NBER WORKING PAPER SERIES

THE REFUGEE ADVANTAGE: ENGLISH-LANGUAGE ATTAINMENT IN THE EARLY TWENTIETH CENTURY

Ran Abramitzky
Leah Platt Boustan
Peter Catron
Dylan Connor
Rob Voigt

Working Paper 31730
http://www.nber.org/papers/w31730

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023

We acknowledge the excellent research assistance of Victoria Angelova, Harriet Brookes Gray, Sarah Frick, Myera Rashid, and Noah Simon. Sima Biondi, Alicia Liu, Lori Mitrano, Lorenzo Rosas, Antigone Xenopoulos and Adam Zhang helped to collect variables from the oral history interviews. Jared Grogan, Bailey Palmer and James Reeves coded the interviews for accented speech. Tom Zohar oversaw audio transcription of missing transcripts. We appreciate suggestions from audiences at Universitat Automoma de Barcelona, UC-Berkeley, European Social Science History Association, Harvard, University of Nottingham, Pompeu Fabra, University of Chicago, and University College, London. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Refugee Advantage: English-Language Attainment in the Early Twentieth Century
Ran Abramitzky, Leah Platt Boustan, Peter Catron, Dylan Connor, and Rob Voigt
NBER Working Paper No. 31730
September 2023
JEL No. J15,N32

ABSTRACT

The United States has admitted more than 3 million refugees since 1980 through official refugee resettlement programs. Scholars attribute the success of refugee groups to governmental programs on assimilation and integration. Before 1948, however, refugees arrived without formal selection processes or federal support. We examine the integration of historical refugees using a large archive of recorded oral history interviews to understand linguistic attainment of migrants who arrived in the early twentieth century. Using fine-grained measures of vocabulary, syntax and accented speech, we find that refugee migrants achieved a greater depth of English vocabulary than did economic/family migrants, a finding that holds even when comparing migrants from the same country of origin or religious group. This study improves on previous research on immigrant language acquisition and refugee incorporation, which typically rely on self-reported measures of fluency. Our findings are consistent with the hypothesis that refugees had greater exposure to English or more incentive to learn, due to the conditions of their arrival and their inability to immediately return to their origin country.

Ran Abramitzky
Department of Economics
Stanford University
579 Jane Stanford Way
Stanford, CA 94305
and NBER
ranabr@stanford.edu

Leah Platt Boustan
Princeton University
Industrial Relations Section
Louis A. Simpson International Bldg.
Princeton, NJ 08544
and NBER
lboustan@princeton.edu

Peter Catron
University of Washington
pcatro@nber.org

Dylan Connor
School of Geographical Sciences
and Urban Planning
Arizona State University
Tempe, AZ 85282
dsconnor@asu.edu

Rob Voigt
Northwestern University
robvoigt@northwestern.edu

From its founding, the United States has been an immigrant-receiving nation. More than one hundred million immigrants have settled in the US since 1850, and their settlement has (re)shaped US society. In recent decades, 15-20 percent of immigrants have arrived as refugees or asylum seekers in response to persecution or violence in their home countries (U.S. Department of Homeland Security 2016). Before the establishment of the modern refugee admission system in 1980, many immigrants moved to the US due to credible fear of persecution or violence, but were not officially labeled as 'refugees' (Arar and FitzGerald 2022).

Contemporary research often highlights the disadvantages that refugees face upon arrival in a destination country (the "refugee gap"). After settlement, refugees are able to lessen these disparities over time, but often fail to achieve parity with other immigrants and natives, especially in most European countries (Akresh 2008; De Vroome and Van Tubergen 2010; Connor 2010; Aydemir 2011; Cortes, 2004; Chin and Cortes, 2015; Bakker, Dagevos, and Engbersen 2017; Evans and Fitzgerald 2017; Kosyakova and Kogan 2022; Ruiz and Vargas-Silva 2018; Zwysen 2019; Fasani, Frattini, and Minale 2022). The United States, however, is a notable exception in achieving rapid refugee integration. Refugees in the U.S. are able to overcome their initial disparities and sometimes surpass other immigrants on a range of measures (Brell, Dustmann and Preston, 2020). In this article, we provide the first historical evidence on refugees' linguistic proficiency in the U.S., finding that refugees achieved higher rates of English attainment relative to other immigrants in the U.S. in the past as today.

Adding historical evidence provides a useful context for understanding why refugees are particularly successful in the U.S. today. The dominant account on refugee success in the U.S. argues that the integration of refugee populations is facilitated by access to resettlement programs and governmental assistance, and by the benefits associated with legal permanent residence and

eventual citizenship, two factors that emerged only in recent decades (Bloemraad 2006; Jiménez 2011; Waters and Pineau 2015). By contrast, other scholars emphasize that integration of refugees may differ from economic migrants even without governmental support due to fundamental aspects of the refugee experience – namely, the conditions of refugee arrivals and their inability to engage in imminent return (Friedberg, 2000; De Vroome and van Tubergen 2010; van Tubergen 2010; Dustmann and Gorlach 2016; Becker et al. 2020; Adda, et al. 2020). Because of their different migration motives and integration strategies, refugees may be more likely than economic migrants to be exposed to the native-born in neighborhoods, workplaces and other points of contact, or to make local investments in human and social capital (Jasso and Rosenzweig 1990).

This paper explores the process of refugee integration in a period – the early twentieth century United States – when refugees did not receive governmental support, nor did they enjoy an advantage in access to legal permanent residency or citizenship (Abramitzky and Boustan, 2017). Furthermore, governmental resettlement assistance was nonexistent, and what help there was came from ethnic or religious volunteers, who provided low levels of aid that was not tied to refugee status (Holman, 1996; Abramitzky, Boustan and Connor, Forthcoming). Instead, this private aid generally tied to class status, allowing poor and working class economic migrants and refugees to benefit equally. Given that we find rapid refugee integration in the U.S. in this earlier period, we argue that the government assistance provided to refugees today may not be a *necessary condition* for success, although it certainly may be a contributing factor.

Specifically, we compare the English language attainment of refugees to other immigrants in the United States in the early twentieth century. Despite being a multilingual nation, English is by far the dominant language in the US, and learning English provides important advantages for immigrant incorporation. We adapt measures of language attainment derived from the field of

linguistics to measure both complexity of speech (in vocabulary and sentence structure) and fluency (in accentedness and speech rate). We use a new dataset of 1,200 oral history transcripts and audio files of immigrants who arrived in the US through Ellis Island in the early twentieth century; overall, the transcripts of these interviews comprise more than 1.2 million words. Moreover, with the exception of accent, which we measure with subjective human ratings, we use computational methods to automatically extract these linguistic measures from oral history transcripts. Our approach adds to recent computational analyses used in the subfield of international migration, which have primarily focused on online digital trace data (e.g., Zagheni et al. 2017; Florès 2017; Spörlein and Schlueter 2021; Drouhot et al. 2022), and greatly expands on the recent introduction of linguistic competency measures into the literature, which focus on single dimensions of linguistic ability (Dollmann et al. 2020; Edele et al. 2015).

We find that immigrants who reported leaving Europe in response to war, violence or persecution ("refugee migrants") achieved a greater depth of English vocabulary than immigrants who came to join family or find better job opportunities ("economic/family migrants"), despite the lack of governmental support at the time.[1] This refugee advantage holds even after controlling for country of origin, religion and decade of arrival. The refugee advantage is also present after controlling for father's occupation and urban status during childhood to address concerns that refugee migrants may have come from higher than average socioeconomic backgrounds (see Guichard 2020; Spörlein, et al. 2020; Aksoy and Poutvaara, 2021; Abramitzky, Baseler and Sin,

[1] Throughout the paper, we use the terms "non-refugees" and "economic/family migrants" interchangeably to refer to immigrants who report coming to the US for reasons other than persecution or violence. The majority of this group (85 percent) report immigrating either to find economic opportunity or to follow family members.

2022). Jewish immigrants are over-represented in our refugee subsample, but the pattern of English attainment is, if anything, stronger in the non-Jewish immigrant population.

Our results derive from a new dataset of oral history transcripts and recordings, each roughly an hour in length. The interviews were conducted in the 1970s and 1980s in English, for a sample of immigrants who arrived in the US between 1893 and 1957. The sample was originally assembled by the Statue of Liberty – Ellis Island Foundation (SOLEIF). Oral histories provide a few unique advantages for understanding immigrant integration. First, they allow us to use actual immigrant speech derived from transcripts and audio files to construct more objective and detailed measures of English attainment taken from the field of linguistics, rather than the coarse measures of English fluency available on surveys. Prior work on linguistic attainment, both in the past and today, is based on self-assessed scales (e.g., answering that one speaks English "very well") from the census or other sources (Portes and Hao 1998; Espenshade and Fu 1997; Waters and Pineau 2016; Fischer and Hout 2006), or assess competency scores of children when they are still in school (Edele et al. 2015). Second, oral histories retrospectively capture aspects of an immigrant's socio-economic context before migration that usually remain unobserved, including religion, urban status in childhood, and father's occupation in the home country. Third, oral histories provide details on stated reason for migration, allowing us to classify migrants into refugees and economic/family migrants based on self-perception rather than relying on proxies for refugee status like nationality and year of arrival, which are often used to study refugees in the US (see Chin and Cortes 2015; Donato and Ferris 2020).[2] Importantly, in our context, immigrants from the

[2] FitzGerald and Arar (2018) have emphasized that the lack of information on refugee status in standard datasets has "analytically hobbled" research on refugee populations. Garip (2016) provides another example of creating a typology of immigrants using information from qualitative interviews and survey data. Two surveys identify modern refugees in the US by visa status – the New Immigrant Survey, used in this article below and the Annual Survey of Refugees. However,

same country of origin, arrival period and religious group report different reasons for migration, allowing us to make comparisons *within rather than across groups*.

Arar and FitzGerald (2022) argue that migration occurs on a "continuum of compulsion and freedom." On one end of the spectrum are immigrants who move to improve their quality of life, and on the other are immigrants escaping life and death circumstances. Classifying immigrants on a binary as economic/family migrants or refugees can often miss the gradations of compulsion that underlie the decision to leave home. We use double coding of each oral history to overcome this limitation, defining immigrants who are classified as refugees by both rounds of coders as "refugees" and immigrants who are classified as refugees by only one round of coders as "mixed reason" (persecution-economic) movers. The most common reasons for classification as a "mixed reason" mover are stage migration (an immigrant who flees imminent danger for a third country before moving to the US) or family migration (an immigrant who hastens to join earlier economic migrants due to unexpected conflict).

We end the paper with a comparison of our historical findings to the best available source of modern data on the refugee population – the New Immigrant Survey (NIS). Unlike the refugees in our historical dataset, modern refugees enjoy both legal status and government relocation supports, allowing us to compare the attainment of English proficiency across two legal regimes. We find a similar pattern in the modern data, whereby refugees are more likely than other legal permanent residents to report speaking some English. Access to language classes is one of the most common forms of refugee assistance today. Notably, we find that refugees enjoy greater language ability today even after controlling for taking a recent English class or an English class

---

both of these sources are available for a limited number of years (see Akresh 2008; Tran and Lara-Garcia 2020). European administrative data often contains more detailed information on refugee status from legal/visa category.

before arrival. Both in the past and today, we find evidence that refugees achieve more complete linguistic assimilation even without governmental supports.

Background and conceptual framework: Refugee integration and linguistic attainment

Refugee integration

The integration of refugees has been of considerable interest in recent years with worldwide flows from Syria, Afghanistan, and Ukraine among other conflict zones. One hundred million people have been forcibly displaced from their homes due to ongoing conflicts throughout the world (UNHCR 2022). Although many refugees today remain internally displaced or live in refugee camps, some are resettled in other countries.

Scholars argue that resettled refugees integrate more quickly in the United States than non-refugee migrants because they benefit from programs that act as effective integration policies (Bloemraad 2006; Jiménez 2011; Waters and Pineau 2015; Zolberg 1988; Menjivar 2000; Tran and Lara-Garcia 2020). Both the *segmented assimilation* and *neo-assimilation* approaches emphasize that refugees benefit from favored treatment in formal institutions and law, including government resettlement assistance and access to legal permanent residence (Portes and Rumbaut 2001; Alba and Nee 2003). Refugees often have access to occupational training, English-language instruction, and support through community organizations (Luthra, Soehl, and Waldinger 2018). In some cases, refugees are offered classes in English and in other labor market skills – e.g., how to search for employment – while housed in temporary facilities even before arrival (FitzGerald and Cook-Martin 2014). In the modern period, refugees also benefit from the fact that they enter the U.S. with legal permanent residence (Portes and Rumbaut 2001; Alba and Nee 2003). By contrast, some economic migrants enter the U.S. without papers; overall, one quarter of the

immigrant population is undocumented today (Lopez, et al. 2021). Although, in some cases, financial assistance to refugees is minimal given structural and resource constraints of resettlement agencies (Fee 2019; Fee and Arar 2019; Gowayed 2019), the positive governmental reception and legal status alone may facilitate refugee integration.

