CENTRE FOR ECONOMIC HISTORY
THE AUSTRALIAN NATIONAL UNIVERSITY
DISCUSSION PAPER SERIES

AGE AT ARRIVAL AND ASSIMILATION IN THE AGE OF MASS MIGRATION

ROHAN ALEXANDER
AUSTRALIAN NATIONAL UNIVERSITY

ZACHARY WARD
AUSTRALIAN NATIONAL UNIVERSITY

DISCUSSION PAPER NO. 2018-03

FEBRUARY 2018

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

Age at Arrival and Assimilation during the Age of Mass Migration[†]

Rohan Alexander
Australian National University

Zachary Ward
Australian National University

February 2018

Abstract: We estimate the effect of age at arrival for immigrant outcomes with a new dataset of arrivals linked to the 1940 United States Census. Using within-family variation, we find that arriving at an older age, or having more childhood exposure to the European environment, led to a more negative wage gap relative to the native born. Infant arrivals had a positive wage gap relative to natives, in contrast to a negative gap for teenage arrivals. Therefore, a key determinant of immigrant outcomes during the Age of Mass Migration was the country of residence during critical periods of childhood development.

JEL Classification: N31, F22, J61

Keywords: Age at arrival, Assimilation, Childhood environment

[†] Rohan Alexander is a Ph.D. student at the Australian National University - Research School of Economics, HW Arndt Building 25A, College of Business and Economics, The Australian National University, Canberra, ACT 2600, Australia. E-mail: rohan.alexander@anu.edu.au. Zachary Ward is a Lecturer at the Australian National University - Research School of Economics, HW Arndt Building 25A, College of Business and Economics, The Australian National University, Canberra, ACT 2600, Australia. E-mail: zach.ward@anu.edu.au.

For comments on earlier stages of this project, we thank Tim Hatton, Priti Kalsi, Ed Kosack, Amber McKinney and seminar participants at La Trobe University. We also thank Bill Collins and two anonymous referees, whose comments substantially improved the paper. An earlier version of this paper was circulated as "The Paramount Importance of Childhood Environment during the Age of Mass Migration." We acknowledge financial support from the College of Business and Economics at the Australian National University. We thank Lee Alston for helping us to access the data used in this study.

It is increasingly apparent that where one was born and the quality of one's childhood environment are key determinants of life-long outcomes.[1] By definition, immigrants are born in a different environment than natives; therefore, immigrants are exposed to a different educational, cultural and health setting during critical periods of development. How much of the economic gaps between immigrants and natives during the Age of Mass Migration can be attributed to growing up in different environments? There are many other factors that may explain the gaps between immigrants and natives besides the where one grew up, such as the direction of selection into immigration, the degree of discrimination from natives, or the extent of sorting into different enclaves (Biavaschi et al. 2017; Borjas 1987; Cutler et al. 2008).

To estimate the importance of growing up abroad we exploit variation in the length of childhood exposure to source country conditions, as measured by the migrant's age at arrival. By comparing the adult outcomes of older child arrivals to younger child arrivals, we can uncover the extent to which outcomes in the United States depended on where one spent his infancy or adolescence (Chetty and Hendren 2017a; Chetty and Hendren 2017b). This method also allows us to identify critical ages for when a move improved migrant outcomes the most; prior research on child arrivals misses variation within the group by treating all children as a single category (Hatton 1997; Minns 2000). While one might expect that younger arrivals would do better in the long run since they lived in the United States environment during critical periods of development, it is also possible that younger arrivals were penalized since immigrants often lived in high-mortality urban areas (Eriksson and Niemesh 2016).

We take advantage of the complete digitization of immigration records to construct a sample of brothers arriving at Ellis Island between 1892 and 1924, which we then link forward

[1] The effect of childhood environment on economic outcomes is a long-standing question in the economics literature. For recent literature reviews, see Almond, Currie and Duque (2017) and Cunha et al. (2006) on the importance of environment during early stages of childhood. Also see the work by Chetty and Hendren on the importance of childhood environment past age eight (Chetty and Hendren 2017a; Chetty and Hendren 2017b; Chetty, Hendren and Katz 2016).

to the full-count 1940 Census. With a linked dataset of over 50,000 brothers, we then estimate the effect of age at arrival by comparing brothers who immigrated at different ages. This strategy controls for household-invariant unobservable characteristics such as parental income and education that may be correlated with both age at arrival and migrant outcomes (Böhlmark 2008; van den Berg et al. 2014; Clarke 2016).

We find that an older age at arrival, and thus longer exposure to the childhood environment in Europe, had a large and negative effect on the native-immigrant gap in outcomes such as wage income and occupational status. For 16-year old arrivals, the native-immigrant wage gap was 17 log points more negative compared with the gap for those who arrived at age one – an effect that is equal in size to two fewer years of education. The size of this effect is larger than the overall wage gap between teenage arrivals and white natives; therefore, we show that infant arrivals had a positive wage gap relative to natives, in contrast to a negative gap for teenage arrivals.

After establishing that arriving at an older age had a large negative effect on the native-immigrant gap in economic outcomes, we explore potential channels for this effect. One mechanism is through educational attainment: 16-year-old arrivals acquired one less year of schooling than infant arrivals. However, a one-year difference in education does not explain the entire income effect, suggesting that other mechanisms besides educational attainment were important. We show that older arrivals were also penalized because potential foreign labor market experience was not rewarded in the United States, implying that pre-migration human capital did not transfer perfectly across borders. Older arrivals were also less socially assimilated, as measured by their rate of marriage to a native-born spouse, which may have penalized them in the labor market (Abramitzky et al. 2016; Biavaschi et al. 2017). While we cannot pinpoint which channel was most important, we consistently show that longer exposure to the European environment during critical periods of development was strongly correlated with a variety of migrant outcomes during the Age of Mass Migration.

Our study contributes to the growing literature on immigrant assimilation during the Age of Mass Migration using newly digitized records (Abramitzky et al. 2014; Abramitzky et al. 2016; Biavaschi et al. 2017; Ward 2018). The current understanding in the literature is that the average immigrant's position in the *occupational* distribution was fixed and did not change relative to natives throughout the life cycle; note that this does not imply *income* convergence did not occur, but incomes are unobserved prior to 1940. We show that a male immigrant's position in the occupational distribution depended strongly on his age at arrival. These results suggest that while human capital acquired during adulthood, such as English fluency post arrival, had a smaller impact on the native-immigrant gap in occupations, human capital acquired during childhood had a larger impact (Ward 2018). Our results also add to the growing literature on age-at-arrival effects by showing that they were large and important during the Age of Mass Migration, a time period when the economic gap between source countries and the United States was smaller than the economic gap between source countries and the United States today (Abramitzky and Boustan 2017; Böhlmark 2008; van den Berg et al. 2014).

Our study also complements the literature on the intergenerational assimilation of immigrants (Abramitzky et al. 2014; Borjas 1994; Card et al. 2001). Child immigrants are sometimes called the "1.5" generation since they bridge the gap between adult arrivals in the first generation and native-born individuals in the second generation. Our results suggest that the second generation should have improved on the first generation's relative position with natives since the second generation spent their entire childhood within the United States. Yet the intergenerational assimilation literature also documents that convergence of occupational status for descendants from different source countries was not complete for the children and grandchildren of European migrants, even though these generations were raised in the same country. A lack of convergence across generations from different sources is consistent with there being large differences in the quality of childhood environment across areas within the United States where immigrants from different sources settled, a potential topic for future research.

AGE AT ARRIVAL: HISTORICAL SETTING AND RELATED LITERATURE

The Age of Mass Migration (1850-1913) is often split into two sub-eras based on the geographical shift of flows from Northern and Western Europe ("Old" sources) to Southern and Eastern Europe ("New" sources) in the late 1880s. At around the same time there was also a shift in family composition from intact households (including many children) to unattached males, lowering the fraction of child arrivals (Baines 1995; Hatton and Williamson 1998). Illustrating this shift in the late 19th century, Michael J. Greenwood (2007) reports that the percentage of those under 14 in the inflow from major European sources dropped from a high of 25 percent in 1884 to a low of 11 percent in 1895. This shift away from family and child migration is associated with younger males taking advantage of the decreasing costs of travel due to the diffusion of steam technology and migration networks (Cohn 2009; Gould 1980).

Our data cover migrants who entered through Ellis Island between 1892 and 1924, a period when child arrivals were slowly making up a larger share of arrivals (see Figure 1).[2] Children were still a small, but increasingly important, part of the inflow: overall, the fraction of child arrivals increased from 10 to 14 percent before World War I to slightly above 15

[2] The data in the series are from the Annual Reports of the Commissioner General of Immigration (1899-1932). One caveat to Figure 1 is that both the definition of an immigrant and a child arrival changed during the early 20th century (Hutchinson 1958). Prior to 1903, any entrant, excluding the cabin class, was counted as an immigrant. For the following two years (1904 and 1905), the definition changed to include the cabin class. From 1906 onward immigrants were those who intended to stay for more than one year and had been outside of the United States for more than one year. Besides the definition of immigrant, the definition of a child arrival also changed from those under the age of 14 prior to the 1917 literacy test to those under the age of 16 afterwards. A final caveat is that the Annual Reports may have underestimated the number of arrivals due to careless compiling of ship manifests by the Bureau of Immigration (Bandiera et al. 2013). However, since undercounting is mostly due to entire ships missing from the totals, it is unclear how it would bias the fraction of children in the arrival flow.

percent in the following decade. Some of this increase is due to several shocks to the immigration system, such as the cut-off of flows during World War I, the Literacy Act of 1917 and the immigration quotas laws of 1921 and 1924. While United States policy significantly restricted the overall flow, child arrivals were favored under these policies since those under 16 were not subject to the literacy test, and children joining a naturalized family member were given preference under the quota system.[3] Consistent with policies favoring children over single adults, the countries that were more restricted under the quotas and literacy test (in Southern and Eastern Europe) had a relative increase of children in their flow.

While child arrivals were less than 20 percent of arrivals in the early 20th century, they were about 30 percent of the migrant stock, partially because they were more likely to remain rather than return home.[4] This can be directly seen in return flow records where children were underrepresented on out-going ships relative to the migrant stock; moreover, arrival records show that families with children were more likely to plan to stay in the United States permanently than single arrivals (Ward 2017). Yet not all young migrants arrived with family members; this is indirectly seen in the distribution of age at arrival in the migrant stock in Figure 2.[5] While there were about the same proportion of arrivals at age one as for age twelve, there were much more arrivals aged thirteen and above, perhaps because teenagers were more likely to migrate by themselves. If older arrivals came individually while younger arrivals came as part of a family, then older teenagers could be selected from a different part of the source country human capital distribution; therefore, it will be important to estimate the effect of age

[3] Both the 1921 and 1924 immigration quotas allowed child immigrants to join naturalized family members even if the quota for the country was full.

[4] This 30 percent number is based on 1899 to 1930 arrivals in the 1900-1930 IPUMS samples (Ruggles et al. 2017).

[5] This figure is created using the 1900-1930 IPUMS samples (Ruggles et al. 2017) and keeping those who arrived between 1899 and 1930 to match Figure 2. A random sample of ships to Ellis Island from Ward (2017) confirms that older arrivals tended to travel alone, where about 20.5 percent of 14-year olds, 23.6 percent of 15 year-olds and 52.9 percent of 16 year olds entered the United States without a family member (defined by same surname) on the ship.

at arrival with an empirical strategy that accounts for changing unobservables across the arrival age distribution.

Early 20th century officials recognized the importance of age at arrival for successful assimilation; of special concern was whether older arrivals were falling behind in school. The 1910 Dillingham Commission Reports on *The Children of Immigrants in Schools* showed that 43 percent of children who arrived under age 6 were behind their grade level, compared with 92 percent of those who arrived at 10 years or older. The authors argued that "the child who comes to this country before he reaches school age often has an opportunity to adjust himself to his new surroundings and in some cases learn the language through contact with other children before entering school" (Immigration Commission 1910, pg. 51). In response to this trend of child arrivals being poorly educated, states passed compulsory schooling to educate immigrant children who arrived from countries without a compulsory educational system (Bandiera et al. 2017).

Despite early 20th century officials' interest in age at arrival, the Congressional Report is one of the only studies that separates historical migrant outcomes by arrival age.[6] Others that account for age at arrival often group all child arrivals into a single category. Both Chris Minns (2000) and Timothy Hatton (1997) show that those who arrived under the age of 16 had higher income levels and better-paid occupations than adult arrivals, consistent with a negative effect of age at arrival and longer exposure to the European environment. On the other hand, Ran Abramitzky, Leah Boustan and Katherine Eriksson (2014) show that assimilation rates were similar whether one keeps or drops those who arrived under the age of 10, but since they do not isolate the sample to only child arrivals, the difference in assimilation for child arrivals is

[6] Ward (2018) estimates the effects of age at arrival on English proficiency using the 1900 to 1930 United States cross sections, and indicator variables for each arrival age. Ward is primarily interested in using the estimates to verify the quality of the English proficiency variable rather than to directly analyze the effect of age of arrival on occupational outcomes.

unclear. We improve on this limited literature by estimating the effect of age at arrival across all ages, rather than grouping all child arrivals together. We also use an empirical strategy that controls for household-invariant unobservables that are correlated with age at arrival and with economic outcomes, which is important in today's studies on age at arrival (Clarke 2016).

In contrast to the scarcity of historical studies, several modern-day studies estimate the effect of age at arrival on adult outcomes with high-quality data.[7] The most credible method to identify the age-at-arrival profile uses sibling fixed effects. This requires a large amount of data and therefore has been primarily studied using Swedish and Norwegian administrative records (Böhlmark 2008; van den Berg et al. 2014). Outside of Northern Europe, there are few studies that identify the effect of age at arrival with siblings. Raj Chetty and Nathaniel Hendren (2017a, 2017b) use United States tax records to show that variation in age at migration across counties has a large effect on adult outcomes, implying that childhood environment varies widely across counties in the United States. We follow this sibling fixed effects approach to estimate the importance of childhood environment for immigrants from the past.

LINKING ELLIS ISLAND RECORDS TO THE 1940 CENSUS

The main dataset used for estimation comes from linking two large data sources: Ellis Island records from 1892 and 1924 and the preliminary full-count 1940 Census available at IPUMS (Ruggles et al. 2017). The Ellis Island records have been digitized and are searchable online; note that this source is the same one used by Bandiera et al. (2013) and Yannay Spitzer and Ariell Zimran (2017).[8] While the clear advantage of the Ellis Island records is that they include millions of observations, there are a few disadvantages. One is that not every variable

[7] Friedberg (1992) is the seminal study of age at arrival on adult outcomes. Several outcomes besides income have been explored, including human capital outcomes such as language acquisition and educational attainment (Bleakley and Chin 2004; Böhlmark 2008; Schoellman 2016), social outcomes such as intermarriage or living in an ethnic enclave (Åslund et al. 2015; Bleakley and Chin 2010), and health outcomes such as height (van den Berg et al. 2014).

[8] Many arrival records prior to 1897 were lost in a fire, so coverage prior to 1897 is not complete (Spitzer and Zimran 2017).

in the arrival records is digitized, such as occupation, relationship status, or height. Moreover, the Ellis Island records include both immigrants and non-immigrants; non-immigrants are other entrants such as business travelers, tourists or those traveling through to another country. However, we are only interested in those who we can locate in the 1940 Census, and thus those who have stayed permanently (and survived) until 1940.

Our population of interest in the Ellis Island records is brothers who are single and arrived between the ages of zero and twenty. For this population (who are primarily European), we collect first name, last name, age, date of arrival, place of last residence and ethnicity. We identify brothers as immigrants who are listed next to each other with the same last name and are less than ten years apart in age, although we do not have their relationship listed in the data. The key variable of interest from these records is age, which we wish to attach to their adult observation in the 1940 Census. For a discussion of the assumptions we made in cleaning the data, please see Online Appendix B. After cleaning, we have 397,003 brothers who can be linked to the 1940 Census.

We link these brothers to the 1940 U.S. Census using a match on first name, last name, country of birth, and year of birth in a 3-year range.[9] We find potential matches based on having an exact NYSIIS match on first and last name; however, we choose the best match based on the smallest sum of the absolute difference in year of birth, Jaro-Winkler distance in first name and Jaro-Winkler distance in last name.[10] Catherine Massey (2017) shows that this method of ranking matches is reasonable for improved match rates and reduced false positives. For more detail on the linking process, see Online Appendix C.

[9] We access the 1940 Census on the National Bureau of Economic Research (NBER) server due to restrictions on observing the first name and last name in the public-use dataset.

[10] NYSIIS, or the New York State Identification and Intelligence System, is a phonetic algorithm to standardize similar sounding names. The Jaro-Winkler algorithm measures the distance between strings based on the number of matching characters. Using the actual first and last name strings to gauge the quality of match is recommended by Bailey et al. (2017), rather than treating all matches with the same NYSIIS code as of equal quality. We show that results are robust to using a method related to Feigenbaum (2016) in Appendix E.

