professional	867	199	599	857	106	607	3,235
		26.8	6.15	18.52	26.49	3.28	18.76	100
	Sales/ Clerical	656	375	528	747	69	488	2,863
		22.91	13.1	18.44	26.09	2.41	17.05	100
	Semi-Skilled	1,957	618	6,360	4,388	633	3,487	17,443
		11.22	3.54	36.46	25.16	3.63	19.99	100
	Unskilled Service/ Operative	3,369	1,216	6,508	13,343	1,326	8,408	34,170
		9.86	3.56	19.05	39.05	3.88	24.61	100
	Farmer	129	26	261	327	1,040	507	2,290
		5.63	1.14	11.4	14.28	45.41	22.14	100
	Laborer	3,955	1,573	11,156	18,927	3,484	21,873	60,968
		6.49	2.58	18.3	31.04	5.71	35.88	100
	Total	10,933	4,007	25,412	38,589	6,658	35,370	120,969
		9.04	3.31	21.01	31.9	5.5	29.24	100

Table A2: Occupational group transitions for migrants who always knew how to speak English (1910-1920)

		Occupational Group in 1920						
		Professional	Sales/ Clerical	Semi-skilled	Unskilled Manual/Oper	Farmer	Laborer	Total
Occupational Group in 1910	Professional	6,140	1,201	2,555	2,802	529	1,282	14,509
		42.32	8.28	17.61	19.31	3.65	8.84	100
	Sales/ Clerical	4,383	2,916	2,422	2,963	364	1,136	14,184
		30.9	20.56	17.08	20.89	2.57	8.01	100
	Semi-Skilled	6,594	1,984	19,880	8,264	2,303	4,952	43,977
		14.99	4.51	45.21	18.79	5.24	11.26	100
	Unskilled Service/ Operative	8,094	2,981	11,083	19,227	2,556	8,084	52,025
		15.56	5.73	21.3	36.96	4.91	15.54	100
	Farmer	433	150	887	758	3,547	1,083	6,858
		6.31	2.19	12.93	11.05	51.72	15.79	100
	Laborer	3,664	1,532	9,653	11,815	5,305	13,633	45,602
		8.03	3.36	21.17	25.91	11.63	29.9	100
	Total	29,308	10,764	46,480	45,829	14,604	30,170	177,155
		16.54	6.08	26.24	25.87	8.24	17.03	100

Table A3: Occupational group transitions for migrants who never learned to speak English (1910-1920)

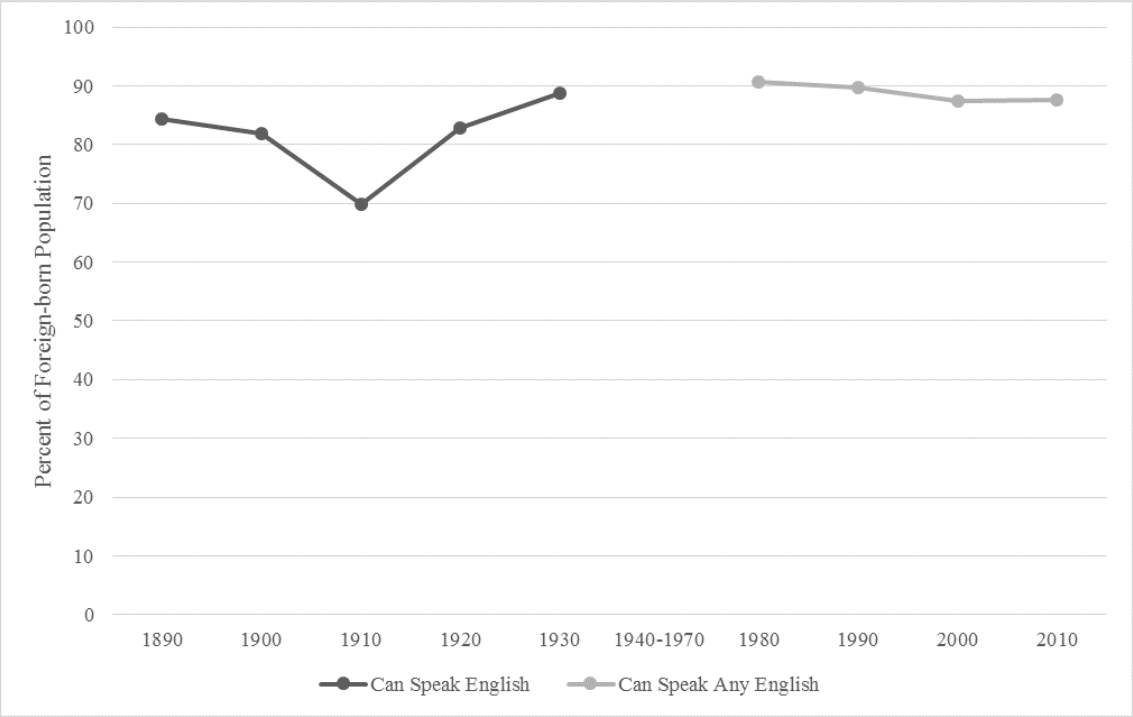
Occupational Group in 1910		Occupational Group in 1920						
		Professional	Sales/ Clerical	Semi-skilled	Unskilled Manual/Oper	Farmer	Laborer	Total
	Professional	63	15	45	105	9	129	366
		17.21	4.1	12.3	28.69	2.46	35.25	100
	Sales/ Clerical	39	44	37	69	10	87	286
		13.64	15.38	12.94	24.13	3.5	30.42	100
	Semi-Skilled	94	36	645	597	57	752	2,181
		4.31	1.65	29.57	27.37	2.61	34.48	100
	Unskilled Service/ Operative	197	97	662	2,108	174	2,116	5,354
		3.68	1.81	12.36	39.37	3.25	39.52	100
	Farmer	7	9	26	41	186	135	404
		1.73	2.23	6.44	10.15	46.04	33.42	100
	Laborer	304	155	1,429	3,341	556	7,317	13,102
		2.32	1.18	10.91	25.5	4.24	55.85	100
	Total	704	356	2,844	6,261	992	10,536	21,693
		3.25	1.64	13.11	28.86	4.57	48.57	100

