*Online Appendix*

APPENDIX A – ROBUSTNESS CHECKS AND ADDITIONAL ANALYSIS

Appendix Figure A1. Generational assimilation is similar when limiting to families with children

****

Sources: 1880, 1910 and 1940 Censuses (Ruggles et al. 2019).

Note: This figure shows that assimilation was similar when limiting the 30-44 year-old adults to those with a children at most 14 years of age. These results are suggestive evidence that return migration does not influence assimilation patterns since return migration was less likely for families than for single males. Second, the results show that the assimilation profile is similar when more precisely linking families across censuses, since the 0-14 year olds in the prior census would be 30-44 year olds in the next census.

Appendix Figure A2. Alternative Definitions of “Mexican” to include former territories of Mexico



Sources: 1880, 1910 and 1940 Censuses (Ruggles et al. 2019).

Note: This figure shows that profile was mostly flat when including former Mexican territories in the “Mexican” group. We perform this check since it is ambiguous who is of Mexican origin in the 1880 Census.

Appendix Figure A3.Limiting sample to those born after the Mexican Cession does not alter results.



Sources: 1880, 1910 and 1940 Censuses (Ruggles et al. 2019).

Note: This figure shows that profile was mostly flat when including only those born after the 1848 Mexican Cession. This group is more clearly of Mexican origin in the 1880 Census.

Appendix Figure A4.second and third-generation Mexican Americans have similar intergenerational relationships

****

Sources: Linked data between the 1910 and 1940 Censuses (Ruggles et al. 2019).

Notes: Standard errors are in parenthesis. The difference in intercepts is insignificant (p-value=0.197) but the difference in slopes is significant (p-value=0.005).

Appendix Figure A5.Hand-linked and machine-linked data produce similar mobility estimates for third-generation Mexican Americans

****

Sources: Linked data between the 1910 and 1940 Censuses (Ruggles et al. 2019).

Notes: Standard errors are in parenthesis. The difference in intercepts is insignificant (p-value=0.595) and the difference in slopes is insignificant (p-value=0.378). See section “Hand-linked versus Machine-Linked data” in Appendix B for more information. In particular, the underlying regression results are in Table B7.

Appendix Figure A6. Income Score Rank Distribution in 1880, 1910 and 1940 Census

1. Distribution in 1880 Census (first-gen Mexican American)



1. Distribution in 1910 Census (second-gen. Mexican American)



1. Distribution in 1940 Census (third-gen. Mexican American)



Sources: Data are from 1880 Census (Figure A), 1910 Census (Figure B) and 1910-1940 Linked data (Figure C, outcomes in 1940). (Ruggles et al. 2019)

Notes: Population is males 30-44 years old. Income score as described in Appendix C.

Appendix Figure A7.Likelihood of son holding a professional job, conditional on father’s occupation

****

Sources: Linked data between the 1910 and 1940 Censuses (Ruggles et al. 2019).

Notes: Data show mobility gaps across groups, based on the likelihood of belonging to the highest skilled occupational group of professionals.

Appendix Figure A8.Likelihood of son holding an unskilled job, conditional on father’s occupation

****

Sources: Linked data between the 1910 and 1940 Censuses (Ruggles et al. 2019).

Notes: Data show mobility gaps across groups, based on the likelihood of belonging to one of the lowest skilled occupational groups of unskilled workers. Note that unskilled workers include laborers and farm laborers.

Table A1. Occupational distributions of non-Mexican white, African and Mexican Americans between 1880 and 1940

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   |   | 1880 |   |   |   | 1910 |   |   |   |   | 1940 |   |   |
|   | Non-Mex White | African Amer. | Mex., first gen |   | Non-Mex White | African Amer. | Mex., second-gen | Mex., first-gen |   | Non-Mex White | African Amer. | Mex., third-gen | Mex., second-gen |
| *Panel A. Entire country* |  |  |  |  |  |  |  |  |  |  |  |
| WC, professional | 12.9 | 1.7 | 3.1 |  | 13.8 | 2.9 | 5.0 | 2.7 |  | 17.7 | 3.4 | 9.4 | 7.6 |
| WC, sales/clerical | 5.6 | 0.4 | 1.0 |  | 9.8 | 0.9 | 3.0 | 1.1 |  | 16.0 | 2.4 | 8.9 | 7.9 |
| Semi-Skilled | 13.4 | 3.9 | 5.8 |  | 17.3 | 5.0 | 6.7 | 4.6 |  | 17.2 | 5.5 | 12.1 | 12.5 |
| Unskilled (non- laborer) | 12.3 | 13.2 | 16.8 |  | 14.5 | 19.5 | 15.0 | 15.3 |  | 22.4 | 29.9 | 21.4 | 22.4 |
| Farmer, owner | 42.1 | 30.1 | 15.1 |  | 19.3 | 7.8 | 7.3 | 1.5 |  | 6.3 | 3.5 | 4.1 | 3.1 |
| Farmer, tenant |  | 11.6 | 28.0 | 14.4 | 7.0 |  | 6.4 | 14.7 | 6.1 | 4.5 |
| Farm laborer | 5.8 | 21.8 | 22.0 |  | 6.3 | 13.6 | 26.6 | 21.5 |  | 4.6 | 11.9 | 19.1 | 19.0 |
| General laborer | 8.0 | 28.7 | 36.2 |  | 7.3 | 22.2 | 21.9 | 46.2 |  | 9.3 | 28.7 | 19.0 | 23.1 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Panel B. Gap with non-Mexican White* |
| WC, professional |  | -11.1 | -9.8 |  |  | -11.0 | -8.8 | -11.1 |  |  | -14.2 | -8.3 | -10.0 |
| WC, sales/Clerical |  | -5.1 | -4.5 |  |  | -8.9 | -6.9 | -8.7 |  |  | -13.7 | -7.2 | -8.1 |
| Semi-Skilled |  | -9.5 | -7.6 |  |  | -12.3 | -10.6 | -12.7 |  |  | -11.7 | -5.1 | -4.8 |
| Unskilled (non-laborer) | 1.0 | 4.6 |  |  | 5.0 | 0.5 | 0.8 |  |  | 7.5 | -1.0 | 0.0 |
| Farmer, owner |  | -12.0 | -27.1 |  |  | -11.5 | -11.9 | -17.7 |  |  | -2.9 | -2.2 | -3.3 |
| Farmer, tenant |  |  |  | 16.4 | 2.8 | -4.6 |  |  | 8.2 | -0.3 | -2.0 |
| Farm laborer |  | 16.0 | 16.2 |  |  | 7.3 | 20.3 | 15.2 |  |  | 7.3 | 14.4 | 14.3 |
| General laborer |  | 20.7 | 28.2 |  |  | 14.9 | 14.6 | 38.9 |  |  | 19.4 | 9.7 | 13.8 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Observations | 2,692,353 | 438,598 | 11,771 |   | 5,265,894 | 774,402 | 9,077 | 33,504 |   | 3,235,260 | 161,764 | 5,088 | 5,381 |

