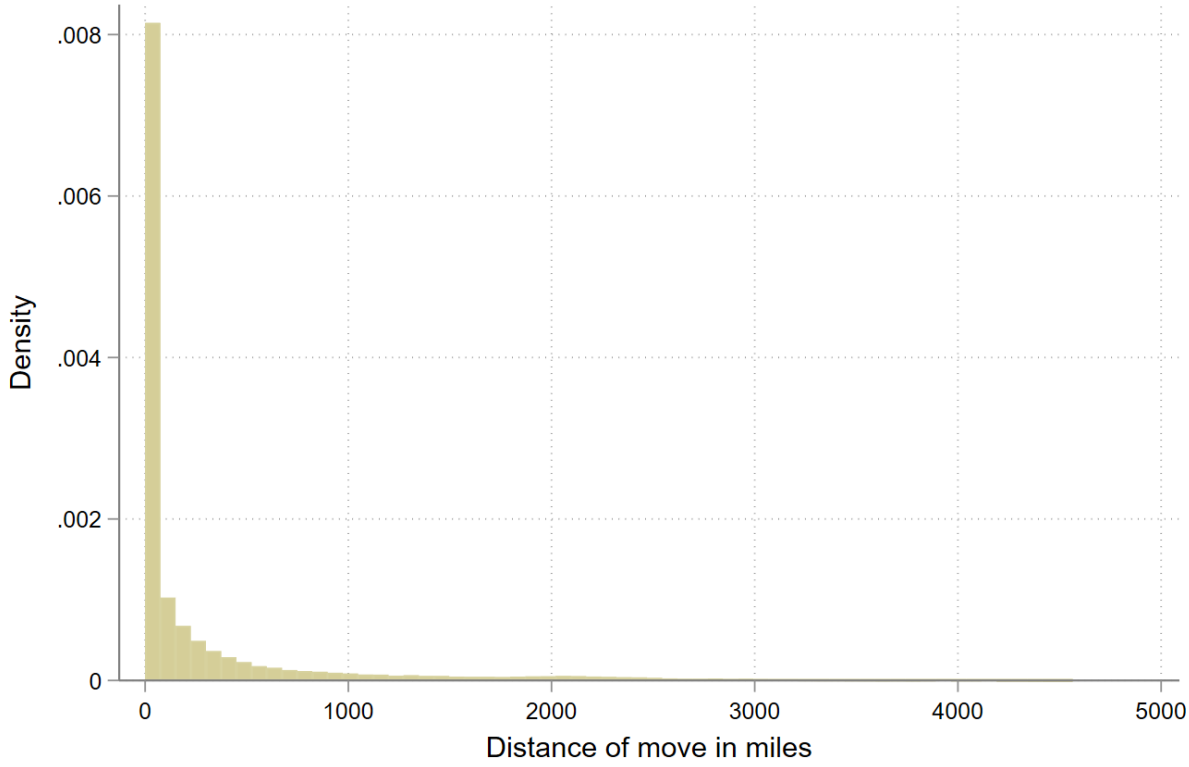
Internal Migration, Education, and Intergenerational Mobility: Evidence from American History

Zachary Ward

Baylor University

Online Appendix

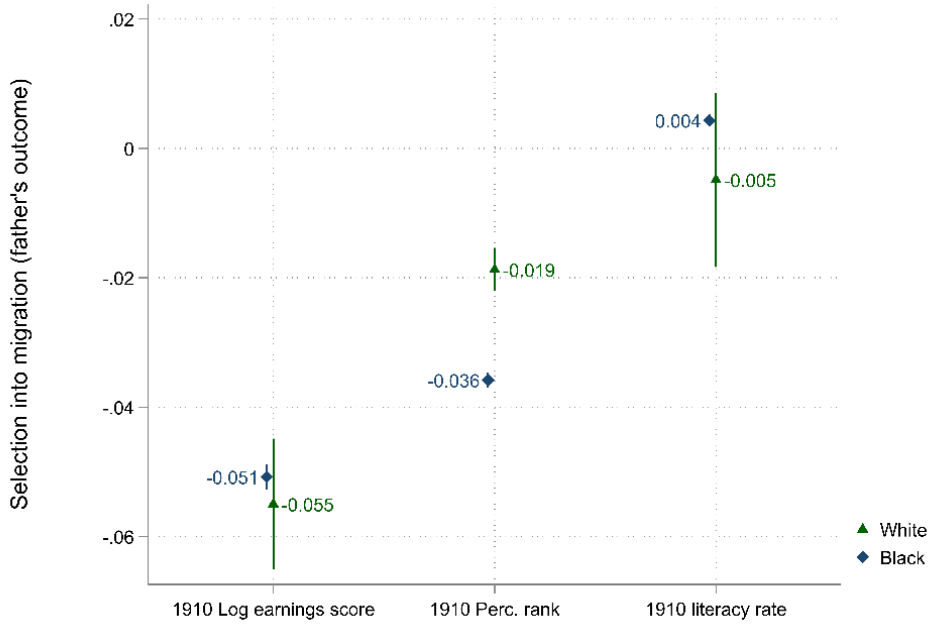
Figure A1. Histogram of migration distances