Table A4: Association between Speaking English and outcomes, 2008-2012 ACS

	I Log (Income)	II Log (Occ Score)
Speak English	0.250***	0.118***
	(0.00580)	(0.00285)
Education (years)	0.0580***	0.0296***
	(0.000453)	(0.000206)
Fraction of own Migrants in County	-0.0796***	0.0156***
	(0.0115)	(0.00576)
Log (County Pop)	0.0118***	0.00971***
	(0.00112)	(0.000501)
Country of Birth FE	Y	Y
Age FE	Y	Y
Observations	370,939	370,939
R-squared	0.194	0.252

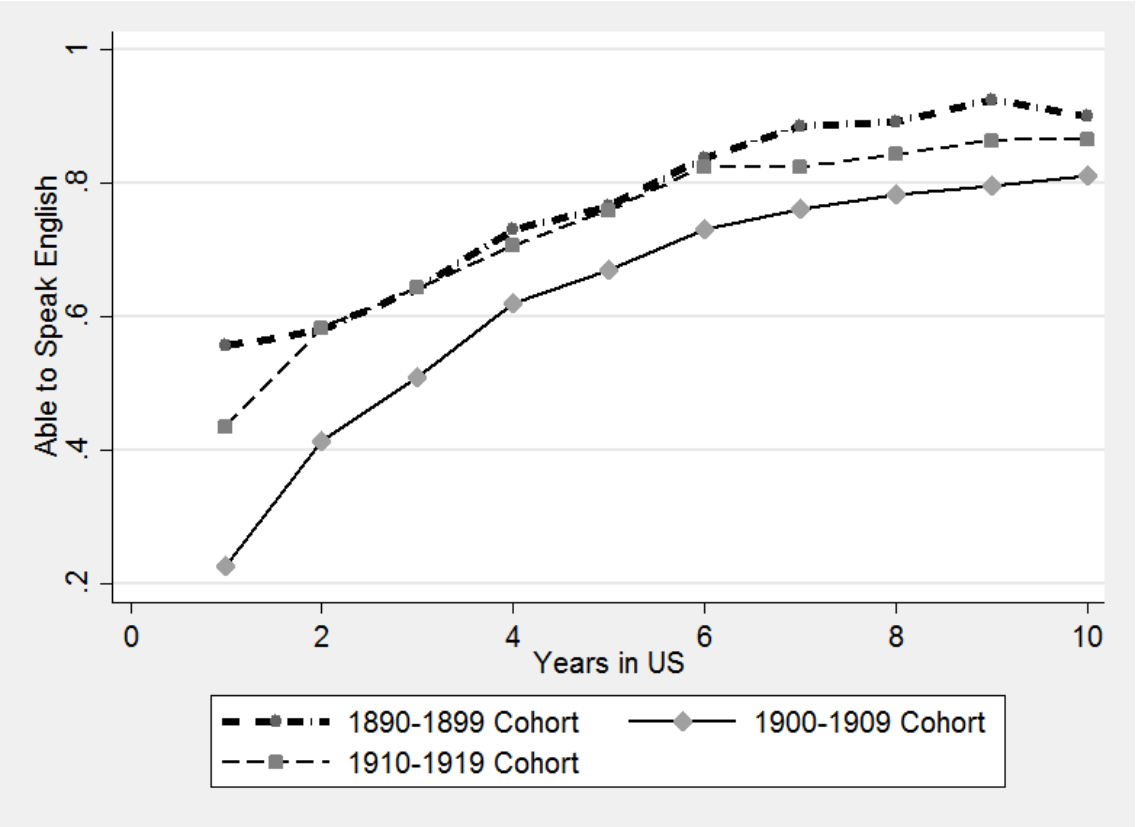
Notes: Data is from the 2008-2012 ACS. Speaking English is coded as 1 if a migrant is able to speak any English, whether not well or very well. Sample is male migrants from non-English speaking countries aged 25 to 60. Occupational score is based on the median wage earnings by occupation for all males in the 2008-2012 ACS.