Sources: 1880 and 1910 full-count censuses, and 1910-1940 linked sample (Ruggles et al. 2019).

Notes: Target population is 30-44 year old males who claim an occupation. non-Mexican white and African Americans are native born; first-generation Mexican Americans are born in Mexico, second-generation are native born to at least one Mexican-born parent, and third-generation are native born to two native-born parents and at least one Mexican-born grandparent. Professional jobs have an occ1950 code that starts with 0 or 2, Sales/Clerical job have an occ1950 code that starts with 3 or 4, Semi-skilled jobs have an occ1950 code that starts with a 5, Unskilled (non-laborer) jobs have occ1950 codes that start with a 6, 7, 8 or 9, excluding farm laborers and general laborers. Farmer owners have an occ1950 code that starts with a 1 and own their home. Farmer tenants have an occ1950 code that starts with a 1 and do not own their home. Farm laborers are those with an 820, 830 or 840 code. General laborers are those with a 970 code.

Appendix Table A2.Generational progress when using 1950 occupational income score

|  |  |  |  |
| --- | --- | --- | --- |
|   | Ratio of Log Occupational Score w/ Anglo Americans |   | Difference in Occupational Score Rank |
|   | 1880 | 1910 | 1940 |   | 1880 | 1910 | 1940 |
| *Panel A: Entire country* |  |  |  |  |
| Mexican (1880 cohort) | 0.843 | 0.756 | 0.791 |  | -11.79 | -16.46 | -14.25 |
| Mexican (1910 cohort) |  | 0.807 | 0.794 |  |  | -13.46 | -14.32 |
| African American | 0.774 | 0.755 | 0.724 |  | -15.95 | -15.91 | -22.66 |
|  |  |  |  |  |  |  |  |
| *Panel B: Only border* |  |  |  |  |
| Mexican (1880 cohort) | 0.895 | 0.767 | 0.748 |  | -8.687 | -15.72 | -18.15 |
| Mexican (1910 cohort) |  | 0.808 | 0.752 |  |  | -13.46 | -18.27 |
| African American | 0.794 | 0.756 | 0.656 |  | -13.20 | -15.77 | -28.58 |
|  |  |  |  |  |  |  |  |
| *Panel C: Non-Border* |  |  |  |  |
| Mexican (1880 cohort) | 0.961 | 0.872 | 0.888 |  | -5.242 | -9.019 | -9.094 |
| Mexican (1910 cohort) |  | 0.928 | 0.957 |  |  | -6.183 | -3.658 |
| African American | 0.773 | 0.754 | 0.732 |  | -16.10 | -15.94 | -21.99 |
|  |  |  |  |  |  |  |  |
| *Panel D: Only Texas* |  |  |  |  |
| Mexican (1880 cohort) | 0.908 | 0.785 | 0.748 |  | -9.841 | -15.20 | -17.60 |
| Mexican (1910 cohort) |  | 0.829 | 0.756 |  |  | -12.78 | -17.14 |
| African American | 0.856 | 0.817 | 0.693 |  | -10.19 | -11.98 | -24.34 |
|  |  |  |  |  |  |  |  |
| *Panel E: Only California* |  |  |  |  |
| Mexican (1880 cohort) | 0.866 | 0.766 | 0.843 |  | -8.635 | -15.29 | -11.00 |
| Mexican (1910 cohort) |  | 0.712 | 0.776 |  |  | -20.42 | -17.25 |
| African American | 0.815 | 0.757 | 0.715 |  | -14.67 | -18.11 | -26.11 |
|  |  |  |  |  |  |  |  |
| *Panel F: Control for state* |  |  |  |  |
| Mexican (1880 cohort) | 0.903 | 0.801 | 0.801 |  | -8.690 | -13.65 | -14.11 |
| Mexican (1910 cohort) |  | 0.844 | 0.792 |  |  | -11.39 | -15.15 |
| African American | 0.848 | 0.817 | 0.746 |  | -11.84 | -12.25 | -20.85 |
|  |  |  |  |  |  |  |  |
| *Panel G: Control for county* |  |  |  |  |
| Mexican (1880 cohort) | 0.867 | 0.793 | 0.802 |  | -7.991 | -13.82 | -14.40 |
| Mexican (1910 cohort) |  | 0.817 | 0.783 |  |  | -12.84 | -16.15 |
| African American | 0.811 | 0.780 | 0.725 |   | -13.72 | -14.76 | -22.88 |

Sources: 1880 and 1910 full-count censuses, and 1910-1940 linked sample (Ruggles et al. 2019).

Notes:This table recreates Table 2 from the main paper when using the *occscore* variable from IPUMS, which reflects the median earnings by occupation in the 1950 census.

APPENDIX B – FURTHER LINKING DETAILS BETWEEN 1910 AND 1940

We create a few new linked datasets between the 1910 and 1940 United States Censuses: first, hand-linking third-generation Mexican American children from 1910 to 1940, and second, predicting links between 1910 and 1940 for the other groups (non-Mexican white and black, and second-generation Mexican). We describe the linking process for each of these datasets in more detail below.