Even without access to resettlement programs, migrants who arrive fleeing war or persecution may have distinct integration strategies that differ from migrants who arrive seeking economic opportunity (Cortes, 2004; Becker et al. 2020). Thus, even absent formal assistance, the conditions of refugee migration may shape how refugees respond to social and economic incentives and how they make constrained decisions about their future. First, many refugees do not expect to return to their home country in the years after their arrival given the political conditions that prompted their initial move, whereas economic migrants may return whenever they choose and often plan for a shorter or temporary stay (Cortes 2004; Kosyakova et al. 2022; Gould 1980; Bandiera, et al. 2013). Immigrants who plan to return home make minimal investments in US-specific human capital because they are unlikely to realize the long-term gains (Piore 1979). By contrast, historical refugee migrants expressed more willingness to invest in their new country by, for example, developing stronger social networks and opening local bank accounts (Anbinder, Ó Gráda, and Wegge, 2019). Second, today's refugees often arrive without a large base of co-ethnics and therefore have greater exposure to the native-born, although it is not clear that the same pattern was true in the past (Alba and Nee 2003; De Vroome and Van Tubergen 2010). Indeed, many countries today intentionally resettle refugees in dispersed locations to prevent the development of ethnic enclaves, which can slow integration processes (Edin, Fredriksson, Aslund, 2003, Arar and FitzGerald 2022). Economic migrants, by contrast, often arrive with the aid of their social networks and find employment and places to live within enclaves at first arrival (Massey et

al. 1987). Refugees who live outside of enclaves may enjoy increased exposure to the native-born, which encourages speaking English and finding employment alongside the native-born (Laliberté, 2019).

The literature suggests that modern refugees may integrate quickly for two main reasons: access to governmental supports and distinct refugee integration strategies. We argue that, if government support is the crucial factor determining refugee success today, we should not see as rapid assimilation for refugee immigrants in the past who did not benefit from governmental assistance or advantages in attaining legal permanent residency. The fact that we do see rapid immigrant assimilation in the past speaks against the central role of governmental supports in supporting refugee assimilation and instead points toward distinct integration strategies deployed by immigrants fleeing persecution.

Linguistic assimilation

Destination-language acquisition is one of the first steps in the assimilation process. Immigrants must use destination-language in different domains to either get by or get ahead (Fishman 1972, Lieberson 1981, Estrada 2007, Kasinitz et al. 2008). Despite the fact that US policy has never explicitly made English a requirement for entry, the US "is a great destroyer of languages, as sooner or later, English almost always reigns supreme" (Luthra, Soehl, and Waldinger 2018: 211).

Language attainment begins in the immigrant generation. Modern studies find that learning English accelerates upward socioeconomic mobility for immigrants from non-English speaking countries (Bean and Stevens 2003). Destination-language acquisition facilitates skill transfers from the sending country and improves interactions with the local population.

Contemporary immigrant populations who report being able to speak the destination-country language fluently enjoy higher wages. For instance, Jasso and Rosenzweig (1989) find that, in the US, the ability to speak English is associated with a 10 percent gain in wages relative to immigrants who cannot speak English. Similarly, Dustmann and Fabbri (2003) find that English proficiency is associated with 18-20 percent higher earnings in the UK. Ward (2020) documents that the ability to speak English was also associated with economic success in the early 20th century; immigrants who learned English were 8 percentage points less likely to work in low-paid laborer positions. As such, the first generation learns enough English to get by in various settings and is an important marker of assimilation and integration (Portes and Hao 1998; Espenshade and Fu 1997; Waters and Pineau 2016).

Achieving linguistic fluency is a complex process determined by exposure to the new language, active efforts at language learning, and interactions with personal attributes (including age). Exposure matters: immigrants who live in an enclave receive less exposure to the dominant language and therefore may be less likely to become fluent (Chiswick and Miller 1996, 2001). By contrast, immigrants with a US-born child or US-born spouse may hasten their fluency if these close family members serve as a language teacher (Pagnini and Morgan, 1990; Bean and Stevens 2003; Kuziemko 2014). Even within enclaves, immigrants who know English may enjoy greater status since they are able to bridge contact with the English-speaking world (Morawaska 2004). Beyond exposure, however, the ability to learn a new language is also associated with age at arrival because younger individuals are cognitively better able to learn new languages than are older individuals (Espenshade and Fu 1997; Chiswick and Miller 2001; Bean and Stevens 2003; Bleakley and Chin 2004). School attendance or enrollment in formal language classes in the host

society also assists with fluency (Sassler, 2006; Lleras-Muney and Shertzer 2015; Gowayed 2019; Carter 2009; Arendt, et al. 2020).

Refugees may take different pathways to English proficiency than do other immigrants. Several initial factors affecting refugees may place them at an initial disadvantage in language attainment (Kosyakova et al. 2022). Compared to economic migrants, refugees tend to have poorer mental health (van Tubergen and Kalmijn 2005) and may have less pre-migration exposure to the destination-language (Kristen and Seuring 2021). In addition, the labor market orientation of economic migrants may make them more likely to learn a new language to get ahead in the labor market. These factors may depress the likelihood that refugees achieve fluency because they can reduce exposure to the destination-language (Kosyakova et al. 2022; Espenshade and Fu 1997).

As noted above, however, refugees are often able to catch-up with economic migrants on a number of social indicators, including language attainment in the U.S. Refugees may follow distinctive investment strategies because they are less likely to return to their home country. In particular, refugees may experience a stronger incentive to learn a new language because they expect to remain in their new destination for the foreseeable future. Because refugees to the US plan to remain in the country over a longer time horizon, they may be more likely to invest in US-specific human capital such as learning English (Cortes 2004). In addition, modern refugees may enjoy heightened exposure to the destination language if they live in more dispersed locations away from ethnic enclaves. Recent research also suggests that refugees are less likely than other immigrants to maintain cross-national ties, thereby reducing their usage of their mother tongue (Luthra, Soehl, and Waldinger 2017; Morawska 2004). Because refugees may have greater incentives to learn English and are less reliant on their mother-tongue, they may outpace economic migrants in language attainment.

Refugees in the Age of Mass Migration

Most research on refugees in the United States has been on modern arrivals who entered the country through the official refugee system, which was established in 1980. The Refugee Act of 1980 standardized resettlement support and broadened the scope for admitting more refugees on humanitarian grounds (Zolberg 1988). Before that time, Congress passed some special refugee acts, including the Displaced Persons Act of 1948 (pertaining to refugees from World War II), the Refugee Relief Act of 1953 (pertaining to Eastern Europe) and the Indochina Migration and Refugee Assistance Act of 1975.

Our focus is on an earlier period – the early twentieth century – when the United States did not maintain a formal refugee system, nor did it use periodic refugee acts to authorize access to the country. Instead, (European) immigrants fleeing persecution could enter the country through entry points like Ellis Island much like immigrants arriving for economic reasons or to join family, even if they were not formally designated as refugees. Refugees in this period fled several major events in Europe. Russian Jews fled discrimination and the pogroms, Armenians fled genocide, and dissidents (e.g., communists, Irish nationalists) fled political persecution in many Western European countries. Events such as World War I, the Balkan Wars, and the collapse of the Prussian, Ottoman, and Russian empires also greatly contributed to displacement from Europe (Zolberg 1988). Displaced persons made up a higher percentage of the total world population in this historical period than today (Gatrell 2013).

In the past, refugees were not offered targeted English classes by the federal government and thus were required to learn English – if at all – through inter-personal contacts and personal investments in the same manner as economic migrants. At the time, some immigrants relied on

family members and children to learn English, while others would learn on-the-job or through night-classes. In addition, some employers, especially large manufacturing firms in the Midwest, required their immigrant workforce to take English classes while on-the-clock, but these classes were not offered differentially by refugee status (Catron 2016). In practice, English was also a requirement to become a citizen in order to answer civics questions posed by a judge; receiving citizenship conferred further economic advantages (Catron 2019). This article seeks to understand refugee linguistic attainment and economic success in a period with little federal support.[3]

Data and Methods

We gather information on refugee status and English attainment from a novel dataset based on 1,200 oral histories of immigrants who arrived in the US in the early twentieth century, originally conducted by the Statue of Liberty – Ellis Island Foundation (SOLEIF). SOLEIF identified interview subjects by placing advertisements in national magazines and newspapers. In addition, visitors of the Ellis Island Museum who identified themselves or family members as Ellis Island immigrants were given forms to collect basic information. SOLEIF then selected immigrants from this pool to produce the final data set of oral histories. Respondents were interviewed by one of several professional oral historians in a semi-structured manner and prompted to include information about life before and after arrival to the United States. The average interview lasted 54 minutes. Varricchio (2011) reports that all interviews (100 percent) covered "reasons for emigrating," whereas only 80 percent covered employment in the US and 50

[3] For more on immigrant assimilation during the Age of Mass Migration, see Morgan, Watkins and Ewbank (1993), Perlmann and Waldinger (1997), Catron (2016, 2023), Catron and Vignau Loria (2021), Connor (2020), Abramitzky, Boustan and Eriksson (2014, 2020) and Goldstein and Stecklov (2016).

percent covered education in the US (see Figure 2). All interviews were in English, which introduces some degree of selection. However, we note that, in the 1980 Census, 91 percent of immigrants who arrived in the US before 1949 report speaking English either 'well' or 'very well,' or speaking English exclusively. In addition, Lieberson (1981) notes similar levels of English attainment when analyzing European immigrants in the first half of the twentieth century. Our results therefore speak to individuals who learned English, but not necessarily to the entire population.

Each interview asked questions about immigrants' lives, including their life in their home country before migration, experiences going through Ellis Island, and life in the United States after migration. The interviews were conducted when interview subjects were in their 70s and 80s, and therefore collected information on the full life course. The timing of the oral history may introduce recall bias on experiences of migration; we discus robustness to this concern below.

Figure 1 presents a flow chart of our sample construction. The SOLEIF website has posted information about 1,889 individuals; however, 479 of these cases have neither a transcript or audio file extant and 198 interviews were conducted with US-born interviewees (e.g., border agents), leaving us with 1,212 possible individual records.[4] We collected 972 interviews that had complete audio files and transcriptions directly from the Ellis Island Foundation website, and used an external contractor to transcribe 240 interviews that had audio files but no written transcript. After dropping cases of 22 interviews with multiple respondents (e.g., siblings), our final sample contains 1,190 oral histories.

[Figure 1 Here]

[4] The files were retrieved from https://heritage.statueofliberty.org/oral-history-library

Assigning refugee status and premigration characteristics from oral histories. We assign refugee status based on the stated reason for migration of each individual from their oral history. We note that this approach bears some similarity to the 1951 Refugee Convention and the current definition of a refugee adopted by the United Nations, which includes persons who have a "well-founded fear of persecution" in their home country, and who have been "forced to flee his or her country because of persecution, war or violence" (UNHCR 2021). An advantage of assigning refugee status from a person's perceived experience is that we are able to observe any differences in outcomes between refugees and economic/family migrants within the same religion or country-of-origin group. Conditions of migrant waves often produce both a refugee flow and an economic flow of people leaving from the same time period and place (Holland and Peters 2020). Assigning refugee status to the individual rather than the group allows us to better understand refugee integration and avoid conflating refugee status with national origin.

To collect data on refugee status and other pre-migration attributes, we started with an extensive template for 250 interviews that included all potential variables of interest. From these, we created a compressed template that included only well-populated variables. These included: year of arrival in the US, country of origin, religion, urban status before migration, father's occupation in the home country, and stated reason for migration. Religion categories include Catholic, Protestant, Eastern Orthodox, Jewish, and Other/Missing. We employed five research assistants to code these variables in a first round of coding, following pre-defined instructions.

We then worked with two additional research assistances to recode the stated reason for migration variable into two categories (refugee/non-refugee). Double-coding our main variable of interest improves accuracy and allows us to differentiate immigrants who were coded as refugees

in both rounds ("refugees") or who were only coded as a refugee in one round ("mixed reason"). We closely supervised our team of research assistants to ensure the integrity of our template and to deal with challenges that arose during the coding process. This supervision involved regular meetings and dialogue between the authors and research assistants. We detail the coding scheme provided to our research assistants in **Appendix Table 1**. **Appendix Figure 1** provides a sample of the transcriptions made by the research assistants in the standardized template.