Figure A1: Able to Speak English, 1900 to 2010



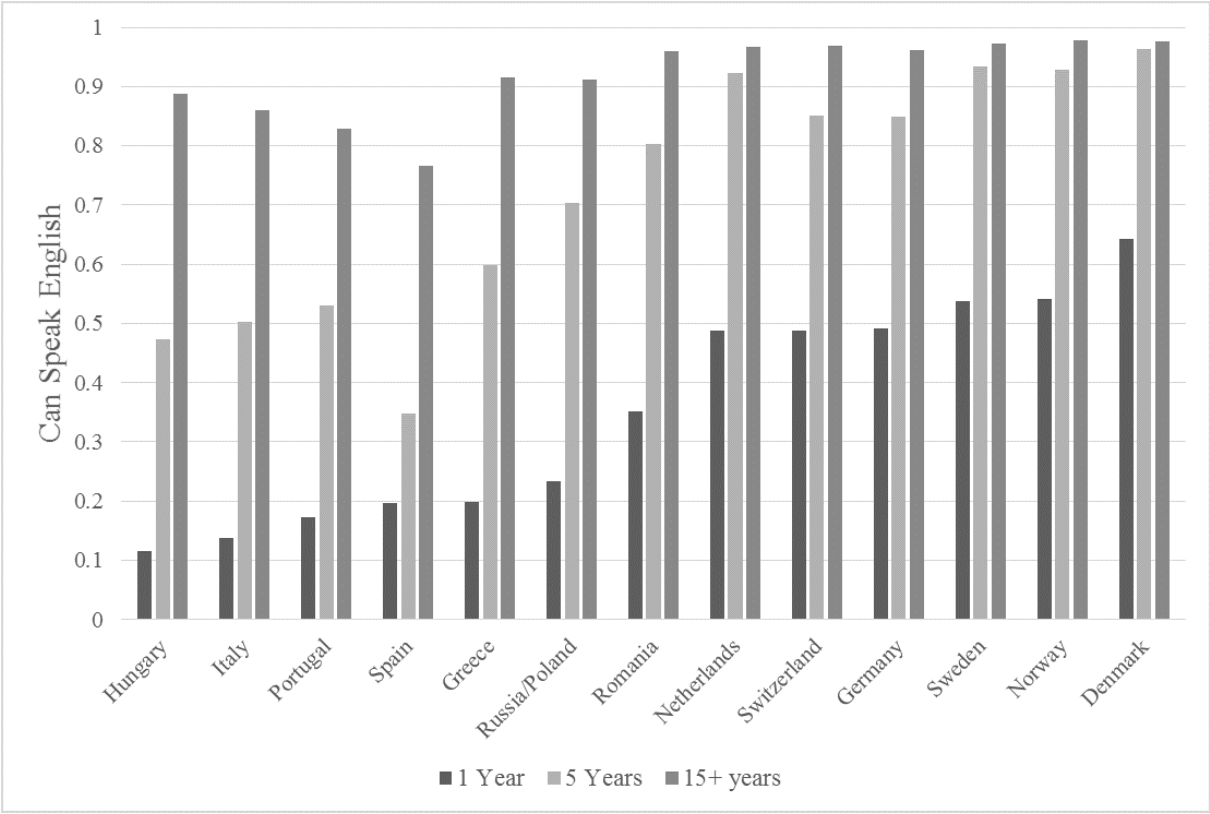
Notes: Data is from IPUMS (1900-1930; 1980-2010).