*Hand-linking third-generation Mexican Americans*

To create this dataset, we first draw the entire population of native-born males aged 0 to 14 from the 1910 preliminary count full-count census available from IPUMS and accessed at the National Bureau for Economic Research (NBER). We keep only those who have native-born mothers and native-born fathers according to the child’s responses to the mother and father’s place of birth question. Then we attach the characteristics of the parents to the child’s observation using the relationship status variables; we are primarily interested in attaching the father’s occupation and their own parents’ places of birth. With this information, we define children as third-generation Mexican Americans for those who have at least one grandparent who was born in Mexico. This leaves us with a data set of 18,656 third-generation Mexican Americans to link forward to the 1940 Census.

To link these children to the 1940 census, we first search for a set of potential links that meet the criteria of being close in first name, last name and year of birth; we do this to make the linking process more feasible. Explicitly, we only keep potential links in 1940 as those who have the following characteristics with the 1910 children:

1. Either an exact match on the first letter of the first name OR exact match on the first letter of the last name
2. Jaro-Winkler distance of less than or equal to 0.25 for the first name
3. Jaro-Winkler distance of less than or equal to 0.20 for the last name
4. Exact match on race and state of birth
5. Absolute value of year of birth difference less than 3

From this set of potential links, we keep the top 25 closest matches based on the predicted match scores from Feigenbaum (2016). This leaves us with 234,279 potential matches for 16,845 children; note that we do not find *any* potential match for 1,811 children, which may be due to death between 1910 and 1940 or other errors in the linking process. Therefore, each third-generation Mexican American has on average 13.9 potential matches to choose from. Rather than hand-linking a subsample of this data and then predicting the hand-linking process for the rest of the data, as done by Feigenbaum (2016), we instead hand-link the entire dataset. We do this to maximize match rates and minimize false positives (Bailey et al. 2019). Ultimately, we are able to link 5,875 of those with potential links, for a linking rate of 31.4 percent. We do not achieve a perfect linking rate since it is still difficult to determine which link is the best given the limited set of information we can link on (that is, first name, last name, year of birth, race and state of birth). Note that the final sample of analysis in the main text is less than 5,875 due to missing information on key variables such as occupation, income, or father and mother’s observables in the 1910 census.

*Linking non-Mexican whites, non-Mexican blacks, and second-generation Mexican Americans*

We compare the intergenerational mobility of Mexican Americans with that of non-Mexicanblackand white Americans. Instead of hand-linking the entire censuses like we do for the third Mexican Americans (which would be highly costly), we instead pursue the strategy laid out by Feigenbaum (2016) where we hand-link a set of training data, model the hand-linking process using a probit model, and then use the predicted coefficients from the probit model to predict who would be the best link for the rest of the data. Note that we do this machine-learning process for second-generation Mexican Americans as well since they are a larger group than third-generation Mexican Americans.

 To create the set of potential links, we first keep the entire native-born population of children aged 0-14. We search the potential set of matches in the 1940 Census that meet the following conditions

1. Both an exact match on the first letter of the first name AND exact match on the first letter of the last name
2. Jaro-Winkler distance of less than or equal to 0.20 for the first name
3. Jaro-Winkler distance of less than or equal to 0.20 for the last name
4. Exact match on race and state of birth
5. Absolute value of year of birth difference less than 3

From these set of potential matches, we hand-linked three subsamples of 2,000, one of non-Mexican white sons, one of non-Mexican black sons, and one of second-generation Mexican American sons. We linked separate groups in case the models that predict the best match varied across populations, perhaps due to variation in names, the likelihood of a potential match, or errors from enumeration or transcription.

The details for linking rates across the different datasets are given in Table B1. There are fewer potential links in 1940 for each African American that meet the above criteria than for non-Mexican white Americans. This may simply be because the African American population was smaller than the white population. The number of potential links is similar for second-generation Mexicans as it is for African Americans. Yet at the same time, the hand-linking rate for non-Mexican whites was 56.1 percent, which was higher than for the African American sample at 43.1 percent and for the second-generation Mexican sample at 43.2 percent. This may be due to higher mortality rates for Mexican and African Americans between 1910 and 1940, higher rates of underenumeration, or more reporting / transcription error.

Table B1. Details on the numbers for linking

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Non-Mexican whiteAmericans | second-generation Mexican | third-generation Mexican | Non-Mexican blackAmericans |
|  |  |  |  |  |
| Random sample in 1910 with at least one potential link in 1940 | 2,000 | 2,000 | 16,845 | 2,000 |
| Potential links in 1940  | 16,248 | 10,298 | 234,279 | 9,695 |
| Successfully linked | 1,121 | 864 | 5,875 | 862 |
| Hand-linking rate for training data | 56.1 | 43.2 | 34.9 | 43.1 |

Sources: 1910 and 1940 censuses (Ruggles et al. 2019).

Note: third-generation Mexican Americans are not a sample but the full-population.

The hand-linked datasets form our training data, from which we model the hand-linking process of picking the true link with a probit. We model the hand-linking process based on our experience linking datasets, including variables such as the Jaro-Winkler string distance for first name and last name, the difference in year of birth, whether there is a NYSIIS match for first name or last name, and interacted variables. The results of the probit are shown in Table B2. In general, the point coefficients for the entire population do not dramatically differ from the non-Mexican white, second-generation Mexican and African American groups; however, there appears to be less information in the year of birth differences for African Americans than for non-Mexican white Americans. It is possible that we should model the linking process differently across races by including or excluding different variables; however, we leave a full exploration of differences in linking groups to another paper.