It is possible that our linking methodology incorrectly links some people, which would induce measurement error and bias our sibling fixed effects estimates toward OLS estimates (Bailey et al. 2017). However, as we will show in robustness checks, our results do not change when limiting our sample to higher-quality links in terms of closer matches in first and last name strings and year of birth. The results are also robust in an alternative sample where links are chosen based on a predicted match score calculated from a hand-linked sample of immigrants, a method that is related to the linking strategy described by James Feigenbaum (2016).[11] Overall, we are confident that our results are not driven by link quality.

We take one extra step when linking the datasets because we start with arrival records unlike others who link from census to census. We are concerned that some immigrants may have changed their first name to be more "American" after arrival; for example, from Giuseppe to Joseph or Pietro to Peter. Biavaschi et al. (2017) show that name changes occur for 32 percent of their sample of naturalization records in New York and that name changes were more common for Southern and Eastern Europeans. To account for this possibility, we Americanize the first names in our dataset of arrival records and the first names in the Census records. This will allow us to match Giovanni at arrival to John in the Census, but also to match Giovanni (Americanized to John) at arrival to Giovanni (Americanized to John) in the Census in case Americanization did not occur. We do this with a list of over 28,000 variants of first names based on information at behindthename.com.[12] The Americanization process improves our linking rates by about 35 percent, but, as we will show in a later robustness check, our results do not substantially change if we do not Americanize first names.

[11] We use the hand-linked samples from Ward (2018) to predict the best link among the set of potential links. While this method is related to Feigenbaum (2016), it is not exactly the same since our "training sample" is from immigrants linked between 1920-1930 US Censuses rather than Ellis Island records to the 1940 Census.

[12] For some names, there are multiple American-sounding variants. We choose the variant that is most popular for years of birth prior to 1930, data which is available from the Social Security Administration at http://www.ssa.gov/oact/babynames/names.zip.

The starting sample of 397,137 brothers is successfully linked for 103,005 individuals in the 1940 Census using our main linking approach, or 25.9 percent of arrivals. Since the main empirical strategy exploits variation within brothers, we drop individuals where one brother was linked and another was not. This restriction gives us a final sample of 53,129, or 13.4 percent of our original set of brothers.

Table 1 shows the linking rates by country of birth and demonstrates a common pattern in the literature where we are less likely to link Southern and Eastern Europeans relative to Northern and Western Europeans (Abramitzky et al. 2014; Ward 2018). Clearly, our linked sample is not a random sample of foreign-born brothers. We are not able to test for the representativeness of the sample on occupation or literacy compared to all Ellis Island arrivals since these variables are not digitized. Yet we would rather test for representativeness according to the 1940 Census since the Ellis Island records include many non-immigrants and thus any difference between our linked sample and those in the Ellis Island data would reflect both selection into permanent migration and selection into the linked sample. However, we also cannot test for representativeness according to the 1940 Census because it does not separate immigrants by cohort or age of arrival, once again making it unclear whether any differences are due to biases from the linking process or because the linked data has younger arrivals.[13]

Most linked samples that use a similar linking methodology are found to be slightly higher skilled than the underlying population and only show a strong bias in country of birth (Abramitzky et al. 2014). New source countries are less likely to be linked to the 1940 Census than Old sources because of common names, return migration, misspelled names, or names that are not captured in our list of Americanized names. To account for this bias, we reweight

[13] We can compare our sample to the 1940 migrant stock, which we do in Appendix B. Our linked sample is higher skilled, more highly educated, and less likely to be from a New source country.

the sample to reflect the migrant stock by country of birth in 1940, although this reweighting does not drive our results.[14]

Table 2 shows the descriptive statistics of the linked sample of brothers and the one percent sample of white natives in the 1940 census, illustrating the gaps in economic and social outcomes between immigrants and natives. We compare our immigrants to white natives on several outcomes, including years of completed education, occupation and wage income. Note that whenever we use wage income, in this table or in later regressions, we exclude self-employed workers since business and farm income are not included in the 1940 Census.

The migrants in our sample have been in the United States for an average of 30 years. Therefore, those who arrived between age zero and five are on average 32 years old in 1940, while those who arrived between ages 16 and 20 are on average 49 years old in 1940. Considering these differences in age, it will be important to adjust for age when examining differences between immigrants and natives, which we show in the bottom half of the table. After adjusting outcomes based on white natives' life-cycle profile, Table 2 shows that there is a strong negative gradient to age at arrival for many variables.[15] For example, zero-year-old arrivals earned 9.5 percent *more* than natives, while 16-20 year old arrivals earned 7.9 percent *less*. Note that children arriving early enough have earnings that are higher than natives of the same age, perhaps implying that childhood environment may explain the entire native-immigrant wage gap for older arrivals. Yet part of the reason why younger arrivals earned more than white natives overall was because they located in urban areas. Table 2 shows that when

[14] Reweighting to match the 1940 stock is done with males who were born between 1872 and 1924 to reflect our sample of 0 to 20-year-old brothers who arrived between 1892 and 1924. We alternatively reweighted to match the 1930 Census distribution of country of birth, a census which includes year of arrival and thus we can reweight to match those who arrived between 1892 and 1924. Either weighting to match the 1930 or 1940 migration distribution yields the same results.

[15] To adjust for age, we use the standard method in the assimilation literature and run a first regression of the outcome on the full-range of age fixed effects with our sample of white natives, and then calculate the residuals for the sample of immigrants based on predicted values from natives. See Equation (1) and the dependent variable of Equation (2) in the next section.

limiting the sample to those only in urban areas, then zero-year-old arrivals earned 4.3 percent *less* than natives. Nevertheless, when limiting the sample to urban areas, the same pattern holds where older arrivals had a larger wage gap with natives compared with younger arrivals.

One explanation for the change in income gap across age at arrival is that older arrivals were exposed to source country conditions for a longer period. On the other hand, the change may be due to selection bias such that older arrivals had worse outcomes because they came from lower income or educated families. Instead of estimating the age-at-arrival profile using variation across families, we will estimate the profile using variation within family to control for unobserved family-invariant variables (such as parental education and income) as described in more detail in the next section.

EMPIRICAL STRATEGY AND IDENTIFICATION

The first challenge when estimating the effect of age at arrival on immigrant outcomes is a standard one of collinearity: it is not possible to simultaneously estimate the effect of age at arrival, age, and years in the United States because they are linearly dependent.[16] We follow the standard practice of using natives to identify the life-cycle profile (or aging effect), and then estimate whether age at arrival influences deviations from this profile (Borjas 1985).[17] We take the two-step approach used by Joseph Schaasfma and Arthur Sweetman (2001): first, we estimate an auxiliary regression to identify the age-earnings profile using only white native-born individuals (superscript *nb*) from the same birth cohorts as our immigrant sample:

$$\ln(income_i) = \lambda_a^{nb} + \varepsilon_i \qquad (1)$$

[16] That is *Age at Arrival = Age – Years in the United States.*

[17] We use the preliminary full-count 1940 Census to estimate the life-cycle profile. We use male white native-born who are aged 15 to 69 to match the immigrant sample. Wage income is top coded at 5,000.

The income-age profile is modelled using a full-range of age fixed effects. We then estimate whether deviations from the native age-earnings profile are related to the immigrant's age at arrival, while controlling for other factors such as years in the United States:

$$\ln(income_i) - \widehat{\lambda_a^{nb}} = g(Age\ At\ Arrival_i) + h(Years\ in\ US_i) + \nu_i \qquad (2)$$

Therefore, this equation shows that the estimated effect of age at arrival on outcomes is the effect of age at arrival on the native-immigrant gap.[18] The method essentially estimates the effect of differences in age at arrival on the difference between natives' and immigrants' log income by age.

The second problem with estimating the effect of age at arrival is selection bias: immigrants who arrive at older ages may differ from those who arrive at younger ages in unobservable ways. For example, families with a strong preference for improving their child's education may have immigrated to the United States with children at younger ages prior to school entry. In this case, an estimated age-at-arrival effect may capture family preferences for investment into children and lead to a negative age-at-arrival profile if families with younger children are positively selected relative to families with older children. Indeed, in present-day data, those immigrating with younger children also tend to have higher education levels than those migrating with older children (Clarke 2016).

To address issues of selection bias when comparing immigrants across families, we compare immigrants within the family (that is, we compare brothers). The regression therefore changes to

$$\ln(income_{ih}) - \widehat{\lambda_a^{nb}} = g(Age\ At\ Arrival_{ih}) + \phi_h + \nu_{ih} \qquad (3)$$

[18] Note that when estimating this equation, the years in the United States function is a mix of assimilation and cohort effects since we only have a cross section (Borjas 1985).

where the key addition is the sibling fixed effect ϕ_h. Therefore, we relate the variation of native-immigrant income gap within siblings to the variation in age at arrival within siblings. Including household fixed effects controls for many household-invariant factors such as parental preferences for education or childhood investment, parental wealth, father and mother's education, culture, family structure, and country of origin. Note that this strategy compares individuals from the same arrival cohort with the same number of years in the United States, so these other variables of interest in the assimilation literature are dropped from the equation.

We use a non-parametric approach and code age at arrival into two-year bins (arrived between zero and one, between two and three, etc.) up until arrival at age 18.[19] This specification allows us to capture a variety of slopes in the profile such as the age-at-arrival profile being flat until ages 8 to 10 and decreasing afterwards, reflecting language acquisition or other effects of this critical period (Bleakley and Chin 2004; van den Berg et al. 2014). Alternatively, the slope could be steepest for arrival ages under five, reflecting the importance of human capital development at very young ages (Almond et al. 2017). Note that we do not control for any post-arrival outcome, such as geography or marital status, because location could be an outcome of age at arrival (Bleakley and Chin 2010).

For the regression to estimate a causal relationship, the identifying assumption is that age at arrival is not correlated with unobservables that vary within the family and also affect income. Unfortunately, we are unable to include other control variables that may bias our estimates due to the limited information in arrival records. The primary concern is birth order: birth order may affect adult outcomes through general birth order effects, and birth order is

[19] We code the bins to the floor of the two ages such that ages zero to one are bin zero, two and three are in bin two, up to bin 18, which includes 18, 19 and 20 year olds. We include 20-year-old arrivals in this bin due to a small number of observations. We show in a robustness check that excluding ages 16 and up from our sample does not change the results. Moreover, using 1-year bins does not change the qualitative conclusions, but increases the noisiness of estimates.

also correlated with age at immigration. It is also possible that parents may time immigration to be optimal for the younger child (for example, immigrate just prior to school entry) such that younger arrivals would have better outcomes due to unobservable parental investment into younger children. Unfortunately, we only observe birth order according to the arrival records, which ignores older siblings who may have stayed in the source country and inherited the farm (Abramitzky et al. 2013). Since birth order is not exactly observed, we do not control for it in our main specification; nevertheless, we show in a robustness check that the results are robust to controlling for birth order.

There are a few threats to the external validity of our estimates, where they may not apply to all child immigrants during the Age of Mass Migration. First, we only identify the age-at-arrival profile with brothers rather than single arrivals; these estimates may differ if there is an extra disruptive effect for older brothers who have to take care of their younger brothers, though the sizes of our effects are likely too large for this to be driving our result. Another possible bias is that we mismeasure the effect of age at arrival for the entire population due to selective return migration (Abramitzky et al. 2014; Ward 2017). Our sample only consists of brothers who remained in the United States; however, if one brother stayed and another returned home, then they would not be included in our sample. If older arrivals were more likely to return home and older arrivals also earned less income, then we would understate the negative effect of age at arrival. A similar story and bias would apply in the case of selective mortality. Finally, our estimates do not apply to all arrivals since we only link those who arrived through Ellis Island, which primarily misses entrants from Asia, Canada or Mexico.[20]

[20] The 1907 Annual Report of the Commissioner General of Immigration lists that about 80 percent of immigrant arrivals were to New York. This percent may have decreased following the immigration quotas when more immigrants entered via land borders.

Therefore, the reader should keep in mind that our results come from brothers who arrived at Ellis Island between 1892 and 1924 and survived until the 1940 Census.

THE EFFECT OF AGE AT ARRIVAL ON ECONOMIC OUTCOMES

We estimate the effect of age at arrival with the brothers fixed effects specification and plot the coefficients in Figure 3. The plotted coefficients estimate the difference in native-immigrant wage gap relative to the native-immigrant wage gap for zero- to one-year-old arrivals. The results show a strong negative slope for age at arrival such that the native-immigrant wage gap was 17 log points (or 15.6 percent) more negative for an 16-year-old arrival compared with the wage gap for an infant arrival. If one assumes that the return to education was 6.5 to 7.9 percent in 1940 (Clay et al. 2016), then having a full childhood in the United States was equivalent to receiving more than two years of schooling.

This result is consistent with the U.S. childhood environment yielding a higher return than the childhood environment in the source country. The pattern could be due to a higher-quality education or health environment in the U.S.; yet at the same time, the quality of the health environment may not have been higher in the United States. For example, Katherine Eriksson and Gregory Niemesh (2016) show that children of black migrants during the Great Migration had higher infant mortality rates relative to children of non-migrants due to poor health conditions in northern cities. However, for states which we have data at the turn of the 20th century, infant mortality rates for foreign-born mothers were less than those for African Americans and similar to the rates in many European sources (Preston and Haines 1991, Tables 2.3 and 3.1). Ultimately, it is unclear whether migrating from Europe led to a lower quality health environment in the United States. Another possibility is that the U.S. environment was not better objectively, but rather that its education system trained individuals specifically for the United States labor market.

Despite the creation of the new linked sample of brothers, the age-at-arrival profile estimated without sibling fixed effects is within the standard errors of the profile estimated with sibling fixed effects.[21] This result suggests that family-invariant unobservables such as parental preferences and education do not strongly bias the age-at-arrival profile estimated with OLS. However, it is also possible that we do not detect a difference in profiles between the methodologies due to errors in the linking process, which would bias the sibling fixed effects result towards the OLS result. This is likely not the case since we find the same result when limiting the sample to higher-quality links. In Online Appendix Table A2, we keep the top 50 percent of links in terms of quality of match based on closeness of name and year of birth and show that even with higher quality links, the sibling fixed effects method estimates a profile that is within the 95 percent confidence interval of the profile estimated with our main sample. We also show that if one links immigrants in a method related to Feigenbaum (2016), then the estimated profile is also within the confidence interval of the profile estimated with our main sample (see Online Appendix E for more detail).

The negatively-sloped age-at-arrival profile appears to be linear; however, standard errors are wide so the profile may not truly be linear. Nevertheless, if one models the age-at-arrival effect to be constant across ages, then the effect of arriving one year later leads to a 1.1 percent more negative wage gap with white natives. A linear age-at-arrival profile would go against expectations in two ways: first, others have found a steepening of the profile around the ages of 8 to 11 due to critical periods of language acquisition or health development (Bleakley and Chin 2004; van den Berg et al. 2014). Our estimate does show a dip in income between age 8 and 14, consistent with a critical period effect, but we cannot statistically detect a break in the slope. We also do not detect a steeper slope for ages under five, which may be surprising given the large returns to improved childhood environment during very young ages

[21] When not using sibling fixed effects, we control for country of birth and years in the United States.

(Almond et al. 2017); yet this may reflect the countervailing effects of a lower quality health environment in the United States relative to Europe.

Not only did arriving at an older age cause the native-immigrant gap to be more negative, but it also caused immigrants to enter lower skilled occupations relative to natives.[22] Table 3 estimates that arriving at an older age increased the likelihood of entering an unskilled job and lowered the likelihood of holding a white-collar job. To provide a summary measure of the effect of age at arrival on occupation, the last two columns estimate the effect on occupational score and show that the native-immigrant gap for 16-year old arrivals was 5 to 12 log points more negative than for infant arrivals.[23] Since the magnitude of the effect on occupation score is less than the magnitude of the effect on income (17 log points), this suggests that age at arrival affected both occupation and income within occupation. Given these effects of age at arrival on income and occupation, it may be that age at arrival also affected other dimensions such as labor supply, but as shown in the Online Appendix, we find no effect of age at arrival on labor force participation or weeks worked (Table A3).[24]

We check the robustness of the age-at-arrival profile when controlling for observed birth order. Recall that observed birth order may not reflect the true birth order since we do not observe family members left behind in the source country. In each specification, the age-at-arrival profile is unchanged, and the birth order effects are statistically insignificant. Another concern is that including immigrants who arrived aged older than 16 may bias results since

[22] The occupational categories are split by occ1950 codes such that professionals (codes starting with 0), managers (1), salesmen (3) and clerical workers (4) are white-collars. Farm owners, tenants and managers (1) are farmers. Craftsmen (5) are skilled workers. Operatives (6), low-skilled service workers (7), farm laborers (8) and laborers (9) are unskilled workers.

[23] We show results based on a created occupational score that reflects mean earnings by occupation and source country in the 1940 Census. Creation of this score is discussed in Appendix D and is largely based on Collins and Wanamaker (2017). We also show the age-at-arrival effect for the 1950 IPUMS variable occscore, the main one presented by Abramitzky et al. (2014). Results across scores differ because the 1940 score reflects a less compressed wage distribution and more adequately reflects immigrant earnings by occupation.

