Online Appendix
The Not-So-Hot Melting Pot: The Persistence of Outcomes for Descendants of the Age of Mass Migration

Zachary Ward*

TABLE A1. COMPARISON OF ETHNIC DIFFERENTIAL ELASTICITIES FROM GROUP-AVERAGED DATA AND MICRODATA.

	Farmer	White Collar	Skilled	Unskilled	ln(Occ. Score), 1890-1950	ln(Occ. Score), 1950
Panel A: Relationship between Second Generation in 1910 and First Generation in 1880						
Mean Convergence (Θ_1),	0.842	0.306	0.366	0.464	0.888	0.976
collapsed at ethnic level	(0.118)	(0.275)	(0.415)	(0.061)	(0.209)	(0.236)
Mean Convergence ($\beta_1+\beta_2$),	0.861	0.328	0.354	0.46	0.920	1.019
from microdata	(0.017)	(0.052)	(0.051)	(0.019)	(0.031)	(0.032)
Panel B: Relationship between Third Generation in 1940 and Second Generation in 1910						
Mean Convergence (Θ_1),	0.465	1.198	0.290	0.334	0.566	0.736
collapsed at ethnic level	(0.027)	(0.282)	(0.066)	(0.160)	(0.038)	(0.048)
Mean Convergence ($\beta_1+\beta_2$),	0.482	1.206	0.314	0.330	0.562	0.737
from microdata	(0.012)	(0.067)	(0.043)	(0.036)	(0.019)	(0.025)
Panel C: Relationship between Third Generation in 1940 and First Generation in 1880						
Mean Convergence (Θ_1),	0.373	0.110	0.197	0.090	0.512	0.739
collapsed at ethnic level	(0.074)	(0.389)	(0.108)	(0.058)	(0.110)	(0.156)
Mean Convergence ($\beta_1+\beta_2$),	0.390	0.128	0.185	0.092	0.508	0.734
from microdata	(0.013)	(0.058)	(0.044)	(0.018)	(0.023)	(0.032)

Notes: Data are from the 1880-1910-1940 linked census data. Each cell is from a different regression that estimates how ethnic differentials converge. The "Mean convergence (Θ_1), collapsed at ethnic level" reports regressions after collapsing the data by ethnicity. This regression is weighted by the number of individuals in the ethnicity. The **"**Mean Convergence ($\beta_1+\beta_2$), from microdata" are the main estimates from the paper. Estimates may differ for a variety of reasons, one of which is that the micro-data includes quartics for age of the son, age of the father, and age of the father interacted with the outcome variable (when normalized to age 40).

*****Ward: Baylor University, Department of Economics, 1621 S 3rd St, Hankamer School of Business, Waco, TX 76710 (email: zachary_ward@baylor.edu).

TABLE A2. PERSISTENCE FROM THE FIRST GENERATION TO THIRD GENERATION, ONLY EUROPEAN SOURCES

	Farmer		White-Collar	
Grandfather's Outcome (β_1)	0.231	0.222	0.201	0.204
	(0.005)	(0.005)	(0.011)	(0.011)
Ethnic Mean (β_2)		0.167		-0.265
		(0.013)		(0.060)
Mean Convergence ($\beta_1+\beta_2$)		0.389		-0.062
		(0.013)		(0.061)
	Semi-Skilled		Unskilled	
Grandfather's Outcome (β_1)	0.031	0.031	0.037	0.036
	(0.007)	(0.007)	(0.007)	(0.007)
Ethnic Mean (β_2)		0.110		0.021
		(0.044)		(0.017)
Mean Convergence ($\beta_1+\beta_2$)		0.140		0.057
		(0.045)		(0.018)
	ln(Occ. Sc.), 1890-1950		ln(Occ. Sc.), 1950	
Grandfather's Outcome (β_1)	0.215	0.207	0.285	0.274
	(0.005)	(0.006)	(0.008)	(0.008)
Ethnic Mean (β_2)		0.260		0.409
		(0.022)		(0.031)
Mean Convergence ($\beta_1+\beta_2$)		0.468		0.683
		(0.022)		(0.032)

Notes: Data is 1880-1910-1940 links. The table recreates Table 4, but I limit the data to only Europeans.

TABLE A3. ALTERNATIVE SPECIFICATIONS FOR PERSISTENCE FROM FIRST GENERATION IN 1880 TO THIRD GENERATION IN 1940

		Grandparental Outcome		Ethnic Mean		Sum		Observations
Row	Alternative Specification	Coef	SE	Coef	SE	Coef	SE	
1	Base sample	0.206	(0.005)	0.303	(0.022)	0.508	(0.023)	96,726
	Occupational Categories							
2	Farmer (owner or operator)	0.216	(0.005)	0.173	(0.013)	0.390	(0.013)	96,726
3	Professional	0.134	(0.010)	-0.057	(0.052)	0.076	(0.053)	96,726
4	Sales or Clerical	0.062	(0.020)	0.314	(0.168)	0.377	(0.168)	96,726
5	Craftsmen	0.030	(0.007)	0.155	(0.043)	0.185	(0.044)	96,726
6	Low-skilled Service Worker	-0.003	(0.011)	0.880	(0.140)	0.877	(0.139)	96,726
7	General or farm laborer	0.008	(0.007)	-0.004	(0.022)	0.003	(0.023)	96,726
8	Operative	0.036	(0.007)	0.164	(0.026)	0.200	(0.027)	96,726
	Alternative Measures of Status							
9	Education (years) in 1940, Literacy in 1880	0.693	(0.209)	13.844	(2.758)	14.53	(2.750)	95,152
10	Wage income in 1940, Occupation income score in 1880	0.198	(0.013)	0.347	(0.057)	0.545	(0.057)	68,912
11	Income score from 1940 Census, allows for regional variation	0.270	(0.009)	0.552	(0.058)	0.822	(0.058)	96,726
12	1901 CLS in 1880, 1910 and 1950 occscore in 1940	0.129	(0.008)	0.141	(0.039)	0.270	(0.040)	96,726
13	1901 CLS in 1880, 1910 and 1940	0.125	(0.007)	0.123	(0.036)	0.248	(0.037)	96,726