Table B2. Modeling the linking process with a probit

|  |  |  |  |
| --- | --- | --- | --- |
|   | White American | Black American | Second-gen. Mexican |
| Jaro-Winkler distance, first name | -4.885\*\*\* | -4.203\*\*\* | -4.858\*\*\* |
|  | (0.576) | (0.635) | (0.532) |
| Jaro-Winkler distance, surname | -13.64\*\*\* | -12.35\*\*\* | -10.24\*\*\* |
|  | (0.853) | (1.113) | (0.834) |
| Year of birth difference = 1 | -0.557\*\*\* | 0.0740 | -0.241\*\* |
|  | (0.114) | (0.198) | (0.121) |
| Year of birth difference = 2 | -0.906\*\*\* | -0.199 | -0.548\*\*\* |
|  | (0.131) | (0.202) | (0.131) |
| Year of birth difference = 3 | -1.426\*\*\* | -0.355\* | -0.847\*\*\* |
|  | (0.157) | (0.203) | (0.139) |
| Number of potential links | -0.114\*\*\* | -0.129\*\*\* | -0.140\*\*\* |
|  | (0.0187) | (0.0235) | (0.0203) |
| Number of potential links squared | 0.00217\*\*\* | 0.00283\*\*\* | 0.00300\*\*\* |
|  | (0.000639) | (0.000922) | (0.000755) |
| Exact surname match and unique surname | 0.619\*\*\* | 0.228 | 0.447\*\* |
|  | (0.235) | (0.303) | (0.220) |
| Exact first and surname string match and unique first and surname string | 0.382\*\* | 0.763\*\*\* | 0.450\*\*\* |
|  | (0.159) | (0.154) | (0.156) |
| Exact first name match and unique first name | -0.391\* | -0.0317 | -0.406\*\*\* |
|  | (0.205) | (0.178) | (0.157) |
| Exact Soundex first name match and unique soundex first name | 0.277 | 0.149 | - |
|  | (0.279) | (0.202) |  |
| Exact Soundex surname match and unique soundex surname | -0.238 | 0.489\*\*\* | 0.972\*\*\* |
|  | (0.192) | (0.168) | (0.123) |
| Exact Soundex first and surname match and unique soundex first and surname | 0.829\*\*\* | 0.622\*\*\* | 0.220 |
|  | (0.188) | (0.158) | (0.160) |
| Exact NYSIIS first name match and unique NYSIIS first name | 0.204 | 0.489\*\* | 0.698\*\*\* |
|  | (0.298) | (0.215) | (0.135) |
| Exact NYSIIS surname match and unique NYSIIS surname | 0.445 | -2.197 | 0.927\*\*\* |
|  | (0.320) | (126.1) | (0.271) |
| Exact NYSIIS first and surname match and unique NYSIIS first and surname | 0.0126 | 0.253 | 0.695\*\*\* |
|  | (0.196) | (0.171) | (0.142) |
| Middle initial match, if have one | 1.212\*\*\* | 1.068\*\*\* | 0.761\*\* |
|  | (0.103) | (0.201) | (0.328) |
| NYSIIS last name match and YOB Diff=0 | 1.131\*\*\* | 4.537 | 0.212 |
|  | (0.234) | (126.1) | (0.177) |
| NYSIIS last name match and YOB Diff=1 | 1.066\*\*\* | 4.175 | 0.223 |
|  | (0.243) | (126.1) | (0.170) |
| NYSIIS last name match and YOB Diff=2 | 0.795\*\*\* | 3.757 | -0.280 |
|  | (0.255) | (126.1) | (0.183) |
| 2 potential links with NYSIIS last name match | -0.308\*\* | -0.951\*\* | -0.485\*\* |
|  | (0.147) | (0.407) | (0.197) |
| >2 potential links with NYSIIS last name match | -0.637\*\*\* | -3.921 | 0.0587 |
|  | (0.224) | (126.1) | (0.158) |
| 2 potential links with last name string match | -1.090\*\*\* | -0.682\*\* | -0.681\*\*\* |
|  | (0.181) | (0.273) | (0.182) |
| >2 potential links with last name string match | -1.575\*\*\* | -1.194\*\*\* | -0.816\*\*\* |
|  | (0.123) | (0.162) | (0.116) |
| One potential link | 0.398\*\* | 0.407\*\*\* | 0.444\*\*\* |
|  | (0.173) | (0.143) | (0.115) |
| Constant | 1.370\*\*\* | 0.318 | 0.795\*\*\* |
|  | (0.167) | (0.233) | (0.181) |
|  |  |  |  |
| *Observations* | 16,248 | 9,695 | 10,292 |

Sources: 1910 and 1940 Censuses (Ruggles et al. 2019).

Notes: This table shows a regression of whether one is a true link on observable characteristics. Data set is the set of potential matches in 1940 for a random sample of 2,000 individuals in 1910 with one potential link.

After modeling the hand-linking process, we calculate the predicted linking score for each potential link for the entire match forward from 1910 to 1940. Given this information, we need to make a decision for who to keep in the final linked datasets based on the predicted probabilities. We run a grid search over our training data in order to let the minimum positive predictive value to be 0.90. The positive predictive value is the ratio of true links over total links; that is, it may be that 10 percent of our linked data is to the wrong father. While this is cause for concern, Bailey et al. (2019) show that intergenerational elasticity estimates are not strongly biased by false links when using the Feigenbaum method since false links tend to have similar characteristics as true links. We later directly show with the third-generation data that a machine-linked version produces similar mobility estimates as the hand-linked version.

Based on a minimum PPV level of 0.90, the true positive rate (TPR) varies across the non-Mexican white, second-generation Mexican and African American population (see Table B3). The TPR is the share of true links in our hand-linked data that we actually categorize as a link; that is, it measures the efficiency of the linking method. The TPR rates are not very high: 0.790 for non-Mexican white Americans, 0.641 for second-generation Mexicans and 0.580 for African Americans. It is unclear why the model performs worse for African Americans. Nevertheless, a TPR rate of 0.79 or 0.58 indicates that we will miss a lot of true links based on predicted match scores from the probit model; however, we are unconcerned about this problem since we are linking full to full count censuses and therefore can afford a lower efficiency rate.

Table B3. Critical values for inclusion into linked sample

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cutoff Predicted probability | Ratio of first to second best link | PPV | TPR |
| Non-Mexican white American | 0.335 | 2.6 | 0.901 | 0.790 |
| Non-Mexican black  | 0.784 | 5.8 | 0.901 | 0.580 |
| Second-gen Mex. American | 0.506 | 4.9 | 0.900 | 0.641 |

Sources: Authors’ calculations from 1910-1940 linked data (Ruggles et al. 2019).