[24] We also show in Appendix Table A4 that age at arrival has no qualitative effect on home ownership and a small positive effect on living in a more urbanized location.

older arrivals may have decided on their own to immigrate while younger arrivals had less choice. To account for this, we re-estimate the age-at-arrival effect when dropping those who arrived older than age 15, and find no difference in the estimated income profile. Finally, we test for the robustness of our linking process to the Americanization process by relinking our data without Americanizing names. The estimated effects without the Americanization process have less precision (see Table A7), but reaffirm the negative effect of age at arrival on adult labor market outcomes. These robustness checks are shown in Online Appendix Tables A5-A7.

POTENTIAL MECHANISMS FOR THE AGE-AT-ARRIVAL EFFECT

Arriving at an older age negatively affected labor market outcomes later in life, but through which channels? The potential mechanisms are numerous, but to name a few: the system of education changed across countries; parental resources may have improved after the move and thus investment into the child also improved; and younger arrivals may have socially adapted at a quicker rate. In this section, we estimate how age at arrival was related to these potential mechanisms.

Total Years of Education

First, we test whether age at arrival affected the total years of educational attainment. When we run the same age-at-arrival regression with native-immigrant gap in education as the dependent variable, we find that the gap for older arrivals was larger than the gap for younger arrivals by one year (see Figure 4). The education profile looks similar to the income profile in that they are both negatively sloped, suggesting that education could be a primary channel for the age-at-arrival effect on income. However, in contrast with the income profile, the education profile becomes flat after age 15. The flattening of the education profile reflects that most immigrants left school prior to age 16 whether in the United States or in the source country.

The effect of age at arrival on educational attainment may not have been the same across all source countries. Many Europeans arrived from countries with relatively robust education systems; for example, Germany had compulsory schooling laws dating back to 1717 and one of the best educational systems in Europe (Lindert 2004). On the other hand, many Southern and European sources had less robust education systems; for instance, Italy and Greece had lower enrollment rates for 5 to 14 year olds compared with the enrollment rates in Norway, Ireland and the Netherlands (Bandiera et al. 2017, Figure 1). A reasonable hypothesis is that the effect of age at arrival on education is smaller for higher income countries in Northern and Western Europe compared with lower income countries in Southern and Eastern Europe.[25]

The profiles for income and education separated by New and Old sources are shown in Figure 5. On the one hand, the *income* profiles are similar across sources, where the negative effect of age at arrival is statistically indistinguishable across the two regions. At the same time, the *education* profiles were quite distinct: New sources had a steep profile where older arrivals received 1.7 fewer years of education. On the other hand, the education profile is completely flat for Old sources, showing no penalty for arriving at an older age. The flat education profile is consistent with the relatively high quality educational institutions in Northern and Western Europe.

The difference in education and income profiles across sources reveal a puzzle: why did older arrivals from Northern and Western Europe earn less despite receiving the same total years of education? One reason may be that foreign education did not yield a high return in the United States labor market, and thus extra schooling acquired in the source country at older ages did not boost wages. Besides education, older arrivals also had more potential labor

[25] We define Old source countries, or those from Northern and Western Europe, to be Denmark, Finland, Norway, Sweden, England, Scotland, Ireland, Belgium, France, Netherlands, Switzerland and Germany. We define New source countries, or those from Southern and Eastern Europe, to be Greece, Italy, Portugal, Spain, Austria, Czechoslovakia, Hungary, Poland, Romania, Yugoslavia, Lithuania, and Russia.

market experience in the foreign source country, and foreign experience may not have had a high value in the United States. On the other hand, it may be that the age-at-arrival effect operated through channels other than education or experience, such as social assimilation. We now turn to other possible explanations for the effect of age at arrival.

Return to education and experience separated into domestic and foreign component

Since we observe age of arrival and completed years of education in our dataset, we can test – after making a few assumptions – whether foreign education and experience yielded a small return in the United States. Following Rachel Friedberg (2000), we assume that individuals entered school at age 6, and then allocate the total amount of schooling to the United States or source country based on the age of arrival.[26] Given this assumption, it is straightforward to further separate potential experience into foreign or United States components. See the descriptive statistics in Table 2 for how foreign education and experience increase at higher ages of arrival, while years of US education decreases.

To measure the wage return to education and experience, we use an augmented Mincer equation and regress log income on years of US education, years of foreign education, potential years of US experience and potential years of foreign experience:

$$y_{ih} = \beta_0 + \beta_1 ForEduc_{ih} + \beta_2 USEduc_{ih} + f(ForExp_{ih}) + g(USExp_{ih}) + \phi_h + \nu_{ih} \quad (4)$$

We are interested in whether the return to foreign education was less than the return to United States education, that is if $\beta_1 < \beta_2$. We are also interested in whether the return to experience gained abroad yielded a different return from experience gained in the United States; we model experience as a quadratic. Note that we always include household fixed effects, eliminating

[26] Let Total Education = Foreign Education + US Education, and Experience = Foreign Experience + US Experience. Further assume that Total Experience = Age – Education – 6. To separate total education and total experience into US and foreign components, we assume that individuals attended schooling continuously. That is, let Foreign Education = 0 if Age at Arrival is less than six, and min(Age at Arrival – 6, Education) if greater than or equal to six. Also let Foreign Experience = 0 if Age at Arrival is less than six, and max(Age at Arrival – Foreign Education – 6, 0) if greater than or equal to six.

household-invariant unobservables that are correlated with years of education, experience and income.

The results are presented in Table 4. When pooling New and Old source countries together, the return to being educated in the United States is estimated at 5.3 percent, which is less than the return for native-born workers (6.5-7.9 percent). A different return to US education for foreign-born workers relative to native workers has been found elsewhere in the literature (for example, Chiswick 1978, Baker and Benjamin 1994), and may reflect discrimination against foreign-born workers or the quality of 'years' of education. The return to education earned in the foreign country was even lower at 4.4 percent, although the 0.9 percentage point difference from the return to US education is not statistically significant.

While there is not strong evidence that the location of schooling mattered, there is evidence that where one gained labor market experience mattered. Table 4 shows that the return to potential foreign experience was not statistically distinguishable from zero, although this is not very precisely estimated, whereas U.S.-experience was positively rewarded. A small return for foreign experience may reflect that immigrants entered different industries and occupations after the move to the United States. Immigrants often came from more agrarian countries in Southern and Eastern Europe, and their European experience appears to have had little value in the U.S. (Hatton and Williamson 1998, Ch. 2). However, this is part speculation, unfortunately, we cannot observe the job history of migrants to determine the value of the type of foreign experience. Data that does observe young arrivals' occupations show that most teenagers reported holding either no job or an unskilled job.[27] If skills learned in these jobs in

[27] We can observe occupations using data from a random sample of ship arrivals to Ellis Island between 1917 and 1924, where the sample is limited to 12-17 year old males (Ward 2017). According to this data, most arrivals reported having no occupation (40.6%), with laborer (26.9%), farm laborer (7.6%) and farmer (3.4%) being the top three reported occupations.

the source country did not transfer to the United States labor market, then older arrivals would be penalized for staying longer in the source country.

The return to human capital may have differed between sources closer in development to the United States and sources further behind. We test for this in Columns 2 and 3 after splitting the sample into New and Old sources. The results show that the return to foreign education was indeed higher for Old sources at 6.5 percent than for New sources at 4.1 percent – a statistically significant difference, as shown in the fully interacted model in the last column. This result implies that education acquired in Northern and Western Europe more easily transferred to the United States, perhaps because it was of higher quality or the economies were closer in industrial structure. Yet at the same time, immigrants from Northern and Western Europe did not earn much return to foreign experience, and the return to foreign experience was similar across regions in Europe.

The results from these wage regressions point to foreign experience and its lack of return as a potential mechanism for a downward sloping age-at-arrival and income profile. For example, 16-20 year old arrivals had on average 5.7 years of potential foreign experience (see Table 2), but this human capital yielded little return in the United States labor market. This result may partially explain the result in Figure 4 that Old source immigrants had a negative age-at-arrival effect on income despite a lack of effect on education; however, other channels, such as the extent of social assimilation, may have also been important.

Social Assimilation: Intermarriage and Geography

Besides the traditional measures of human capital such as education and experience, age at arrival may have affected adult earnings through a different channel: social assimilation. This could be due to higher levels of English fluency for younger arrivals or because younger arrivals appeared more "American" and thus experienced less discrimination. We measure the effect of age at arrival on social assimilation by changing the dependent variable to outcomes related to residential segregation and marriage; specifically, the likelihood of living near native-born households and the likelihood of marrying a native-born spouse.[28] English fluency is not observed in the 1940 Census, but we discuss the potential effect of English fluency in the next section.

Age at arrival had a strong effect on intermarriage, as shown in Table 5. A 16-year-old arrival was 37.2 percentage points less likely to marry a native-born spouse than a 1-year-old arrival, a very large effect given that 69 percent of infant arrivals married a native-born spouse. Although we do not estimate the return to intermarriage, others have shown with late-20th century data that intermarriage is associated with higher earnings (for example, Meng and Gregory 2005); therefore, it may be a mechanism for the downward sloping age-at-arrival income profile.

While there is a large effect of age at arrival on intermarriage, there is little evidence that younger arrivals were more spatially integrated with native-born household heads. Table 5 also shows the effect of age at arrival on the fraction of native-born household heads in the county of residence. Note that we use fraction of native-born *household heads* rather than fraction of all *individuals* in the county to ensure that native-born second-generation children in the home do not influence the estimate (that is, "childrearing" bias). We find that age at arrival had no effect on living in a county with more native-born household heads. However, this county-level measure may mask segregation within a county. Given that we have the entire 1940 Census, we can further narrow the geography from county to the immediate neighborhood, as proxied by the fraction of native-born household heads on the same census

[28] When estimating these equations, we do not do the two-step process of predicting residuals for immigrants based on the native life-cycle profile and then regressing these residuals on age at arrival. This is because there may not be a well-defined relationship between native's age and marrying a native-born spouse or having a native-born neighbor. Therefore, we present results based on the simple age-at-arrival effects on the levels of having a native-born spouse or native-born neighbor. However, the results are qualitatively the same if we use the residuals after predicting the lifecycle profile with natives.

page (Logan and Parman 2017). However, even using this measure suggests that age at arrival had little impact on spatial assimilation for our dataset of brothers.

Other unobserved but potential channels: English fluency and parental investment

An important indicator of social assimilation and human capital that is not included in the 1940 Census is English proficiency. It is certain that English proficiency was lower for older arrivals from non-English-speaking countries, given the robust evidence for the critical period of language acquisition from Hoyt Bleakley and Aimee Chin (2004, 2010). An indirect way to uncover the effect of English skills on adult outcomes is to test whether the age-at-arrival income profile steepens at older ages for non-English-speaking sources relative to English-speaking sources (see Bleakley and Chin (2004) for a further discussion). However, we do not find that the age-at-arrival income profile for non-English-speaking sources steepens relative to the profile for English-speaking sources after the critical period of language acquisition ends, which is consistent with the argument that acquiring English fluency was relatively unimportant for improving one's occupation in the early 20th century compared with the late 20th century (Ward 2018). Yet standard errors are wide when splitting the sample into English-speaking and non-English-speaking sources (see Figure A1). Therefore, while a lower level of English fluency for older arrivals may have contributed to the negatively sloped age-at-arrival and income profile, this mechanism cannot be conclusively confirmed.

Finally, it is also possible that family real wages increased during the move; if so, then younger arrivals may have benefitted by receiving more parental inputs during critical stages of development. An increase to family income almost certainly occurred after migration: Abramitzky et al. (2012) estimate the return to immigration at 70 percent from Norway to the United States, and the return was probably even larger for immigrants from Southern and Eastern European sources. Jeffrey G. Williamson (1995) estimates that United States real wages were 67 percent higher than in Great Britain in 1905, and over three times higher than real wages in Italy. Therefore, an additional mechanism for the age-at-arrival profile is likely that household investment into children increased and the effectiveness of this investment was higher at younger ages, but we do not observe the change in real income before and after the move.

Overall, we interpret the age-at-arrival effect as the effect of changing various environmental attributes at critical stages of childhood development. On average, older arrivals were penalized relative to younger arrivals because they had fewer total years of education, less valuable labor market experience, and were less socially assimilated. While we cannot precisely show which mechanism was most important, we do show that the effect of age at arrival was large enough to erase the negative native-immigrant wage gap that older arrivals experienced.

CONCLUSION

Using a new dataset of brothers linked from Ellis Island records to the 1940 Census, we show that there was a large wage and educational return to arriving at a younger age in the United States. Spending one's childhood in the United States rather than in Europe significantly improved immigrants' long-run economic outcomes. The variation in immigrant outcomes based on their age at arrival complements prior research that finds that occupational-based earning differentials between migrants and natives were fixed throughout the life cycle. The difference in results suggests that while human capital acquired during childhood led to a large occupational return, the human capital acquired during adulthood after arrival did not (Abramitzky et al. 2014; Ward 2018).

While the results suggest that the United States childhood environment was advantageous relative to that of Europe, the results are limited by the lack of data on childhood location in the United States. In particular, one question that is left unanswered is the effect of childhood environment when living in or outside an ethnic enclave. Given the intergenerational literature's result that the source country's position in the occupational distribution persists across generations of immigrants despite our result of childhood environment having a large effect, persistence of skill levels across generations must be due to other factors such as immigrants sorting into different quality childhood environments in the United States (Abramitzky et al. 2014, Borjas 1994).

Finally, the results show that immigrants' position in the skill distribution was not fixed by inherited or genetic factors, as many nativists at the time claimed. For example, Francis Walker, one-time president of the American Economic Association, charged that New source immigrants had "none of the inherited instincts and tendencies which made it comparatively easy to deal with the immigration of the olden time. They are beaten men from beaten races; representing the worst failures in the struggle for existences. Centuries are against them" (Walker 1896). This pessimistic view of immigrants has been reprised throughout time, right up to today's wave of nativism against Muslims and Hispanics (Higham 1955; Huntington 2004). Our research shows that these "beaten men from beaten races" in the past did remarkably well with a change to their location at young ages, reaffirming the paramount importance of childhood environment for long-run outcomes.

REFERENCES

Abramitzky, Ran, and Leah Boustan. “Immigration in American economic history.” *Journal of Economic Literature* 55, no. 4 (2017): 1311-45.

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. “Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration.” *American Economic Review* 102, no. 5 (2012): 1832-56.

———. “Have the poor always been less likely to migrate? Evidence from inheritance practices during the Age of Mass Migration.” *Journal of Development Economics* 102 (2013): 2-14.

———. “A nation of immigrants: Assimilation and economic outcomes in the age of mass migration.” *Journal of Political Economy* 122, no. 3 (2014): 467-506.

———. “Cultural assimilation during the age of mass migration.” NBER Working Paper No. 22381, Cambridge, MA, July 2016.

Almond, Douglas, Janet Currie, and Valentina Duque. “Childhood Circumstances and Adult Outcomes: Act II.” NBER Working Paper No. 23017. Cambridge, MA, January 2017.

Åslund, Olof, Anders Böhlmark, and Oskar Nordström Skans. “Childhood and Family Experiences and the Social Integration of Young Migrants.” *Labour Economics* 35 (2015): 135-144.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. “How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth.” NBER Working Paper No. 24019. Cambridge, MA, November 2017.

Baines, Dudley. *Emigration from Europe 1815-1930*. Vol. 11 of New Studies in Economic and Social History. Cambridge University Press, 1995.

Baker, Michael, and Dwayne Benjamin. "The performance of immigrants in the Canadian labor market." *Journal of Labor Economics* 12, no. 3 (1994): 369-405.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo. “The Making of Modern America: Migratory Flows in the Age of Mass Migration.” *Journal of Development Economics* 102 (2013): 23-47.

Bandiera, Oriana, Myra Mohnen, Imran Rasul, and Martina Viarengo. *Nation-Building Through Compulsory Schooling During the Age of Mass Migration*. No. 057. Suntory and Toyota International Centres for Economics and Related Disciplines, LSE, 2016.

Bleakley, Hoyt, and Aimee Chin. “Language Skills and Earnings: Evidence from Childhood Immigrants.” *Review of Economics and Statistics* 86.2 (2004): 481-496.

———. “Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants.” *American Economic Journal. Applied Economics* 2.1 (2010): 165.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique. “The Economic Payoff of Name Americanization.” *Journal of Labor Economics* 35, no. 4 (2017): 1089-1116.

Böhlmark, Anders. “Age at Immigration and School Performance: A Siblings Analysis Using Swedish Register Data.” *Labour Economics* 15, no. 6 (2008): 1366-1387.

Borjas, George J. “Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants.” *Journal of Labor Economics* 3, no. 4 (1985): 463-489.

———. “Self-Selection and the Earnings of Immigrants.” *The American Economic Review* (1987): 531-553.

———. “Long-run Convergence of Ethnic Skill Differentials: The Children and Grandchildren of the Great Migration.” *ILR Review* 47, no. 4 (1994): 553-573.

Card, David, John DiNardo, and Eugena Estes. “The More Things Change: Immigrants and the Children of Immigrants in the 1940s, the 1970s, and the 1990s.” *Issues in the Economics of Immigration*. University of Chicago Press, 2000. 227-270.