Coders found classifying reason for migration to be a relatively straightforward task because refugees shared harrowing details about the conditions of their exit. For example, Emilie Adams, a World War I refugee from France, remembers at the age of 5 five avoiding being shot by soldiers by having to "take… a white flag, and…we had to put that flag out before we could come out of the house." Another refugee, Wadih Zogby, lived in Lebanon through the same war. "One third of the population of Lebanon died of starvation or diseases that came after starvation," he recalls. "Once the [sea lanes] opened, about 1920, my brother, who was here [in the United States], kept on writing to us, 'Please come, please come.' And we were very happy to be out, because we suffered like blazes during the First World War."

Non-refugees, on the other hand, often discussed opportunities that the US offered. Lillian Amundson from Finland notes, "then my cousin who had been living in New York City, she came over to Finland to visit her mother and father. And she was telling me how wonderful it is in United States and how easy it is for people to make money when they are willing to work. That's how I got the idea. 'Mother,' I said, 'in Finland I can't help you, but if I go to United States, I will be able to send you money.' And that is how I came..."

However, in some cases, there were ambiguities about whether individual stories should be classified as refugees or non-refugees. For instance, some immigrants in our sample fled their

home countries due to violence or persecution to settle in a second country, and then moved from there to the US at a later date for economic reasons. For instance, during the Russian Revolution in 1917, Samuel Rosen recalls when he was 4 years old, "…you know what a sound, a sewing machine makes? Ta-ta-ta-ta-ta-ta-ta. One day early in morning I asked my mother, 'Why are they using a sewing machine so early in the morning.' She says, 'That's not sewing machines, my boy. That's machine guns being used.' And there was a war going on right, all around us. We stayed sequestered in our houses, we never went out, and we really had a tough time." Samuel's family was able to flee to Romania in 1919 where they lived for five years before making the decision to move to the US for more opportunities and to be closer to family. The 1924 Quota Act stopped that move so they went to Israel instead. Samuel's parents, however, "wanted [the family] to come to the Goldene Medina" and they eventually moved in 1930.

Other immigrants in our sample provide multiple reasons for their move to the US. Theodora Pellegrino describes her first move from Italy to the US in 1912 in economic terms, "Me come this country because everybody in the town talk the America, and they told America nice, America pay more." Theodora returned to Italy to help her sister who lost her husband in WWI. After the Fascist takeover of Italy in 1922, a neighbor falsely accused Theodora and her family of being socialist. As Theodora recalls "Because, see, that time Mussolini give white papers kill…the socialists…See, the Fascists come. One night, see, knock on the door…And twelve o'clock, see, it's bad. And after open the door me, 'What do you want?' 'I want this and them.'" Theodora convinced the police that she was not a socialist, but then the police began harassing her husband for not enlisting in the military. The family made the decision to move back to the United States soon after. Still others in our sample had already decided to move to the US for economic reasons, but wars or persecution sped up their plans to leave.

Because of these ambiguities, some immigrant stories were classified as 'refugees' by the first round of coding but as 'non-refugees' by the second round of coding. Of the 511 immigrants who arrived as adults and thus form our main analysis sample, there was coder agreement on reason for migration for 411 cases (= 80 percent) and disagreement for the remaining 100 (= 20 percent) of cases. In various specifications, we consider "refugee" and "mixed reason" immigrants together, focusing on any immigrant who lists flight from persecution as one reason for their move, or we analyze the two groups separately.

Figure 2 presents two word clouds that may provide insight on the potential topics and words that led coders to classify immigrants as refugees or non-refugees from the entire oral history dataset in our first round of coding. In particular, the figure shows which words are statistically disproportionately associated with refugee and non-refugee oral histories, calculated using Monroe et al. (2008)'s log-odds ratio, informative Dirichlet prior method. For example, refugees are more likely to emphasize words linked to political events such as "war," "revolution," "military" or words related to cultural or religious persecution like "concentration," "camp," and "prison." Non-refugees are more likely to mention words related to economic activity like "farm," "housework" or "cement," or words related to family relations like "dad," and "kids." The words associated with refugee status also make clear that the refugees are more likely to be Jewish ("synagogue" and "kosher" versus "church" and "Christmas").

[Figure 2 Here]

Our sample has good coverage across countries of origin and socio-economic background. **Table 1** reports mean attributes of the sample overall, and separately for refugee/mixed reason and

non-refugees. For the full sample, 10-20 percent of respondents were born in each of the large European sending countries (Ireland, Germany, Italy, Russia/Poland). The remainder (40 percent) hail from a set of smaller sending countries. Our sample is fairly evenly split by father's occupation, including white-collar positions, manual skilled workers (e.g., tailors, factory workers), farmers and laborers. Refugees tend to arrive in the US at older ages, are more likely to be male, and are more likely to come from urban and white-collar backgrounds.

[Table 1 Here]

We divide the sample into four arrival periods: 1893-1914, a period of nearly open entry to European immigration; 1915-1923, which covers World War I and the large backlog of immigrants who arrived in the years after the war; 1924-1933, the years after the imposition of strict immigration quotas and before the rise of the Nazi regime in Europe, and 1934-1957, a period marked by flight from Nazism, World War II and the Cold War. **Figure 3** illustrates the exact year of arrival for refugee and non-refugee immigrants in our sample. Note that 11 percent of our sample immigrated to the US after 1933. Some of these immigrants would have been screened for entry by the 1948 Displaced Persons Act and the 1953 Refugee Relief Act. Although these refugees did not receive formal resettlement assistance, we present robustness results that excludes immigrants who arrived after 1933 below.

[Figure 3 Here]

Members of our sample must survive until at least the 1970s or 1980s in order to be interviewed. As a result, our sample is weighted toward immigrants who arrived at the end of the Age of Mass Migration and who came to the US at younger ages. Because immigrants who arrived as children may have learned English more readily, our main analysis sample includes immigrants who arrived after early childhood (after age 12 or age 14).[5] Furthermore, 10 percent of our sample hailed from Britain or Ireland, two English-speaking countries, and most of these immigrants would have been fluent in English upon arrival. We drop immigrants from these countries-of-origin in our main analysis.

Relying on retrospective oral histories may produce recall bias. As immigrants age, they may reconstruct their memory of hardship around political oppression or anti-Semitism. American Jewish identity was particularly shaped around collective memories of flight from pogroms and oppression (Zipperstein, 2013). As a result, we present results that split the sample between Jewish and non-Jewish migrants below.

Measures of linguistic attainment from computational linguistics and second-language acquisition research. We adapt methods from existing linguistics research to estimate English attainment from oral history transcripts. Each oral history contains nearly an hour of actual speech from recorded interviews available both in audio files and as a transcription. After filtering the transcripts to omit very common words (known as "stop words" in computational linguistics,

[5] Our interview subjects who arrived as young children often remark on how quickly they could pick up English. For example, Lucy Attarian who arrived at age 5 recalls "my mother came to visit me in the class a few days later, and the teacher said to her, 'She's really picking it up very well. Just watch her.' Well, she'd asked the children to do something, and of course I'd see them doing it, and I would do the same thing. Not because I understood, but I was following. But it didn't take me long. It didn't take me long at all to pick up the language."

including function words like “the” and “and,” as well as pronouns and prepositions), we observe over 1.2 million distinct word tokens across our database of respondents.

Research in second-language acquisition has demonstrated that language proficiency is multifaceted, including aspects of complexity, accuracy, and fluency (Skehan 1998; Ellis and Barkhuizen 2005; Housen, Kuiken, and Vedder 2012). The proficiency of speech can be marked along various dimensions of linguistic structure, including lexical (vocabulary), morphosyntactic (grammar), phonological (sound system), and prosodic (rhythm and intonation). Higher attainment on any of these dimensions may signal to native-speakers a lower degree of foreignness, which may in turn affect treatment in the labor market and other social settings. The use of transcripts and especially the inability to directly interview historical persons limits what we may measure, but here we aim to measure a meaningful subset of the complex variability that constitutes linguistic proficiency and attainment. Our measures have been evaluated by other scholars for their intrinsic associations with attainment, and have been applied in settings such as automated written and spoken assessment (Attali and Burstein 2006; Crossley and McNamara 2012; Kyle et al. 2018; Chen et al. 2018).

Our first measure addresses lexical complexity (vocabulary), relying on the average “age of acquisition” (AoA) of each spoken word, which captures the average age at which any given word is typically learned by native English speakers. The AoA is a measure based on a dictionary from Kuperman et al. (2012), in which each of 30,121 English lemmas (dictionary form of a word) are assigned estimated values of the ages they were learned based on subjective judgments from a large survey. We calculate a person-specific measure as the mean AoA value of all words spoken by the immigrant during their interview. Our second measure captures syntactic complexity by simply calculating the mean length of sentences uttered in an interview, since more complex

sentence structures that use dependent clauses and complex phrases tend to be longer. Prior work demonstrates that the use of complex vocabulary and longer sentences are both associated with more proficient speech (Chen and Zechner 2011; Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012; Kyle and Crossley 2015).

Our third measure, accentedness, captures phonological accuracy and fluency via subjective perceptions of independent evaluators, as coded by three pre-doctoral research assistants who are native speakers of American English. Finally, to supplement this perceptual measure of accentedness, we also calculate immigrants' speech rate in syllables per minute of speech (Riggenbach 1991; Ginther, Dimova, and Yang 2010). Faster speech rate is characteristic of native language speech and increased proficiency, and may interact with perceptions of accentedness in complex ways as speakers negotiate conversational interactions in a non-native language (Derwing 1990; Munro and Derwing 1998; Guion et al. 2000).

We report summary statistics for these four dimensions of speech in **Appendix Table 2**. In the average transcript, the "age of acquisition" for the average word is close to 5 years of age, and the average sentence is 12 words long. These measures are positively correlated with each other, but the degree of correlation is not high (around 0.3 to 0.4), suggesting that we are picking up independent aspects of language proficiency (**Appendix Table 3**). These oral histories were collected once for each subject when interviewees were late in life, so we treat these measures as "lifetime attainment." Though proficiency in a language is multidimensional and cannot be fully captured by the four measures used here, our approach offers a much more naturalistic and detailed view of linguistic attainment than typical measures of self-reported fluency.

Figure 4 presents two extracts from the oral history database, reproducing a set of sentences spoken by two Jewish respondents: Morris Helzner ("MH") and Paul Deutsch ("PD").

To demonstrate how we quantified specific dimensions of speech within our framework, we provide descriptive statistics on depth of vocabulary and sentence length. Morris Helzner uses words that earn high vocabulary scores ("brewing", "hectic") and a longer, more sophisticated sentence structure. Paul Deutsch in contrast, uses shorter and more punctuated sentences with words associated with a lower depth of vocabulary ("father", "stay").

[Figure 4 Here]

We validate our measures of English attainment by documenting the presence of common patterns of proficiency by nationality and age of arrival in the US. First, immigrants who knew English before arrival are likely to have greater English proficiency in our interview sample. **Figure 5** shows that immigrants who hailed from English-speaking countries like Great Britain have higher measured English attainment by our four measures. Second, as we noted above, there is a well-known pattern by which immigrants who arrive in the US at earlier ages achieve greater English proficiency, either because they are educated in American schools or because they arrive at a "critical age" for language acquisition. The downward slope of each panel suggests that immigrants who arrived at early ages achieved greater English proficiency by our measures than immigrants who arrived later.

[Figure 5 here]

We note some important limitations of our linguistic measures here. First, we only have one interview per immigrant, and so cannot measure individual changes over time. Second,

interviews were conducted late in life (median age at interview = 83), and linguistic ability can diminish with general cognitive function (Denis 2016). Third, we rely on speech transcripts to measure vocabulary and syntax. Transcription alone is a complex process of filtering data (Juncos-Rabadan and Iglesias 1994). Transcribers are particularly likely to be inaccurate in coding speech errors (Ochs 1979), so we do not aim to directly quantify errors such as grammatical agreement mismatches. Fourth, linguistic competence can be task dependent. Accuracy and fluency are foregrounded in personal information exchange tasks like a casual conversation (Skehan 1998), whereas complexity is foregrounded in narrative tasks. Oral history interviews share features with both of these settings but may not reflect a speaker's full linguistic competence. Fifth, our sample of interviews was conducted in English. Although our sample may exclude immigrants who did not achieve a sufficient level of English competency to participate in an interview, we note that the vast majority of immigrants – more than 90 percent – both in the past and today achieve "some" ability to speak English (Lieberson 1981; Alba et al. 2002). Note that many of the interview subjects in our sample speak very halting or imperfect English, and so our results are broadly applicable (but not fully so). Despite these limitations, projecting these methods onto historical data allows researchers to study language use for populations that are no longer alive, and at a scale that would be infeasible with oral proficiency evaluations even for contemporary populations.