Figure A2: Rate of English Acquisition Raw Data, 1890-1919 Cohorts



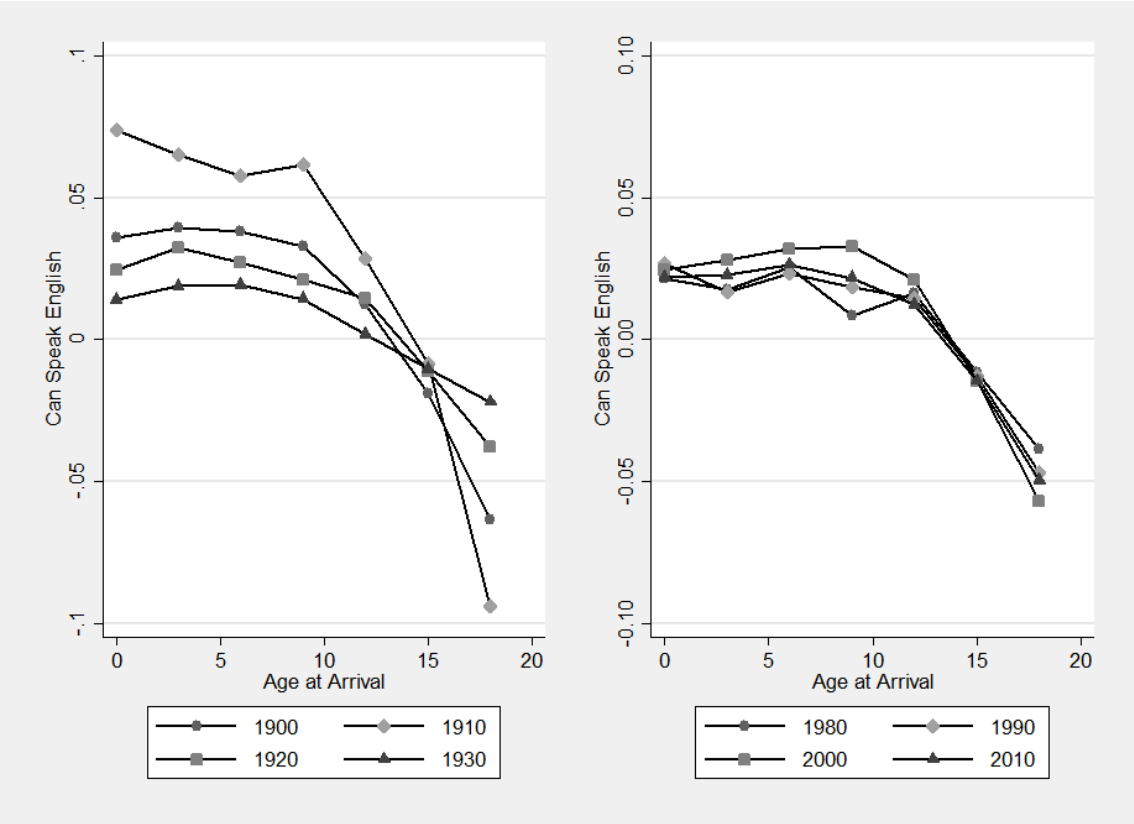
Notes: Data is from linked data from 1900-1910, 1910-1920, and 1920-1930.