Notes: Data is from 1880-1910-1940 link. Row 2 are occ1950 codes starting with 1; Row 3 are occ1950 codes starting with 0 or 2; Row 4 are occ1950 codes starting with 3 or 4; Row 5 are occ1950 codes starting with 5; Row 6 are occ1950 codes starting with 7; Row 7 are codes starting with 8 or 9 (excluding non-occupational responses); Row 8 are occ1950 codes starting with 6. Row 10 uses wage income for wage workers in 1940, and main occupational score in 1880. Row 11 uses mean income by occupation and census region from 1940 census, following Collins and Wanamaker's fix for farmer earnings (2017). Row 12 uses 1901 Cost of Living Survey in 1880 and the 1950 occupational score in 1950. However, the CLS does not estimate farmer or farm laborer income. In 1880, I use the farmer income estimates by state created by Ager, Boustan and Eriksson (2019). These are created from the 1880 Census of Agriculture on revenue from output and costs from labor, taxes, fertilizer and maintenance. Assumptions for calculating earnings (subtracting costs from revenue) are given in Online Appendix Table 3 in Abramitzky, Boustan and Eriksson (2012). Also, I use the farm laborer estimates by state as given by Ager, Boustan and Eriksson (2019) and originally found in the Young report (1871). Row 13 uses 1901 Cost of Living Survey in 1880 and 1940. I use the 1880 farmer and farm laborer estimates by state throughout the period as in row 13.

TABLE A4. CORRELATION WITH ETHNIC MEAN FOR OCCUPATIONAL CATEGORIES

	Farmer		White Collar	
Grandfather Outcome	0.216	0.119	0.205	0.138
	(0.005)	(0.006)	(0.010)	(0.011)
Ethnic mean	0.173	0.065	-0.077	-0.106
	(0.013)	(0.015)	(0.057)	(0.073)
Mean Convergence	0.390	0.184	0.128	0.031
	(0.013)	(0.017)	(0.058)	(0.073)
1880 Neighborhood FE	N	Y	N	Y
	Semi-Skilled		Unskilled	
Grandfather Outcome	0.030	0.022	0.042	0.043
	(0.007)	(0.008)	(0.007)	(0.008)
Ethnic mean	0.155	-0.020	0.050	0.029
	(0.043)	(0.061)	(0.017)	(0.023)
Mean Convergence	0.185	0.002	0.092	0.072
	(0.044)	(0.062)	(0.018)	(0.024)
1880 Neighborhood FE	N	Y	N	Y

Notes: Data are from the 1880-1910-1940 linked census data. This table recreates the specification in Table 7, but instead uses occupational categories as the outcome rather than occupational income.

TABLE A5. CORRELATION WITH ETHNIC MEAN WHEN MEASURING THE ETHNIC MEAN AT THE ENUMERATION-DISTRICT LEVEL

	Farmer		White Collar	
Grandfather Outcome	0.112	0.117	0.128	0.131
	(0.006)	(0.007)	(0.011)	(0.012)
Ethnic mean (local level)	0.181	0.022	0.375	0.092
	(0.007)	(0.016)	(0.015)	(0.026)
Mean Convergence	0.293	0.139	0.503	0.222
	(0.005)	(0.015)	(0.016)	(0.027)
1880 Neighborhood FE	N	Y	N	Y
	Semi-Skilled		Unskilled	
Grandfather Outcome	0.014	0.019	0.034	0.040
	(0.007)	(0.008)	(0.007)	(0.008)
Ethnic mean (local level)	0.095	0.031	0.032	0.042
	(0.010)	(0.018)	(0.008)	(0.017)
Mean Convergence	0.109	0.050	0.066	0.082
	(0.011)	(0.019)	(0.009)	(0.017)
1880 Neighborhood FE	N	Y	N	Y
	ln(Occ. Sc.), 1890-1950		ln(Occ. Sc.), 1950	
Grandfather Outcome	0.095	0.094	0.123	0.117
	(0.006)	(0.007)	(0.009)	(0.009)
Ethnic mean (local level)	0.229	0.053	0.324	0.068
	(0.007)	(0.016)	(0.009)	(0.021)
Mean Convergence	0.325	0.147	0.447	0.185
	(0.006)	(0.016)	(0.009)	(0.022)
1880 Neighborhood FE	N	Y	N	Y

Notes: Data are from the 1880-1910-1940 linked census data. This table recreates the specification in Table 7, but instead of measuring the ethnic mean at the national level, instead measures the ethnic mean at the enumeration district level in 1880.