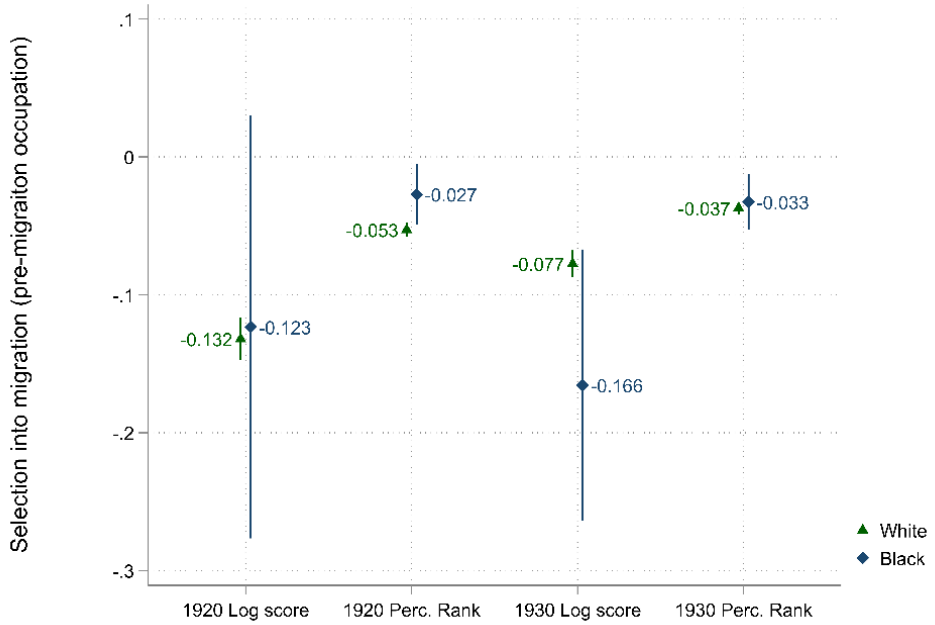
Notes: Distance between the 1910 and 1940 county centroids as measured by a straight line. Distance is mapped for intercounty migrants.

Figure A2. Observable selection into internal migration by race

Panel A. Based on father's earnings score and literacy



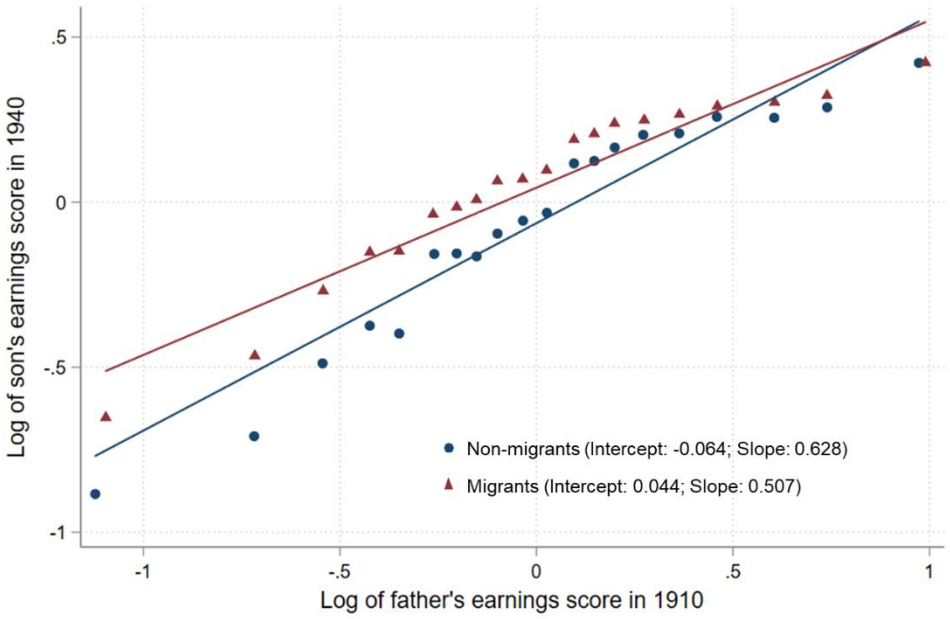
Panel B. Based on pre-migration earnings score