New Immigrant Survey Data. We use data from the New Immigrant Survey (NIS) to compare the English attainment of refugees in the past with refugees in the modern period. The NIS is a nationally representative multi-year survey of new legal immigrants to the US. We focus on the first survey wave, a random sample of adults receiving legal permanent residence between May

and November of 2003. The survey includes immigrants who arrived in the US as refugees or asylum seekers; we classify both of these arrival types as 'refugees.' From the NIS, we obtain information on whether an immigrant speaks any English or speaks English "well." Note that, as is typical in modern datasets, we do not have access to the same detailed level of information on English proficiency in the NIS as we do for the historical oral histories. Thus, we view the NIS results as a pilot study for the modern period and encourage future work collecting and analyzing transcripts of immigrant speech. We can control in the NIS data for country of origin, year of departure from country of origin, gender, age, religion, years of schooling prior to immigration, and number of English classes taken within the last year and prior to arrival in the US.

Appendix Table 4 reports summary statistics for immigrants in the NIS data by refugee status. Refugees are more likely to speak any English at the time of survey (85 percent vs. 74 percent), but they are less likely to speak English well (33 percent vs. 39 percent). Refugees are also more likely to have taken an English class in the past year, perhaps because of government supports, but they are less likely to have taken an English class before moving to the US. Refugees are reasonably well balanced relative to non-refugee migrants on attributes like gender, age and education, but are notably more likely to have moved from Europe/Central Asia or from Russia, Ukraine or Poland (total = 62 percent vs. 16 percent). This geographic pattern reflects refugee priorities for the US in the immediate aftermath of the Cold War and the aftermath of the Bosnian War as immigrants in the sample arrived in the US in the late 1990s and early 2000s.

Estimation strategy. Our analysis rests on a series of multiple regression models. Each observation is an immigrant drawn from the Ellis Island oral histories.

We start by considering the relationship between refugee status and English proficiency. In particular, we estimate models of the following form:

$$y_i = \beta_1 REFUGEE_i + \beta_k X_k + \epsilon_i \quad (1)$$

where the dependent variables y_i are one of the four linguistic measures (vocabulary, syntax, accentedness or syllables per minute) and the main right-hand side variable is an indicator for whether the immigrant reports leaving their sending country for non-economic reasons, such as a flight from persecution or violence ($REFUGEE_i$). In our first model, we include any immigrant who mentions flight from persecution as one of their reasons for moving into the *REFUGEE* indicator, combining both refugees and "mixed reason" movers. In our second model, we create two indicator variables ($REFUGEE^1_i$ and $REFUGEE^2_i$), with $REFUGEE^1_i$ equal to one for refugees classified as such in both rounds of coding, and $REFUGEE^2_i$ equal to one for immigrants who had mixed reasons for moving and were classified as a refugee by either round 1 or round 2 of coding, but not both. Note that, in both cases, the comparison group includes any immigrant who was never marked as a refugee; that is, for immigrants in the comparison group, both round 1 and round 2 coders consider this immigrant to have moved to the US for economic or family reasons.

The regression also includes a vector of controls (X) that include age and age squared, gender, an indicator for arrival period (before 1924, 1924-33, 1934-after), an indicator for country of birth, and – in some specifications – controls for other aspects of pre-migration environment, including indicators for religion, father's occupation in the sending country and urban status. For this analysis, we focus on immigrants who arrived after early childhood (after age 12) because the process of language acquisition is different – and often more immediate – for immigrants who

arrive as young children. We also consider robustness for different cutoffs of when 'early childhood' ends.

Results

Refugees achieved higher levels of English proficiency than economic/family migrants in the past. We start by estimating the relationship between refugee status and English attainment in our historical dataset.

We find that immigrants who left Europe for the US under duress in the early twentieth century achieved a higher level of English proficiency by the end of life, particularly in terms of building a deeper English vocabulary. **Figure 6** presents coefficients on an indicator variable that includes immigrants coded as a refugee by at least one coder (both refugees and "mixed reason" immigrants). **Figure 7** instead presents coefficients on two indicators: one for being coded as a "refugee" and one for being coded as moving for "mixed reasons." The solid colored diamonds reflect results from a set of basic regressions that control only for a quadratic in age, and indicators for arrival period, country of birth and gender.

In both models, immigrants fleeing from persecution achieved a greater depth of vocabulary, scoring 0.4 standard deviations higher on our "Age of Acquisition" score. This relationship holds for both refugees and "mixed reason" immigrants in Figure 7. Refugee immigrants also have more complex sentence syntax, although this relationship is not present for "mixed reason" immigrants and appears to be driven by pre-migration attributes of refugees, who are more likely to hail from white collar and urban backgrounds. After controlling for these characteristics, the association between refugee status and mean sentence length substantially weakens.

Refugee migrants are no more able than economic/family migrants to eliminate their accent and do not utter more syllables per minute. This pattern is consistent with existing results that while grammar and lexis (vocabulary) are subject to fewer maturational constraints, learning how to pronounce new phonemes becomes markedly more challenging with age and is thus less mutable or responsive to investment, commonly called "the Conrad phenomenon" in the field of linguistics (Patowski 1990; Bongaerts et al. 1997; Moyer 1999).

We reject the hypothesis that the initial selection of immigrants fleeing persecution might explain why they achieve a greater depth of English vocabulary by the end of life. The open colored diamonds present coefficients from regressions that add an expanded set of controls, including indicators for childhood religion, being raised in an urban area, and father's occupation. Refugees and "mixed reason" immigrants in our sample are more likely to be Jewish, more likely to be raised in an urban area, and more likely to hail from a household headed by a white-collar worker, as opposed to a farmer or laborer. Adding these controls moderates the positive association between refugee migrants and English vocabulary to a small degree, but we still observe a strong refugee advantage on this measure. We report the coefficients from the regressions underlying Figure 6 in **Appendix Table 5**.

[Figure 6 Here]

One concern is that refugees have greater measured vocabulary simply because talking about persecution requires a more complex set of words. That is, the construction of our Age of Acquisition measure could be systematically biased by the use of words that align with specific topics of conversation. Most notably, a migrant's description of fleeing persecution or of a

particular historical event could evoke the use of more sophisticated words (e.g. "affidavit," "revolution" "government"; see Figure 2). If true, this relationship could bias our analysis toward finding a positive correlation between the Age of Acquisition measure and refugee status. To address this issue, we create an alternative vocabulary measure that is calculated after dropping persecution-related words from interview transcripts. We defined persecution related words through the manual inspection of thousands of the most common words in our transcripts. A second set of estimates in Figure 6 and Figure 7 (orange diamonds) confirms that our analyses are robust to the omission of a large list of words that are specifically tied to persecution and conflict, which otherwise could artificially distort the association between refugee status and English proficiency (see the list of words in **Appendix Table 6**).

[Figure 7 here]

We emphasize that our results do not depend on comparing English attainment across groups (e.g., Jews and non-Jews). Rather, we are comparing the linguistic proficiency of refugees and economic/family migrants from the same country-of-origin and religious group. For example, our specification compares a Jewish migrant leaving the Russian Empire in the 1900s under threat of persecution to another Jewish migrant from the Russian Empire in this period who reports migrating to find employment or to reunite with family. We also make similar contrasts within the non-Jewish population (e.g., a German Catholic fleeing from war versus a German Catholic moving for economic opportunity).

In order to confirm that our results are not driven by the Jewish population, we conduct the analysis on sub-sets of the data. **Figure 8** splits the sample into Jewish and non-Jewish immigrants.

We find a similar association between refugee status and English attainment in both groups, illustrating that the findings are not being driven by one group alone. If anything, the association between refugee status and vocabulary is stronger for the non-Jewish sample. As we mentioned above, American Jewish identity has developed around a collective memory of persecution in the old country (Zipperstein, 2013). Recall bias on the part of Jewish migrants may have overemphasized the harsh conditions leading to their departure. Alternatively, all/most Jewish immigrants in this period, even those who report economic concerns as their proximate reason for migration, may have been somewhere on the continuum of coercion given the intensity of anti-Jewish violence at the time. Either force could be attenuating the results in the Jewish subsample.

[Figure 8 here]

Our results are not dependent on specifics of sample selection. **Appendix Figure 2** show that results are unchanged if we conduct our analysis on the full sample, including immigrants from English speaking countries (e.g., Britain). One particularly distinctive refugee flow is the select group of German and Eastern European refugees who fled from the Nazi regime and the advancement toward war in the 1930s. This group included many scientists, authors and industrialists who may have arrived with particularly high levels of education or capacity to learn English. **Appendix Figure 3** show that our vocabulary results are robust to excluding immigrants who arrived after the rise of Nazism in 1933, some of whom were accepted under special refugee acts in 1948 and 1953. However, the association between refugee status and syntax (mean sentence length) disappears. As we noted above, this association seems to be driven by pre-migration characteristics of refugees (higher initial socio-economic status). In **Appendix Figure 4**, we adjust

the definition of arriving in adulthood from arriving after age 12 to arriving after age 14. Results are qualitatively similar.

The English proficiency of refugee migrants today. We next compare our historical findings to the best available source of modern data on the refugee population – the New Immigrant Survey (NIS). We estimate the relationship between refugee status and two measures of English attainment – whether an immigrant speaks any English, and whether an immigrant speaks English well. Note that we do not have detailed information on English proficiency in the modern data.

We find a similar pattern today as in the past, whereby refugees are more likely than other legal permanent residents to report speaking some English. Panel A of **Table 2** shows the coefficients on the indicator variable for being a refugee migrant. We start in column (1) by including a basic set of demographic controls (age, gender, year of migration to the US and country of origin). In column (2), we introduce socio-economic controls, including years of schooling prior to immigration, whether an immigrant previously lived in a rural or urban area, and religion. After we adjust for differences in socio-economic background, we find that refugee migrants are 8 percentage points more likely to speak some English relative to similar non-refugee migrants.

[Table 2 Here]

In the modern context, refugee migrants may receive more government supports, including subsidized or free English classes (indeed, we find in Appendix Table 4 that refugees are more likely to have taken a recent English class). In column (3) we control for taking a recent English class or an English class before arrival. We continue to find that refugees are more likely than non-

refugees to speak some English, suggesting that the heightened attainment achieved by refugees in the modern period occurs above and beyond any access to common government supports.

Lastly, given that our historical analysis examines immigrant English attainment that occurred over several decades, we restrict our modern sample to immigrants who have lived in the US for at least two years, shown in column (4). We find a similar result, that refugees are more likely to speak some English relative to non-refugee migrants. This result aligns with other studies that find refugees in the US today experience rapid improvements in (self-reported) English proficiency with time spent in the country (Cortes 2014; Chin and Cortes 2015).

Panel B repeats the analysis for speaking English 'well,' rather than speaking some English. Before adding socio-economic controls, it appears that refugees are *less* likely than non-refugee migrants to speak English well. Some of these differences in English ability may reflect pre-migration differences in exposure to English (refugees are less likely to report taking an English class before immigration; see Appendix Table 4). Indeed, after adding controls for pre-migration background, we find no significant relationship between refugee status and the likelihood that an immigrant speaks English well.

Discussion/Conclusion

The special case of refugee integration has been of considerable importance to theories of assimilation. Scholars often point to contemporary refugees to highlight the positive impact of governmental programs on assimilation and integration. However, many immigrants today (and certainly in the past) arrived having fled similar conditions, but are not labeled as official refugees and thus do not qualify for assistance (Garcia et al. 2021; Hamlin 2021). Because prior studies assign refugee status to those who meet strict legal definitions of refugee status, little is known about those who do not benefit from federal assistance, but otherwise may fear returning home.

Our analysis reveals that, even in the past, a period without government support for refugees, immigrants who moved to the US to escape from persecution or war achieved greater English proficiency as measured by depth of vocabulary than immigrants who moved to pursue economic opportunity. This pattern holds even after controlling for country of origin, arrival period, and religion, thereby comparing refugees to economic/family migrants within the same national origin or ethnic group. Our analysis of modern refugees reinforces this point, demonstrating that refugees have greater English proficiency today even after controlling for opportunities to enroll in English classes. Taken together, this pattern implies that the success of refugee immigrants in the United States is a long-run feature of the country's history.

Our study demonstrates that the refugee advantage occurred in a historical period without resettlement assistance and a modern setting after controlling for one aspect of government support (language courses). This finding suggests that refugees themselves engage in behaviors, either intentionally or unintentionally, that improve their linguistic proficiency (Alba and Nee 2003). Many refugees do not expect to be able to return home in the near term. Faced with this situation, refugees may therefore be more likely to invest in US-specific human capital, relying on the American educational system and the open labor market, to find employment, housing, and social networks, leading ultimately to greater English-language attainment.