Carneiro, Pedro Manuel, Sokbae Lee, and Hugo Reis. “Please call me John: name choice and the assimilation of immigrants in the United States, 1900-1930.” Mimeo (2015).

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. “The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment.” *The American Economic Review* 106, no. 4 (2016): 855-902.

Chetty, Raj, and Nathaniel Hendren. “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects.” NBER Working Paper No. 23001. Cambridge, MA, May 2017a.

———. “The Impacts of Neighborhoods on Intergenerational Mobility II: County-level Estimates.” NBER Working Paper No. 23002. Cambridge, MA, December 2017b.

Chiswick, Barry R. “The effect of Americanization on the Earnings of Foreign-Born Men.” *Journal of Political Economy* 86, no. 5 (1978): 897-921.

Clarke, Andrew. “Age at Immigration and the Educational Attainment of Foreign-born Children in the United States: The Confounding Effects of Parental Education.” *International Migration Review* (2016).

Cohn, Raymond L. *Mass migration under sail: European immigration to the antebellum United States*. Cambridge: Cambridge University Press, 2009.

Collins, William J., and Marianne H. Wanamaker. “Selection and economic gains in the great migration of African Americans: New Evidence from Linked Census Data.” *American Economic Journal: Applied Economics* 6, no. 1 (2014): 220-252.

———. “Up from Slavery? African American Intergenerational Economic Mobility Since 1880.” NBER Working Paper No. 23395. Cambridge, MA, May 2017.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. “Interpreting the Evidence on Life Cycle Skill Formation.” *Handbook of the Economics of Education* 1 (2006): 697-812.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. “Is the Melting Pot Still Hot? Explaining the Resurgence of Immigrant Segregation.” *The Review of Economics and Statistics* 90, no. 3 (2008): 478-497.

Eriksson, Katherine, and Gregory Niemesh. “Death in the Promised Land: The Great Migration and Black Infant Mortality.” Unpublished Manuscript (2016).

Feigenbaum, James J. “Automated Census Record Linking: A Machine Learning Approach.” Unpublished Manuscript (2016).

Friedberg, Rachel M. “The Labor Market Assimilation of Immigrants in the United States: The Role of Age at Arrival.” Unpublished Manuscript (1992).

———. “You Can't Take It With You? Immigrant Assimilation and the Portability of Human Capital.” *Journal of Labor Economics* 18, no. 2 (2000): 221-251

Greenwood, Michael J. "Modeling the age and age composition of late 19th century US immigrants from Europe." *Explorations in Economic History* 44, no. 2 (2007): 255-269.

Gould, John D. "European International Emigration: The Role of Diffusion and Feedback." *Journal of European Economic History* 9, no. 2 (1980): 267-315.

Hatton, Timothy J. “The Immigrant Assimilation Puzzle in Late Nineteenth-Century America.” *The Journal of Economic History* 57, no. 1 (1997): 34-62.

Hatton, Timothy J., and Jeffrey G. Williamson. *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press on Demand, 1998.

Higham, John. *Strangers in the Land: Patterns of American Nativism, 1860-1925*. Rutgers University Press, 1955.

Huntington, Samuel P. *Who Are We?: The Challenges to America's National Identity*. Simon and Schuster, 2004.

Hutchinson, Edward P. “Notes on Immigration Statistics of the United States.” *Journal of the American Statistical Association*53, no. 284 (1958): 963-1025.

Lindert, Peter H. *Growing Public: Volume 1, The Story: Social Spending and Economic Growth Since the Eighteenth Century*. Vol. 1. Cambridge University Press, 2004.

Massey, Catherine G. "Playing with matches: An assessment of accuracy in linked historical data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50, no. 3 (2017): 129-143.

Meng, Xin, and Robert G. Gregory. "Intermarriage and the Economic Assimilation of Immigrants." *Journal of Labor Economics* 23, no. 1 (2005): 135-174.

Minns, Chris. "Income, Cohort Effects, and Occupational Mobility: A New Look At Immigration to the United States at the Turn of the 20th century." *Explorations in Economic History* 37, no. 4 (2000): 326-350.

Logan, Trevon D., and John M. Parman. "The National Rise in Residential Segregation." *The Journal of Economic History* 77, no. 1 (2017): 127-170.

Preston, Samuel H., and Michael Haines. *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Cambridge: National Bureau of Economic Research, 1991.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 7.0* [dataset]. Minneapolis: University of Minnesota, 2017. http://doi.org/10.18128/D010.V7.0.

Schaafsma, Joseph, and Arthur Sweetman. "Immigrant Earnings: Age at Immigration Matters." *Canadian Journal of Economics* 34, no. 4 (2001): 1066-1099.

Schoellman, Todd. "Early Childhood Human Capital and Development." *American Economic Journal: Macroeconomics* 8, no. 3 (2016): 145-74.

Spitzer, Yannay, and Ariell Zimran. "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925." *Mimeo* (2017).

U.S. Immigration Commission "The Children of Immigrants in Schools, Vol. 1." *Reports of the Immigration Commission, Sixty-First Congress (3rd Session)*. Washington DC: US Government Printing Office, 1910.

Van den Berg, Gerard J., et al. "Critical Periods During Childhood and Adolescence." *Journal of the European Economic Association* 12, no. 6 (2014): 1521-1557.

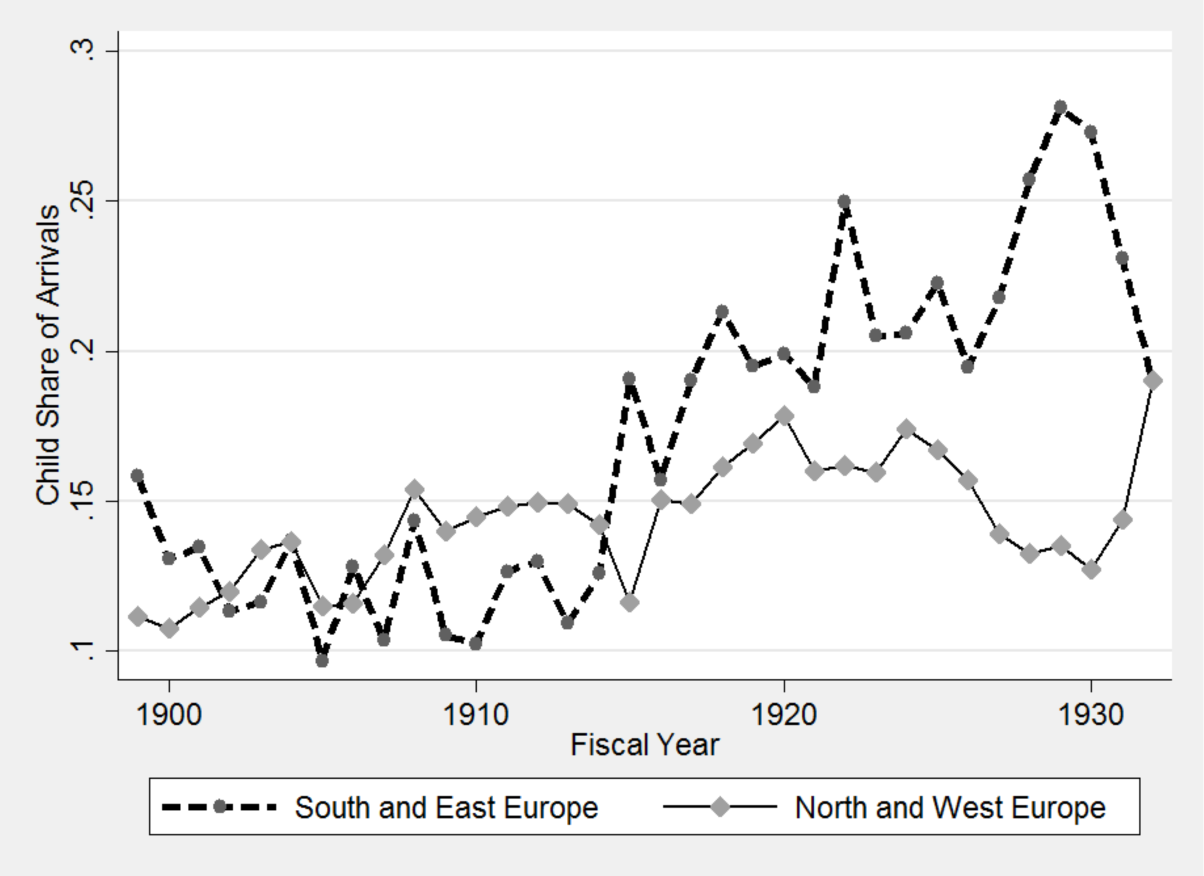
Walker, Francis A. "Restriction of Immigration." *Atlantic Monthly* 77, no. 464 (1896): 822-829.

Ward, Zachary. "Birds of Passage: Return Migration, Self-Selection and Immigration Quotas." *Explorations in Economic History* 64 (2017): 37-52.

———. *Have Language Skills Always Been So Valuable? The Low Return to English Fluency During the Age of Mass Migration.* 2018

Williamson, Jeffrey G. "The Evolution of Global Labor Markets Since 1830: Background Evidence and Hypotheses." *Explorations in Economic History.* 32, no. 2 (1995): 141-196.

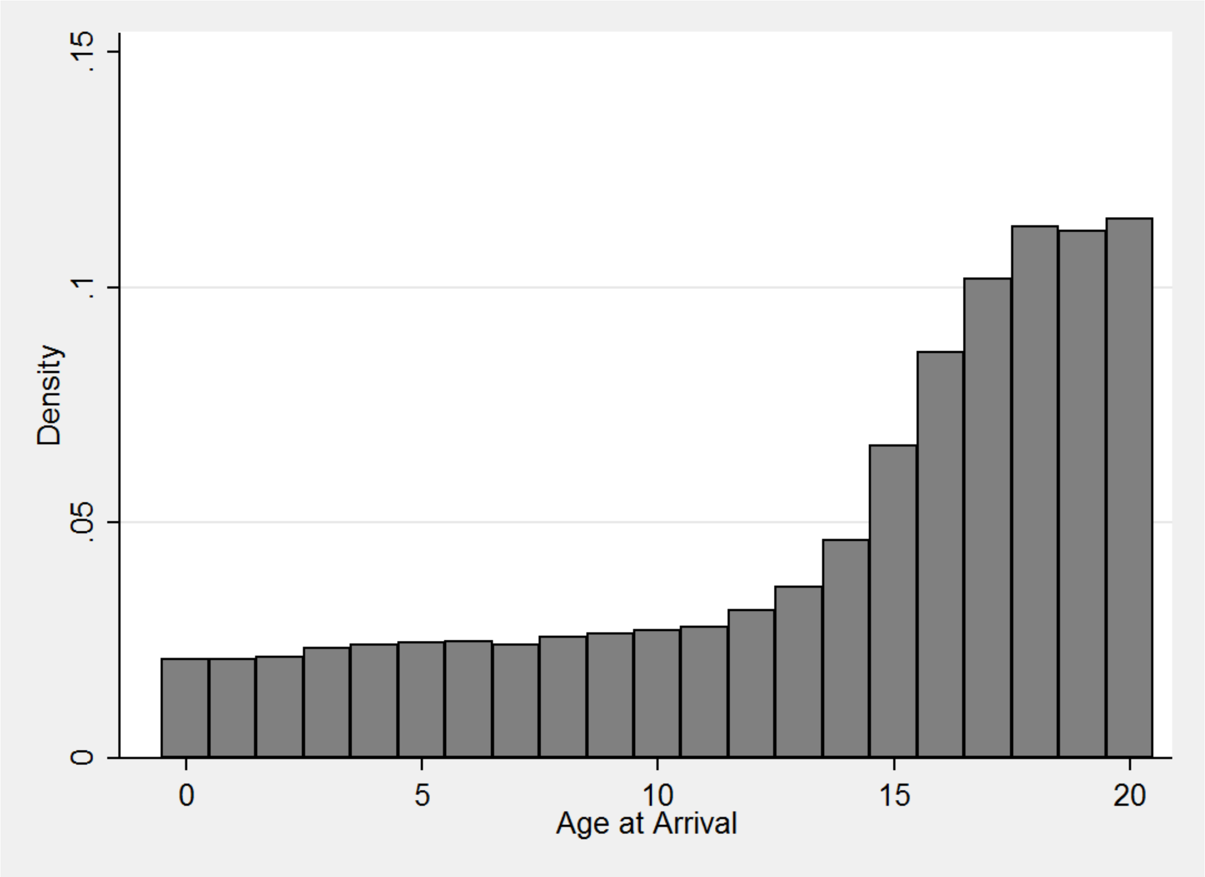
Figure 1. Child Share of Immigrant Inflows to the U.S., 1899-1932



Notes: Fiscal years are between July 1st and June 30th. Child arrivals are those under the age of 14 between 1899 and 1916 and under the age of 16 between 1917 and 1932. See footnote 4 for definition of arrival.

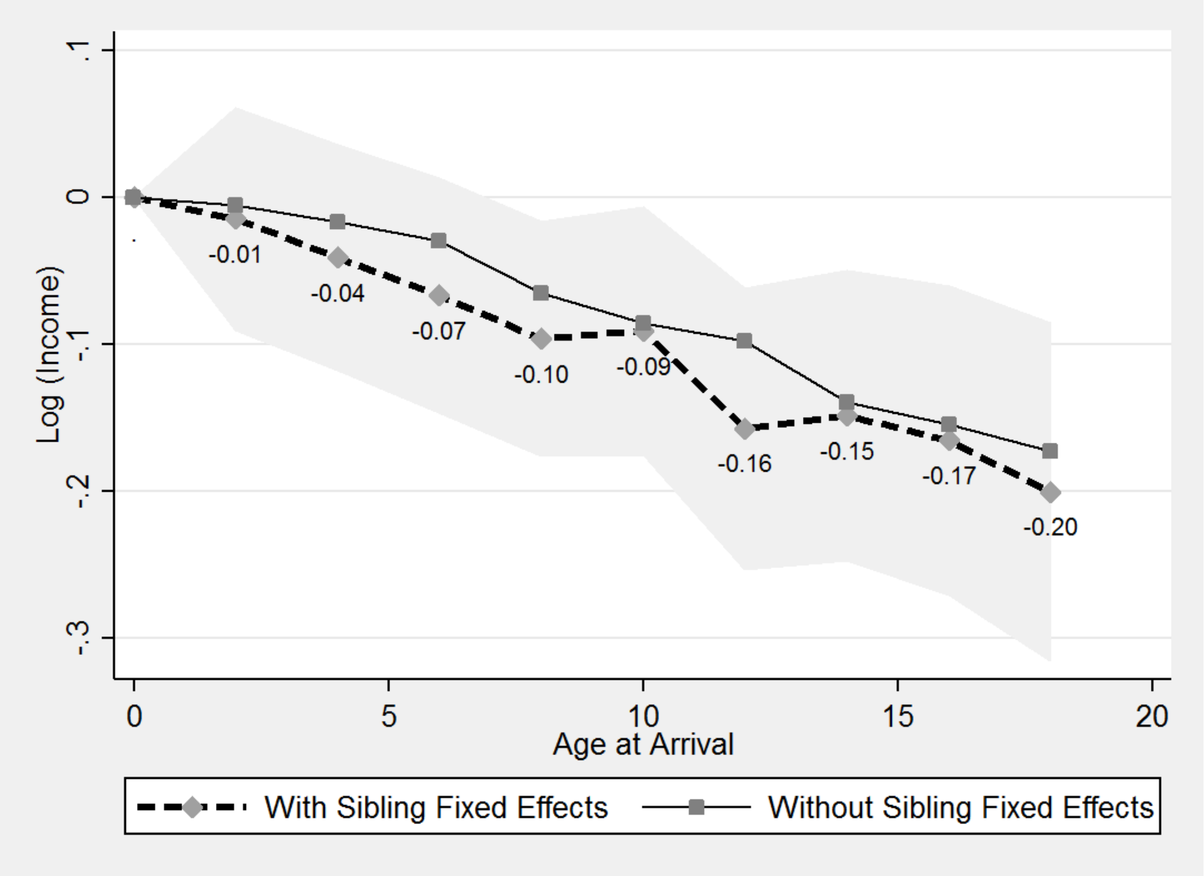
Sources: Annual Reports of the Commissioner General of Immigration, 1899-1932.

