*Online Appendix*

*Appendix A*



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. Also see the text of the main document.

Table A1

AGE AT ARRIVAL IN CENSUS, 1899 TO 1930 ARRIVALS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country of Birth | Age at Arrival | Percent 0–15 Arrivals | Percent 16–45 Arrivals | Percent 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.

*Source*: See the text of the main document.

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.

*Source*: See the text of the main document.

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 changes 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.

*Source*: See the text of the main document.

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. The number of observations changes 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.

*Source*: See the text of the main document.

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 |
| Second born |  |  | –0.00879 |  |  | 0.0902 |
|  |  |  | (0.0213) |  |  | (0.0613) |
| Third born |  |  | –0.0429 |  |  | 0.0898 |
|  |  |  | (0.0453) |  |  | (0.130) |
| Fourth 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.

*Source*: See the text of the main document.

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.

*Source*: See the text of the main document.

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.

*Source*: See the text of the main document.

*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 needs 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.[[1]](#footnote-1) 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 is a second ethnicity listed for these sources, but if not, then we dropped 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 1 January 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 U.S. 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 U.S.-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 U.S. 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, five, and one 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, five, and one 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 U.S. 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 (e.g., Abramitzky, Boustan, and Eriksson 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, Jӧrg 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 one 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 (zero 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) |
|  Sample: | Linked Sample | Non-Brothers | Brothers | 1940 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.

*Source*:Linked sample of brothers from Ellis Island records to 1940 Census, and also 1940 Census (Ruggles, Genadek, Goeken, et al. 2017).

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, Giulietti, and Siddique 2017; Carneiro, Lee, and Reis 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 U.S. Census.

*Source*: Linked sample of brothers from Ellis Island records to 1940 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 U.S. 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 two 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 the 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 WWII, and therefore understates wage gaps for data prior to WWII (Goldin and Margo 1992). Moreover, neither score reflects 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 U.S. Census. Our approach follows William J. Collins and Marianne H. Wanamaker (2014, 2017) in that we impute income separately by group; but instead of groups separated by race and region as in Collins and Wanamaker, 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 Collins and Wanamaker (2017) to fix for self-employed earnings and non-monetary compensation for farm laborers and farmers.

 First, we take the full-count 1940 U.S. 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 Collins and Wanamaker (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 Collins and Wanamaker (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 James J. Feigenbaum (2016). In the main text, our method of picking the best link is based on Catherine G. 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 U.S. Census.

Related to our study on immigrants, Zachary 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 (e.g., Polish, Italian, German, etc.), hand-links them to the 1930 U.S. Census, and estimates a model to predict a match score for each immigrant.[[2]](#footnote-2) 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 Census, 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 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, andthe 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.[[3]](#footnote-3) 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 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 figures 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 |

*Source*: 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 |

*Source*: 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 |

*Source*: 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 |

*Source*: 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 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.

*Source*: Ward (2018).

Table E6

ROBUSTNESS OF TABLE 3: EFFECT OF AGE AT ARRIVAL ON OCCUPATIONS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | White-Col. | Skilled | Farmer | Unskilled | Log (Occ. Score) | Log (Occ. Score) |
|  |  |  |  |  | 1940 Census | 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 |
| R2 | 0.591 | 0.554 | 0.677 | 0.584 | 0.712 | 0.624 |

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

Table E7

ROBUSTNESS OF TABLE 4: THE RETURN TO EDUCATION AND EXPERIENCE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | Full Sample | Only NW Europe | Only SE Europe |   | Full Sample |
|   |   |   |   |   |   |
| US Educ. | 0.0611\*\*\* | 0.0705\*\*\* | 0.0501\*\*\* |  | 0.0705\*\*\* |
|  | (0.0102) | (0.0117) | (0.0181) |  | (0.0120) |
| US Educ. x SE Europe |  |  |  |  | –0.0203 |
|  |  |  |  |  | (0.0211) |
| Foreign Educ. | 0.0468\*\*\* | 0.0614\*\*\* | 0.0354\*\* |  | 0.0614\*\*\* |
|  | (0.00856) | (0.0105) | (0.0141) |  | (0.0107) |
| Foreign Educ. x SE Europe |  |  |  |  | –0.0259 |
|  |  |  |  |  | (0.0173) |
| Foreign Exp. | 0.0289 | 0.0438 | 0.0290 |  | 0.0438 |
|  | (0.0235) | (0.0490) | (0.0318) |  | (0.0501) |
| Foreign Exp. x SE Europe |  |  |  |  | –0.0148 |
|  |  |  |  |  | (0.0587) |
| (Foreign Exp./10)2  | –0.170 | –0.607 | –0.206 |  | –0.607 |
|  | (0.274) | (0.820) | (0.329) |  | (0.839) |
| (Foreign Exp./10)2 x SE Europe |  |  |  |  | 0.402 |
|  |  |  |  |  | (0.897) |
| US Exp. | 0.0792\*\*\* | 0.0943\*\*\* | 0.0574\*\* |  | 0.0943\*\*\* |
|  | (0.0123) | (0.0138) | (0.0232) |  | (0.0141) |
| US Exp. x SE Europe |  |  |  |  | –0.0369 |
|  |  |  |  |  | (0.0264) |
| (US Exp./10)2  | –0.128\*\*\* | –0.172\*\*\* | –0.0751\* |  | –0.172\*\*\* |
|  | (0.0232) | (0.0259) | (0.0419) |  | (0.0265) |
| (US Exp./10)2 x SE Europe |  |  |  |  | 0.0973\*\* |
|  |  |  |  |  | (0.0482) |
|  |  |  |  |  |  |
| Observations | 14,703 | 8,980 | 5,723 |  | 14,703 |
| R-squared | 0.715 | 0.705 | 0.720 |   | 0.717 |