Immigrants who were pushed from Europe by persecution or war were drawn from higher socio-economic status backgrounds than immigrants who left to seek economic opportunity. This pattern may suggest that refugee migrants arrive with a higher aptitude to acquire English skills, as has been found in many refugee populations (Birgier et al. 2016; Aksoy and Poutvaara 2019; Guichard 2020; Spörlein, et al. 2020). However, we document that the association between refugee

migration and depth of vocabulary changes very little after controlling for father's pre-migration occupation and urban status.

Some forces of migrant selection could push in the opposite direction. Economic/family migrants who are unsuccessful in their new destination may be more likely than refugees to return home. Previous research finds that return migrants were "negatively selected," in the sense that they were drawn from lower-paying occupations (Borjas and Bratsberg 1994; Abramitzky, Boustan, and Eriksson 2019). As a result, the economic/family migrants who remain in our sample may have been higher-skilled or more successful than the full population of economic migrants, and thus their linguistic attainment is likely biased upward. Yet, despite this possible upward bias, we continue to find that refugees achieve more English proficiency, suggesting that refugees would likely look even more accomplished relative to the full population of economic/family migrants.

As in other studies of refugees in the U.S., our finding of rapid integration for refugees in the Ellis Island period deviates from the slower rates of assimilation found for refugee populations in Europe and other OECD countries today. In destination countries like Australia, Germany and Canada, refugees' lack initial destination language skills and have a much longer attainment process in the modern era (Chiswick, Lee, and Miller 2006; Kosyakova, Kristen, and Spörlein, 2022; Kristen and Seuring, 2021; Chiswick & Miller, 2001). The U.S. has always been exceptional in integrating immigrants compared to other receiving nations (Mollenkopf and Hochschild 2009). In most European countries, labor markets are highly regulated and have strong social welfare states that impose restrictions on immigrant job opportunities and even greater job restrictions are placed on refugees (Hainmueller, et al. 2016; Marbach, et al. 2018; Fasani, et al. 2021, 2022). Indeed, many European countries resettle refugees where there may be abundant or low-cost

housing but a struggling labor market, and then impose residency obligation rules that prohibit regional mobility that slows integration (Damm 2009; Kosyakova and Kogan 2022).

In the United States, by contrast, labor markets are relatively open to refugee participation and there are fewer governmental supports, which heightens the incentives of refugees to gain language skills. Further, the U.S. is more open to foreigners than most European countries, which may increase willingness to learn the host language among immigrants (Mollenkopf and Hochschild 2009). As we argue above, refugees may be particularly responsive to this setting because the conditions of their arrival imply that they may not be able to return home imminently.

Our paper builds on existing work in historical sociolinguistics utilizing oral histories as a source of data (Heller and Mumma 2020). We are able to analyze multiple levels of linguistic structure, which is a large improvement from previous studies that analyze language using Likert scales or use competency scales that capture only one aspect of language proficiency (Dollmann et al. 2020; Edele et al. 2015). We suggest new ways to apply measures developed in linguistic research to the study of historical records at a large scale. Our approach can be extended to other sources of qualitative immigration data including historical archives of immigrants' letters, diaries, and other oral history collections as well as more contemporary sources of qualitative data. These rich data sources may help future researchers explore other aspects of language acquisition, including reading and writing abilities.

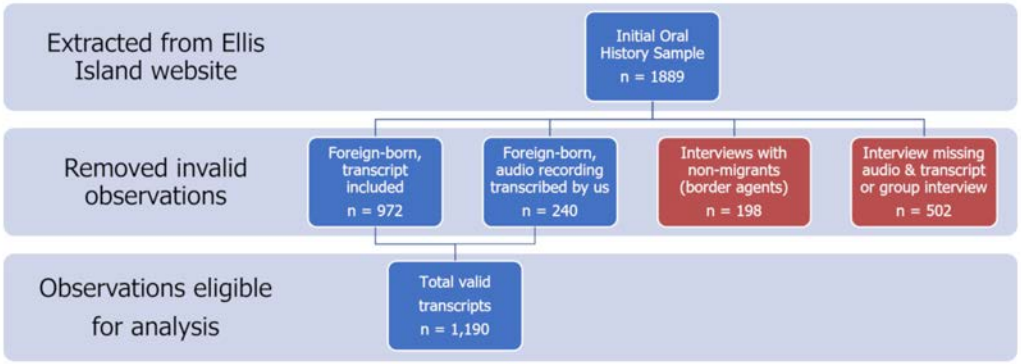


Figure 1. Sample construction flowchart. The flowchart demonstrates the construction of the samples used in the analysis of the paper. We start with a sample of 1,899 oral histories collected from the Ellis Island Foundation website. After accounting for missing or incomplete files and interviews with non-immigrants, we end up with a sample of 1,190 oral histories.

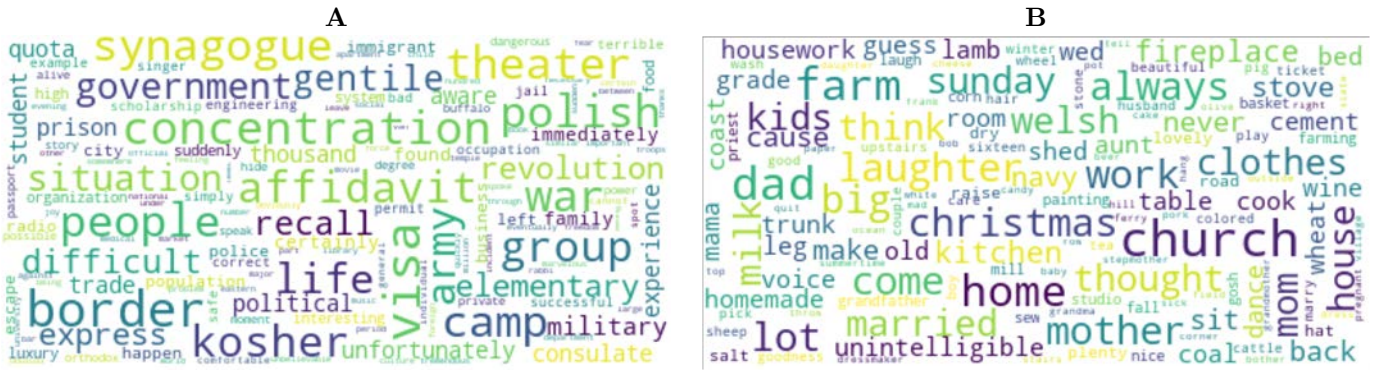


Figure 2. Interview word clouds by refugee status. A depicts a word cloud highlighting frequent words appearing in interviews with refugees or those who immigrated for mixed reasons. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. **B** depicts a word cloud highlighting words that frequently appear in economic migrant interviews. Frequency of words is calculated using a log-odds ratio.

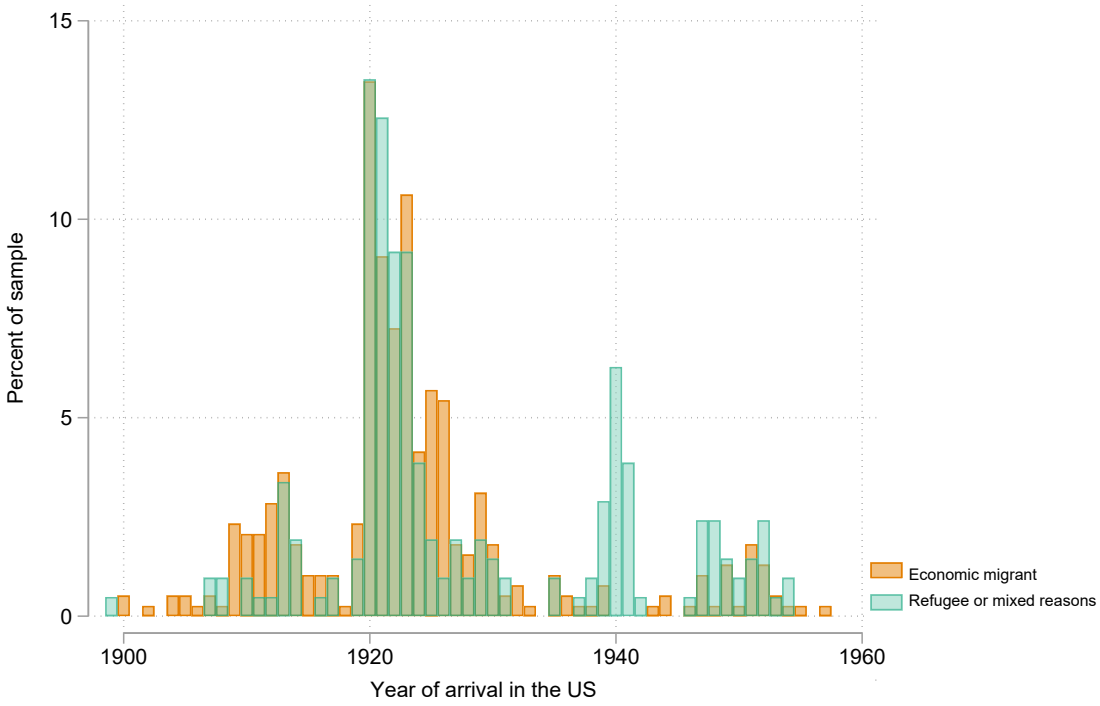


Figure 3. The distribution of immigrant arrival year by refugee status for immigrants arriving to the US after age 12. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons.

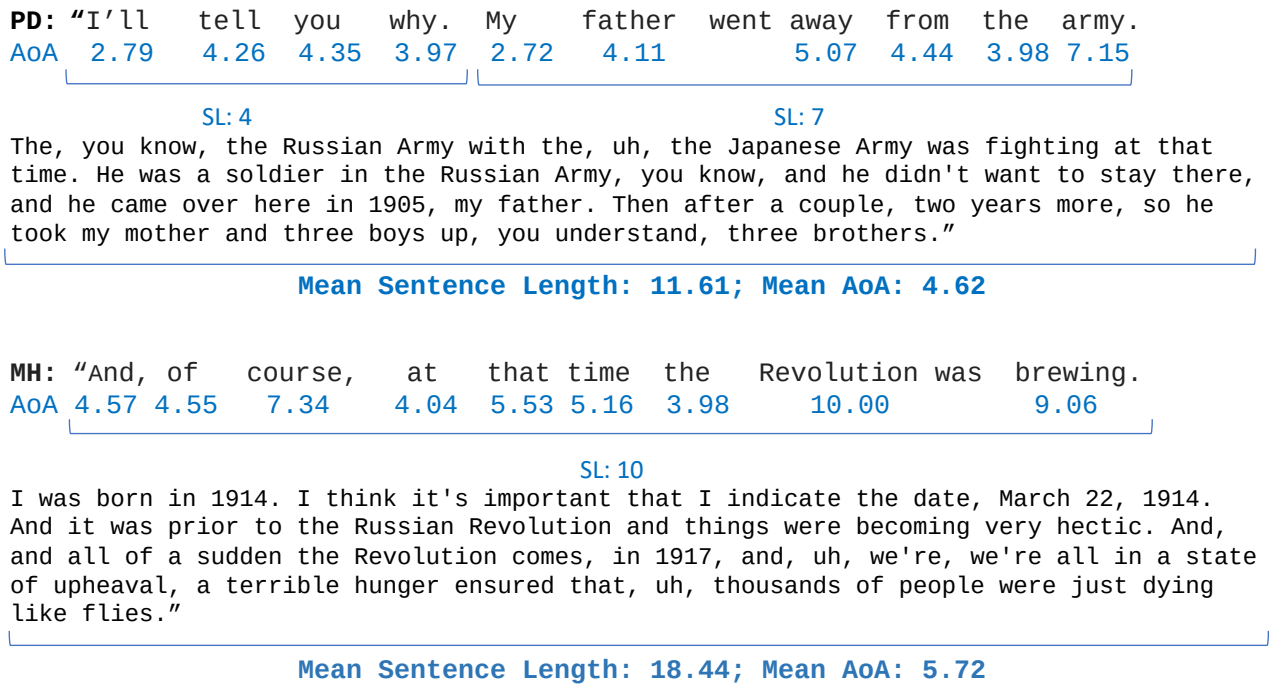


Figure 4. Examples of Age of Acquisition and Mean Sentence Length. This figure lists examples of Age of Acquisition (AoA) and Mean Sentence Length (MSL) from transcripts of two different migrants, Paul Deutsch (PD) and Morris Helzner (MH). We also note that Morris Helzner had an accent closer to that of natives as compared to Paul Deutsch. Morris was assigned an accent z-score of 0.38 while Paul had a score of -0.36, where a more positive score indicates an accent closer to that of natives.