Figure 2. Distribution of immigrant age at arrival in the 1900-1930 U.S. Censuses



Notes: The sample is limited to those who arrived between 1899 and 1930 to match with Figure 1. Distribution is estimated after applying the person weight available from IPUMS. *Sources:* 1% samples of the 1900-1920 Censuses, 5% sample of the 1930 Census (Ruggles et al. 2017)

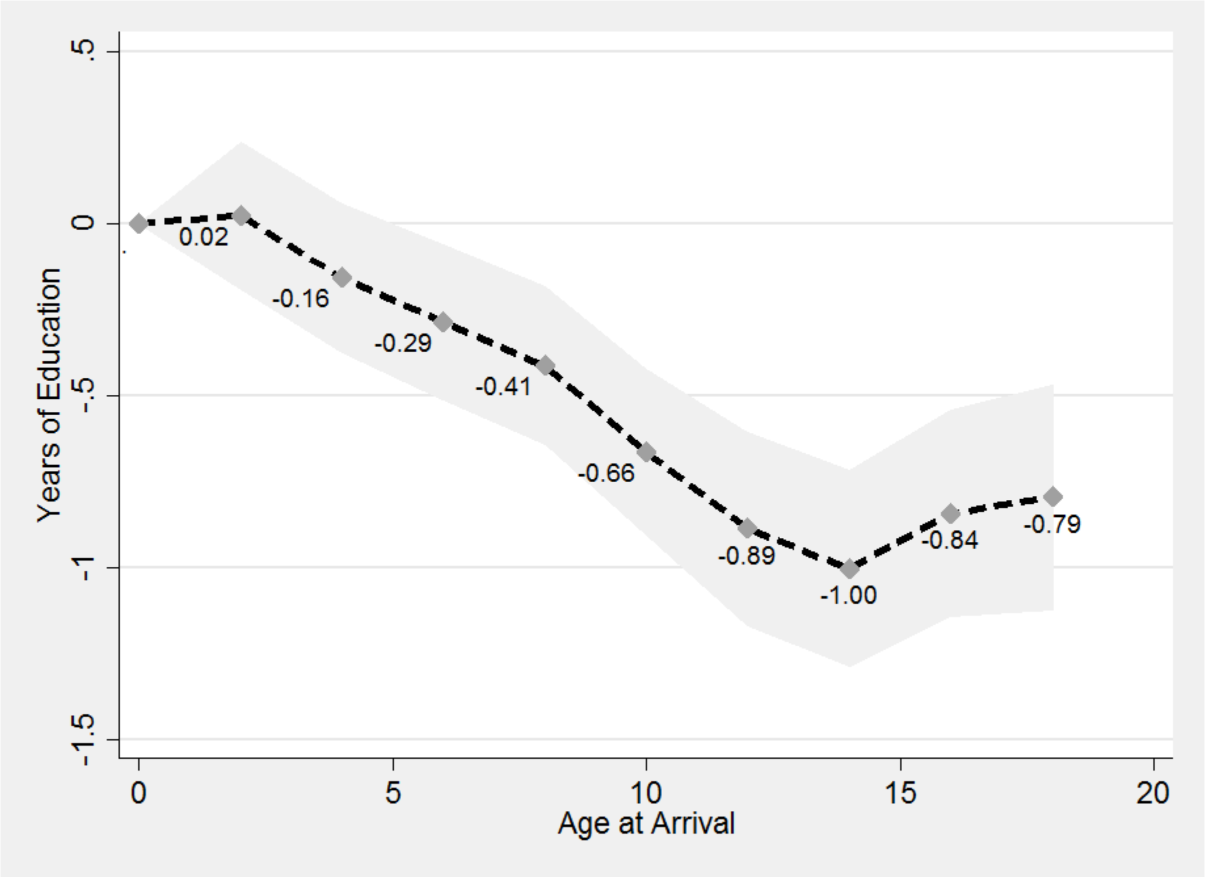
Figure 3. The negative effect of age at arrival on the native-immigrant gap in wage income in 1940



Notes: The dependent variable is age-adjusted gap in log wage income between immigrants and natives. Self-employed workers are dropped. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval when using sibling fixed effects. Standard errors are clustered at the household level.

Sources: Sample of brothers linked from Ellis Island records to the 1940 Census.

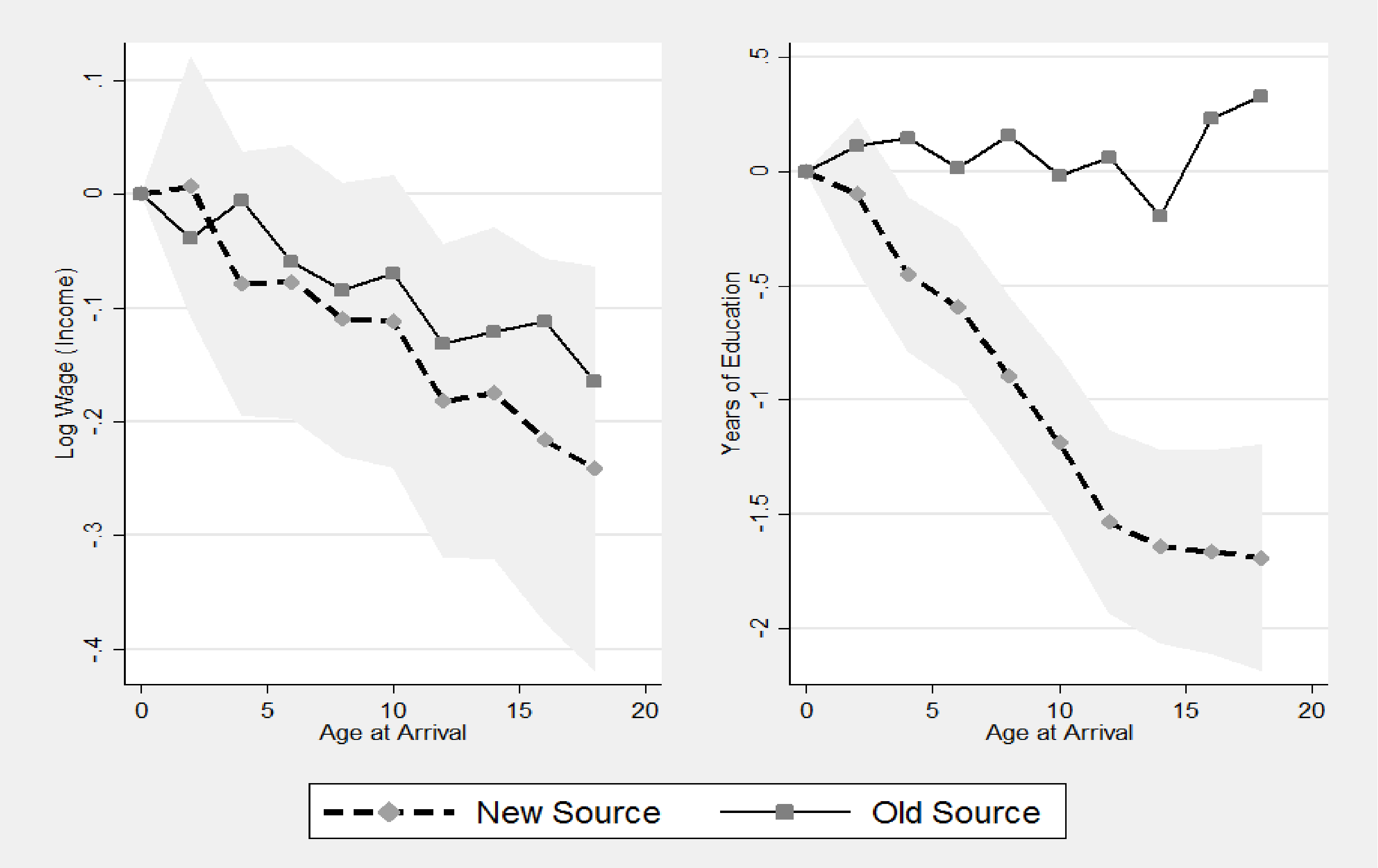
Figure 4. The negative effect of age at arrival on the native-immigrant gap in years of education in 1940



Notes: The dependent variable is age-adjusted gap in years of education between immigrants and natives. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval. Standard errors are clustered at the household level.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Figure 5. The age-at-arrival profiles were differently sloped across New and Old sources



Notes: The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval for the New source group. Standard errors are clustered at the household level. New source countries are in Southern and Western Europe and Old source countries are in Northern and Western Europe.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Table 1. Birthplace composition in Ellis Island Data linked to the 1940 Census

Country of birth	Brothers at Arrival	Linked to Census	2+ Brothers Linked	2+ Brothers Link Rate
Old source countries:	96,790	34,877	21,421	22.1
Denmark	3,876	1,558	961	24.8
Finland	3,807	677	269	7.1
Norway	7,468	2,363	1,306	17.5
Sweden	8,621	3,226	1,816	21.1
England	23,695	9,601	6,091	25.7
Scotland	9,739	5,109	3,512	36.1
Ireland	8,154	4,433	3,130	38.4
Belgium	3,694	653	272	7.4
France	5,668	843	332	5.9
Netherlands	11,752	2,680	1,369	11.6
Switzerland	2,848	898	528	18.5
Germany	7,468	2,836	1,835	24.6
New source countries:	275,205	65,952	30,799	11.2
Greece	9,411	1,139	298	3.2
Italy	150,476	49,183	24,224	16.1
Portugal	2,376	740	445	18.7
Spain	2,812	590	291	10.3
Austria	14,789	2,251	895	6.1
Czechoslovakia	9,835	2,309	1,157	11.8
Hungary	8,736	1,204	450	5.2
Poland	16,717	2,204	794	4.7
Romania	6,757	665	197	2.9
Yugoslavia	3,309	572	205	6.2
Lithuania	2,759	122	22	0.8
Russia	47,228	4,973	1,821	3.9
Other (Asia, Canada, Mexico)	25,142	2,176	909	3.6
Total	397,137	103,005	53,129	13.4

Notes: The empirical strategy uses sibling fixed effects, so we only keep sets of brothers where at least two are successfully linked (2+ Brother Linked column).

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Table 2. Descriptive Statistics of Linked Sample of Brothers

		Age at Arrival			
	Native-Born	0-5	6-10	11-15	16-20
Age	35.76	34.80	39.50	43.52	48.58
	(14.58)	(7.815)	(7.752)	(7.815)	(7.161)
Years in the United States		31.91	31.47	30.76	30.85
		(7.654)	(7.652)	(7.755)	(6.983)
Southern and Eastern European		0.544	0.601	0.625	0.623
(New source)		(0.498)	(0.490)	(0.484)	(0.485)
Years of US education	9.286	8.169	5.615	1.565	0.165
	(3.280)	(3.392)	(3.616)	(2.378)	(0.789)
Years of foreign education		0	1.861	5.157	5.987
		(0)	(1.389)	(2.327)	(3.369)
Potential US labor	20.49	20.60	25.83	29.17	30.66
market experience	(15.50)	(9.515)	(9.363)	(8.405)	(7.078)
Potential foreign labor		0	0.163	1.605	5.741
market experience		(0)	(0.652)	(2.418)	(3.570)
		Age-Adj. Difference from White Native Born			
Log (Income), if wage	6.712	0.0945	0.0344	-0.00257	-0.0792
Worker	(0.960)	(0.686)	(0.697)	(0.683)	(0.722)
Log (Income), if wage	6.905	-0.0428	-0.108	-0.153	-0.225
worker and urban	(0.878)	(0.677)	(0.687)	(0.671)	(0.704)
Self employed	0.226	-0.0587	-0.0585	-0.0631	-0.0870
	(0.418)	(0.359)	(0.389)	(0.410)	(0.418)
White collar	0.291	-0.0437	-0.0472	-0.0786	-0.107
	(0.454)	(0.444)	(0.447)	(0.430)	(0.408)
Farmer	0.129	-0.0832	-0.102	-0.109	-0.123
	(0.336)	(0.180)	(0.182)	(0.210)	(0.231)
Unskilled	0.418	0.0822	0.105	0.137	0.187
	(0.493)	(0.501)	(0.502)	(0.502)	(0.502)
Semi-skilled	0.162	0.0447	0.0444	0.0504	0.0432
	(0.368)	(0.411)	(0.418)	(0.426)	(0.423)
Urban area	0.532	0.255	0.256	0.261	0.255
	(0.499)	(0.395)	(0.396)	(0.399)	(0.413)
Native-born spouse,	0.960	-0.317	-0.391	-0.483	-0.628
if married	(0.196)	(0.477)	(0.494)	(0.498)	(0.468)
Fraction of HH in county	0.838	-0.145	-0.148	-0.148	-0.149
which are native born	(0.162)	(0.141)	(0.138)	(0.137)	(0.137)
Observations	372,870	14,979	15,693	11,367	11,090

Notes: Native born are white males restricted to the same birth cohorts are the immigrant sample. Total education and potential labor market experience is split under the assumption that individuals enter school at age 6 and continuously attend for their full years of schooling (see footnote 18). The outcomes are age-adjusted residuals after predicting life-cycle variation with the native born (see Equation (1) in text). There is missing information for some of these variables. Specifically, 17,152 do not have a positive log income, 1,538 have missing education, 5,226 have missing self-employment, and 3,225 have a blank occupation. HH stands for household heads.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census and 1% sample of 1940 Census.

Table 3. Effect of Age at Arrival on Occupations

	White-Col.	Skilled	Farmer	Unskilled	Log (Occ. Score) 1940 Census	Log (Occ. Score) 1950 Occscore
Age at arrival						
2 to 3	-0.0198	0.00332	-0.0155**	0.0320	-0.0275**	-0.0103
	(0.0174)	(0.0170)	(0.00679)	(0.0196)	(0.0118)	(0.0137)
4 to 5	-0.00308	2.55e-05	-0.0176**	0.0206	-0.0478***	-0.0108
	(0.0175)	(0.0167)	(0.00697)	(0.0196)	(0.0120)	(0.0139)
6 to 7	-0.00276	-0.00804	-0.0317***	0.0425**	-0.0681***	-0.0140
	(0.0180)	(0.0174)	(0.00726)	(0.0203)	(0.0123)	(0.0142)
8 to 9	-0.0220	0.00812	-0.0332***	0.0472**	-0.0790***	-0.0175
	(0.0181)	(0.0175)	(0.00740)	(0.0203)	(0.0123)	(0.0144)
10 to 11	-0.0339*	-0.00117	-0.0379***	0.0730***	-0.0895***	-0.0288*
	(0.0190)	(0.0184)	(0.00776)	(0.0215)	(0.0131)	(0.0151)
12 to 13	-0.0332	0.000934	-0.0430***	0.0754***	-0.112***	-0.0383**
	(0.0211)	(0.0207)	(0.00873)	(0.0242)	(0.0144)	(0.0168)
14 to 15	-0.0720***	0.0184	-0.0428***	0.0963***	-0.113***	-0.0453***
	(0.0216)	(0.0213)	(0.00930)	(0.0246)	(0.0146)	(0.0173)
16 to 17	-0.0558**	0.00519	-0.0597***	0.110***	-0.122***	-0.0470**
	(0.0227)	(0.0227)	(0.00976)	(0.0260)	(0.0154)	(0.0183)
18 to 20	-0.0590**	0.0164	-0.0542***	0.0968***	-0.118***	-0.0437**
	(0.0244)	(0.0244)	(0.0112)	(0.0279)	(0.0167)	(0.0199)
N	49,904	49,904	49,904	49,904	49,904	49,904
R^2	0.545	0.511	0.603	0.550	0.700	0.557

Notes: *** p<0.01, **p<0.05, * p<0.10. The occupational score in the second to last column is the logged occupation based on mean wages in the 1940 Census (see Appendix D). The last column is the log occupational score based on the variable *occscore* in IPUMS. The excluded group is arrivals age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Table 4. The Return to Education and Experience Separated by US and Foreign Component

Sample:	Full Sample	Only NW Europe	Only SE Europe	Full Sample
US education	0.0528*** (0.00692)	0.0645*** (0.00917)	0.0415*** (0.0105)	0.0645*** (0.00957)
US educ. x SE Europe				-0.0233* (0.0138)
Foreign education	0.0439*** (0.00535)	0.0583*** (0.00758)	0.0347*** (0.00772)	0.0583*** (0.00793)
Foreign educ. x SE Europe				-0.0236** (0.0108)
Foreign experience	0.00534 (0.0125)	0.0109 (0.0208)	-0.000811 (0.0164)	0.0107 (0.0218)
Foreign exp. x SE Europe				-0.0113 (0.0269)
(Foreign experience /10)^2	-0.0130 (0.124)	-0.0712 (0.261)	0.00117 (0.154)	-0.0710 (0.274)
(Foreign exp./10)^2 x SE Europe				0.0690 (0.311)
US experience	0.0686*** (0.00889)	0.0796*** (0.0107)	0.0564*** (0.0150)	0.0795*** (0.0112)
US experience x SE Europe				-0.0232 (0.0181)
(US experience /10)^2	-0.117*** (0.0152)	-0.143*** (0.0192)	-0.0956*** (0.0245)	-0.143*** (0.0201)
(US exp./10)^2 x SE Europe				0.0469 (0.0309)
Observations	35,229	15,881	19,348	35,229
R-squared	0.678	0.663	0.675	0.679

Notes: *** p<0.01, **p<0.05, * p<0.10. The dependent variable is log wage income. We assume that individuals enter school at age six and stay in school continuously in order to separate totals years of education and potential experience into foreign and United States components. Brothers fixed effects are included in each column.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Table 5. Effect of Age at Arrival on Social Outcomes