Figure A3: Speed of Language Acquisition by Country of Birth, 1910 to 1920



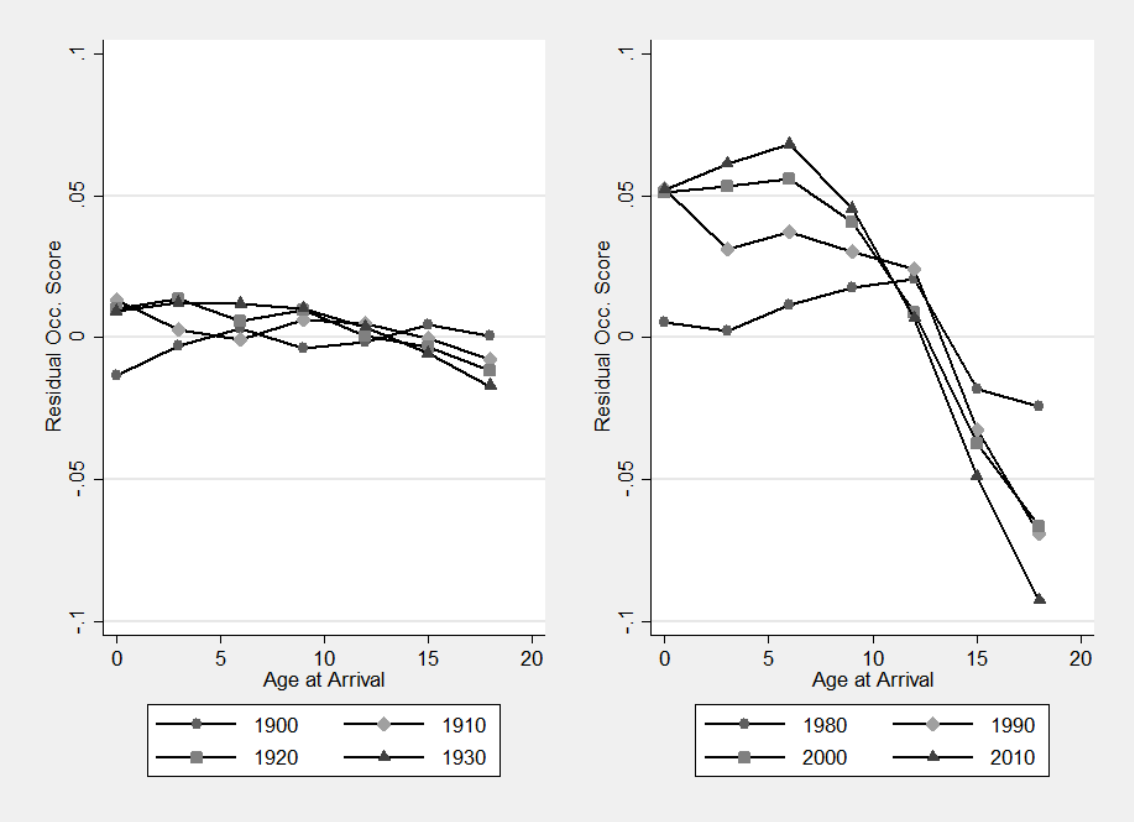
Notes: Data is from linked panel data from 1910 to 1920. The figure shows the mean ability to speak English in the years after arrival.

Figure A4: Age-at-Arrival and English Ability as an Adult, 1900 to 2010



Notes: Data is from IPUMS (1900-1930; 1980-2010). The figure shows the residuals of the ability to speak English after removing the effects of age, sex and country of birth. The right hand side graph treats migrants as able to speak English if they speak not well, well or very well.

Figure A5: Age-at-Arrival and Occupation Score as an Adult, 1900 to 2010



Notes: Data is from IPUMS (1900-1930; 1980-2010). The figure shows the residuals of the ability the log occupational score after removing the effects of age, sex and country of birth. The right hand side graph treats migrants as able to speak English if they speak not well, well or very well.

B Linking Process

I create a few new linked datasets in this paper: one tracks individuals from the 1900 5% sample to the full-count 1910 census, another tracks the 1910 full-count to 1920 full-count census, and the other tracks from the 1920 full-count to the 1930 full-count census. Note that I do not use the full-count census for year 1900 because the English variable has yet to be digitized.

I mostly follow a linking process based on probabilistic matching techniques that has been used by Massey (2016), Mill and Stein (2016) and Pérez (2016). These papers aim to improve on prior linking techniques by Abramitzky, Boustan and Eriksson (2012, 2014) and Ferrie (1996). Before linking, I drop numerous individuals from the sample including:

- Native born and migrants born in English-speaking countries
- Females
- Those under the age of 10 and over the age of 40 in the earlier census. Therefore, the oldest one should be when linked to a census ten years later is 50 years old. This is to reduce bias from deaths.
- Those who arrived outside of the previous ten years (e.g., drop those who arrived prior to 1890 in the 1900 Census).

After making these restrictions, I pre-process the first name strings to get ride of diminutives and nick names. For example, the name William may appear as Will, Willy, Willie, Wm, Billy, Bill, or Billey. I convert nicknames to full names.

Then I link migrants from Census X to Census Y (e.g., 1900 Census to 1910 Census) in the following manner:

1. I drop migrants in each Census who have the same first name (fixed for nicknames and diminutives), last name, year of birth, country of birth, and year of arrival. These individuals have the same set of linking variables making it impossible to separate which person would be the true link. Note that the year of arrival variable is an additional variable to link on, something not used by Abramitzky, Boustan and Eriksson (2014); the benefit of this is that there are fewer common names dropped in this step.
2. I group migrants into blocks for each Census. The blocks are all observations with the same first name (standardized by the NYSIIS algorithm), last name (also standardized by NYSIIS), year of birth, country of birth and year of arrival.