B. Iterating the AR(1) model with ethnic means

A common interest in the multigenerational literature is to compare how an iterated AR(1) model compares with a multigenerational model (e.g., Stuhler, 2012). Here are I interested in how an iterated AR(1) model holds for ethnic differentials, rather than just between grandfather and grandson. It is straightforward to iterate this ethnic-mean model in Equation (3) to predict outcomes for the third generation. Suppose the relationship between first and second generation of immigrants (G1-G2) is as follows:

$$\mathbf{y_{i,c,g-1}} = \delta_0 + \delta_1 \mathbf{y_{i,c,g-2}} + \delta_2 \bar{\mathbf{y}}_{\mathbf{c,g-2}} + \varepsilon_{i,c,g-1} \quad \text{(B1)}$$

and let Equation (3) be the relationship between the second and third generations (G2-G3).[1] After plugging Equations (B1) and (4) into Equation (3) and grouping terms, the relationship between the grandson's skill level, the grandfather's skill level and first-generation ethnic mean is

$$\mathbf{y_{i,c,g}} = c + \beta_1\delta_1 \mathbf{y_{i,c,g-2}} + (\beta_1\delta_2 + \beta_2\delta_1 + \beta_2\delta_2)\bar{\mathbf{y}}_{\mathbf{c,g-2}} + \upsilon_{i,c,g-2} \quad \text{(B2)}$$

where $\beta_1\delta_1$ is the correlation with grandfather's skill and $\beta_1\delta_2 + \beta_2\delta_1 + \beta_2\delta_2$ is the correlation with the ethnic mean in the grandfather's generation.[2] Note that the mean convergence between the first and third generations is the sum of the grandfather and ethnic mean effect $(\beta_1\delta_1 + \beta_1\delta_2 + \beta_2\delta_1 + \beta_2\delta_2)$, which is also the product of mean convergence between the first and second generations $(\beta_1 + \beta_2)$ and second and third generations $(\delta_1 + \delta_2)$ – a reflection of the AR(1) set up.

This modelling assumes an AR(1) process where a generation is only influenced by the immediately prior generation. However, there is growing evidence that a grandson's outcomes are also correlated with the outcome of the grandfather, above and beyond the effect of the father (Mare, 2011; Solon, 2018). This could be due to a variety of theoretical reasons, such as investment from the grandparent into the grandchild or a latent factor that is inherited across generations (Clark, 2014; Solon, 2014). One could make a similar argument for first-generation ethnic mean influencing the grandchild's generation, such as the grandparent's generation

[1] Equations (3) and (5) show that the effect of father's occupation and ethnic mean may differ across generations ($\beta_2 \neq \delta_2$), which may occur if attachment to ethnicity fades or if the second generation are less residentially segregated.

[2] After grouping terms, the constant $c = \beta_0 + \beta_1\delta_0 + \beta_2\delta_0$. The effect of grandfather's skill on grandchild's skill is a familiar term from the multigenerational literature since it is simply the product of the correlation between the grandfather and father (δ_1), and then between father and son (β_1). The effect of the ethnic mean in the grandfather's generation comes from a variety of avenues: first, the product of the ethnic mean effect across generations $(\beta_2\delta_2)$, the indirect effect of ethnic mean in the grandfather's generation and the father's skill level $(\beta_1\delta_2)$, and the indirect effect of ethnic mean in the father's generation and the grandfather's skill level $(\beta_2\delta_1)$.

providing job connections, financial resources, or serving as role models for the grandchild. Therefore, one could extend the ethnic differential model to an AR(2) process:

$$\mathbf{y_{i,c,g}} = \gamma_0 + \gamma_1 \mathbf{y_{i,c,g-1}} + \gamma_2 \mathbf{y_{i,c,g-2}} + \gamma_3 \bar{\mathbf{y}}_{\mathbf{c,g-1}} + \gamma_4 \bar{\mathbf{y}}_{\mathbf{c,g-2}} + \varepsilon_{icg} \quad \text{(B3)}$$

Of particular interest is if the grandfather's occupation or first-generation ethnic mean predicts the grandson's occupation, after controlling for the skill of the father and father's generation. The expected group average for the third-generation Americans from source *c*, abstracting from the constant, would be $(\gamma_1 + \gamma_3)\bar{\mathbf{y}}_{\mathbf{c,g-1}} + (\gamma_2 + \gamma_4)\bar{\mathbf{y}}_{\mathbf{c,g-2}}$. Depending on the coefficients, intergenerational persistence in group averages could converge at a faster or slower rate than that predicted by the AR(1) model.

See Table B1 for the results from Equation (B3). The results show that there is a positive coefficient on both the father's and grandfather's occupational score, which is consistent with a general result in the literature that economic gaps across families converged across three generations at a slower-than-geometric rate (e.g., Long and Ferrie, 2018). At the same time, there is a *positive* coefficient on the second-generation ethnic mean, but a *negative* coefficient on the first-generation ethnic mean. This result suggests that the "ethnic mean effect" converges at a faster-than-geometric rate, perhaps due to social assimilation or a fading attachment to ethnicity across generations. When summing the family and ethnic mean effects together, convergence of ethnic averages in occupational income occurs at a geometric rate. Projecting these results beyond the third generation suggest that ethnic gaps should approach the rate of convergence as family gaps, since ethnic influences fade in importance more quickly. Note that these results are descriptive and that measurement error could be partially due to measurement error.