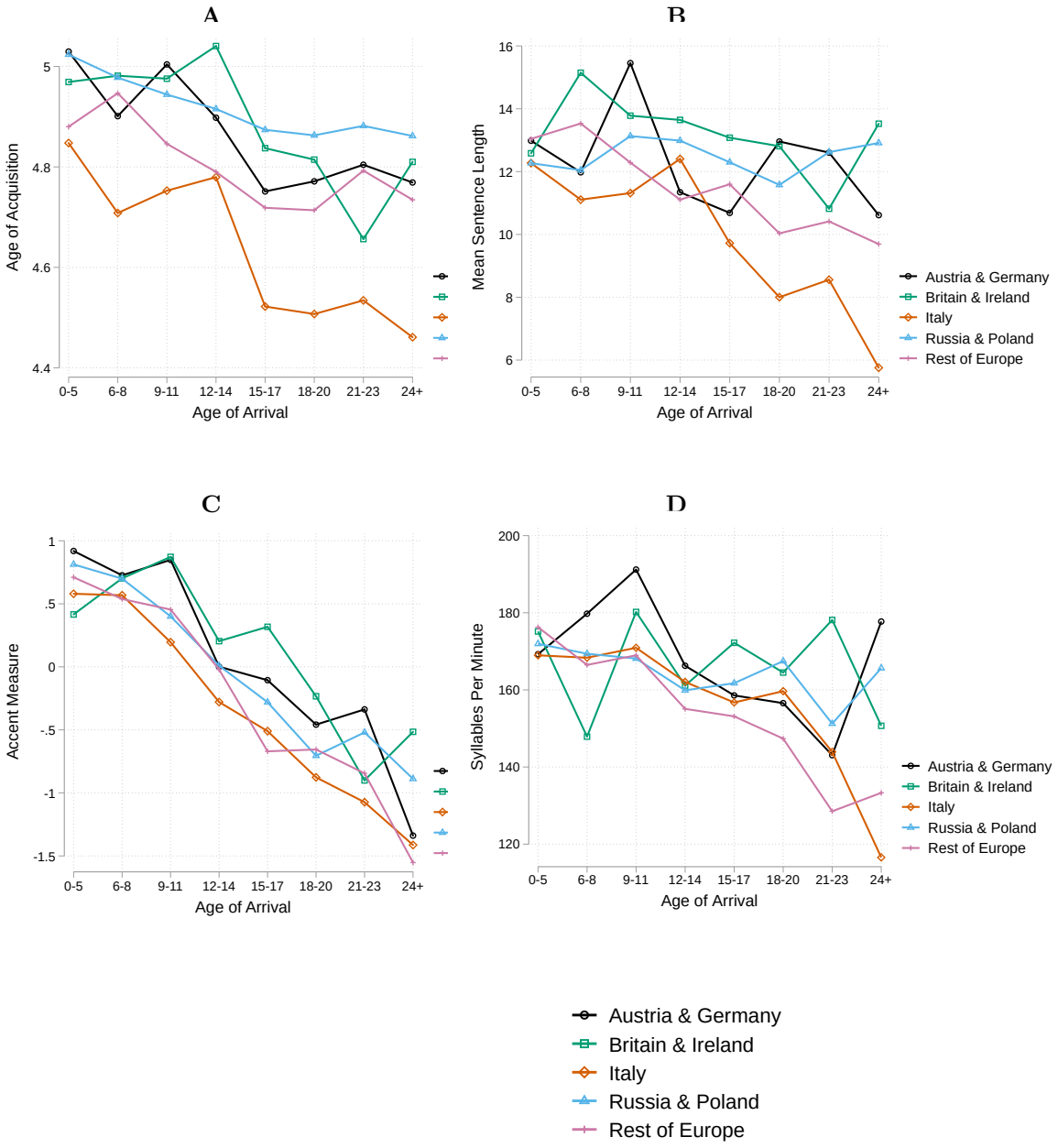


Figure 5. Mean linguistic measures by age of arrival and birthplace. Panel **A** plots mean Age of Acquisition (AoA) (N= 952), Panel **B** plots Mean Sentence Length (N= 952), Panel **C** plots accent measure (where positive indicates accent closer to native) (N= 809), and Panel **D** plots Syllables Per Minute (N=789) by age of arrival and birthplace. Migrants arriving after 1933 are dropped.

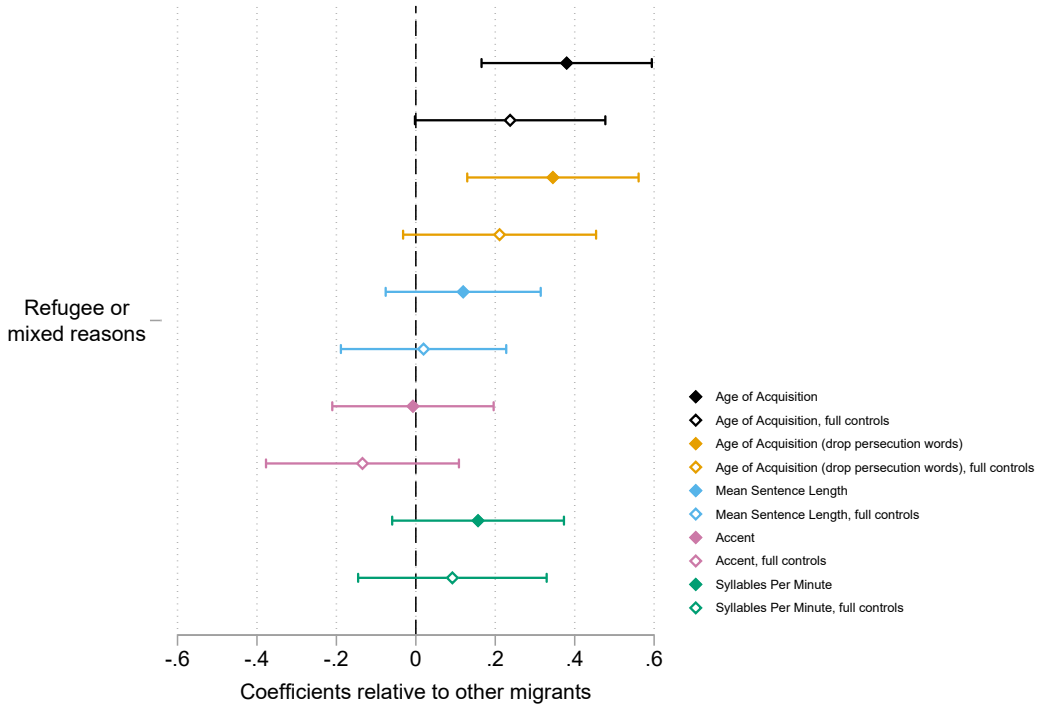


Figure 6. Association between refugee status and linguistic outcomes for immigrants arriving after age 12. The five linguistic measures are: Age of Acquisition (N= 451, N= 359 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 451, N= 359), Mean Sentence Length (N= 451, N= 359), Accent (N= 391, N= 317) and Syllables Per Minute (N =381, N=309). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are not included in this sample. A more positive accent score indicates an accent closer to that of the US born. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.

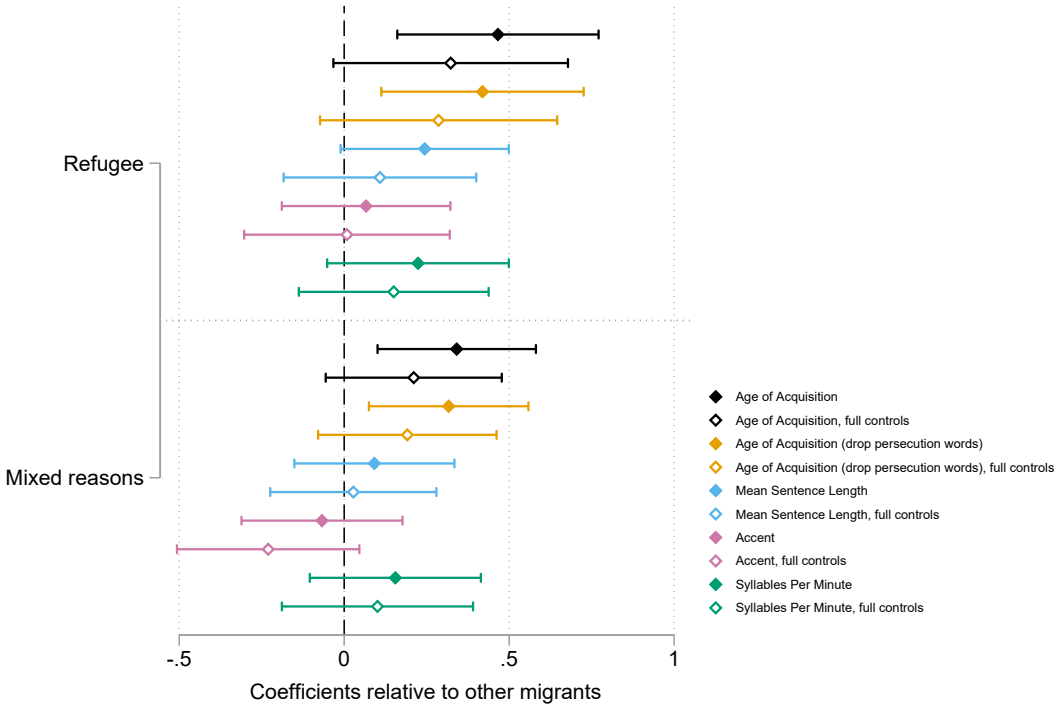


Figure 7. Association between refugee status and linguistic outcomes for immigrants arriving after age 12. We include two main explanatory variables. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding, i.e., both round 1 and round 2 of coding independently agreed the immigrant was refugee. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Coefficients are relative to immigrants who were not coded as a refugee in either round 1 or round 2. The five linguistic measures are: Age of Acquisition (N= 444, N= 350 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 444, N= 350), Mean Sentence Length (N= 444, N= 350), Accent (N= 389, N= 313) and Syllables Per Minute (N =379, N=305). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. English speaking immigrants, those from Britain and Ireland, are not included in this sample. A more positive accent score indicates an accent closer to that of the US born. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.

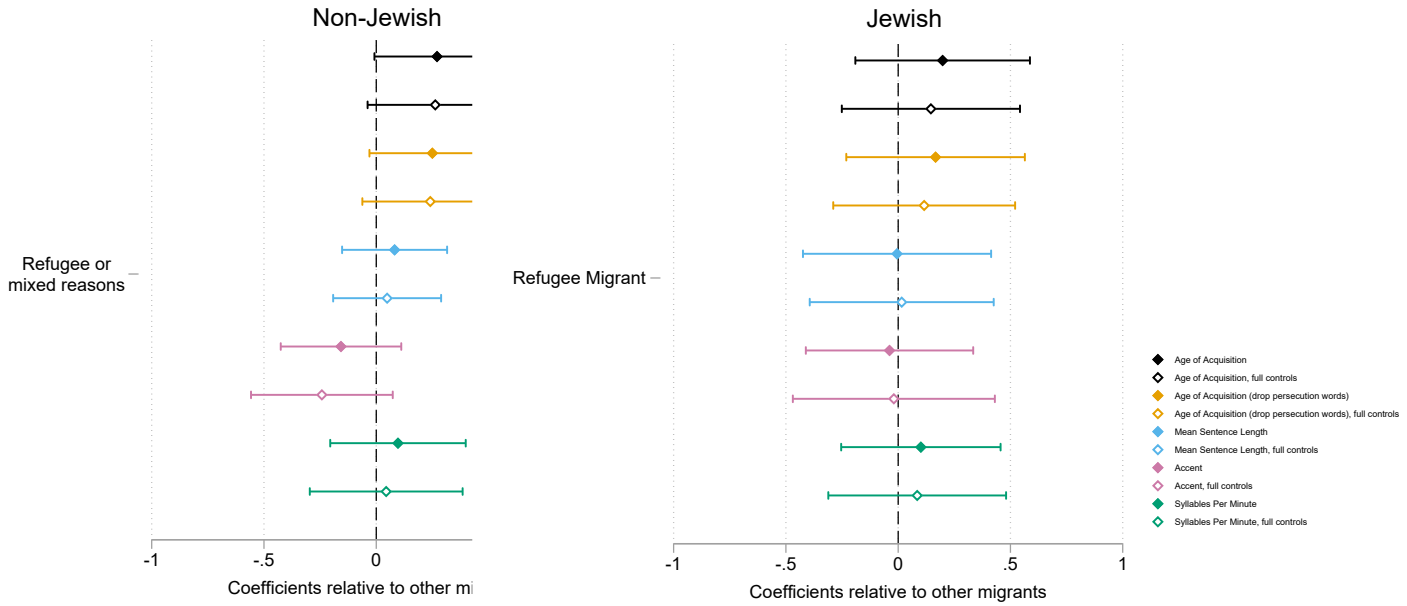


Figure 8. Association between refugee status and linguistic outcomes for Non-Jewish/Jewish immigrants arriving after age 12. The five linguistic measures for the non-Jewish sample are: Age of Acquisition ($N = 305$, $N = 233$ with full controls), Age of Acquisition calculated after dropping persecution related words ($N = 305$, $N = 233$), Mean Sentence Length ($N = 305$, $N = 233$), Accent ($N = 266$, $N = 207$) and Syllables Per Minute ($N = 259$, $N = 201$). The four linguistic measures for the Jewish sample are: Age of Acquisition ($N = 149$, $N = 127$ with full controls), Age of Acquisition calculated after dropping persecution related words ($N = 149$, $N = 127$), Mean Sentence Length ($N = 149$, $N = 127$), Accent ($N = 127, N = 110$) and Syllables Per Minute ($N = 124$, $N = 108$). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are not included in this sample. A more positive accent score indicates an accent closer to that of the US born. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status. Significance is at the 5% level.