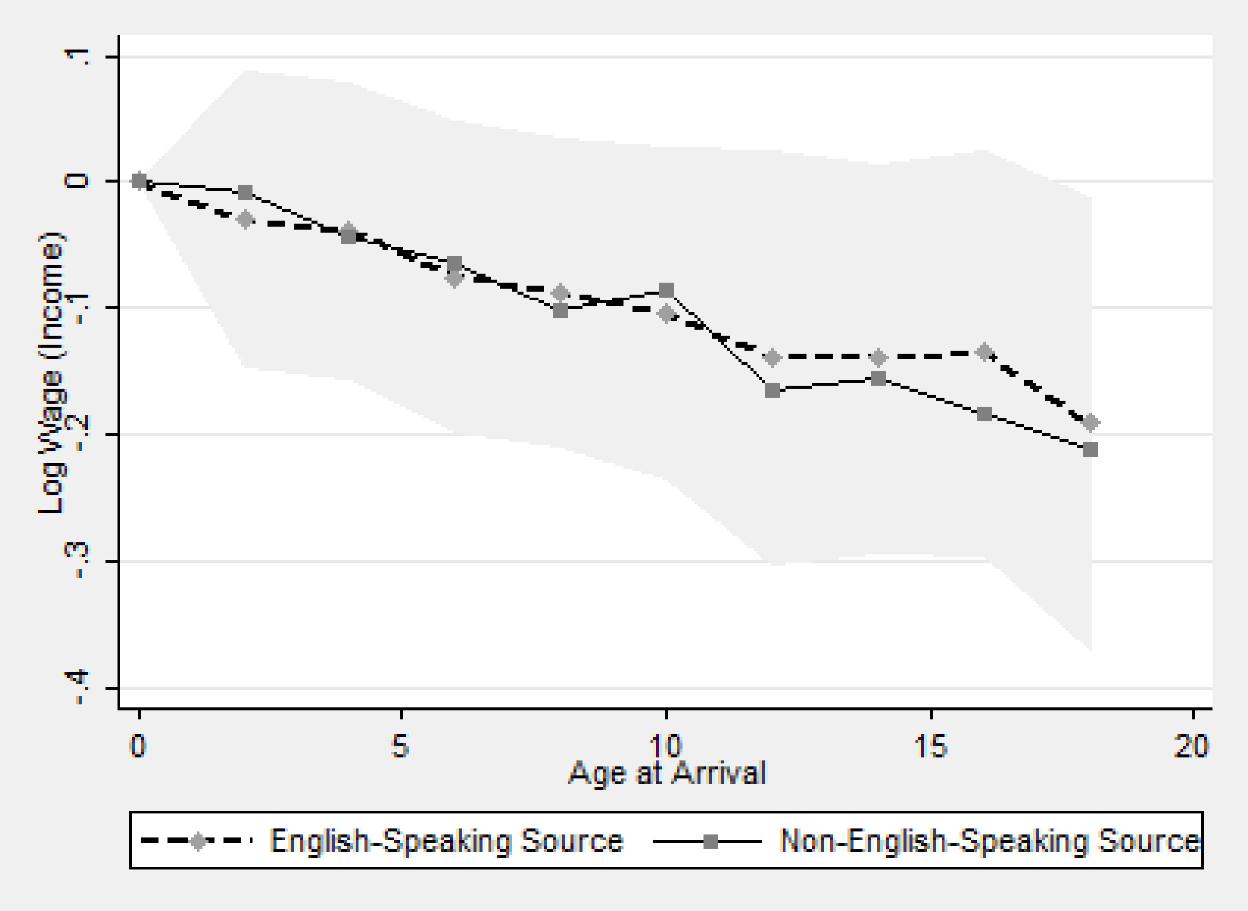
	Intermarriage		Spatial Assimilation	
	Native Spouse	Spouse from Different Source	Fraction of County Native HH	Fraction of Page Native HH
Age at arrival:				
2 to 3	-0.0505*	-0.0501**	-0.00411	-0.00382
	(0.0272)	(0.0254)	(0.00436)	(0.00817)
4 to 5	-0.0953***	-0.0760***	-0.00366	-0.00937
	(0.0274)	(0.0254)	(0.00449)	(0.00826)
6 to 7	-0.107***	-0.0974***	-0.00286	-0.00480
	(0.0280)	(0.0262)	(0.00464)	(0.00872)
8 to 9	-0.172***	-0.161***	0.000386	-0.00620
	(0.0280)	(0.0262)	(0.00463)	(0.00861)
10 to 11	-0.238***	-0.225***	-0.00308	-0.0109
	(0.0288)	(0.0269)	(0.00488)	(0.00901)
12 to 13	-0.274***	-0.261***	-0.00160	-0.00186
	(0.0321)	(0.0304)	(0.00556)	(0.0105)
14 to 15	-0.320***	-0.304***	0.00411	-0.0185*
	(0.0323)	(0.0307)	(0.00558)	(0.0105)
16 to 17	-0.368***	-0.341***	0.00331	-0.00895
	(0.0339)	(0.0323)	(0.00593)	(0.0111)
18 to 20	-0.419***	-0.408***	0.0111*	-0.0151
	(0.0360)	(0.0348)	(0.00662)	(0.0121)
Observations	38,803	38,803	53,129	53,129
R-squared	0.661	0.673	0.600	0.578

Notes: *** p<0.01, **p<0.05, * p<0.10. For the first two columns, we only include people who are married. The first column regresses whether the spouse is native-born on age at arrival, and the second column regresses whether the spouse is from a different source country on age at arrival. The fraction of page that are native household heads is the census page, which reflects immediate neighbors. Brothers fixed effects are included in each column.

Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

Appendix A. Figures and Tables Referenced in Main Text

Figure A1. Age-at-Arrival Profile for English and Non-English Sources



Notes: Data is split by England, Scotland, Ireland and Wales versus all other sources.

Sources: Linked sample of brothers from Ellis Island to the 1940 Census.

Table A1. Age at Arrival in Census, 1899 to 1930 arrivals

Country of Birth	Age at Arrival	% 0-15 Arrivals	% 16-45 Arrivals	% 45+ Arrivals
North and West Europe (Old Source)	22.7	24.6	71.2	4.1
South and East Europe (New Source)	21.1	31.4	65.6	3.0
Russia	20.7	33.0	64.0	3.0
Romania	20.8	36.4	59.9	3.7
Portugal	20.8	33.5	63.4	3.1
Italy	21.4	31.9	64.7	3.4
Finland	21.5	19.6	78.8	1.6
Greece	21.6	27.7	69.9	2.4
Hungary	21.7	28.4	69.2	2.4
Netherlands	21.8	33.2	61.9	4.9
Austria	21.9	27.3	70.1	2.5
Norway	21.9	22.7	73.7	3.5
Sweden	22.0	21.0	75.8	3.2
Denmark	22.2	20.8	76.1	3.1
Spain	22.4	23.7	73.7	2.6
France	22.4	28.2	67.2	4.5
Ireland	22.4	19.8	76.5	3.7
Belgium	22.6	25.9	71.8	2.4
Scotland	23.1	27.2	67.3	5.5
Germany	23.5	24.1	70.9	5.0
England	23.6	27.6	66.4	6.0
Switzerland	23.7	19.7	76.3	4.0

Notes: Data is from the 1900 to 1930 United States Censuses, keeping only 1899 to 1930 arrivals. We keep these years to match with the years of arrival in Figure 1.

Table A2. Robustness to higher quality links

	Income	Income	Income	Education	Education	Education
Sample:	Main	High Quality	Alternative Match Scores	Main	High Quality	Alternative Match Scores
Age at Arrival						
2 to 3	-0.0157	-0.0110	-0.0126	0.0229	0.0667	0.0325
	(0.0385)	(0.0523)	(0.0544)	(0.109)	(0.141)	(0.149)
4 to 5	-0.0421	-0.0234	-0.0392	-0.158	-0.165	-0.0794
	(0.0391)	(0.0520)	(0.0547)	(0.110)	(0.145)	(0.152)
6 to 7	-0.0680	-0.0889	-0.0659	-0.285	-0.325	-0.222
	(0.0407)	(0.0548)	(0.0566)	(0.115)	(0.151)	(0.160)
8 to 9	-0.0976	-0.114	-0.0907	-0.413	-0.407	-0.358
	(0.0406)	(0.0539)	(0.0578)	(0.116)	(0.151)	(0.162)
10 to 11	-0.0924	-0.102	-0.0874	-0.663	-0.602	-0.743
	(0.0429)	(0.0561)	(0.0621)	(0.123)	(0.159)	(0.173)
12 to 13	-0.159	-0.199	-0.121	-0.888	-0.802	-0.940
	(0.0487)	(0.0642)	(0.0692)	(0.143)	(0.185)	(0.209)
14 to 15	-0.150	-0.161	-0.158	-1.003	-0.909	-1.097
	(0.0502)	(0.0662)	(0.0728)	(0.145)	(0.187)	(0.210)
16 to 17	-0.168	-0.190	-0.177	-0.843	-0.744	-0.944
	(0.0536)	(0.0689)	(0.0787)	(0.153)	(0.198)	(0.227)
18 to 20	-0.204	-0.219	-0.117	-0.795	-0.675	-0.854
	(0.0587)	(0.0763)	(0.0946)	(0.167)	(0.220)	(0.263)
Observations	35,978	16,955	14,968	51,591	24,057	21,443
R-squared	0.659	0.634	0.694	0.632	0.628	0.668

Notes: Data is a sample of brothers linked from Ellis Island records to the 1940 Census. High-quality links are determined to be in the better 50 percent of scores for our linked dataset, as determined by the sum of Jaro-Winkler distance in first name, Jaro-Winkler distance in last name and absolute difference in year of birth. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Table A3. Effect of Age at arrival on Labor Supply, Weekly Wages, and Self-Employment

	LFP	Weeks of Work	Log (Weekly Wage)	Self Employed	Self Empl. and not farmer
Age at Arrival:					
2 to 3	-0.00181	0.188	-0.0130	-0.0312**	-0.0170
	(0.00918)	(0.641)	(0.0291)	(0.0151)	(0.0140)
4 to 5	-0.00434	-0.552	-0.0201	-0.0118	0.00648
	(0.00910)	(0.647)	(0.0291)	(0.0152)	(0.0141)
6 to 7	0.000293	-0.388	-0.0385	-0.0304*	0.00122
	(0.00946)	(0.670)	(0.0304)	(0.0159)	(0.0148)
8 to 9	-0.00386	-0.668	-0.0614**	-0.0302*	0.00279
	(0.00949)	(0.684)	(0.0307)	(0.0159)	(0.0149)
10 to 11	-0.00164	-0.481	-0.0739**	-0.0392**	-0.00152
	(0.00998)	(0.714)	(0.0328)	(0.0169)	(0.0156)
12 to 13	0.0132	0.00527	-0.107***	-0.0367*	0.00714
	(0.0110)	(0.805)	(0.0375)	(0.0194)	(0.0181)
14 to 15	0.00231	-0.891	-0.0965**	-0.0372*	0.00590
	(0.0117)	(0.836)	(0.0382)	(0.0199)	(0.0185)
16 to 17	-0.0106	-1.178	-0.138***	-0.0538**	0.00782
	(0.0125)	(0.885)	(0.0405)	(0.0211)	(0.0198)
18 to 20	-0.00472	-1.297	-0.158***	-0.0481**	0.0131
	(0.0140)	(0.959)	(0.0454)	(0.0235)	(0.0219)
Observations	53,129	53,129	35,663	47,901	47,901
R-squared	0.480	0.489	0.655	0.559	0.546

Notes: *** p<0.01, **p<0.05, * p<0.10. Data is a sample of brothers linked from Ellis Island records to the 1940 Census. The number of observations change across columns because only wage workers are included in the third column, and those who have missing information from the self-employed category are dropped in the fourth column. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Table A4. Effect of Age at Arrival on Home Ownership and Location

	Own house	Log(Value of house)	Urban	Urban Population
Age at Arrival				
2 to 3	0.00106	-0.00147	-0.00870	281.8
	(0.0179)	(0.117)	(0.0130)	(318.3)
4 to 5	0.0102	0.00106	-0.0102	568.7*
	(0.0182)	(0.120)	(0.0131)	(322.0)
6 to 7	0.0116	-0.0261	0.00170	459.4
	(0.0188)	(0.126)	(0.0137)	(334.9)
8 to 9	0.0190	-0.0525	-0.00727	427.7
	(0.0189)	(0.125)	(0.0138)	(337.0)
10 to 11	0.0236	-0.0796	-0.0114	622.2*
	(0.0199)	(0.128)	(0.0145)	(353.9)
12 to 13	-0.00468	0.0291	-0.00791	549.8
	(0.0227)	(0.141)	(0.0164)	(400.9)
14 to 15	0.0217	-0.0639	0.00208	478.7
	(0.0230)	(0.142)	(0.0167)	(411.4)
16 to 17	0.0207	-0.0379	-0.00851	937.5**
	(0.0242)	(0.153)	(0.0176)	(430.9)
18 to 20	0.0208	-0.0772	-0.00479	769.8
	(0.0265)	(0.168)	(0.0193)	(469.2)
Observations	51,616	20,746	53,129	53,129
R-squared	0.521	0.788	0.572	0.581

Notes: *** p<0.01, **p<0.05, * p<0.10. Data is a sample of brothers linked from Ellis Island records to the 1940 Census. Number of observations change across columns because missing information is dropped, and only those who own a house are in the second column. The excluded group is arrivals aged zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Table A5. Robustness when Controlling for Birth Order

	Income	Income	Income	Education	Education	Education
Age at Arrival						
2 to 3	-0.0157	-0.0215	-0.0202	0.0229	0.0488	0.0511
	(0.0385)	(0.0398)	(0.0398)	(0.109)	(0.113)	(0.113)
4 to 5	-0.0421	-0.0539	-0.0523	-0.158	-0.107	-0.105
	(0.0391)	(0.0437)	(0.0437)	(0.110)	(0.124)	(0.124)
6 to 7	-0.0680*	-0.0845*	-0.0831*	-0.285**	-0.214	-0.213
	(0.0407)	(0.0491)	(0.0491)	(0.115)	(0.139)	(0.139)
8 to 9	-0.0976**	-0.119**	-0.118**	-0.413***	-0.322**	-0.320**
	(0.0406)	(0.0536)	(0.0536)	(0.116)	(0.154)	(0.154)
10 to 11	-0.0924**	-0.119*	-0.118*	-0.663***	-0.550***	-0.547***
	(0.0429)	(0.0613)	(0.0613)	(0.123)	(0.175)	(0.175)
12 to 13	-0.159***	-0.189***	-0.188***	-0.888***	-0.758***	-0.756***
	(0.0487)	(0.0695)	(0.0694)	(0.143)	(0.202)	(0.203)
14 to 15	-0.150***	-0.185**	-0.184**	-1.003***	-0.854***	-0.849***
	(0.0502)	(0.0764)	(0.0764)	(0.145)	(0.217)	(0.217)
16 to 17	-0.168***	-0.207**	-0.205**	-0.843***	-0.675***	-0.665***
	(0.0536)	(0.0838)	(0.0837)	(0.153)	(0.238)	(0.238)
18 to 20	-0.204***	-0.252***	-0.248**	-0.795***	-0.589**	-0.559**
	(0.0587)	(0.0967)	(0.0967)	(0.167)	(0.279)	(0.280)
Birth Order Linear						
Birth Order		-0.0121			0.0522	
		(0.0195)			(0.0558)	
Birth Order Dummies						
2nd Born			-0.00879			0.0902
			(0.0213)			(0.0613)
3rd Born			-0.0429			0.0898
			(0.0453)			(0.130)
4th Born			-0.00381			-0.138
			(0.0786)			(0.221)
Observations	35,976	35,976	35,976	51,591	51,591	51,591
R-squared	0.659	0.659	0.659	0.632	0.632	0.632

Notes: *** p<0.01, **p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island records to the 1940 Census. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Table A6. Robustness to Dropping Arrivals 16 years and Older

	Income	Income	Education	Education
Age at Arrival				
2 to 3	-0.0157	-0.0155	0.0229	0.0250
	(0.0385)	(0.0396)	(0.109)	(0.113)
4 to 5	-0.0421	-0.0419	-0.158	-0.159
	(0.0391)	(0.0401)	(0.110)	(0.115)
6 to 7	-0.0680*	-0.0667	-0.285**	-0.297**
	(0.0407)	(0.0419)	(0.115)	(0.120)
8 to 9	-0.0976**	-0.0992**	-0.413***	-0.415***
	(0.0406)	(0.0419)	(0.116)	(0.122)
10 to 11	-0.0924**	-0.0968**	-0.663***	-0.667***
	(0.0429)	(0.0446)	(0.123)	(0.129)
12 to 13	-0.159***	-0.164***	-0.888***	-0.933***
	(0.0487)	(0.0516)	(0.143)	(0.154)
14 to 15	-0.150***	-0.157***	-1.003***	-1.066***
	(0.0502)	(0.0547)	(0.145)	(0.161)
16 to 17	-0.168***		-0.843***	
	(0.0536)		(0.153)	
18 to 20	-0.204***		-0.795***	
	(0.0587)		(0.167)	
Observations	35,976	28,982	51,591	40,837
R-squared	0.659	0.675	0.632	0.657

Notes: *** p<0.01, **p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island records to the 1940 Census. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household. Columns 2 and 4 drop those who arrived older than age 16.