TABLE B1. AN AR(2) MODEL OF ETHNIC DIFFERENTIALS

	Farmer	White-Collar	Semi-Skilled	Unskilled	ln(Occ. Sc.), 1890-1950	ln(Occ. Sc.), 1950
G2 Father	0.270	0.267	0.105	0.086	0.228	0.268
Outcome	(0.007)	(0.008)	(0.007)	(0.008)	(0.005)	(0.007)
G2 Ethnic	0.360	1.058	0.180	0.875	0.308	0.420
Mean	(0.039)	(0.070)	(0.047)	(0.083)	(0.027)	(0.037)
G1 Grandfather	0.072	0.130	0.012	0.026	0.097	0.152
Outcome	(0.006)	(0.010)	(0.007)	(0.007)	(0.006)	(0.008)
G1 Ethnic	-0.212	-0.404	0.075	-0.380	-0.068	-0.087
Mean	(0.038)	(0.059)	(0.048)	(0.041)	(0.031)	(0.044)
Sum of G2 / G1 parental and ethnic Mean:						
G2 Mean	0.630	1.326	0.284	0.961	0.536	0.687
Convergence	(0.039)	(0.070)	(0.048)	(0.083)	(0.028)	(0.037)
G1 Mean	-0.140	-0.274	0.087	-0.355	0.029	0.065
Convergence	(0.038)	(0.059)	(0.048)	(0.041)	(0.031)	(0.044)

Notes: Data are from the 1880-1910-1940 linked sample. See Equation (B3) for the specification. There are 96,726 observations in each regression. The dependent variable is the third generation (G3) outcomes, which varies across columns. I control for life-cycle effects with a quartic of grandson's age, quartic of father's age, quartic of grandfather's age, quartic of grandson's age interacted with grandfather's outcome, and quartic of grandson's age interacted with father's outcome; these quartics are normalized to age 40. Standard errors are clustered by G1 grandfather.

TABLE B2. COMPARISON OF AR(2) MODELS FROM GROUP-AVERAGED DATA AND MICRODATA.

	Farmer	White Collar	Skilled	Unskilled	ln(Occ. Score), 1890-1950	ln(Occ. Score), 1950
Second-generation Ethnic Mean	0.663	1.326	0.256	0.945	0.551	0.702
collapsed at ethnic level	(0.057)	(0.198)	(0.038)	(0.211)	(0.060)	(0.080)
First-generation Ethnic Mean	-0.185	-0.296	0.103	-0.348	0.023	0.054
collapsed at ethnic level	(0.047)	(0.164)	(0.054)	(0.085)	(0.062)	(0.100)
Second-generation Ethnic Mean	0.630	1.326	0.284	0.961	0.536	0.687
from microdata	(0.039)	(0.070)	(0.048)	(0.083)	(0.028)	(0.037)
First-generation Ethnic Mean	-0.140	-0.274	0.087	-0.355	0.029	0.065
from microdata	(0.038)	(0.059)	(0.048)	(0.041)	(0.031)	(0.044)

Notes: Data are from the 1880-1910-1940 linked census data. Each cell is from a different regression that estimates how ethnic differentials converge over three generations. The first two rows are "collapsed at ethnic level", which report a regression of the third-generation average ethnic outcome on the second-generation and first-generation average ethnic outcome, after collapsing the data by ethnicity. This regression is weighted by the number of individuals in the ethnicity. The last two rows are estimated "from microdata" and are the estimates from found in Table B1. These estimates are from a regression of grandson's outcome on (1) father's occupation (2) grandfather's occupation (3) the second-generation ethnic mean and (4) the first-generation ethnic mean. See Equation B3.

C. Applying the Feigenbaum (2016) method to link censuses

There are many ways one could link censuses, ranging from the extremes of hand-linking the entire dataset to using a fully automated method. I take a mixed approach described by Feigenbaum (2016) where I first hand-link a sample of individuals from 1880 to 1910 and from 1910 to 1940, and then model the hand-linking process to automate the linking process for the rest of the dataset. This approach is less costly than hand linking and is advantageous relative to automated methods by increasing linking rates and decreasing false positives. Most importantly, Bailey et al. (2017) show that the Feigenbaum method reliably produces intergenerational elasticity estimates compared with the higher standard of a hand-linked dataset. Now I will describe the details behind the linking process. I do not believe that the linking method drives the results since I have also used an automated linking strategy which led to the same results.

For each census, I first draw all native-born males under the age of 14 from the 1880 and 1910 full-count censuses. There are 9,595,033 male children under the age of 14 in 1880, and 14,793,768 in 1910. I wish to link these individuals thirty years later to their adult outcomes in either 1910 or 1940. I draw the entire set of boys (rather than just second or third generation immigrants) so that I can drop those with duplicate first name strings, last name strings, race, year of birth, and place of birth.

I then find the set of potential links in the next census in either 1910 or 1940. To manage computational demands, I restrict the set of potential links to meet the following criteria:

(i) Exact match on the first letter of first name

(ii) Exact match on the first letter of last name

(iii) Exact match on race

(iv) Exact match on state of birth

(v) Year of birth difference at most three years

(vi) Jaro-Winkler distance between the first name strings equal to 0.80 or more

(vii) Jaro-Winkler distance between the last name strings equal to 0.80 or more

These are like the criteria outlined by Feigenbaum (2016), except for (1) and (2); Feigenbaum does not require a match on first letter of the first or last name, but not including this restriction makes the process for a full to full-count census match too computationally intensive.