Table 1. Summary of full sample, refugees and economic migrants.

	Full		Refugee or mixed reasons		Economic	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Gender						
Men	528	44.37	201	56.78	327	39.11
Women	662	55.63	153	43.22	509	60.89
N	1190		354		836	
Arrived as child						
Before age 12	574	48.24	141	39.83	433	51.79
After age 12	616	51.76	213	60.17	403	48.21
N	1190		354		836	
Arrival Period						
1890-1914	254	21.6	55	15.67	199	24.12
1915-1923	534	45.41	170	48.43	364	44.12
1924-1933	241	20.49	43	12.25	198	24
1934 onward	147	12.5	83	23.65	64	7.76
N	1176		351		825	
Country of Birth						
Austria & Germany	143	12.02	56	15.82	87	10.41
Britain & Ireland	119	10	8	2.26	111	13.28
Italy	206	17.31	28	7.91	178	21.29
Russia & Poland	236	19.83	122	34.46	114	13.64
Rest of World	486	40.84	140	39.55	346	41.39
N	1190		354		836	
Religion						
Catholic	302	32.44	43	14.24	259	41.18
Jewish	326	35.02	193	63.91	133	21.14
Orthodox	56	6.02	13	4.3	43	6.84
Protestant	247	26.53	53	17.55	194	30.84
N	931		302		629	
Father Occupation						
Farmer	161	17.61	27	9.25	134	21.54
Laborer	145	15.86	35	11.99	110	17.68
Skilled	361	39.5	111	38.01	250	40.19
White collar	247	27.02	119	40.75	128	20.58
N	914		292		622	
Urban						
Urban	362	30.42	141	39.83	221	26.44
Non urban	828	69.58	213	60.17	615	73.56
N	1190		354		836	

Note: Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. We combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. The "Rest of World" category includes the rest of Europe, Asia and South America.

Table 2. Association between refugee status and English fluency in the New Immigrant Survey

	(1)	(2)	(3)	(4)
Panel A: Outcome: Speaks any English				
Refugee	0.0168	0.0793**	0.0797**	0.0714*
	(0.031)	(0.032)	(0.032)	(0.037)
Recent English class			0.0691***	0.0489*
			(0.015)	(0.028)
Pre-US English class			0.103***	0.0491**
			(0.013)	(0.024)
Adjusted R^2	0.351	0.416	0.433	0.299
Panel B: Outcome: Speaks English Well				
Refugee	-0.108**	-0.0465	-0.0333	-0.0284
	(0.045)	(0.042)	(0.042)	(0.053)
Recent English class			-0.151***	-0.122***
			(0.018)	(0.042)
Pre-US English class			0.156***	0.120***
			(0.018)	(0.038)
Add'l Control	No	Yes	Yes	Yes
Sample Type	Full	Full	Full	In US $\geq$ 2 yrs
Adjusted R^2	0.347	0.391	0.420	0.379
No. of Obs.	2938	2938	2938	654

Note: All regression specifications include fixed effects for departure timing and country of origin, as well as controls for age, age squared, and gender. Additional Controls includes years of schooling prior to immigration, rural/urban, and religion. Specifications 1-3 use the full sample, where as specification 4 includes only immigrants who have been in the US for at least 2 years. Refugee = 1 for immigrants with Refugee or Asylee visa status. Recent English class is defined as a current English class or one taken within the past year. Standard errors in parentheses

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

Appendix Table 1. Coding scheme for variables collected from oral histories.

Variable Name	Label	Description/Coding	Data Type	Coding Notes
id	Numbered list of entries	Number	Integer	
source	Oral history archive (e.g. Ellis Island Foundation)	Organization name	Open String	See "sources" tab
firstname	Respondent's first name at time of interview	Name	Open String	
lastname	Respondent's middle name at time of interview	Name	Open String	
birthplace_general	Country of birth noted at top of oral history	Country	Open String	
birthplace_specific	Specific place of birth given by respondent when asked where they were born	Town/Area, County/Country	Open String	
birth_year	Birth year	yyyy	Integer	N/A if missing
arrival_year	Year that respondent arrived	yyyy	Integer	N/A if missing
age_at_arrival	Age the respondent was when they came to the US	Age	Integer	N/A if missing
interview_year	Year of interview	yyyy	Integer	N/A if missing
age_at_interview	Age of respondent at time of interview	Age	Integer	N/A if missing
religion	Respondent's religion	Name (i.e. catholic, orthodox, jewish, etc)	Open String	
refugee	Respondent left country for reasons of violence, religious/ethnic/racial persecution, or an event such as famine	Yes/No	Constrained String	
pm_urban_rural	Grew up in rural or urban area before migration	Urban/Rural	Constrained String	Rural = agricultural area or small town or village. Urban = city or large town, Urban-rural = grew up in some mix of urban and rural places
pm_occupation	Last job of respondent in home country before they migrated	Name (i.e. farm laborer, helper, etc)	Open String	This is before they came to the US
pm_father_occ	Respondent's father's job in home country	Name (i.e. farmer, blacksmith, etc)	Open String	
pm_mother_occ	Respondent's mother's job in home country	Name (i.e. farmer, blacksmith, etc)	Open String	

pm_educ_level	Highest level completed before the interviewee migrated	Options: none, elementary school, some high school, high school, some college, trade school, associate's degree, bachelor's degree, master's degree, doctoral or professional degree, N/A	Constrained String	If respondent migrated during school, indicate level at migration (e.g. someone who started high school abroad and finished high school in US would be "some high school")
pm_married	Already married when they migrated?	Yes/No	Constrained String	
language_native	Respondent's native language	Language (e.g. English, Russian etc)	Open String	
pm_english	Did the interviewee know English before migrating?	Yes/No	Constrained String	
migrate_alone	Did the interviewee migrate alone?	Yes/No	Constrained String	"No" if they travelled alone but were meeting someone, or if someone else in their party was following immediately after (within 6 months)
documented	"Yes" if respondent talked about a visa, greencard, etc; "No" if respondent was undocumented when they migrated	Yes/No	Constrained String	"NA" if not discussed
refugee	Was the respondent a refugee?	Yes/No	Constrained String	
family_sponsored	Did respondent's relatives financial support their migration or apply for visas for them?	Yes/No	Constrained String	
org_sponsored	Did respondent receive aid and/or support from a NGO to migrate? (e.g. church sponsoring or hosting refugees; educational scholarship from a university or nonprofit...)	Yes/No	Constrained String	
learn_english	If the respondent did not speak English before migration, how did they learn English? (e.g. school, friends, college course)	Open String	Not a Yes/No	
total_children	number of children the respondent had (including adopted)	0, 1, 2...	Integer	
am_married	Respondent got married after migration	Yes/No	Constrained String	"No" if the respondent got married before migrating and did not remarry

spouse_same _nationality	Spouse born in same country as respondent	Yes/No	Constrained String	Answer can be for a spouse married before or after migration
am_occ_first	Respondent's first job in US	Name of job	Open String	
am_latest_occ	What is the respondent's current job? If retired or unemployed, what was their most recent job?	name of job	Open String	
am_educ_level	Highest educational level completed at time of interview	Options: none, elementary school, some high school, high school, some college, trade school, associate's degree, bachelor's degree, master's degree, doctoral or professional degree, N/A	Constrained String	If the interviewee completed no further school after migration, response should match pm_educ_level answer
am_loc_first	Location of first residence in US	location in US (e.g. Portland, Maine)	Open String	
imm_destination	Was respondent's first location somewhere where they had a community of immigrants from the same country?	Yes/No	Constrained String	
am_loc_current	Location of current residence in US (at time of interview)	Location in US (e.g. Portland, Maine)	Open String	
ever_visit	Has the respondent been back to their home country since migrating?	Yes/No	Constrained String	
citizen	Did respondent ever become an American citizen?	Yes/No	Constrained String	

Note: This table provides the coding scheme used by research assistants to convert transcripts obtained from oral histories to data. "Open String" refers to RAs being able to enter the information that they come across in the oral histories, whereas "Constrained String" refers to RAs picking between the provided options.

Appendix Table 2. Summary statistics of measures of linguistic ability.

	N	Mean	Std Dev	Min	Max
Age of Acquisition	1100	4.86	0.33	4	6
Mean Sentence Length	1100	12.20	4.54	2	42
Accent	915	0	1	-4	3
Syllables Per Minute	884	163.76	39.47	53	274

Note: This table shows the summary statistics of the four measures of linguistic ability: Age of Acquisition (AoA), Mean Sentence Length (MSL), Accent and Syllables Per Minute (SPM).

Appendix Table 3. Correlation between measures of linguistic ability.

	AoA	MSL	Accent	Years of Schooling	Arrival Age	Income	Syllables Per Minute
Age of Acquisition	1.0000						
Mean Sentence Length	0.3812	1.0000					
Accent	0.3199	0.2577	1.0000				
Years of Schooling	0.5394	0.2006	0.3202	1.0000			
Arrival age	-0.1551	-0.1381	-0.5214	-0.2068	1.0000		
Income	0.2431	0.1371	0.0355	0.2389	0.0548	1.0000	
Syllables Per Minute	0.1222	0.1434	0.2652	0.2064	-0.1957	0.0275	1.0000

Note: This table depicts correlation between the four measures of linguistic ability: Age of Acquisition (AoA), Mean Sentence Length (MSL), Accent and Syllables Per Minute (SPM) and other characteristics of migrants from our data such as years of schooling, arrival age and income. N= 438.

Appendix Table 4. Summary of non-refugee and refugees in the National Immigrant Survey

	Non-refugee		Refugee	
	Observations	Mean	Observations	Mean
Speaks English	2,994	.74	169	.85
		(.44)		(.36)
Speaks English well	2,994	.39	169	.33
		(.49)		(.47)
English class, current or within the last year	2,871	.28	169	.33
		(.45)		(.47)
English class before arrival in US	2,830	.34	168	.26
		(.47)		(.44)
Age	3,148	40.36	171	39.93
		(14.75)		(14.62)
Female	3,157	.54	171	.51
		(.50)		(.50)
Years of schooling	3,148	12.10	171	11.37
		(4.72)		(4.11)
Rural	3,151	.42	171	.41
		(.49)		(.49)
Catholic	3,157	.32	171	.12
		(.47)		(.33)
Christian Orthodox	3,157	.11	171	.18
		(.31)		(.38)
Protestant	3,157	.15	171	.26
		(.35)		(.44)
Muslim	3,157	.11	171	.12
		(.32)		(.33)
Other Religion	3,157	.14	171	.17
		(.35)		(.38)
No Religion or declined to answer	3,157	.12	171	.15
		(.33)		(.35)
Year of departure	3,125	2000	168	1997
		(6.73)		(5.34)
Europe and Central Asia	3,157	.10	171	.26
		(.30)		(.44)
Russia, Ukraine and Poland	3,157	.03	171	.36
		(.18)		(.48)
Middle East and North/Sub-Saharan Africa	3,157	.16	171	.16
		(.37)		(.37)
Rest of World	3,157	.71	171	.22
		(.46)		(.41)
N	3,157		171	

Note: This table presents the characteristics of the 2003 cohort of surveyed immigrants in NIS, split by non-refugee and refugee status. This cohort is a random sample of adults receiving legal permanent residence between May and November of 2003. Refugee = 1 for immigrants with Refugee or Asylee visa status. The standard deviation of each variable is in parentheses.

Appendix Table 5. Association between refugee status and linguistic outcomes, immigrants arriving after age 12.