3. I create all of the pairwise combinations between Census X and Census Y that match on the blocking variables. I expand the window for year of birth to allow for matches that have a difference in year of birth by three years, and a difference for year of arrival by three years. This is because both age and year of arrival are often rounded to the nearest zero or five.

4. I calculate the Jaro-Winkler distance for the first name and last name. The distance ranges from 0 to 1 where 0 represents an exact match between strings and 1 measures no match.

5. I create a match score for each of the pairwise combinations between X and Y. This is the sum of the difference in year of birth, difference in year of arrival, the Jaro-Winkler distance for the first name and Jaro-Winkler distance for the last name. Note that a given observation in Census X may be matched to multiple observations in Y at this point because of the window for year of birth and year of arrival.

6. I keep the minimum matching score for each block. I also drop observations that have a matching score too close to the second minimum match score, as suggested by Mill and Stein (2016) and Pérez (2016); the cut-off I use is that the match score for the minimum observation may be at most 80% of the minimum for the second least observation. This rule is arbitrary, but the results are not sensitive to this restriction. If an observation in X is linked to multiple in Y because of the year of birth or year of arrival window, I keep the one with the minimum score

Note that this method of relying on a match score may be improved in the future by the use of training data as in Feigenbaum (2016) and Parman (2015); however, this method is less costly and it has been shown to be the best amongst various linking algorithms in historical papers in terms of the balance of accuracy and linking rates (Massey, 2016).

The linking rates are shown in Table B1. The linking rates are low in absolute value (17-22 percent), but they are similar to other papers that use this method. Linking rates are not 100 percent for a variety of reasons: death, return migration and under enumeration necessarily lowers the linking rate; further, historical censuses are handwritten and may be difficult for data entrants to key in the correct name, which may not have been written down by the enumerator properly in the first place.

C Discussion of Data Choices

C.1 Coding of Countries

I group the following countries into one country to maintain consistency across the 1900 to 1930 data:

- Russia includes Russia, Poland, Latvia, Lithuania, Estonia, and any Baltic state
- Hungary includes Hungary, Czechoslovakia and Yugoslavia

The following are how I code "Old" source countries, "New" source countries, and English-speaking countries.

- Old source countries include: Canada, Denmark, Finland, Iceland, Norway, Sweden, England, Scotland, Wales, Ireland, Belgium, France, Luxembourg, Netherlands, Switzerland, Australia and New Zealand. New source countries are all others, including those from Eastern Europe, Asia, Africa and Central/South America.
- English-speaking countries in 1900 are Australia, (English) Canada, England, Ireland, Scotland, India, British West Indies colonies (e.g., Antigua, Barbados, etc.), New Zealand, the Philippines and Wales. The British and American colonies are a very small number of the overall migrant total, so coding them as non-English does not qualitatively affect results. For the years 1910 to 1930, English is coded as whether the mother's tongue is English or one is born in an English-speaking countries for 1900. Between 1980 and 2010, English countries are coded to follow Bleakley and Chin (2010).

D Recreating Bleakley and Chin's (2004) Empirical Strategy

D.1 Age-at-Arrival Analysis

One way to measure the English premium is to exploit a well-defined relationship between age at arrival and English fluency as an adult. In particular, second language acquisition is easier at younger ages during the so-called "critical period;" after age 8-11, the ability to acquire a second language decreases at a linear rate (Bleakley and Chin, 2004). To fix ideas for how one can test this empirically, researchers typically estimate a variation of the following regression using a sample of adults who arrived as children:

$$SpeakEng_{ij} = a_0 + a_1 1[AgeArrival \geq 8]_{ij} + a_2 NonEngSpeakCntry_{ij} + \mathbf{a_3} 1[AgeArrival \geq 8] \times NonEngSpeakCntry_{ij} + \gamma_j + \Pi' X_{ij} + \upsilon_{ij} \quad (3)$$

where $SpeakEng_{ij}$ is an indicator that equals one if individual i from country of birth j can speak English as an adult. After controlling for country of birth (γ_j) and various individual-level observables (X_{ij}), the equation fundamentally estimates two age-at-arrival profiles: one for migrants from English-speaking countries and one for migrants from non-English-speaking countries. Using English fluency as the dependent variable, a_1 should equal zero as English-speaking migrants already know English no matter the age at arrival, and a_3 should be negative, reflecting the critical period of language acquisition.