The initial search between the 1880 and 1910 census turns up a result where 6.7 million in 1880 have 121.6 million potential matches in 1910 (see Table C1); for the 1910-1940 match, there are 8.5 million in 1910 who have 142.5 potential matches in 1940. Therefore, there are about 16.6 to 18 potential matches for each person to choose from. After going through the linking process described in the next few paragraphs, I will be able to link 19.9 percent of the original 1880 sons, and 24.7 percent of the original 1910 sons. Now I will describe the linking process in detail.

TABLE C1. LINKING THE 1880-1910 AND 1910-1940 CENSUSES

	1880-1910	1910-1940
Starting population of native-born males ≤14	9,685,804	14,952,246
Drop unidentifiable name strings	9,634,436	14,935,066
Number with at least one potential match that meet criteria	6,754,259	8,572,538
Number of potential matches in ending year that meet criteria	121,621,147	142,470,757
Number of successfully linked based on predicted scores, critical values, and dropping duplicates	1,924,410	3,695,419
Overall linking rate from starting population	19.9%	24.7%
Linking rate given 1 potential match in ending census	28.5%	43.1%

Notes: This table shows the linking rates between the 1880-1910 and 1910-1940 censuses.

From the set of potential matches, I randomly sample 2,000 1880 individuals and 1910 individuals, and keep all their potential links in the second census. I choose which individual is the true link amongst the set of potential links based on closeness in first name, last name and year of birth. If there are two close matches, then I do not choose any as the match. After going through this process, I hand-link about 50.5 percent of the 2,000 in 1880, and 55.9 percent of the 2,000 in 1910. The primary reason why I fail to match 100 percent is because there are no potential link in the second census that has a similar name and year of birth combination.

I then model my hand-linking process with a probit model where I predict which of the potential links is the true link based on observable characteristics. Observable characteristics include the Jaro-Winkler distance in first name, Jaro-Winkler distance in last name, whether the middle initial matches, difference in year of birth, etc. I also rely on variables where one of the potential links both matched and was unique. For example, if "Zach Ward" in 1880 came up with 12 possible hits in 1910, but only one of the possible hits had the surname "Ward," then I term this hit as a unique and exact match. The full probit model is presented in Table C2, which

shows how each observable difference contributes to the prediction for the full match.

TABLE C2. PROBIT MODEL FOR LINKING THE 1880-1910 AND 1910-1940 CENSUSES

	1880-1910	1910-1940
First name Jaro-Winkler distance	-0.859	-4.516***
	(0.627)	(0.560)
Last name Jaro-Winkler distance	-10.98***	-13.16***
	(0.730)	(0.845)
One year of birth difference	-0.260**	-0.568***
	(0.113)	(0.112)
Two years of birth difference	-0.411***	-0.924***
	(0.117)	(0.130)
Three years of birth difference	-0.561***	-1.466***
	(0.126)	(0.155)
Hits	0.0257	-0.110***
	(0.0203)	(0.0185)
Hits squared	-0.00178***	0.00206***
	(0.000670)	(0.000637)
Unique AND exact last name string match	0.00832	0.774***
	(0.230)	(0.239)
Unique AND exact first and last name string match	0.688***	0.396**
	(0.200)	(0.157)
Unique AND exact first name string match	-0.210	-0.417**
	(0.259)	(0.196)
Unique AND exact first name SOUNDEX match	-0.0126	0.279
	(0.336)	(0.265)
Unique AND exact last name SOUNDEX match	0.856***	-0.259
	(0.223)	(0.179)
Unique AND exact first and last name SOUNDEX match	0.191	0.887***
	(0.226)	(0.173)
Unique and exact NYSIIS first name match	0.110	0.273
	(0.359)	(0.271)
Unique and exact NYSIIS last name match	0.580**	0.532*
	(0.294)	(0.303)
Unique and exact NYSIIS first and last name match	0.187	0.0423
	(0.229)	(0.182)
Middle init match, if have one	0.731***	1.141***
	(0.139)	(0.109)
Nysiis Last name match AND year of birth match	0.987***	1.009***
	(0.200)	(0.218)
NYSIIS last name match AND 1 year of birth difference	0.603***	1.032***
	(0.201)	(0.227)
NYSIIS last name match AND 2 year of birth difference	0.596***	0.781***
	(0.198)	(0.240)
1 hit with NYSIIS last name match	-0.487***	-0.340**
	(0.133)	(0.147)
2+ Hits with NYSIIS last name match	-0.700***	-0.573***
	(0.164)	(0.207)
1 hit with exact last name match	-0.337***	-1.155***
	(0.124)	(0.189)

2+ Hits with exact last name string match	-1.109***	-1.524***
	(0.109)	(0.122)
One hit	0.401	0.482***
	(0.258)	(0.167)
Unique and mother's birth place match	-0.112	
	(0.170)	
Unique and father's birth place match	0.745***	
	(0.131)	
Unique and mother and father's birth place match	0.990***	
	(0.169)	
Mother's birth place match	0.775***	
	(0.146)	
Father's birth place match	0.829***	
	(0.127)	
Mother and Father's birth place match	-0.829***	
	(0.170)	
NYSIIS last name and mother's fathers birth palce match	0.242**	
	(0.123)	
Mother's birth place match and non-US	0.201*	
	(0.107)	
Father's birth place match and non-US	0.692***	
	(0.129)	
Constant	-1.463***	1.312***
	(0.207)	(0.165)
Observations	19,041	15,847

Notes: Data is from the training data between the 1880-1910 or 1910-1940 linked datasets.