Table A7. Age-at-Arrival Profiles are Robust to Americanization Process

	Income	Income	Education	Education
Linking:	Main	Non-Americanized	Main	Non-Americanized
Age at Arrival				
2 to 3	-0.0157	-0.0419	0.0229	-0.131
	(0.0385)	(0.0536)	(0.109)	(0.147)
4 to 5	-0.0421	-0.0254	-0.158	-0.280*
	(0.0391)	(0.0529)	(0.110)	(0.152)
6 to 7	-0.0680*	-0.0511	-0.285**	-0.360**
	(0.0407)	(0.0566)	(0.115)	(0.158)
8 to 9	-0.0976**	-0.102*	-0.413***	-0.670***
	(0.0406)	(0.0552)	(0.116)	(0.157)
10 to 11	-0.0924**	-0.0364	-0.663***	-0.774***
	(0.0429)	(0.0600)	(0.123)	(0.169)
12 to 13	-0.159***	-0.131*	-0.888***	-1.178***
	(0.0487)	(0.0693)	(0.143)	(0.197)
14 to 15	-0.150***	-0.169**	-1.003***	-1.326***
	(0.0502)	(0.0746)	(0.145)	(0.199)
16 to 17	-0.168***	-0.168**	-0.843***	-1.012***
	(0.0536)	(0.0769)	(0.153)	(0.215)
18 to 20	-0.204***	-0.201**	-0.795***	-1.062***
	(0.0587)	(0.0836)	(0.167)	(0.235)
Observations	35,977	18,052	51,591	25,712
R-squared	0.659	0.678	0.632	0.662

Notes: *** p<0.01, **p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island Records to the 1940 Census. This table tests the robustness of results when not Americanizing names in our dataset. See Appendix C for more detail. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.

Appendix B. Further Details on Data Creation

Information about Ellis Island arrivals was downloaded from http://www.jewishgen.org/databases/EIDB/ellisgold.html. The data collection focused on single males, aged 0-20, who arrived at Ellis Island between 1892 and 1924. The data fields that were collected were: first and last name; city and country of last residence; arrival day, month and year; age at arrival; departure port; ship name; passenger id; and ethnicity. Sex and marital status were also collected, but just to restrict the sample to male and single. Passenger id is a unique identifier for each entry into Ellis Island and is numbered such that those next to each other on the ship manifest are next to each other for passenger id. Since families are listed together in ship manifests, we can identify brothers as those listed next to each other who have the same surname, after sorting by ship name and passenger id. Since we do not collect females, we still capture brothers even if brothers were not immediately next to each other on the original manifests because the brothers appear next to each other in our data of only males. This leaves us with 447,540 potential brothers.

Next, we clean the residence field. From that field we needed a city and a country of origin; however, the initial origin field need a significant amount of cleaning. For instance, the field contains abbreviations, inconsistent spelling, and differing amounts of information (for instance, just the city name, or the city name and the state, or the city name, state and country). We clean the country of birth variable for origins that have ten or more observations, or 319,510 of our initial sample of 447,540 brothers.[29] For records that do not identify the country of birth, we assume the country of birth based on the reported ethnicity. This is mostly straightforward, but we cannot match for ethnicities such as Jewish, Arabian, or Black. Most of the time there

[29] Of the approximately 320,000 observations that have more than 10 entries, about 0.5% of countries of origin could not be identified. We assume that the country of origin matches one's ethnicity.

is a second ethnicity listed for these sources, but if not, then we drop those (12,620 observations) from our dataset.

Next, we also wish to link on year of birth, but the Ellis Island records only have age and date of arrival rather than year of birth. Therefore, we need to back out year of birth, which is typically done with the formula: Year of Observation – Age. However, this implies that an arrival who listed their age as 10 and arrived on January 1st 1910 would be born in 1900, but this arrival instead was likely born in 1899. Therefore, we back out the year of birth as *Year of Arrival - Age* for those who arrived in the second half of the year, and *Year of Arrival - Age - 1* for those who arrived in the first half of the year.

Third, we drop those who have missing letters in their first or last names, which is identified by strings such as "…" or "?", which drops 4,653 individuals. Fourth, if an individual lists an initial as the first name but then a longer second name, then we keep the second name as the main name; however, we drop those who only report an initial for the first name and give no second name.

Fifth, we Americanize the names. The first names found in the data were anglicized to increase the likelihood of matching. For instance 'Giuseppe' was changed to 'Joseph'. Each name was run through http://www.behindthename.com/ to provide a list of related names. To anglicize a name required at most a many-to-one relationship between the original name and the anglicized one. The issue was that this was found to be a many-to-many mapping; for instance, Joseph maps to both Joe and Guiseppe, but both of those also map back to Joseph. This meant that the mapping was circular and would depend on the order that the names were processed. Additionally, it was not clear from behindthename.com which name was best considered the anglicized version - should Joseph be changed to Guiseppe or vice versa?

To address these issues the database of first names at birth from US censuses that occurred before 1930 were obtained and combined to give a ranking of the popularity of each first name, as defined by the number of US-born children with that name. For each mapping, grouped by the initial name, say Guiseppe to Guisep and Guiseppe to Joseph, the script provided a preferred choice, based on which of the possible names is the most popular in the US census dataset. With this, we created a data file that included two primary variables: the first name string as observed in the Ellis Island name, and the Americanized name. We then merged our Ellis Island dataset with this file to attach the Americanized name to our dataset.

Finally, we drop potential brothers who are next to each other and are more than ten years apart. We do this in case those with the same surname that are more than ten years apart are not truly brothers but represent a father-son relationship or uncle-nephew relationship. Note that this does not drop sets of brothers where the oldest and youngest are more than ten years apart. For example, if there are a 14, 5 and 1 year-old who are identified as potential brothers, we keep them since none are more than ten years apart; but if there is a 20 year old, 5 and 1 year old, we drop the 20 year old from the dataset. Keeping those more than ten years apart does not lead to a qualitative change in results. Ultimately, we are left with 397,137 potential brothers to link.

Appendix C. Linking Methodology

We link our cleaned dataset of 397,137 brothers to white males in the 1940 United States Census by searching for the best match among the potential set of matches on Americanized first name, last name, year of birth (within a range of three years) and country of birth. Our process follows the same idea as others in the literature (for example, Abramitzky et al., 2014) with a few modifications. The main difference in our methodology is that we first Americanize all foreign-born names in the 1940 Census in case an immigrant changed his name from, for example, Jörg to George. Another difference is that we rate the quality of potential matches by determining the differences in string similarity via the Jaro-Winkler algorithm. The steps to our linking process are as follows:

1) "Americanize" the first names of Ellis Island and Census records with a list of 28,000 name variants from behindthename.com. Names that do not have an American equivalent are unchanged.
2) Standardize the first name resulting from step 1) and the last name with the NYSIIS algorithm. Drop observations that have the same Americanized first name string, last name string, year of birth and country of birth in both the Ellis Island records and 1940 Census.
3) Find all possible matches on NYSIIS Americanized first name, NYSIIS last name, country of birth and exact year of birth. Repeat this step but expand the window for difference in year of birth to allow up to a three-year difference.
4) Calculate a match score for each potential match, which is the sum of the Jaro-Winkler distance in Americanized first name string, Jaro-Winkler distance in last name string and difference in year of birth (0 for exact year of birth match). Note that this method does not actually treat all NYSIIS names equally, but only uses the NYSIIS algorithm to find potential matches.

5) Keep the minimum match score for each observation in the Ellis Island records, and then keep the minimum match score for each observation in the 1940 Census.

This process leads to linking 103,005 individuals from the set of 397,137 brothers in the Ellis Island records, a match rate of 25.9 percent, a reasonable rate for a single match. Since the empirical strategy requires the use of siblings, we drop individuals who do not have another matched sibling, which leads to our final sample of 53,129 brothers used in the main text.

In Table C1, we show differences between our entire linked sample and the sample of brothers we use in the main text. The primary difference between samples is that people in the brothers sample are 12 percentage points less likely to be from Southern and Eastern Europe, which is unsurprising since these sources had lower linking rates and thus there are fewer sets of two brothers linked than single individuals linked. This also leads our brothers sample to be slightly higher skilled than the overall linked sample by 0.2 years of education and by earning 2.9 percent more wage income.

While it is well known that linking may bias the representativeness of the sample, we cannot directly test the representativeness because the 1940 Census does not include year of arrival, and thus we cannot compare our sample to those from the same arrival cohort and those with the same arrival age. However, we can show how our linked dataset of brothers has different attributes than the European migrant stock with the same years of birth in the 1940 Census. The difference is shown in Columns III and IV in Table C1; note that differences between our linked sample of brothers and the 1940 migrant stock may result from biases in the linking process, because we have a specific migrant cohort, or because we only keep those who arrived at young ages. As expected, our linked sample of brothers is higher skilled and earns more income than the 1940 Census as a whole, partially because we have younger arrivals and younger arrivals have higher earnings later in life.

Table C1. Characteristics of our linked sample of brothers in the 1940 Census

	I	II	III	IV	(III-II)	(III - VI)
	Linked	Non-		1940		
Sample:	Sample	Brothers	Brothers	Stock	Difference	Difference
Age	40.88	40.84	40.91	48.04	0.0677	-7.123
	(9.040)	(8.913)	(9.152)	(11.56)	(0.0571)	(0.0654)
Education	7.171	7.051	7.277	6.860	0.227	0.417
	(3.705)	(3.762)	(3.650)	(3.951)	(0.0238)	(0.0241)
Log Occ.	6.902	6.879	6.924	6.910	0.0451	0.0137
Score	(0.335)	(0.327)	(0.341)	(0.330)	(0.00218)	(0.00223)
Self Employed	0.201	0.210	0.193	0.233	-0.0170	-0.0393
	(0.401)	(0.408)	(0.395)	(0.422)	(0.00267)	(0.00273)
Log Income						
Wage	6.951	6.936	6.965	6.887	0.0295	0.0778
	(0.709)	(0.711)	(0.707)	(0.777)	(0.00547)	(0.00581)
South and	0.642	0.712	0.580	0.617	-0.132	-0.0372
East Europe	(0.479)	(0.453)	(0.494)	(0.486)	(0.00300)	(0.00309)
Age arrival	2.332	0	4.412			
diff. in family	(3.194)	(0)	(3.181)			
N	100,476	47,353	53,123	47,667		

Notes: This table shows descriptive statistics of the linked sample, the linked sample split into brothers and non-brothers, and then the 1940 Census. Note that not all individuals have observed wage income, years of education or self-employment status.

One way in which our sample may be unrepresentative is because we Americanize names and this introduces a bias in our linking process. In our dataset, 36 percent of matches are matched due to the Americanization process. Given that about 30 percent of immigrants switched their first names at the naturalization stage according to data from New York, and that arrival records had more foreign-sounding names than census records, we believe that 36 percent is a reasonable number (Biavaschi et al., 2017; Carneiro et al., 2015). In Table C2, we list the top 25 names that were Americanized in our dataset of linked brothers. At the top of the list are primarily Italian names such as Giuseppe, the alternative (and misspelled) Guiseppe, Giovanni and Antonio. There are also non-Italian names that are Americanized, such as Josef, Johann, and Wilhelm.

Table C2. Top 25 Americanizations in Linked Dataset of Brothers

Rank	First name Arrival	First name 1940	N
1	Giuseppe	Joseph	2,227
2	Giovanni	John	1,567
3	Antonio	Anthony	1,362
4	Luigi	Louis	860
5	Vincenzo	Vincent	858
6	Guiseppe	Joseph	835
7	Pietro	Peter	715
8	Michele	Michael	585
9	Josef	Joseph	545
10	Domenico	Dominick	453
11	Jan	John	439
12	Nicola	Nicholas	299
13	Paolo	Paul	257
14	Johann	John	240
15	Carlo	Charles	185
16	Johan	John	183
17	Wilhelm	William	178
18	Johannes	John	164
19	Jose	Joseph	162
20	Heinrich	Henry	159
21	Andrea	Andrew	155
22	Janos	John	151
23	Georg	George	142
24	Filippo	Philip	141
25	Tommaso	Thomas	128

Notes: This table lists the top 25 Americanizations in our linked dataset of brothers, where the arrival name is the one listed in the Ellis Island records, while the 1940 name is the one listed in the 1940 United States Census.

Americanizing names is a not a standard process when linking individuals and therefore may somehow drive our results. We perform a robustness check in which we link the arrival records with the 1940 United States Census without Americanizing any of the names in the Ellis Island records or the 1940 Census. Not Americanizing names leads to a smaller set of linked individuals of 67,427, a drop of about 30 percent. This leads to an even smaller set of 2 successfully linked brothers of 26,412, which is unsurprising since we do not link those who changed their name. The smaller set of observations leads to noisier estimates, but our qualitative results hold when using the non-Americanized dataset, suggesting that the Americanization process does not drive the results in the main text. Table A7 shows the results for log wage income and years of education when not Americanizing our data compared with our main results.

Appendix D: Creation of Immigrant-Specific Occupational Score

In this section, we provide further details on the creation of immigrant-specific occupational score used in text. We create this score to improve on the standard occupational scores used in the literature, such as the 1950 *occscore* from IPUMS and the 1901 Cost of Living Survey score. There are important limitations when using these commonly used scores; for example, the 1901 Cost of Living Score is only representative for married urban families and therefore does not provide an accurate estimate for rural or single workers. The 1950 occupational score reflects earnings after World War II, and therefore understates wage gaps for data prior to World War II (Goldin and Margo, 1992). Moreover, neither score reflects earnings both scores do not reflect earnings that are specific to immigrants and thus they understate any difference between immigrants and natives, a key interest for this paper.

We create an alternative occupational score that is based on income reported in the full-count 1940 United States Census. Our approach follows Collins and Wanamaker (henceforth 'CW') (2014, 2017) in that we impute income separately by group; but instead of groups separated by race and region as in CW, we impute income separately by country of birth. Therefore, the occupation score is essentially the average earnings in each occupation / country of birth cell. We provide further details on how we create the score below, but we follow Appendix I.b of CW (2017) to fix for self-employed earnings and non-monetary compensation for farm laborers and farmers.

First, we take the full-count 1940 United States Census and top-code income to $5,000 for wage workers. For self-employed workers, we ignore their reported wage income since this is not consistently reported, but we instead impute their income. To do this, we follow the strategy laid out by CW (2017) where we take the ratio of self-employed earnings to wage-worker earnings by occupation in the 1960 census, assume this ratio from 1960 is a good proxy for the ratio in 1940, and multiply the ratio with the mean wage income by occupation and country of birth. This leads to an imputed income for each self-employed person that varies by occupation and country of birth. Then we collapse the 1940 data by detailed occupation code and country of birth to get an average income for each occupation, which forms the occupational score for the large majority of our data.

We do not take the above approach for farm laborers and farmers because they may receive compensation in kind which is not recorded in the income data. We take a few extra steps to estimate their incomes. Starting with farm laborers and once again following CW (2017), we increase farm laborers' mean wage income in the 1940 census by 26 percent to reflect in-kind compensation, which is based on the 1957 USDA report *Major Statistical Series of the U.S. Department of Agriculture*. The next step is to estimate income for farmers. First, we assume that the perquisite rate of farmers in the 1960 census is 35 percent (also based on the USDA report), and we scale up their reported (wage and business) income by this factor. To create the final estimate for farmer income in 1940, we assume that the ratio between farm laborers and farmer income (inclusive of perquisites) in 1960 is the same as in 1940. Therefore, we need to estimate farm laborers' income in 1960, which we boost their income by 19 percent to reflect in-kind compensation.

Appendix E. Robustness of results to a linking approach related to Feigenbaum (2016)

An alternative approach to linking.

In this section, we discuss an alternative method of linking immigrants from Ellis Island records to the 1940 Census that is related to Feigenbaum (2016). In the main text, our method of picking the best link is based on Massey (2017) where we rate matches by summing the difference in year of birth, Jaro-Winkler distance in first name, and Jaro-Winkler distance in last name. Rather than rating matches based on these values, we could instead use training data to estimate the penalty for having deviations in year of birth, first name and last name, as well as other variables. Feigenbaum (2016) uses this approach when linking children from the 1915 Iowa Census to the 1940 United States Census.

Related to our study on immigrants, Ward (2018) applies the Feigenbaum method to immigrants during the Age of Mass Migration in his study of English fluency in the 1910 to 1930 Censuses. Concerned that the penalty for deviations in name may vary by the source of immigrants, Ward draws random samples of 2,000 from 16 different ethnicities in 1920 (for example, Polish, Italian, German, etc.), hand-links them to the 1930 United States Census, and estimates a model to predict a match score for each immigrant.[30] We use Ward's (2018) hand-linked data on immigrants between the 1920 and 1930 Censuses as "training data" for our sample of Ellis Island arrivals linked to the 1940 Census. While linking Ellis Island records to the 1940 Census is different than linking the 1920 Census to 1930 Censuses, it will serve as a good quality check on our main results in sample. We do not use the data created from this linking process as our main sample because the "training data" is not specific to our linked data

[30] Ward (2018) discusses linking 15 different ethnicities: German, Jewish, Dutch, Swedish, Danish, Norwegian, Italian, French, Romanian, Greek, Russian, Czech/Slovak, Polish, Finnish and Hungarian. Ward additionally links immigrants from English-speaking sources (that is, England, Ireland and Scotland), but does not report this since his study is on the acquisition of English skills for immigrants from non-English-speaking sources.

between Ellis Island arrival records and the 1940 Census; however, qualitative results from this dataset are consistent with results in the main paper.