	AoA	AoA (drop words)	MSL	Accent	SPM	AoA	AoA(drop words)	MSL	Accent	SPM
Refugee or mixed reasons	0.386***	0.352***	0.127	-0.00634	0.160	0.241**	0.215*	0.0227	-0.134	0.0921
	(0.098)	(0.098)	(0.100)	(0.101)	(0.110)	(0.110)	(0.112)	(0.110)	(0.121)	(0.126)
Laborer						-0.138	-0.149	-0.124	0.490***	0.00876
						(0.176)	(0.178)	(0.175)	(0.189)	(0.199)
Skilled						0.0840	0.0763	0.0973	0.250	0.0398
						(0.141)	(0.142)	(0.140)	(0.153)	(0.161)
White Collar						0.321**	0.310**	0.149	0.156	-0.137
						(0.150)	(0.152)	(0.150)	(0.164)	(0.173)
Urban						0.175*	0.173*	-0.0446	-0.0444	0.0222
						(0.104)	(0.105)	(0.103)	(0.113)	(0.120)
Catholic						-0.265	-0.272	-0.441***	-0.0983	-0.433**
						(0.166)	(0.167)	(0.165)	(0.175)	(0.182)
Jewish						0.0941	0.0525	-0.0847	0.220	-0.108
						(0.163)	(0.165)	(0.162)	(0.174)	(0.182)
Protestant						-0.247	-0.252	-0.380**	-0.108	-0.694***
						(0.162)	(0.164)	(0.161)	(0.172)	(0.182)
Orthodox						-0.514**	-0.533**	-0.615***	-0.337	-0.0177
						(0.223)	(0.225)	(0.221)	(0.242)	(0.252)
Outcome mean	-0.118	-0.127	-0.119	-0.513	-0.145	-0.109	-0.120	-0.157	-0.525	-0.111
R^2	0.346	0.334	0.151	0.123	0.0941	0.384	0.369	0.197	0.162	0.170
N	454	454	454	393	383	360	360	360	317	309

Note: This table reports the underlying coefficients for Figure 5. Linguistic measures have been standardized. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are not included in this sample. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Results are unweighted.
$p < 0.1$; **$p < 0.05$; ***$p < 0.01$; standard errors are shown in parentheses.

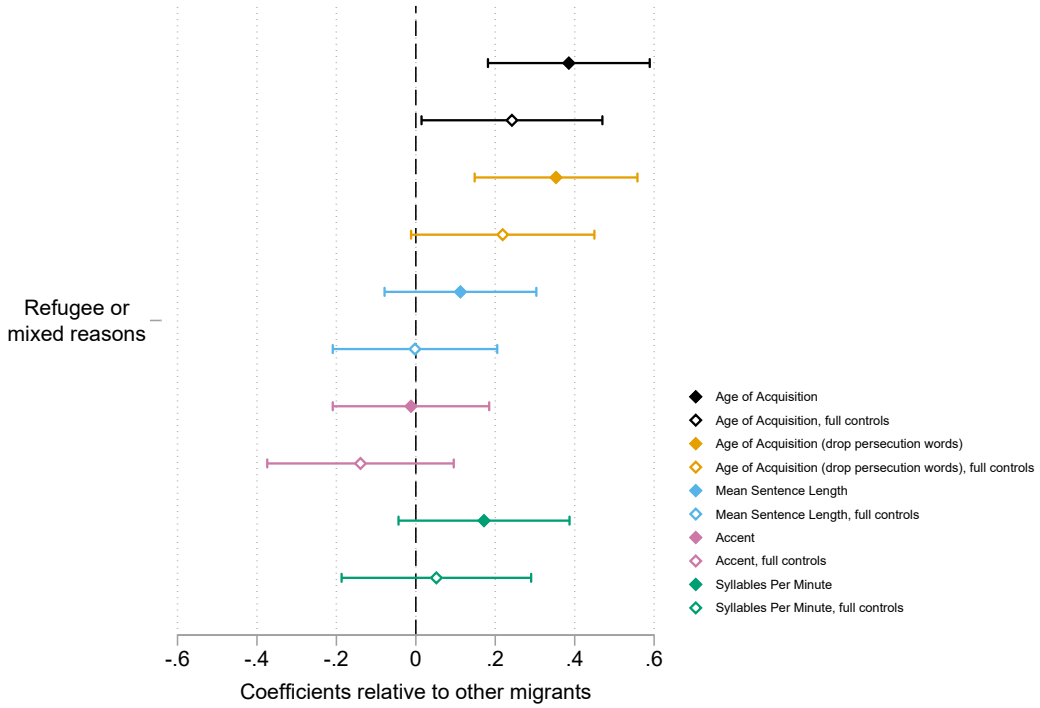
Appendix Table 6. Persecution related words in Age of Acquisition measure.

Word	Age of Acquisition	Total Usage
annexation	13.85	1
anti	9.37	170
armistice	14	9
armistices	14	1
army	7.15	861
attack	6.58	65
battalion	11.95	29
bomb	8	54
camps	5.78	70
communism	11.74	14
communist	13.22	54
communists	13.22	48
duty	7.15	76
fascism	14.33	3
fascist	14.68	8
freedom	7.05	175
genocide	13.2	23
ghetto	10.15	65
kill	6.35	217
killed	6.35	458
navy	7.15	127
oven	5.67	177
pogrom	14.33	33
pogroms	14.33	53
recession	13.74	11
refugees	10.56	107
revolution	10	68
socialist	13.61	25
socialists	13.61	16
survive	7.11	73
survived	7.11	121

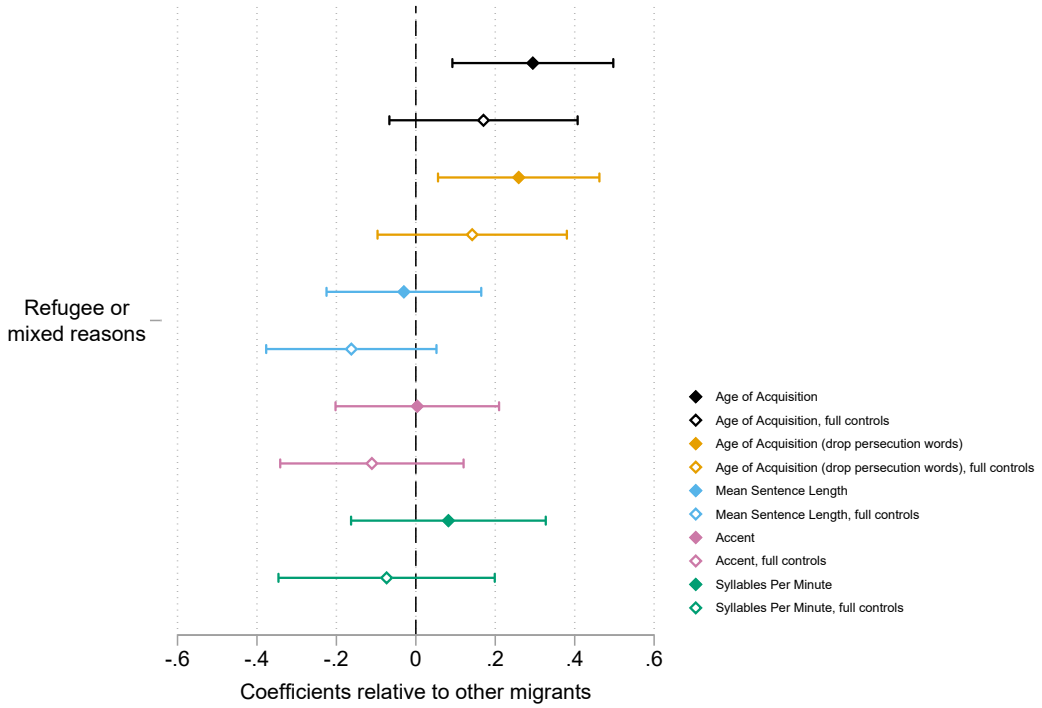
Note: This table lists persecution related words used in oral histories. These words were dropped to create a robust Age of Acquisition measure, referred to as "Age of Acquisition (drop persecution words)" in main figures.

ID_base	row	first_name	middle_name	last_name	birthplace_general	birthplace_specific	birth_date	interviewer	interview_location	interview_year
1167	1	Jenny	NA	Bohsung	Germany	Westhausen, Germany	NA	Margo Nash	NA	1973
1184	2	John (Jack)	Brenden	Brady	Ireland	Ballyhaise, Ireland	7/29/23	Paul E. Sigrest	Ellis Island Recording Studio	1995
1190	3	Rose	NA	Breci	Italy	Carletini, Italy	NA	Dr. Willa Appel	Ellis Island Recording Studio	1985
1371	4	Tilda	NA	De Mello	Brazil	Manaus, Brazil	4/12/17	Janet Levine	Albertson, New York	1992
1494	5	Rita	Costa	Finco	Italy	Asiago, Italy	7/16/16	Janet Levine	Cudahy, Wisconsin	1996
1642	6	Julia	Barlas	Groulx	Greece	NA		Nancy Dallet	NA	1989

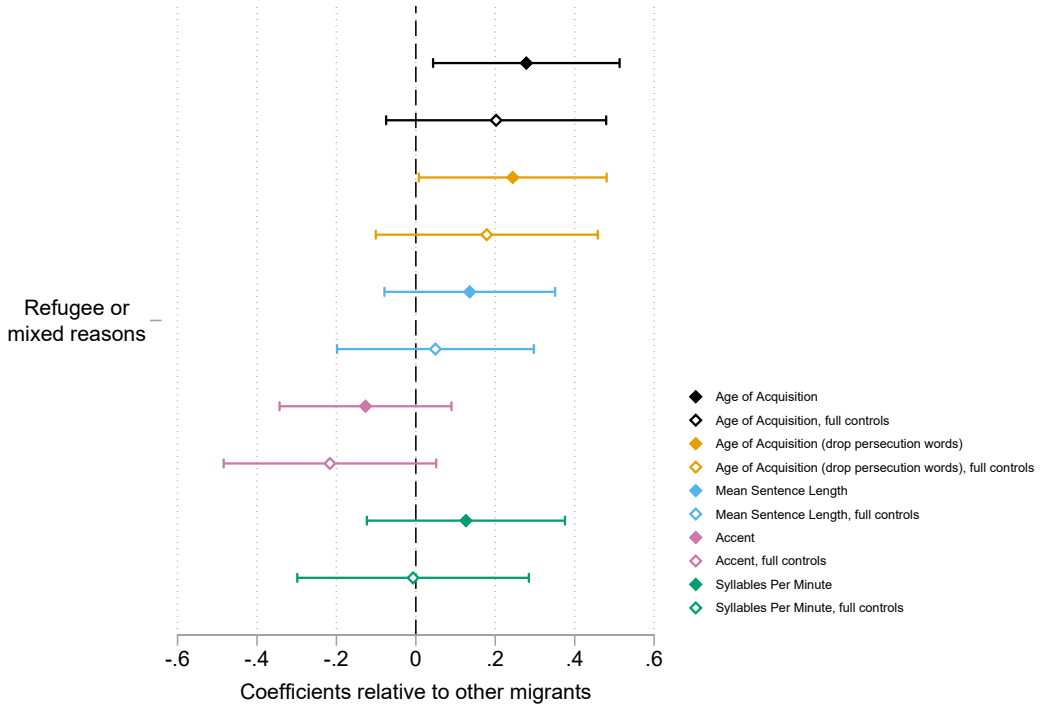
Appendix Figure 1. Image of standardized template used to manually extract data from oral histories. Image of part of template used by coders to fill in information for individuals in oral histories.



Appendix Figure 2. Association between refugee status and linguistic outcomes for immigrants arriving after age 12, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The five linguistic measures are: Age of Acquisition (N= 512, N= 413 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 512, N= 413), Mean Sentence Length (N= 512, N= 413), Accent (N= 449, N= 368) and Syllables Per Minute (N =438, N=359). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. A more positive accent score indicates an accent closer to that of the US born. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.



Appendix Figure 3. Association between refugee status and linguistic outcomes for immigrants arriving after age 12, dropping migrants arriving after 1933, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The four linguistic measures are: Age of Acquisition (N= 424, N= 340 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 424, N= 340), Mean Sentence Length (N= 424, N= 340), Accent (N= 372, N= 302) and Syllables Per Minute (N =362, N=294). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Results are unweighted. Significance is at the 5% level.



Appendix Figure 4. Association between refugee status and linguistic outcomes for immigrants arriving after age 14, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The five linguistic measures are: Age of Acquisition (N= 415, N= 331 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 415, N= 331), Mean Sentence Length (N= 415, N= 331), Accent (N= 363, N= 296) and Syllables Per Minute (N =355, N=288). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.