To relate speaking English with labor market outcomes, one can estimate the same equation as above, but instead using wages as the dependent variable:

$$log(Wages_{ij}) = b_0 + b_1 1[AgeArrival \geq 8]_{ij} + b_2 NonEngSpeakCntry_{ij} + \mathbf{b_3} 1[AgeArrival \geq 8] \times NonEngSpeakCntry_{ij} + \eta_j + \Gamma' X_{ij} + \nu_{ij} \quad (4)$$

Given that those arriving at older ages from non-English-speaking countries generally have less English fluency, one would expect that these same migrants would also have lower wages; in other words, the expectation is that b_3 is negative. Indeed, this is exactly what studies find in present-day settings (Bleakley and Chin, 2004; Guven and Islam, 2015). To retrieve a precise estimate of the return to speaking English, studies use a two-stage least squares strategy where Equation 3 predicts English ability in the first stage with the interaction $1[AgeArrival \geq 8] \times NonEngSpeakCntry_{ij}$ as the excluded instrument. The results from this first stage are then used in the second stage for a regression of wages on predicted English ability.[29] The 2SLS estimate of the premium for English is essentially a ratio of the reduced form and the first stage ($\widehat{\beta_{IV}} = \frac{\widehat{b_3}}{\widehat{a_3}}$). In the typical study, the instrument is not simply the interaction between a dummy variable for age at arrival above eight, but rather a linear interaction for past age 8 and being born in a non-English-speaking country. This reflects the linear decrease in English-speaking ability observed in the data (Bleakley and Chin, 2004). Finally, note that this IV estimates the effect for only those who are affected by the instrument, or the LATE.

[29] Thus, the reduced form equation of the outcome on the excluded instrument is Equation 4.

D.2 Regression Evidence

I run the age-at-arrival regression specifications for years 1900 to 1930 from Equation 3 with three dependent variables: the ability to speak English, ability to read and write, and occupational score. The control variables are purposely parsimonious − I only control for country of birth, current age, and sex − as other controls like current neighborhood or marital status could be considered outcomes of age at arrival. The results are shown in Table D1, which statistically verifies some of the patterns observed in the figures, but also reveals new findings. As expected from the critical period hypothesis, the coefficient on the interaction between arriving at or older than age 8 and being from a non-English-speaking country is negative. The effect is consistently negative across all four decades and estimates that arriving at an older age leads to about a 1.3 to 2.2 percentage point drop in the likelihood of speaking English as an adult.[30]

The second panel shows the relationship between age at arrival and literacy. Interestingly, despite older-arriving non-English speakers having lower English fluency for 1900 and 1910, these same migrants had similar literacy rates. Literacy reflects general literacy skills as the variable is defined as the ability to read and write in any language. By 1920 and 1930, those who arrived at older ages also had lower literacy rates, by 2.4 to 2.8 percentage points. This provides a first indication that the birth country composition of the *stock* for non-English sources is shifting toward more poor countries which had worse schooling environments.

The third panel shows the effect of a migrant's age at arrival on occupation. The results, especially in the years 1900 and 1910, show a perplexing correlation: migrants from non-English-speaking countries who arrived older than age 8, despite speaking English at lower rates, held jobs that paid about 2.5% *more* than otherwise similar migrants who arrived under the age of 8. The effect is similarly positive in 1910, and 1920 but is statistically indistinguishable from zero. Interestingly, 1920 is the same year that older arriving migrants had lower literacy rates in addition to lower English fluency rates. However, despite disadvantages that should seemingly penalize migrants in the labor market, they had the same occupational score. Finally, by 1930, the expected results hold: migrants who arrived at older ages had (a) lower English fluency rates (b) lower literacy rates and (c) worse-paying occupations.

The results for lower English fluency and occupation are not being driven by literacy: when controlling for literacy in the bottom two panels, the main results on English fluency

[30]However, this estimate likely masks the effect on the fluency of a migrant; for example, in the late 20th century, the variable speaking English was recorded on a scale of 0 to 3, 0 being not able to speak English at all and 3 being able to speak English very well. A lot of variation exists in between these measures between zero to three, which is missed in the binary measure from the early 20th century.

and occupation do not change. Finally, for completeness, in Panel A of Table D2 I show the 2SLS estimates, recreating the Bleakley and Chin (2004) specification (i.e., using a linear decrease in English-speaking ability rather than a dummy variable). The 2SLS strategy estimates an unreasonably large negative premium for English in 1900 (240 log points) and an unreasonably large positive premium in 1930 (91 log points). Clearly, the age-at-arrival analysis does not isolate the English premium in the early 20th century.