Notes: This table shows a regression of whether one is a true link on observable characteristics. Data set is the set of potential matches in 1940 for a random sample of 2,000 individuals in 1910 with one potential link.

 For the set of 1940 potential matches for each individual in 1910, we keep the highest score match that also meets the cutoffs in Table B3. Even with this process, we may two separate individuals in 1910 to the same 1940 person; if this occurs at all, we drop both from the dataset. This also lowers the general efficiency of the linking process. In the end, the set of individuals we are able to link are shown in Table B4, with final linking rates of 11.9 percent for African Americans, 14.0 percent for second-generation Mexican Americans, 29.8 percent for non-Mexican white Americans, and 31.5 percent for the (hand-linked) third-generation Mexican Americans. A major reason for the lower African American linking rate is because we fail to find a potential link for 40 percent of the African American children in 1910, while we fail to find a potential link for only 19 percent of non-Mexican whites. For those who wish to improve the linking rates for African Americans, future research may want to relax the linking criteria for African Americans either by allowing wider string differences, year of birth differences, or state of birth differences. However, this is not the only reason for a lower linking rate for African Americans; even when we find a potential match for either non-Mexican white or black Americans, the cut-off values (in Table B3) are more restrictive for African Americans, leading to a lower linking rate.

Table B4. Applying hand-linking results to full 1910-1940 link, details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   |  Non-Mexican white American | Third-gen. Mexican | Second-gen. Mexican | Non-Mexican black Americans |
| Starting group in 1910 | 12,567,861 | 18,656 | 48,714 | 1,851,076 |
| Starting group in 1910 with a potential link in 1940 based on criteria | 10,180,244 | 16,845 | 28,857 | 1,094,394 |
| Potential links in 1940 | 136,372,727 | 234,279 | 231,158 | 6,085,262 |
| Linked based on predicted match scores | 4,576,067 | 5,875 | 7,698 | 284,615 |
| Unique match in 1940 amongst links | 3,748,917 | 5,875 | 6,839 | 220,145 |
| Linking rate | 29.8 | 31.5 | 14.0 | 11.9 |
| Linking rate given potential match | 36.8 | 34.9 | 23.7 | 20.1 |
| Kept in main sample (observe grandparents’ COB, father’s outcomes, has occupation) |  3,235,260 | 5,088 | 5,381 | 161,820  |

Sources: 1910 and 1940 censuses (Ruggles et al., 2019)

Notes: We hand-linked third-generation Mexican Americans, so we do not use predicted match scores for this group.

The final step prior to running analysis is to apply proper weights. Selection into the linked sample is not random; rather, those with uncommon names are more likely to be included in the linked dataset, as well as those in less populated states. Following the recommendation of Bailey et al. (2019), we use inverse probability weighting. To do this, we first append the base dataset in 1910 with our linked dataset of 1910 children; then we run a probit where the dependent variable is a successful link on a variety of observable characteristics such as length of names, whether the father is literature, whether one lives in an urban area, state of birth fixed effects and other measures shown in Table B5. The linked samples are indeed biased, including the hand-linked sample; for example, those with literate, farmer or white-collar fathers are more likely to be linked. For Mexican Americans, we include a measure for whether one has a Spanish-sounding last name, which actually *increases* the probability of link (conditional on other observables). After running this probit, we then use the formula for weights from Bailey et al. (2019) in our main analysis.[[1]](#footnote-1),[[2]](#footnote-2) After weighting, the linked samples are more representative on observables (see Table B5).

Table B5. Weighting the linked sample based on 1910 observables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | non-Mexican white American | Second-gen Mexican | Third-Generation Mexican | African American |
|   |   |   |   |   |   |   |   |   |
| Weighted | No | Yes | No | Yes | No | Yes | No | Yes |
|  |  |  |  |  |  |  |  |  |
| Length of last name | 0.0184\*\*\* | 0.00197\*\*\* | 0.0137\*\*\* | 0.00120 | 0.0100\* | 0.00188 | -0.00128 | -0.00158\* |
|  | (0.000216) | (0.000228) | (0.00451) | (0.00488) | (0.00542) | (0.00564) | (0.000823) | (0.000866) |
| Length of first name | 0.0144\*\*\* | 0.000287 | 0.0195\*\*\* | 0.00396 | 0.00876 | 0.00137 | 0.00756\*\*\* | 0.000356 |
|  | (0.000216) | (0.000229) | (0.00520) | (0.00597) | (0.00599) | (0.00642) | (0.000884) | (0.000932) |
| Father is literate | 0.145\*\*\* | 0.00780\*\*\* | 0.171\*\*\* | 0.00698 | 0.132\*\*\* | 0.0115 | 0.0725\*\*\* | 0.000378 |
|  | (0.00160) | (0.00174) | (0.0160) | (0.0176) | (0.0212) | (0.0225) | (0.00293) | (0.00309) |
| Urban | -0.0415\*\*\* | -0.00458\*\*\* | 0.00593 | 0.00976 | 0.0151 | 0.00740 | 0.0143\*\*\* | 0.00128 |
|  | (0.000989) | (0.00106) | (0.0179) | (0.0197) | (0.0231) | (0.0243) | (0.00429) | (0.00449) |
| Father has white collar job | 0.0444\*\*\* | 0.000860 | 0.0470 | -0.00571 | 0.0499 | 0.0107 | 0.0380\*\*\* | 0.00340 |
|  | (0.00123) | (0.00130) | (0.0340) | (0.0359) | (0.0410) | (0.0424) | (0.0104) | (0.0108) |
| Father is farmer | 0.0271\*\*\* | 0.000146 | -0.0467 | -0.0198 | -0.00956 | 0.00693 | 0.0519\*\*\* | -0.00155 |
|  | (0.00121) | (0.00128) | (0.0295) | (0.0320) | (0.0360) | (0.0374) | (0.00686) | (0.00722) |
| Father is unskilled | -0.0534\*\*\* | -0.00271\*\* | -0.0816\*\*\* | -0.00154 | -0.0716\*\* | 0.00639 | -0.0139\*\* | -0.00483 |
|  | (0.00112) | (0.00120) | (0.0254) | (0.0273) | (0.0328) | (0.0341) | (0.00668) | (0.00703) |
| Age | -0.000940\*\*\* | 0.000243 | 0.00665 | 0.000276 | 0.00945 | 9.31e-05 | -0.0118\*\*\* | -0.00167 |
|  | (0.000310) | (0.000330) | (0.00623) | (0.00694) | (0.00764) | (0.00807) | (0.00114) | (0.00120) |
| Age squared | 0.000188\*\*\* | 4.13e-06 | 0.000215 | 4.75e-05 | -0.000726 | 2.00e-06 | 0.000707\*\*\* | 0.000114 |
|  | (2.19e-05) | (2.33e-05) | (0.000454) | (0.000503) | (0.000559) | (0.000591) | (8.03e-05) | (8.45e-05) |
| Has a Spanish-sounding last name |  |  | 0.294\*\*\* | 0.0321\* | 0.212\*\*\* | 0.0157 |  |  |
|  |  |  | (0.0163) | (0.0182) | (0.0193) | (0.0204) |  |  |
| State of birth FE | X | X | X | X | X | X | X | X |
|  |  |  |  |  |  |  |  |  |
| *Observations* | 14,778,857 | 14,778,857 | 47,433 | 47,433 | 23,563 | 23,563 | 1,565,782 | 1,565,782 |