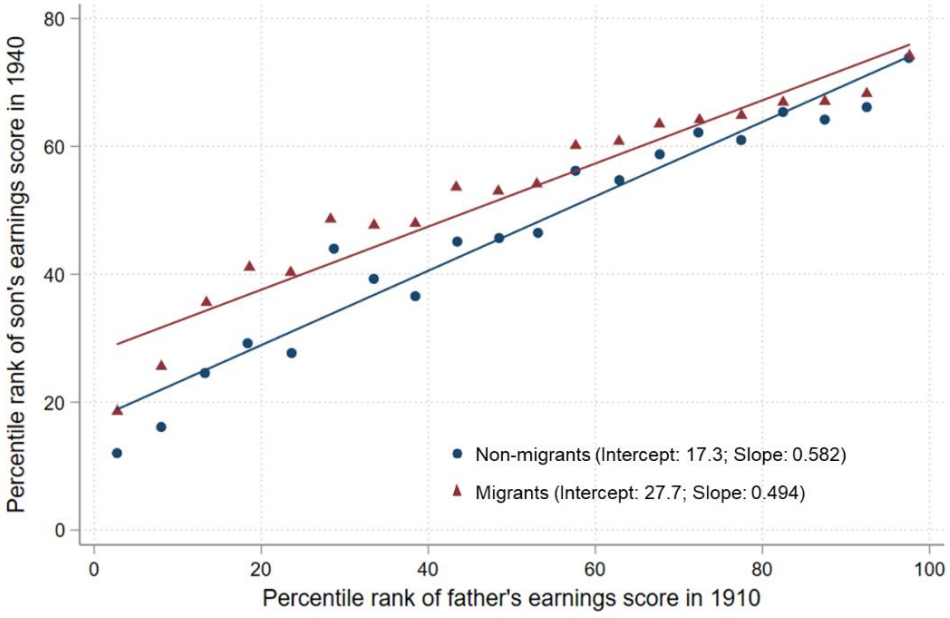
Notes: Data are from a linked sample between 1910 and 1940. Earnings scores are imputed based on the occupation, race and state (see Appendix C). Panel A estimates the raw difference in outcomes across migrants' and stayers' fathers by race (like Table 1). Panel B estimates the across-household difference in outcomes across migrants and stayers for pre-migration outcomes (like in Table 2).