The probit coefficients yield a predicted probability for each potential link in the complete-count to complete-count match, but it does not tell me who to keep in the linked dataset. To decide who to keep in the match, I must choose a cut off for the predicted probability that one is a true link. I also set a parameter where the first-best link must have a predicted probability that is *b* times the predicted probability of the second-best link; this is to eliminate close matches where I am uncertain which one is the true match. To decide these cut off values, I set the PPV (positive predictive value or the rate of true positives) in my training data to be at least 0.90; in order words, I set the rate of false positives at 0.10. These cut off values are listed in Table C3. After predicting the probabilities, setting the cut off values, I also drop any duplicate match where, for example with the 1880-1910 link, both persons A and B in 1880 were linked to person Z in 1910.

After the linking process is completed, I am left with 1.9 million links between 1880 and 1910, and 3.7 million links between 1910 and 1940 (see Table C1). My linking rate is 19.9 percent of the original 1880 children, and 24.7 percent of the original 1910 children. These rates are less than in Feigenbaum (2016), which may be because linking Iowan children is more efficient than linking non-Iowan children. Indeed, his efficiency rate of 0.88 (ratio of predicted true links to all possible true links) is higher than ours of between 0.69-0.76. Further, I choose a more restrictive positive predictive value of 0.90, while Feigenbaum's (2016, Table 6) rate is more lenient (0.85) (See Table C3).

TABLE C3. STATISTICS FROM THE TRAINING DATA

	1880-1910	1910-1940
Sample of starting year with one potential match	2,000	2,000
Number of potential matches in ending year	19,043	15,890
Successfully linked	1,010	1,118
Linking rate given one potential match	50.5%	55.9%
Predicted probability needs to be above:	0.453	0.363
1^{st}-best predicted score is $\geq$ x times 2^{nd}-best predicted score	3.6	3.4
True positive rate	0.688	0.761
Predictive positive value	0.901	0.901

Notes: This table shows the linking rates for the training data for the 1880-1910 and 1910-1940 censuses. It also gives the tuning parameters to determine who remains in the sample

The actual sample used in the paper is the intersection between the 1880-1910 link and 1910-1940 link, where the son linked from 1880-1910 is also the father on the son linked from 1910-1940. Table C4 shows how I narrow down to my main sample of G1-G2-G3 links, when starting with the G2-G3 links. For example, the interaction of the two datasets is 512,717 (that is, I have 512,717 G1-G2-G3 links). However, since I am only interested in the descendants of immigrants, I drop

native-born G1, which leaves me with 136,420 G1-G2-G3 links. Then I make an age restriction where G1 is between 30 and 55 years old, which leaves me with 108,713G1-G2-G3 links. Then I keep only those with occupations and an observable enumeration district in 1880 since I use enumeration district fixed effects; I also drop those with less than 30 G1 grandfathers. I am left with a final sample of 96,726 G1-G2-G3 links. Based on information from the 1880 sample, this is 4.0 percent of the original population of native-born sons to foreign-born fathers.

TABLE C4. LINKING UP THE 1880-1910 AND 1910-1940 SAMPLE

Linked population of G3 between 1910 and 1940	3,695,419
G3 between 1910-1940 is son of linked G2 1880-1910 (Intersection of linked datasets)	512,717
G3 between 1910-1940 is son of linked G2 1880-1910 AND G1 is foreign-born	136,420
Keep G1 between 30 and 55 years old	108,713
Keep if all G1, G2 and G3 lists an occupation and G1 has observable enumeration district	96,726

The linking algorithm does not successfully find people for a random subset of sons in 1880 to 1910 or 1910 to 1940; rather, a link can only be found if someone has a unique combination of first name, last name, state/country of birth and age. Therefore, in cases where people cannot be distinguished from each other, such as for populous states or countries of birth, it is less likely to find someone. Moreover, if an ethnicity has more common names rather than unique names, it will also be less likely to find someone.

I check the biases of the sample by comparing the fathers of successfully linked sons to other fathers in the full-count censuses. I am interested if there are biases by country of birth, skill level, or region in the United States. Table C5 displays the characteristics of the children who are successfully linked between 1880 and 1910 and the universe of children with foreign-born fathers aged 30-55 (thus matching the restrictions of linked data). There are indeed a few significant differences

between the linked sample and the universe. First, Germans are much more likely to be linked while Irish are much less likely to be linked. It is unclear what is driving this result, but it may be that Irish were from populous states like New York and Massachusetts, which makes it difficult to determine a unique link. Indeed, Northeast residents are less likely to be matched while Midwest residents are more likely to be matched. Further, sons with farmer fathers are more likely to be linked than sons with unskilled fathers. Some of these biases are correlated: for example, having more farmer fathers in the Midwest is consistent with linking Irish at lower rates.

The directions of these biases for origin, occupational group and region are similar for the 1910 and 1940 link (See Table C6). For this representativeness check in 1910, I compare my set of linked third-generation sons with other sons who have similarly aged second-generation fathers in the 1910 census (that is, native-born fathers aged 30-44 with a foreign-born father). Now origin is not defined by grandfather's country of birth. Once again, those with German grandfathers are more likely to show up in my linked sample, while those with Irish grandfathers are less likely. Farmers are overrepresented and unskilled workers are underrepresented; similarly, the Northeast is underrepresented, and the Midwest is overrepresented.