We cannot directly use the estimated probit coefficients from Ward's paper since he predicts scores based on year of arrival, a variable that is unavailable in the 1940 Census; therefore, we re-estimate a probit for each of the 16 ethnicities after removing the year of arrival variables from the model. The results for each probit model are shown in Tables E1 – E4, and show generally that having smaller deviations in Jaro-Winkler distance and year of birth predicts a match. We can then use the coefficients from this model to predict the probability that each potential link would be a match. Potential links between the Ellis Island data and the 1940 Census are chosen such that they have a Jaro-Winkler distance in first name of less than 0.20, Jaro-Winkler distance in last name of less than 0.25, a year of birth distance of less than 3, an exact match on country of birth, the same first letter of the first name and the same first letter of the last name.

At this point, we have predicted match probabilities for each potential link; now we have to determine parameters for who is included in the dataset. We choose the meta-parameters shown in Table E5 following Ward's (2018) conservative strategy such that the PPV (predictive positive value, or estimated share of true positives to overall positives) is 0.90.[31] This method of being conservative to increase the number of true positives (or reduce the number of false positives) leads to a significantly lower linking rate than the training sample and compared to our main method of linking immigrants.

This linking process leads to a smaller sample of brothers of 21,994 compared with our main sample of 53,129. This is partially because it is difficult to predict the best link from observable variables in hand-linked data; it also may be because the data we use to predict links

[31] The meta-parameters are the cut-off of predicted probability for keeping an immigrant in the sample, and the minimum ratio between highest match score and second-highest match score (to drop close second matches).

is not specific to the Ellis Island records matched to the 1940 Census. Being more restrictive about who is kept in the sample may also lead to biases in representativeness, but once again this is difficult to determine given the lack of census or representative sample that observes immigrant outcomes in 1940 in addition to year/age of arrival. We follow the same weighting process in the main section and weight to ensure that our sample is representative on country of birth.

All results qualitatively hold with alternative linked sample

We recreate all Tables and Figure from the main text with this linked sample (See Tables E6-E8; Figures E1-E3). We show that all results are qualitatively the same as in the main text: the age-at-arrival and income profile are similarly sloped with or without brothers fixed effects, older arrivals experienced a larger native-immigrant wage gap than younger arrivals, older arrivals acquired fewer years of education than younger arrivals, and older arrivals were less likely to marry a native-born spouse.

Table E1. Probit Coefficients, Part 1

	English	German	Yiddish, Jewish	Dutch
Year of birth difference = 1	-0.728***	-0.454***	-0.738***	-0.782***
	(0.0825)	(0.0865)	(0.0743)	(0.103)
Year of birth difference = 2	-1.104***	-0.705***	-0.974***	-1.201***
	(0.0962)	(0.0986)	(0.0842)	(0.135)
Year of birth difference = 3	-1.163***	-1.025***	-1.180***	-1.395***
	(0.104)	(0.116)	(0.0986)	(0.146)
Jaro-Winkler distance in first name string	-6.630**	-2.495	-3.574	1.330
	(2.892)	(2.366)	(2.611)	(2.029)
Jaro-Winkler distance in last name string	-9.190***	-9.705***	-8.584***	-12.13***
	(0.848)	(0.723)	(0.756)	(0.925)
Exact First name match (NYSIIS)	-0.204	0.102	0.0417	0.601*
	(0.404)	(0.313)	(0.347)	(0.314)
Exact first and last name match (NYSIIS)	-0.342***	-0.401***	-0.280***	-0.533***
	(0.105)	(0.120)	(0.0926)	(0.173)
Total number of hits	-0.171***	-0.202***	-0.165***	-0.266***
	(0.0190)	(0.0189)	(0.0220)	(0.0246)
Total number of hits squared	0.00413***	0.00521***	0.00347***	0.00653***
	(0.000655)	(0.000673)	(0.000717)	(0.000916)
First letter of last name match	0.230	-0.116	0.121	-0.487***
	(0.173)	(0.119)	(0.153)	(0.167)
First letter of first name match	0.171	0.557***	1.596***	0.385
	(0.291)	(0.181)	(0.528)	(0.262)
More than two hits have NYSIIS last name match	0.533***	0.535***	0.636***	-1.041***
	(0.129)	(0.165)	(0.116)	(0.266)
One hit has NYSIIS last name match	1.223***	0.848***	1.383***	1.958***
	(0.126)	(0.163)	(0.119)	(0.253)
Jaro-Winkler distance in NYSIIS first name	-1.548**	-3.213***	-2.540***	-5.875***
	(0.758)	(0.637)	(0.719)	(0.949)
Jaro-Winkler distance in NYSIIS last name	-1.255**	-0.308	-0.0312	-0.534*
	(0.584)	(0.257)	(0.122)	(0.319)
Middle initial match, if have one	1.116***	1.624***	0.414	1.011***
	(0.126)	(0.318)	(0.727)	(0.315)
Constant	0.837	1.133***	-0.625	2.500***
	(0.530)	(0.408)	(0.672)	(0.474)
Observations	12,975	11,227	25,691	6,651

Sources: Ward (2018)

Table E2. Probit Coefficients, Part 2

	Swedish	Danish	Norwegian	Italian
Year of birth difference = 1	-0.824***	-0.722***	-0.685***	-0.445***
	(0.0693)	(0.0733)	(0.0822)	(0.0652)
Year of birth difference = 2	-1.060***	-1.140***	-1.161***	-0.822***
	(0.0803)	(0.0916)	(0.105)	(0.0808)
Year of birth difference = 3	-1.342***	-1.446***	-1.459***	-1.212***
	(0.0973)	(0.114)	(0.123)	(0.107)
Jaro-Winkler distance in first name string	-4.822***	-4.209***	-2.601*	0.906
	(1.466)	(1.552)	(1.512)	(1.243)
Jaro-Winkler distance in last name string	-6.467***	-5.573***	-7.379***	-10.49***
	(0.828)	(0.889)	(0.793)	(0.592)
Exact First name match (NYSIIS)	0.228	0.268	0.403*	0.462**
	(0.226)	(0.217)	(0.222)	(0.188)
Exact first and last name match (NYSIIS)	-0.0560	-0.0728	-0.361***	0.00105
	(0.0954)	(0.0947)	(0.107)	(0.0932)
Total number of hits	-0.175***	-0.218***	-0.215***	-0.0654***
	(0.0182)	(0.0200)	(0.0194)	(0.0246)
Total number of hits squared	0.00366***	0.00485***	0.00472***	0.000384
	(0.000608)	(0.000670)	(0.000684)	(0.000787)
First letter of last name match	0.217	0.464**	0.354**	-0.00628
	(0.133)	(0.182)	(0.150)	(0.143)
First letter of first name match	0.446***	0.956***	0.740***	0.151
	(0.158)	(0.198)	(0.187)	(0.107)
More than two hits have NYSIIS last name match	0.690***	0.0203	-0.0275	0.686***
	(0.121)	(0.153)	(0.156)	(0.118)
One hit has NYSIIS last name match	1.229***	1.389***	1.547***	0.713***
	(0.128)	(0.157)	(0.153)	(0.123)
Jaro-Winkler distance in NYSIIS first name	-1.651**	-1.819**	-1.756**	-3.688***
	(0.738)	(0.848)	(0.693)	(0.593)
Jaro-Winkler distance in NYSIIS last name	-0.765***	-0.695***	-0.855***	-0.0696
	(0.253)	(0.243)	(0.272)	(0.144)
Middle initial match, if have one	1.275***	1.661***	1.042***	-
	(0.131)	(0.123)	(0.226)	
Constant	0.109	-0.528	0.0449	0.526
	(0.324)	(0.378)	(0.350)	(0.322)
Observations	21,648	18,690	13,893	29,591

Sources: Ward (2018)

Table E3. Probit Coefficients, Part 3

	French	Romanian	Greek	Russian
Year of birth difference = 1	-0.641***	-0.265*	-0.520***	-0.209*
	(0.132)	(0.146)	(0.0841)	(0.110)
Year of birth difference = 2	-1.040***	-0.521***	-0.968***	-0.530***
	(0.158)	(0.153)	(0.103)	(0.118)
Year of birth difference = 3	-0.899***	-0.602***	-1.023***	-0.559***
	(0.158)	(0.161)	(0.115)	(0.124)
Jaro-Winkler distance in first name string	-5.319*	-8.447	-0.214	-2.643
	(2.841)	(5.189)	(1.898)	(3.739)
Jaro-Winkler distance in last name string	-12.31***	-9.571***	-8.833***	-9.095***
	(1.081)	(0.980)	(0.716)	(0.785)
Exact First name match (NYSIIS)	-0.106	-1.240*	0.228	0.366
	(0.358)	(0.712)	(0.287)	(0.558)
Exact first and last name match (NYSIIS)	-1.031***	-0.652***	-0.137	-0.887***
	(0.200)	(0.221)	(0.111)	(0.151)
Total number of hits	-0.349***	-0.240***	-0.209***	-0.223***
	(0.0351)	(0.0306)	(0.0245)	(0.0217)
Total number of hits squared	0.0114***	0.00620***	0.00486***	0.00555***
	(0.00164)	(0.00134)	(0.000802)	(0.000790)
First letter of last name match	0.185	0.103	0.256	0.161
	(0.208)	(0.184)	(0.198)	(0.153)
First letter of first name match	1.283***	0.943*	0.472**	-0.130
	(0.414)	(0.518)	(0.225)	(0.280)
More than two hits have NYSIIS last name match	-0.852***	-0.534	0.895***	0.0920
	(0.323)	(0.362)	(0.140)	(0.258)
One hit has NYSIIS last name match	1.969***	1.759***	0.750***	1.073***
	(0.309)	(0.363)	(0.147)	(0.254)
Jaro-Winkler distance in NYSIIS first name	-3.694***	-2.865***	-3.269***	-5.015***
	(1.024)	(0.846)	(0.670)	(0.820)
Jaro-Winkler distance in NYSIIS last name	0.0116	-0.0180	-0.558***	-0.239
	(0.210)	(0.265)	(0.208)	(0.236)
Middle initial match, if have one	1.179**	-	0.864*	2.620**
	(0.469)		(0.461)	(1.041)
Constant	1.273**	1.849**	0.617	1.320**
	(0.614)	(0.925)	(0.456)	(0.672)
Observations	3,190	2,899	21,761	10,481

Sources: Ward (2018)

Table E4. Probit Coefficients, Part 4

	Czech	Polish	Finnish	Hungarian
Year of birth difference = 1	-0.570***	-0.438***	-0.580***	-0.425***
	(0.101)	(0.0742)	(0.0919)	(0.0998)
Year of birth difference = 2	-1.009***	-0.625***	-0.911***	-0.739***
	(0.123)	(0.0861)	(0.109)	(0.111)
Year of birth difference = 3	-1.229***	-0.700***	-0.950***	-1.106***
	(0.148)	(0.0996)	(0.111)	(0.128)
Jaro-Winkler distance in first name string	-6.374**	-1.801	1.632	-7.916***
	(3.125)	(2.694)	(1.783)	(2.525)
Jaro-Winkler distance in last name string	-12.24***	-11.45***	-8.022***	-8.046***
	(0.898)	(0.645)	(0.738)	(0.777)
Exact First name match (NYSIIS)	-0.827*	0.223	1.030***	-0.593*
	(0.432)	(0.364)	(0.294)	(0.356)
Exact first and last name match (NYSIIS)	-0.274*	-0.582***	-0.673***	-0.460***
	(0.162)	(0.112)	(0.122)	(0.132)
Total number of hits	-0.227***	-0.115***	-0.279***	-0.236***
	(0.0265)	(0.0243)	(0.0209)	(0.0225)
Total number of hits squared	0.00560***	0.00154*	0.00784***	0.00572***
	(0.000899)	(0.000793)	(0.000780)	(0.000822)
First letter of last name match	-0.159	0.237*	-0.0230	-0.0458
	(0.157)	(0.140)	(0.135)	(0.154)
First letter of first name match	0.450	0.370*	0.244	0.326
	(0.300)	(0.215)	(0.175)	(0.318)
More than two hits have NYSIIS last name match	0.264	0.300*	0.106	-0.107
	(0.221)	(0.162)	(0.165)	(0.203)
One hit has NYSIIS last name match	1.054***	1.188***	1.396***	1.627***
	(0.222)	(0.165)	(0.159)	(0.202)
Jaro-Winkler distance in NYSIIS first name	-4.807***	-3.273***	-2.033***	-2.655***
	(0.889)	(0.617)	(0.620)	(0.703)
Jaro-Winkler distance in NYSIIS last name	-0.465*	0.0369	-1.251***	-0.139
	(0.276)	(0.236)	(0.467)	(0.191)
Middle initial match, if have one	-0.464	1.537	1.216***	2.947***
	(2.109)	(4.076)	(0.350)	(1.072)
Constant	2.914***	0.863*	0.304	1.625***
	(0.615)	(0.482)	(0.382)	(0.490)
Observations	16,041	27,298	8,006	9,891

Sources: Ward (2018)

Table E5. Critical Values used to keep links

Language	Probability Threshold	Ratio of first-best score to second-best score	PPV	TPR
English	0.305	1.4	0.904	0.728
German	0.434	1.2	0.901	0.714
Yiddish, Jewish	0.372	1.7	0.901	0.594
Dutch	0.337	1.1	0.901	0.881
Swedish	0.268	3.4	0.901	0.572
Danish	0.356	1.9	0.901	0.611
Norwegian	0.331	1.5	0.902	0.731
Italian	0.521	1.5	0.901	0.432
French	0.313	1.2	0.903	0.871
Romanian	0.402	1.6	0.905	0.643
Greek	0.527	2.2	0.904	0.285
Russian	0.397	4.3	0.904	0.479
Czech/Slovak	0.325	3.1	0.901	0.622
Polish	0.357	9.1	0.904	0.383
Finnish	0.257	2.4	0.900	0.688
Hungarian	0.38	7.7	0.903	0.518

Notes: This table gives the meta-parameters for inclusion in the linked sample. The predicted probability for a match must be above the probability threshold, and the predicted probability must be at least be the multiple (in column 3) of the second-best score. The PPV, or positive prediction value, is the ratio of true positives to all positives; a higher number indicates fewer false positives. The TPR, or the true positive rate, is the ratio of true positives to all possible links; a lower number reflects that the probit does not include all matches from the hand linked data.

Table E6. Robustness of Table 3: Effect of Age at Arrival on Occupations

	White-Col.	Skilled	Farmer	Unskilled	Log (Occ. Score) 1940 Census	Log (Occ. Score) 1950 Occscore
Age at Arrival						
2 to 3	-0.00301	-0.00558	-0.00739	0.0160	-0.0477***	-0.0137
	(0.0251)	(0.0235)	(0.0103)	(0.0275)	(0.0174)	(0.0195)
4 to 5	-0.00722	-0.00405	-0.00397	0.0152	-0.0528***	-0.0160
	(0.0255)	(0.0236)	(0.0106)	(0.0280)	(0.0184)	(0.0203)
6 to 7	-0.0175	-0.00943	-0.0113	0.0382	-0.0930***	-0.0294
	(0.0269)	(0.0250)	(0.0112)	(0.0300)	(0.0185)	(0.0210)
8 to 9	-0.0260	0.0161	-0.0158	0.0258	-0.111***	-0.0254
	(0.0271)	(0.0250)	(0.0112)	(0.0297)	(0.0188)	(0.0213)
10 to 11	-0.0362	0.00340	-0.0143	0.0471	-0.118***	-0.0521**
	(0.0280)	(0.0265)	(0.0123)	(0.0314)	(0.0201)	(0.0227)
12 to 13	-0.0237	0.0111	-0.0203	0.0329	-0.140***	-0.0539**
	(0.0323)	(0.0306)	(0.0134)	(0.0362)	(0.0226)	(0.0258)
14 to 15	-0.0722**	0.0317	-0.0211	0.0616*	-0.151***	-0.0730***
	(0.0323)	(0.0318)	(0.0141)	(0.0369)	(0.0224)	(0.0256)
16 to 17	-0.0773**	0.00595	-0.0185	0.0899**	-0.156***	-0.0834***
	(0.0348)	(0.0339)	(0.0163)	(0.0389)	(0.0242)	(0.0290)
18 to 20	-0.0703*	0.0328	-0.0303	0.0678	-0.148***	-0.0610*
	(0.0394)	(0.0379)	(0.0195)	(0.0441)	(0.0276)	(0.0316)
N	20,715	20,715	20,715	20,715	20,715	20,715
R^2	0.591	0.554	0.677	0.584	0.712	0.624

Notes: This table recreates Table 3 from the main text but with the sample linked using the Feigenbaum (2016) method.