Why do the results change so drastically between 1900 and 1930? In the second row, I show what happens when the birth country composition from 1900 to 1930 is fixed at 1900 levels. For example, Germany was 27.5% of the sample in 1900, but only 8.9% of the sample in 1930; I reweight the Germans from 1910 to 1930 to be 27.5% of the sample. The results from this weighting show that there is a significant negative premium from 1900 to 1930 using the age-at-arrival analysis. Therefore, one reason for the difference in premium from 1900 to 1930 in the unweighted sample (from negative to positive) is that the composition of non-English-speaking sources shifted to poorer countries who had worse outcomes. This suggests that the instrumental variable is correlated with other aspects of the child's environment rather than just English ability.

In summary, the results from this section show that migrants held higher paying jobs in 1900 and 1910 *despite* having lower English skills. This implies that English was relatively unimportant in 1900 and 1910 compared to other non-language human capital that is correlated with age at arrival. By 1930, this relationship is no longer true, which could either be because English was more important or that non-English sources came from poorer countries. Ultimately, the age-at-arrival analysis likely does not isolate the English premium throughout the entire 1900-1930 time period because of the violation of the exclusion restriction due to English-speaking sources having different childhood environments from non-English-speaking sources.

Table B1: Linking Rates for Non-English-Speaking Source Countries

Base Year	Observations	Linked Year	Linked Sample Size	Linking Rate
1900	47,088	1910	10,422	22.1
1910	2,194,927	1920	367,036	16.7
1920	1,534,982	1930	307,529	20.0

Notes: The linked sample sizes for 1900 to 1910, 1910 to 1920, and 1920 to 1930 censuses. Note that I only link non-English-speaking source countries.

Table D1: Age-at-arrival Effects, 1900 to 1930

	I 1900	II 1910	III 1920	IV 1930
Outcome: Can Speak English				
Arrived Older than 8 ×	-0.0136***	-0.0218***	-0.0189***	-0.0174***
Non-English Speaking Country	(0.00162)	(0.00457)	(0.00543)	(0.00135)
Arrived Older than 8	-0.000368	-0.00480***	0.000313	0.00146**
	(0.000515)	(0.00177)	(0.00414)	(0.000699)
Outcome: Can Read and Write				
Arrived Older than 8 ×	0.00394	-0.00431	-0.0239***	-0.0282***
Non-English Speaking Country	(0.00333)	(0.00560)	(0.00526)	(0.00189)
Arrived Older than 8	-0.00990***	-0.00776**	-0.00316	-0.000504
	(0.00218)	(0.00336)	(0.00347)	(0.00118)
Outcome: Occupational Score				
Arrived Older than 8 ×	0.0254***	0.0164	0.00769	-0.00951*
Non-English Speaking Country	(0.00534)	(0.0114)	(0.0114)	(0.00499)
Arrived Older than 8	-0.00913**	-0.0127	-0.0258***	-0.0303***
	(0.00393)	(0.00957)	(0.00995)	(0.00449)
Outcome: Can Speak English, Control for Literacy				
Arrived Older than 8 ×	-0.0142***	-0.0209***	-0.0141***	-0.0117***
Non-English Speaking Country	(0.00158)	(0.00449)	(0.00537)	(0.00131)
Arrived Older than 8	0.00112**	-0.00321*	0.000947	0.00156**
	(0.000556)	(0.00179)	(0.00414)	(0.000707)
Outcome: Occupational Score, Control for Literacy				
Arrived Older than 8 ×	0.0249***	0.0170	0.0104	-0.00596
Non-English Speaking Country	(0.00532)	(0.0114)	(0.0113)	(0.00498)
Arrived Older than 8	-0.00788**	-0.0116	-0.0255**	-0.0303***
	(0.00392)	(0.00955)	(0.00994)	(0.00449)
N	68,227	16,184	20,342	110,265

Notes: Data is from IPUMS (1900-1930). The dependent variable is listed for each panel. Each column represents the same regression but for a different census. Sex, current age, and country of birth are also controls in the regression. Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01

Table D2: Recreating Bleakley and Chin (2004) 2SLS Estimate

Normal Weights				
	1900	1910	1920	1930
Can Speak English	-2.399***	-0.997*	-0.193	0.908***
	(0.489)	(0.525)	(0.414)	(0.286)
1900 Country Composition				
Can Speak English	-2.399***	-1.933**	-4.074	-3.177***
	(0.489)	(0.923)	(3.312)	(1.226)
Observations	68,226	16,182	20,320	109,978

Notes: Data is from IPUMS (1900-1930). The dependent variable is the logged occupational score. The second row reweights the 1910 to 1930 samples to have the same birth country composition as in the 1900 sample. *p<0.10, **p<0.05, ***p<0.01.