Fortunately, it is relatively straightforward to fix these biases in representativeness by weighting the sample to match the universe's characteristics. Given that I have full-count censuses in both 1880 and 1910, I have plenty of observations to match the distribution in my sample and provide accurate weights. Of course, I can only match on observable characteristics; whether my weights match on unobservable characteristics is unknown. Given that the main biases appear for region, country of birth and occupational group, I weight my linked samples to match these characteristics. That is, I calculate the proportion of the population that is in each region / country of birth / occupation group cell from the

full-count census; then I calculate the same proportion in my linked sample, and finally reweight my linked sample to match the population distribution. I group countries with less than 1000 individuals in the linked sample as an "other" country because of issues with small cells.

The above process creates weights for one of the linked sets between either the 1880 and 1910 Censuses or the 1910 to 1940 Censuses. These are shown in the "Single Weighted" or "Weighted" columns in Table C5 and C6. The outcomes for occupational categories, country of origin and residence are now aligned with the characteristics of the full population.

These weights make the sample representative on observables for one link; however, my main sample is double linked. Thus, I cannot use the single weighed outcomes for my grandfather-grandson dataset. To create weights for the linked 1880 to 1940 sample, I pursue an iterative process. First, I create the weights between 1880 and 1910 such that each has a weight in the 1910 census. Then using these weights, I calculate the proportions in each cell relative to the 1910 census in the same way as described above. The resulting characteristics of the 1880 census when applying these "Double Weights" are show in Table C5, where the linked sample is still representative of fathers in the 1880 Census.

TABLE C5. REPRESENTATIVENESS OF CHILDREN IN 1880 (GENERATION 2 OF G1-G2-G3 LINKS)

	Linked	Universe	Single Weighted	Double Weighted
Age of G2 Son	7.154	6.570	7.165	7.132
	(3.964)	(4.115)	(3.976)	(3.967)
Characteristics of G1 Father				
Age	41.99	41.37	41.74	41.77
	(6.678)	(6.629)	(6.644)	(6.665)
White-Collar	0.117	0.132	0.132	0.124
	(0.321)	(0.339)	(0.339)	(0.330)
Unskilled	0.270	0.388	0.388	0.360
	(0.444)	(0.487)	(0.487)	(0.480)
Skilled	0.155	0.180	0.180	0.174
	(0.362)	(0.384)	(0.384)	(0.379)
Farmer	0.458	0.300	0.300	0.341
	(0.498)	(0.458)	(0.458)	(0.474)
Northeast	0.236	0.388	0.388	0.328
	(0.425)	(0.487)	(0.487)	(0.470)
Midwest	0.617	0.495	0.495	0.550
	(0.486)	(0.500)	(0.500)	(0.497)
South	0.0820	0.0709	0.0709	0.0757
	(0.274)	(0.257)	(0.257)	(0.265)
West	0.0654	0.0469	0.0469	0.0455
	(0.247)	(0.211)	(0.211)	(0.208)
ln(Occ. Score), 1950	9.858	9.938	9.939	9.916
	(0.379)	(0.370)	(0.372)	(0.374)
ln(Occ. Score), 1890-1950	9.234	9.350	9.352	9.318
	(0.532)	(0.510)	(0.512)	(0.518)
Germany	0.466	0.376	0.376	0.439
	(0.499)	(0.484)	(0.484)	(0.496)
Ireland	0.144	0.291	0.291	0.209
	(0.351)	(0.454)	(0.454)	(0.407)
England	0.134	0.0952	0.0952	0.113
	(0.341)	(0.294)	(0.294)	(0.317)
Canada	0.0865	0.0656	0.0656	0.0657
	(0.281)	(0.248)	(0.248)	(0.248)
France	0.0199	0.0190	0.0190	0.0162
	(0.140)	(0.137)	(0.137)	(0.126)
Netherlands	0.0154	0.00943	0.0182	0.0157
	(0.123)	(0.0966)	(0.134)	(0.124)
Switzerland	0.0196	0.0146	0.0146	0.0129
	(0.139)	(0.120)	(0.120)	(0.113)
Russia	0.00471	0.0118	0.00622	0.00534
	(0.0684)	(0.108)	(0.0786)	(0.0729)
Austria/Hungary	0.0159	0.0176	0.0176	0.0178
	(0.125)	(0.131)	(0.131)	(0.132)
Norway	0.0265	0.0263	0.0263	0.0298
	(0.161)	(0.160)	(0.160)	(0.170)
Sweden	0.0182	0.0208	0.0208	0.0192
	(0.134)	(0.143)	(0.143)	(0.137)
Scotland	0.0280	0.0224	0.0224	0.0277
	(0.165)	(0.148)	(0.148)	(0.164)
Italy	0.00305	0.00378	0.00492	0.00511
	(0.0551)	(0.0614)	(0.0700)	(0.0713)
Observations	71,814	1,769,714	71,814	71,814

Notes: Universe is 1880 full count of second-generation sons. Weighted is to match 1880 universe. Double weighted is also weighting to match 1910 father attributes.