*Notes and Source*: This table recreates Table 4 from the main text, but with the sample linked using the Feigenbaum (2016) method.

Table E8

ROBUSTNESS OF 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 Census Page Native HH |
| Age at Arrival |
| 2 to 3 | –0.0176 | –0.0357 | 0.00321 | 0.000183 |
|  | (0.0483) | (0.0430) | (0.00520) | (0.0110) |
| 4 to 5 | –0.0649 | –0.0583 | –0.00333 | –0.00929 |
|  | (0.0502) | (0.0447) | (0.00552) | (0.0113) |
| 6 to 7 | –0.0744 | –0.0832\* | –0.00155 | 0.00463 |
|  | (0.0518) | (0.0470) | (0.00584) | (0.0123) |
| 8 to 9 | –0.142\*\*\* | –0.156\*\*\* | 0.00100 | 0.00411 |
|  | (0.0512) | (0.0458) | (0.00576) | (0.0119) |
| 10 to 11 | –0.178\*\*\* | –0.192\*\*\* | 0.00149 | 0.0107 |
|  | (0.0539) | (0.0481) | (0.00617) | (0.0127) |
| 12 to 13 | –0.244\*\*\* | –0.249\*\*\* | 0.00303 | 0.0172 |
|  | (0.0620) | (0.0571) | (0.00703) | (0.0152) |
| 14 to 15 | –0.307\*\*\* | –0.308\*\*\* | 0.00400 | –0.00113 |
|  | (0.0619) | (0.0567) | (0.00709) | (0.0151) |
| 16 to 17 | –0.382\*\*\* | –0.381\*\*\* | 0.00867 | –2.41e-05 |
|  | (0.0670) | (0.0621) | (0.00781) | (0.0164) |
| 18 to 19 | –0.424\*\*\* | –0.443\*\*\* | 0.00893 | –0.0106 |
|  | (0.0735) | (0.0707) | (0.00934) | (0.0187) |
| Observations | 12,503 | 12,503 | 21,994 | 21,994 |
| R-squared | 0.732 | 0.741 | 0.737 | 0.658 |

*Notes and Source*: This table recreates Table 5 from the main text, but with the sample linked using the Feigenbaum (2016) method.



Figure E1

ROBUSTNESS OF FIGURE 3:

THE NEGATIVE EFFECT OF AGE AT ARRIVAL ON THE NATIVE-IMMIGRANT GAP IN WAGE INCOME IN 1940

*Notes*: The dependent variable is the 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 using Feigenbaum (2016) method.



Figure E2

ROBUSTNESS OF 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 the 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 using Feigenbaum (2016) method.



Figure E3

ROBUSTNESS OF 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 using Feigenbaum (2016) method.

REFERENCES

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. “A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration.” *Journal of Political Economy* 122, no. 3 (2014): 467–506.

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

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.

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): 220–52.

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

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

Goldin, Claudia, and Robert A. Margo. "The great compression: The wage structure in the United States at mid-century." *The Quarterly Journal of Economics* 107, no. 1 (1992): 1-34.

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): 129–43.

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

Ward, Zachary. “Have Language Skills Always Been So Valuable? The Low Return to English Fluency During the Age of Mass Migration.*”* Unpublished manuscript, 2018.

1. Of the approximately 320,000 observations that have more than 10 entries, about 0.5 percent of countries of origin could not be identified. We assume that the country of origin matches one’s ethnicity. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. 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). [↑](#footnote-ref-3)