Source: 1910 Census (Ruggles et al. 2019).

Note: The underlying data is the linked dataset appended with dataset of potential links in 1910. We predict the characteristics of the link, and then reweight the linked dataset using inverse probability weights.

*Hand-linked versus Machine-linked data for third-generation Mexican Americans*

A potential concern about the estimated differences across groups is that they might reflect differences in the linking method rather than a true mobility gap. It is possible to “fix” the linking methodology by creating a machine-linked version of the hand-linked data. We mimic the process when creating the other datasets: we take a subsample of 2,000 from the set of potential links, model the hand-linking process with a probit, and then use the probit to create linking scores for the rest of the dataset. We then keep only those who both have a high linking score and have a linking score that is sufficiently higher than the second-best score.[[3]](#footnote-3) Table B6 shows the results from the probit model. We also use the same inverse proportional weighted process as for the other machine-linked dataset.

 Figure B1 shows how the machine-linked data overlap with the hand-linked data. While the hand-linked data contain 5,088 third-generation Mexican Americans, the machine-linked data contain only 3,398. Out of these 3,398 individuals, 2,959 are also in the hand-linked dataset while 439 are not. Therefore, the machine-linked data contain 58 percent of the hand links, but add an additional 15 percent of false links. Since false links exist, one may expect that the rank-rank slope will be attenuated in the machine-linked data relative to the hand-linked data. Further, the set of links that are both in the hand-linked and machine-linked data may have a different rank-rank slope than the overall hand-linked data, perhaps because those who are machine-linked are a select subgroup.

 First, we test whether the mobility estimates in the hand-linked data differ from the machine-linked data. In Column I of Table B6, we recreate our main estimate for rank-rank mobility of third-generation Mexican Americans with the hand-linked data, which leads to an intercept of 11.3 and a slope of 0.417. Column II estimates the same relationship with the machine-linked data and finds a similar intercept of 11.7 and slope of 0.401. Column III pools the machine-linked and hand-linked sample together and estimates a fully interacted model where a dummy variable for machine linked is included to test the difference in intercepts, and an interaction of machine linked and father rank is included to test the difference in slopes. The test shows that there is no statistical difference in intercept or slope across the machine-linked and hand-linked data.

 In Columns IV through VI, we further examine differences in mobility estimates across the hand-linked data and machine-linked data. Column IV limits the same to false positives, or those who are in the machine linked data but not in the hand-linked data set. The rank-rank slope for false positives is 0.299. Therefore, even though the wrong father is attached to the son, the false positives have similar enough characteristics to the true positives that a positive slope is found between the (wrong) father and the son. This result is similar to that found when using the Feigenbaum method for the LIFE-M data (Bailey et al., 2019). At the same time, Column VI shows that the slope for false positives is statistically weaker than the slope for those who are found in both the hand-linked data and machine-linked data.

Table B6.Probit model applied to third-generation Mexican Americans

|  |  |
| --- | --- |
|   | Third- Gen Mex.  |
| Jaro-Winkler distance, first name | -5.983\*\*\* |
|  | (0.394) |
| Jaro-Winkler distance, surname | -11.80\*\*\* |
|  | (0.729) |
| Year of birth difference = 1 | -0.499\*\*\* |
|  | (0.115) |
| Year of birth difference = 2 | -0.887\*\*\* |
|  | (0.138) |
| Year of birth difference = 3 | -1.117\*\*\* |
|  | (0.149) |
| Number of potential links | 0.00516 |
|  | (0.0225) |
| Number of potential links squared | -0.00151\*\* |
|  | (0.000722) |
| Exact surname match and unique surname | 0.473\*\* |
|  | (0.215) |
| Exact first and surname string match and unique first and surname string | 0.209 |
|  | (0.151) |
| Exact first name match and unique first name | -0.0159 |
|  | (0.194) |
| Exact Soundex first name match and unique soundex first name | 0.539\*\*\* |
|  | (0.206) |
| Exact Soundex surname match and unique soundex surname | 0.287\* |
|  | (0.149) |
| Exact Soundex first and surname match and unique soundex first and surname | 1.054\*\*\* |
|  | (0.127) |
| Exact NYSIIS first name match and unique NYSIIS first name | 0.0648 |
|  | (0.227) |
| Exact NYSIIS surname match and unique NYSIIS surname | 0.643\*\*\* |
|  | (0.235) |
| Exact NYSIIS first and surname match and unique NYSIIS first and surname | 1.314\*\*\* |
|  | (0.139) |
| Middle initial match, if have one | - |
|  |  |
| NYSIIS last name match and YOB Diff=0 | 0.207 |
|  | (0.158) |
| NYSIIS last name match and YOB Diff=1 | 0.559\*\*\* |
|  | (0.158) |
| NYSIIS last name match and YOB Diff=2 | 0.00700 |
|  | (0.181) |
| 2 potential links with NYSIIS last name match | -0.349\*\* |
|  | (0.153) |
| >2 potential links with NYSIIS last name match | -0.280\*\* |
|  | (0.141) |
| 2 potential links with last name string match | -0.712\*\*\* |
|  | (0.171) |
| >2 potential links with last name string match | -0.977\*\*\* |
|  | (0.105) |
| One potential link | 0.507\*\* |
|  | (0.210) |
| Constant | 0.996\*\*\* |
|  | (0.202) |
|  |  |
| *Observations* | 28,102 |

Sources: 1910 and 1940 censuses (Ruggles et al. 2019)

Notes: This table recreates Table B2, but for a subsample from the hand-linked data of third generation Mexican Americans.