Figure A3. Mobility measures for ever migrants and non-migrants

A. Log-log



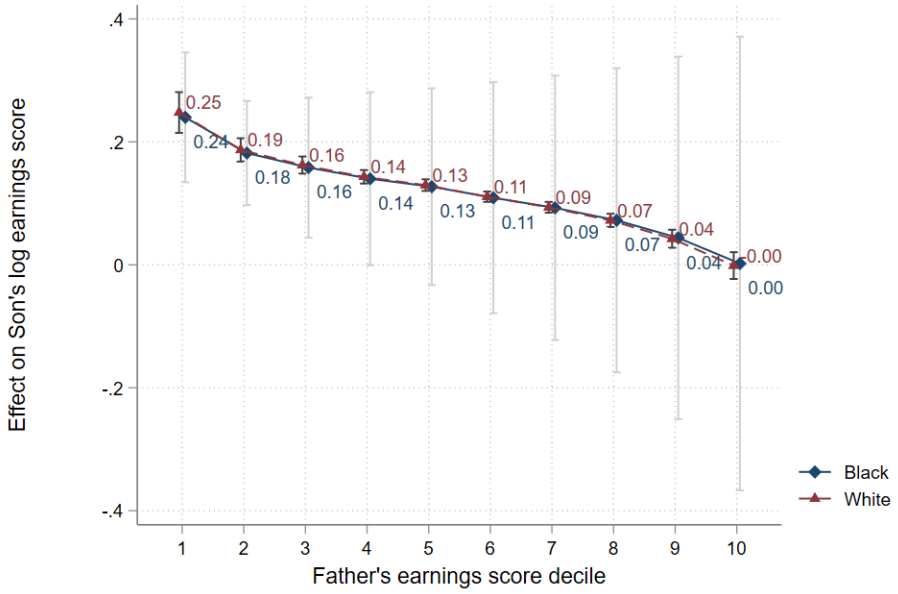
B. Rank-Rank



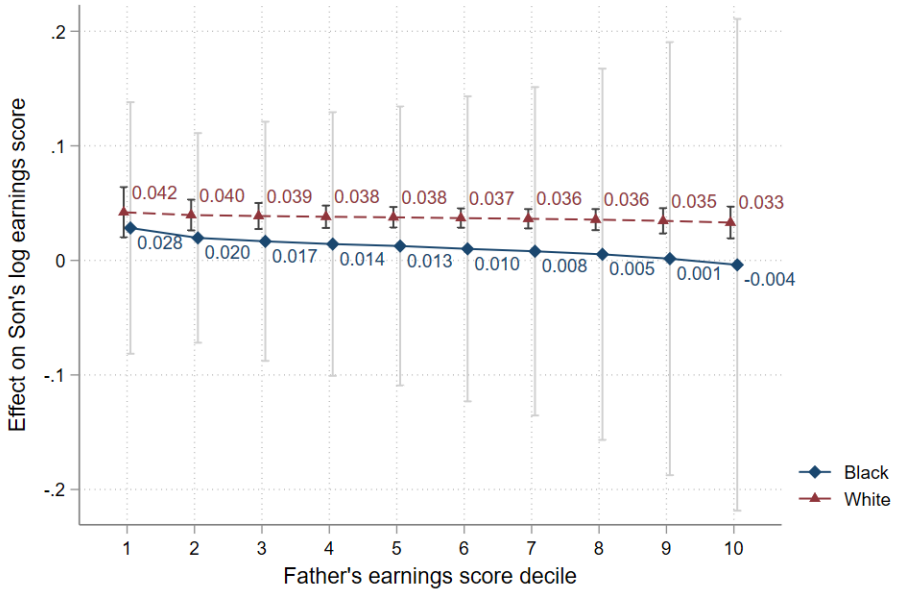
Notes: Data are from the linked sample of individuals between the 1910, 1920, 1930, and 1940 censuses. Son's outcomes are observed in the 1940 census. Migration is defined as having ever migrated across counties according to observation in the 1920, 1930 and 1940 censuses. Earnings scores are based on the father's occupation, race and state (see Appendix C). No controls are included in the above relationships, but Panel A first removes life-cycle effects with a quartic in age for the father and son.