TABLE C6. REPRESENTATIVENESS OF CHILDREN IN 1910 (GENERATION 3 OF G1-G2-G3 LINKS)

	Linked	Universe	Single Weighted
Age of G3 Son	6.191	6.272	6.207
	(4.019)	(4.035)	(4.027)
Characteristics of G2 Fathers			
Age	37.03	37.23	36.99
	(4.036)	(4.022)	(4.044)
White-Collar	0.224	0.209	0.209
	(0.417)	(0.407)	(0.407)
Unskilled	0.211	0.255	0.255
	(0.408)	(0.436)	(0.436)
Skilled	0.184	0.201	0.201
	(0.388)	(0.401)	(0.400)
Farmer	0.380	0.335	0.335
	(0.485)	(0.472)	(0.472)
Northeast	0.211	0.290	0.290
	(0.408)	(0.454)	(0.454)
Midwest	0.587	0.536	0.536
	(0.492)	(0.499)	(0.499)
South	0.0926	0.0935	0.0935
	(0.290)	(0.291)	(0.291)
West	0.109	0.0801	0.0801
	(0.312)	(0.271)	(0.271)
ln(Occ. Score), 1950	9.945	9.966	9.966
	(0.456)	(0.441)	(0.446)
ln(Occ. Score), 1890-1950	9.356	9.381	9.382
	(0.594)	(0.572)	(0.574)
Germany	0.472	0.446	0.446
	(0.499)	(0.497)	(0.497)
Ireland	0.138	0.201	0.201
	(0.344)	(0.401)	(0.401)
England	0.132	0.111	0.111
	(0.339)	(0.315)	(0.315)
Canada	0.0848	0.0653	0.0653
	(0.279)	(0.247)	(0.247)
France	0.0200	0.0162	0.0162
	(0.140)	(0.126)	(0.126)
Netherlands	0.0165	0.0123	0.0165
	(0.127)	(0.110)	(0.128)
Switzerland	0.0203	0.0133	0.0133
	(0.141)	(0.115)	(0.115)
Russia	0.00457	0.00694	0.00509
	(0.0675)	(0.0830)	(0.0712)
Austria/Hungary	0.0168	0.0190	0.0190
	(0.129)	(0.136)	(0.136)
Norway	0.0284	0.0323	0.0322
	(0.166)	(0.177)	(0.176)
Sweden	0.0180	0.0190	0.0190
	(0.133)	(0.136)	(0.136)
Scotland	0.0271	0.0263	0.0263
	(0.162)	(0.160)	(0.160)
Italy	0.00284	0.00449	0.00482
	(0.0532)	(0.0668)	(0.0692)
Observations	96,874	955,894	96,874

Notes: Universe is 1910 full count of third-generation sons. Weighted is to match 1910 universe as described in text.

Linking the 1910-1920 Censuses.

I also link the 1910-1920 censuses to get two observations of the 1910 father's occupation. The linking process is the same as the 1880-1910 and 1910-1940 link where I build potential links between 1910-1920, hand link a subset of 2,000 from the 1910 census, and then predict the hand linking process. Linking the 1910-1920 is done for the white native-born population between 0-47 in 1910 since I am generally interested in building a panel dataset between 1910 and 1920. Below I recreate some of the same tables as in the above section (e.g., the linking rates and the coefficients from the probit model), but now for the 1910-1920 link.

TABLE C7. LINKING RATES

	1910-1920 link
Starting group in base year	18,524,622
Starting group in 1910 with a potential link in ten years later	15,448,111
Potential links ten years later	152,390,867
Match in 1940 amongst links	6,113,276
Overall Linking Rate	33.0
Linking Rate given Potential Match	39.6

Notes: This table shows the linking rates between the 1910-1920 censuses. See Table C1 for counterpart between 1880-1910 and 1910-1940

TABLE C8. PROBIT MODEL FOR 1910-1920 LINK

	1910-1920 link
Jaro-Winkler Distance, First name	-6.566***
	(0.589)
Jaro-Winkler Distance, Last name	-13.57***
	(0.876)
Year of Birth Difference = 1	-0.158
	(0.126)
Year of Birth Difference = 2	-0.760***
	(0.154)
Year of Birth Difference = 3	-1.161***
	(0.168)
No. of potential links	-0.0623***
	(0.0190)
No. of potential links squared	0.00143**
	(0.000655)
Unique and Exact NYSIIS First name match	0.424***
	(0.161)
Unique and Exact NYSIIS Last name match	-0.0570
	(0.266)
Unique and Exact NYSIIS First AND Last name match	0.948***
	(0.132)
Unique Exact Last name String match	1.288***
	(0.219)
Middle initial match, if have one	1.191***
	(0.110)
NYSIIS last name match AND Year of Birth Diff=0	1.031***
	(0.196)
NYSIIS last name match AND Year of Birth Diff=1	0.867***
	(0.182)
NYSIIS last name match AND Year of Birth Diff=2	0.732***
	(0.207)
2 Potential links with NYSIIS last name match	-0.387**
	(0.182)
>2 potential links with NYSIIS last name match	-0.525***
	(0.167)
2 Potential links with last name string match	-1.425***
	(0.194)
>2 Potential links with last name string match	-1.354***
	(0.125)
One potential link	0.790***
	(0.171)
Difference in length of last name strings	-0.336***
	(0.0499)
Mother place of birth match	0.537***
	(0.0799)
Father place of birth match	0.520***
	(0.0781)
Constant	0.458**
	(0.185)
Observations	15,993

Notes: Data is from the training data between the 1910-1920 linked datasets.

TABLE C9. HAND LINKING NUMBERS FOR THE 1910-1920 CENSUSES

	1910-1920
Random sample in base year	2,000
Potential links ten years later	15,993
Successfully linked	1,263
Handlinking Rate for training data (given 1 potential match)	63.2
Cutoff for predicted probability	0.383
Score Ratio of 1st best link to 2nd best	2.1
PPV	0.900
TPR	0.814

Notes: This table shows the training data statistics for the 1910-1920 link.