Figure B1. Overlap of the hand-linked data and the machine-linked data

****

Sources: Linked samples of the 1910 and 1940 censuses (Ruggles et al. 2019)

Table B7.Mobility differences across hand-linked (HL) and machine-linked (ML) data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | I | II | III | IV | V | VI |
| Sample: | HL | ML | Pool HL and ML | ML and not HL | ML and HL | ML |
|   |   |   |   |   |   |   |
| Percentile rank of father | 0.417\*\*\* | 0.401\*\*\* | 0.417\*\*\* | 0.299\*\*\* | 0.412\*\*\* | 0.411\*\*\* |
|  | (0.013) | (0.015) | (0.013) | (0.043) | (0.016) | (0.016) |
| Machine linked |  |  | 0.294 |  |  |  |
|  |  |  | (0.554) |  |  |  |
| Machine linked x perc. rank of father |  |  | -0.017 |  |  |  |
|  |  |  | (0.019) |  |  |  |
| False positive |  |  |  |  |  | 1.574 |
|  |  |  |  |  |  | (1.228) |
| False positive x perc. rank of father |  |  |  |  |  | -0.113\*\* |
|  |  |  |  |  |  | (0.046) |
| Constant | 11.34\*\*\* | 11.72\*\*\* | 11.34\*\*\* | 13.11\*\*\* | 11.52\*\*\* | 11.52\*\*\* |
|  | (0.347) | (0.434) | (0.347) | (1.144) | (0.469) | (0.469) |
|  |  |  |  |  |  |  |
| *Observations* | 5,088 | 3,398 | 8,486 | 439 | 2,959 | 3,398 |
| *R-squared* | 0.208 | 0.193 | 0.203 | 0.092 | 0.207 | 0.195 |

Sources: Linked data between the 1910 and 1940 Censuses (Ruggles et al. 2019)

Notes: HL stands for hand-linked data. ML stand for machine-linked data. (See Appendix Figure B1 for Venn Diagram of samples.) False positives are those who are in the ML data but not the HL data.

APPENDIX C – CREATING INCOME SCORES

Due to the limitations of the census data prior to 1940, we only observe occupation rather than income or wages. A standard method to estimate economic differences across groups is to use the *occscore* variable from IPUMS, which reflects median earnings by occupation in 1950. Instead of using this score, we create a more detailed score that aims to capture earnings differences within occupation across race/ethnicity; otherwise, we may misstate differences in economic status across Mexican, African and non-Mexican white Americans. Here we give further details on how we created the income score; much of our strategy closely follows the decisions made by Collins and Wanamaker (referred to as “CW”, 2017) to create their income score for black and white Americans.

 First, we take the 1940 Census and separate 25-55 year-old males into different groups: non-Mexican white American, African Americans, and first and second-generation Mexicans. Note that the second-generation Mexican Americans are from the sample-line individuals in the full count data. The income score then is essentially the average earnings for each detailed occupation at the 3-digit level, race/ethnic group and region (based on the 4 census regions).

 The average income forms the basis of our income score, but we make a few further corrections that follow the work of CW. First, we scale up income for those who are self-employed in 1940 based on the ratio of total income for self-employed workers to wage workers in the 1960 census – the first census with a large sample that reports wage, business and farm income. This approximates missing business or farm income in the 1940 census. To do this, we first assign self-employed workers the average wage income by occupation, race/ethnic group and region; then we multiply this average income by factor to scale up. The final income score that averages by occupation includes this scaled up income for self-employed workers.

 Second, we scale up farm laborer wage income by 26 percent to account for perquisites that are not reflected in wage income. This is based on information from a 1939 USDA report, as discussed by CW. Third, we estimate farmer income for both farm owners and farmer tenants. To do this, we first take the ratio of farm laborer and farmer income in the 1960 census, and assume this ratio is the same in 1940.[[4]](#footnote-4) Then, we multiply this ratio by the (scaled-up) farm laborer income in 1940 to get an estimate for farmer income.

 After these adjustments, we take the average age-adjusted earnings by occupation/region/group cell. If there are less than 30 people in an occupation/region/group cell, then we replace it with the national average at the occupation/group. If there are less than 30 people in this cell, we replace it with the national average at the 1-digit level for each group. We then apply the income scores for non-Mexican white Americans for all years (1880, 1910 and 1940), the scores for African Americans for all years, for first-generation Mexican Americans in 1880, and second/third generation Mexican Americans in 1910 and 1940.

APPENDIX D – ACCOUNTING FOR MEASUREMENT ERROR IN PERMANENT ECONOMIC STATUS

 Recent evidence suggests that one observation of the father’s occupation does not perfectly capture his permanent economic status, which leads to measurement error attenuating intergenerational associations (Ward 2019). Measurement error could also exist for the son’s outcome as well since we use percentile rank outcomes (Nybom and Stuhler 2017).[[5]](#footnote-5) In effect, measurement error causes us to overstate upward mobility for children from poorer households and overstate downward mobility for children from richer households. It is unclear how measurement error biases our estimates of race/ethnic mobility gaps since all groups would be influenced by error.