Figure A4. The within-brother migration and education premium by race

Panel A. Ever migration premium

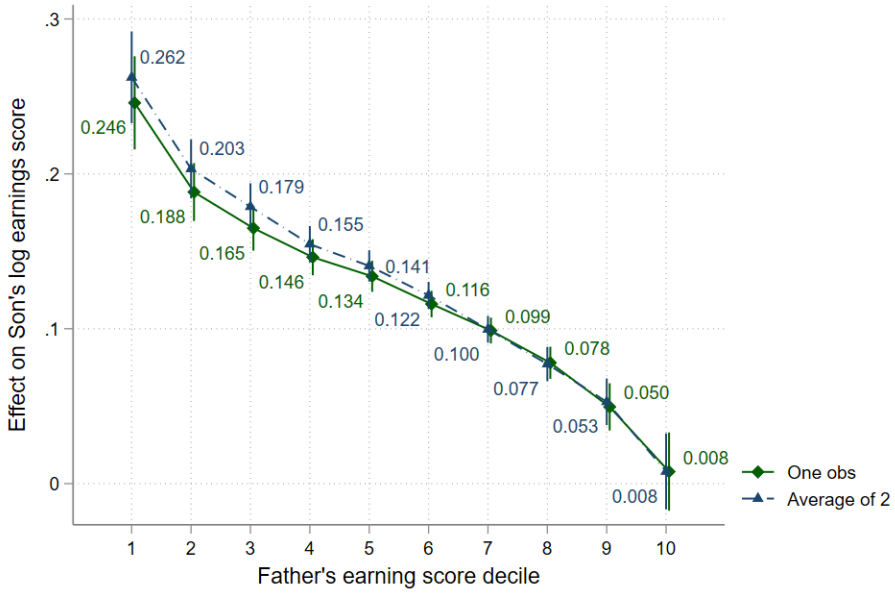


Panel B. Education premium

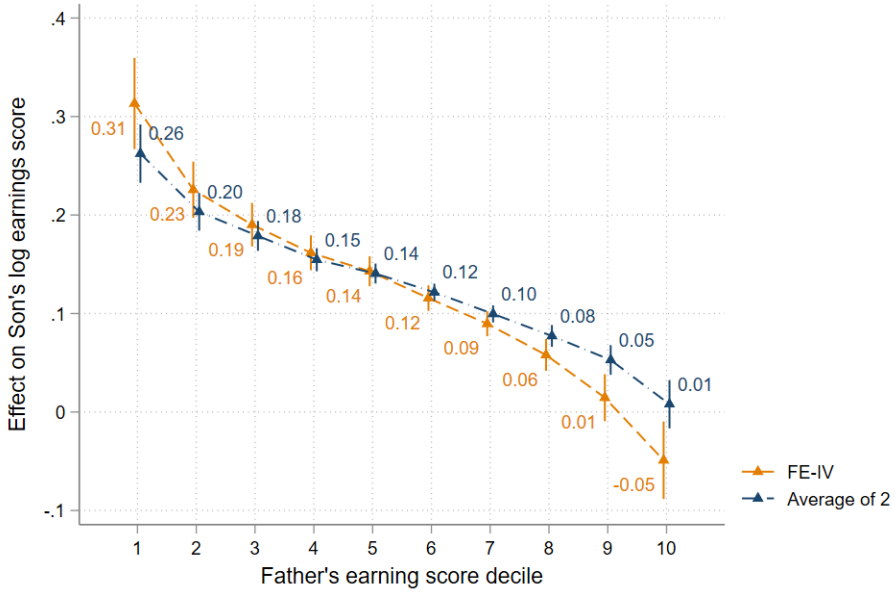


Notes: Data are from the linked sample of individuals between the 1910, 1920, 1930 and 1940 censuses. The education and migration premium are estimated across the distribution (Equation (3)). The figure shows the mean premium within each decile. Earnings scores are based on the father's occupation, race and state (see Appendix C). The sample is weighted to be representative and standard errors are clustered at the 1910 household level. Estimates are done separately by race. See Figure 5 for pooled estimates.

Figure A5. Averaging two observations also increased the migration premium for poorer families
Panel A. One observation versus average of two observations

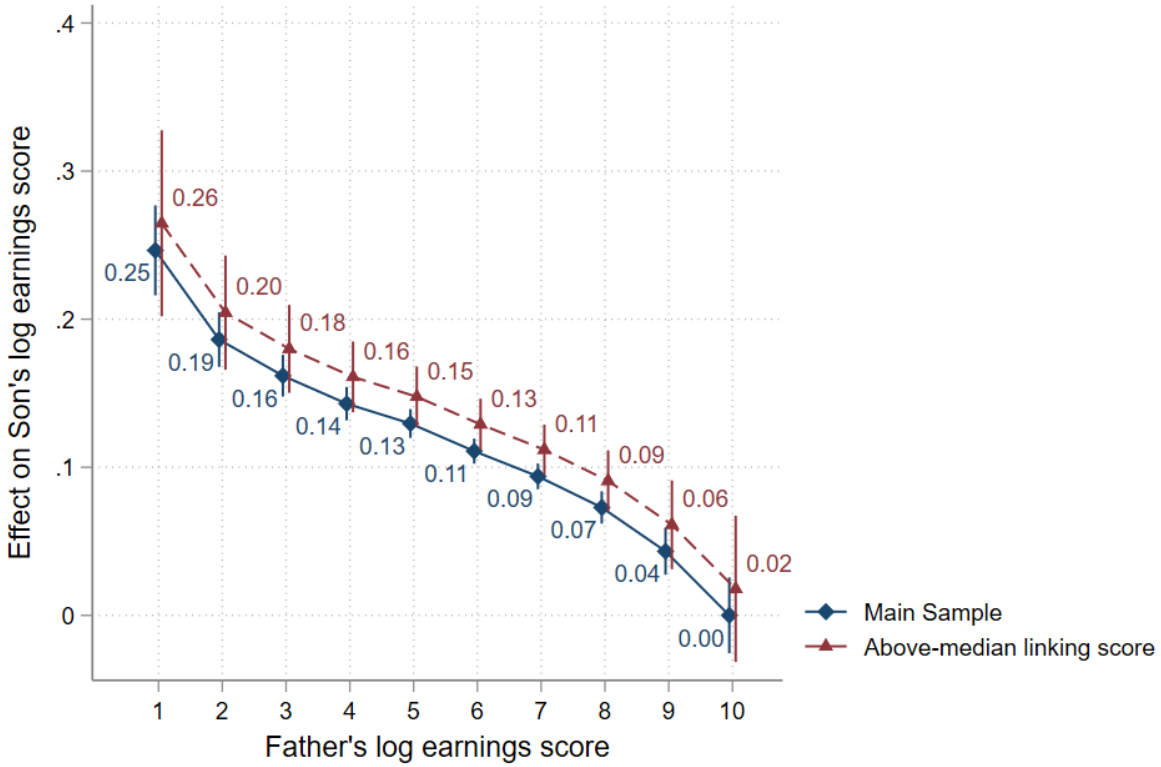


Panel B. FE-IV versus average of two observations



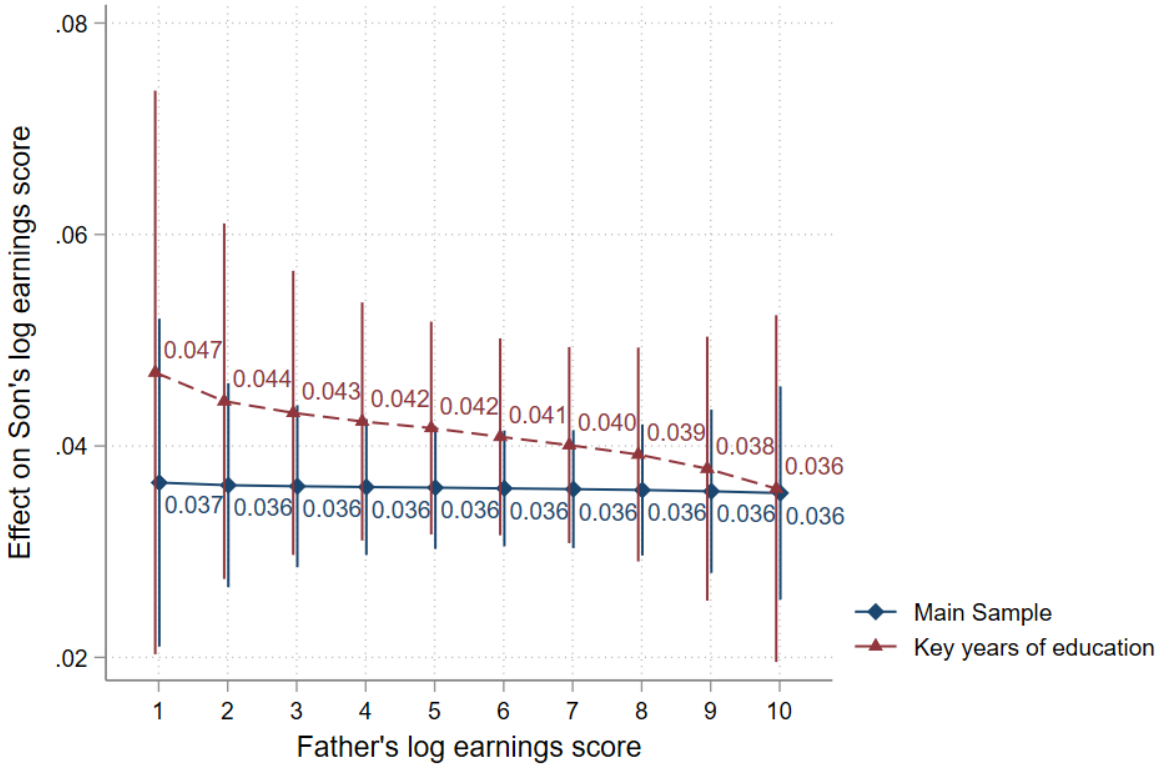
Notes: These figures recreate the analysis from Figures 5 and 6, but shows results when using the average of the son's 1930 and 1940 earnings score and the average of the father's earning scores (from the 1910 and either the 1900 or 1920 census). As expected, the average migration premium is between the one observation and 2SLS estimates because the average only partially addresses measurement error. See Appendix D for a further discussion of measurement error.