 To determine whether measurement error influences our race/ethnic mobility gap estimates, we further link the 1910 fathers to a second observation in the 1920 census. We do so by using the 1910-1920 linked dataset detailed in Ward (2020b), which is also based on the Feigenbaum method. After linking the 1910 fathers to a second observation in 1920, we are left with a subsample of 1,012,336 non-Mexican white sons and fathers (31.2 percent of 1910-1940 link), 26,305 black sons and fathers (16.2 percent), and 2,305 Mexican sons and fathers (22.0 percent). The Mexican group is split between 1,163 second-generation Mexicans (21.6 percent) and 1,142 third-generation Mexicans (22.4 percent). Therefore, connecting to a second observation loses a significant portion of the dataset, as expected.

 In Table D1, we first estimate whether the double-linked sample yields different mobility gaps relative to our main single-linked sample. That is, if the change in representativeness matters. In the double-linked sample (but still using only the 1910-1940 observations), we find that Mexican sons are expected to end up 22.7 percentiles lower than non-Mexican white sons (conditional on the father’s rank). This gap is similar in magnitude to the 22.6 percentile gap for the single-linked sample. However, the double-linked sample finds a smaller black-white gap of 26.7 percentiles compared to the single-linked sample of 31.0 percentiles. This difference suggests that the double-linking process captures more upwardly mobile black families.

 The key question in this section is whether the mobility gap estimates change after accounting for measurement error. In Column III, we use an alternative measure of father’s economic status based on the average income score across the 1910 and 1920 Censuses. (Note that we average the 1910 and 1920 scores and then rank fathers.) Averaging should reduce measurement error and lead to a steeper rank-rank slope. Indeed, the rank-rank slope based on this revised positioning of the father increases from 0.38 to 0.43. Interestingly, the mobility gaps *shrink* after this correction for measurement error. For example, the updated estimate suggests that Mexicans are expected to end up 19.6 percentiles lower than non-Mexican whites, which is 86 percent of the gap estimated with our main sample (22.7 percentiles). When one more properly places children in the economic distribution, then those raised in poorer households had less upward mobility, which leads to a smaller gap across race/ethnic groups. Likewise, the non-Mexican white and black gap falls from 26.7 to 23.0 percentiles.

 Averaging two father occupations may not fully account for measurement error if both observations contain error (Ward 2019). One way to address measurement error is to use an instrumental variables strategy where one father observation is instrumented with a second (Modalsli and Vosters 2019). The assumption for this strategy is that the error components are uncorrelated. One advantage of this method is that it also addresses measurement error in the son’s percentile rank, under the assumption that the error in the son’s rank is of the same magnitude as the error in the father’s (Nybom and Stuhler 2017, footnote 13). Column IV reports the results after instrumenting the 1910 father rank with the 1920 father rank. Based on this specification, the rank-rank slope increases to 0.66, which is much higher than the original rank-rank slope of 0.38. Due to the steeper slope, we now predict that children raised in poorer households also ended up poor; likewise, children raised in richer households also ended up rich (in percentile rank terms). A consequence of the steeper slope is that mobility gaps narrow for both Mexican and African Americans, to about 13.0 percentiles. Therefore, the Mexican and non-Mexican white mobility gap, after accounting for measurement error, is about 57 percent of the original mobility gap with one father observation. While the size of the mobility gap narrows, our main argument holds: Mexican Americans had an intergenerational disadvantage relative to non-Mexican white children raised in the same percentile household.

Table D1. Mobility Gaps after accounting for measurement error

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | I | II | III | IV |
|   | OLS | OLS | OLS | 2SLS |
|   |   |   |   |   |
| Sample | Single linked (1910-1940) | Double linked (1910-1920-1940) | Double linked (1910-1920-1940) | Double linked (1910-1920-1940) |
|  |  |  |  |  |
| African American | -30.973 | -25.994 | -23.017 | -12.963 |
|  | (0.049) | (0.092) | (0.091) | (0.115) |
| Mexican American | -22.583 | -22.741 | -19.646 | -13.023 |
|  | (0.188) | (0.532) | (0.520) | (0.557) |
| Percentile rank of father, 1910 observation | 0.357 | 0.378 |  | 0.658 |
|  | (0.000) | (0.001) |  | (0.002) |
| Percentile rank of father,  |  | 0.430 |  |
|  average of 1910 and 1920 observations |  |  | (0.001) |  |
| Constant | 35.694 | 33.610 | 30.837 | 18.166 |
|  | (0.030) | (0.057) | (0.056) | (0.096) |
|  |  |  |  |  |
| *Observations* | 3,407,493 | 1,040,946 | 1,040,946 | 1,040,946 |
| *R-squared* | 0.328 | 0.303 | 0.334 | 0.243 |

Sources: 1910-1920-1940 linked data (Ruggles et al. 2019).

Notes: The first column are the main estimates from the paper. The second column is the same specification as the first column, but restricts the sample be double linked (sons from 1910 to 1940 and fathers from 1910 to 1920). The third column uses the averages the father outcomes in 1910 and 1920 before ranking. The fourth coumn instruments the 1910 father observation with the 1920 father observation. See Ward (2019) for more detail on measurement error in historical intergenerational associations.

1. Let *p* be the predicted probability, and *q* be the share of linked individuals over the total linkable. The weight we use is ((1-*p)/p)\*(q/(1-q))*. See Bailey et al. (2019). [↑](#footnote-ref-1)
2. The Spanish surname variable is not available in the 1910 Census, so we link people’s last names with the list of Spanish surnames, as taken from the 1940 full-count census when the variable is available. [↑](#footnote-ref-2)
3. The cutoff for predicted probability is 0.531. The ratio of the 1st to 2nd best linking score needs to be at least 3.2. The resulting positive predictive value is 0.901 and the true positive rate is 0.581. [↑](#footnote-ref-3)
4. Once again following CW (2017), in the 1960 census we scale up farmer income by 35 percent in 1960 to reflect perquisites, and we also scale up farm laborer income by 19 percent. [↑](#footnote-ref-4)
5. Classical measurement error in the son’s outcome does not bias intergenerational elasticity estimates since it is measurement error in the dependent variable. However, non-classical measurement error in the son’s outcome may bias results due to life-cycle bias, an issue we address by only using sons in the middle of the lifecycle (Haider and Solon 2006; Nybom and Stuhler 2017). [↑](#footnote-ref-5)