Figure A6. The migration premium is still high when limiting sample to high-quality links



Notes: Data are from the linked sample of individuals between the 1910, 1920, 1930 and 1940 censuses. The education and migration premium are estimated across the distribution (Equation (3)). The figure shows the mean premium within each decile. Earnings scores are based on the father's occupation, race and state (see Appendix C). Sample is weighted to be representative and standard errors are clustered at the 1910 household level. The above-median linking score first sums the 1910-1940, 1920-1930 and 1930-1940 linking scores and then keeps those above the median.

Figure A7. The education may be attenuated by error in education



Notes: Data are from the linked sample of individuals between the 1910, 1920, 1930 and 1940 censuses. The education and migration premium are estimated across the distribution (Equation (3)). The figure shows the mean premium within each decile. Earnings scores are based on the father's occupation, race and state (see Appendix C). The sample is weighted to be representative and standard errors are clustered at the 1910 household level. The "key years of education" sample limits the sample to 8, 12, 16 and 16+ years of education.

Table A1. Overview of Nominal Earnings scores

	National	White	Black	Northeast	Midwest	South	West
Income Scores							
Professional	39,364	40,065	17,749	42,232	37,526	37,941	39,129
Farmers	14,003	14,905	7,744	23,942	15,247	10,655	20,441
Managers and Officials	37,934	38,188	13,999	41,609	37,052	35,113	38,360
Clerical and Kindred	27,008	27,140	22,145	27,436	26,850	26,249	27,922
Sales Workers	28,450	28,606	12,742	30,693	27,739	25,735	28,690
Craftsmen	23,359	23,700	12,507	24,529	23,999	20,519	24,705
Operatives	19,262	19,815	12,494	20,035	20,435	15,888	22,027
Service Workers	19,301	22,315	11,316	22,719	19,745	14,609	22,002
Farm laborers	7,959	8,686	5,296	13,180	7,854	5,783	12,048
Laborers	12,158	13,080	9,427	14,362	13,098	9,070	14,904
Percentile Ranks							
Professional	84	85	35	88	83	79	85
Farmers	25	28	6	58	29	14	48
Managers and Officials	89	89	25	92	88	85	91
Clerical and Kindred	69	70	53	71	69	67	72
Sales Workers	72	73	21	77	71	66	74
Craftsmen	57	58	20	60	59	48	61
Operatives	44	46	20	47	48	32	54
Service Workers	41	51	16	51	43	27	50
Farm laborers	8	10	2	22	7	3	19
Laborers	19	22	10	27	21	9	29

Notes: Data shows the mean earnings scores within broad occupational categories. The means are calculated using the 1940 outcomes of the sons in the linked sample. The categories are separated by the first digit of the *occ1950* codes from IPUMS. See Appendix C for description of earnings scores; generally, they are estimated by one's 3-digit occupation, race and state.

Table A2. Overview of Real Earnings scores

	National	White	Black	Northeast	Midwest	South	West
Earnings Scores							
Professional	33,472	34,064	15,209	35,301	31,445	34,013	32,809
Farmers	13,943	14,799	8,005	22,698	14,928	11,031	20,049
Managers and Officials	32,931	33,154	11,942	35,185	31,753	32,260	32,639
Clerical and Kindred	22,881	23,020	17,763	22,794	22,477	23,309	23,328
Sales Workers	24,207	24,342	10,585	25,679	23,288	23,170	23,971
Craftsmen	20,269	20,572	10,652	20,822	20,509	18,853	21,226
Operatives	16,844	17,325	10,955	17,043	17,496	14,913	19,196
Service Workers	16,252	18,823	9,440	18,620	16,481	12,987	18,365
Farm laborers	7,794	8,470	5,324	12,393	7,618	5,913	11,490
Laborers	10,922	11,818	8,270	12,464	11,558	8,633	13,430
Percentile Ranks							
Professional	83	85	33	87	82	81	84
Farmers	30	33	7	65	34	18	57
Managers and Officials	89	90	21	91	88	88	89
Clerical and Kindred	67	68	47	67	66	68	69
Sales Workers	71	72	16	75	68	68	71
Craftsmen	56	57	17	58	57	50	60
Operatives	43	45	18	44	46	35	54
Service Workers	38	48	12	46	39	27	47
Farm laborers	9	11	2	24	8	4	20
Laborers	17	21	8	23	19	9	28

Notes: Data shows the mean earnings scores within broad occupational categories, after adjusting for rural-urban differences in cost of living. The means are calculated using the 1940 outcomes of the sons in the linked sample. The categories are separated by the first digit of the *occ1950* codes from IPUMS. See Appendix C for description of earnings scores; generally, they are estimated by one's 3-digit occupation, race and state.

Table A3. Occupation Transition Matrix for non-migrants

Father in 1910	Son in 1940					
	White Collar	Semi-skilled	Unskilled	Farmer, Owner	Farmer, Tenant	Total
White Collar	36,562	16,943	6,596	1,318	675	62,093
	(58.88)	*(27.29)*	*(10.62)*	*(2.12)*	*(1.09)*	*(100.00)*
Semi-skilled	44,214	57,037	19,457	1,469	875	123,051
	(35.93)	*(46.35)*	*(15.81)*	*(1.19)*	*(0.71)*	*(100.00)*
Unskilled	18,740	27,768	20,667	2,269	2,240	71,684
	(26.14)	*(38.74)*	*(28.83)*	*(3.16)*	*(3.12)*	*(100.00)*
Farmer, Owner	14,366	18,132	22,855	32,657	20,605	108,615
	(13.23)	*(16.69)*	*(21.04)*	*(30.07)*	*(18.97)*	*(100.00)*
Farmer, Tenant	4,495	8,416	12,122	5,683	11,883	42,599
	(10.55)	*(19.76)*	*(28.46)*	*(13.34)*	*(27.89)*	*(100.00)*
Total	118,376	128,297	81,696	43,396	36,277	408,041
	(29.01)	*(31.44)*	*(20.02)*	*(10.64)*	*(8.89)*	*(100.00)*

Notes: Data are from the 1910, 1920, 1930 and 1940 linked sample. Table only shows results for those who never moved across counties (according to the 1910, 1920, 1930, and 1940 Censuses). Sample is split into occupational categories based on the *occ1950* variable. White collar are professionals, managers, sales and clerical. Semi-skilled are craftsmen and operatives. Unskilled are service workers, farm laborers and laborers. Farmers are separated by owners and tenants based on whether claimed to own a home. Farmers without a farm ownership variable are excluded from the matrix.

Table A4. Occupation Transition Matrix for Ever Migrants

Father in 1910	Son in 1940					
	White Collar	Semi-skilled	Unskilled	Farmer, Owner	Farmer, Tenant	Total
White Collar	57,943	22,985	9,234	1,889	1,461	93,512
	(61.96)	*(24.58)*	*(9.88)*	*(2.02)*	*(1.56)*	*(100.00)*
Semi-skilled	55,091	56,221	20,222	2,478	2,145	136,156
	(40.46)	*(41.29)*	*(14.85)*	*(1.82)*	*(1.58)*	*(100.00)*
Unskilled	25,848	33,293	25,901	2,541	3,669	91,251
	(28.33)	*(36.48)*	*(28.38)*	*(2.79)*	*(4.02)*	*(100.00)*
Farmer, Owner	35,169	40,716	29,442	12,627	14,407	132,362
	(26.57)	*(30.76)*	*(22.24)*	*(9.54)*	*(10.88)*	*(100.00)*
Farmer, Tenant	15,219	26,146	28,228	5,079	13,062	87,735
	(17.35)	*(29.80)*	*(32.17)*	*(5.79)*	*(14.89)*	*(100.00)*
Total	189,269	179,360	113,027	24,615	34,745	541,016
	(34.98)	*(33.15)*	*(20.89)*	*(4.55)*	*(6.42)*	*(100.00)*

Notes: Data are from the 1910, 1920, 1930 and 1940 linked sample. Table only shows results for those who ever moved across counties (according to the 1910, 1920, 1930, and 1940 Censuses). Sample is split into occupational categories based on the *occ1950* variable. White collar are professionals, managers, sales and clerical. Semi-skilled are craftsmen and operatives. Unskilled are service workers, farm laborers and laborers. Farmers are separated by owners and tenants based on whether claimed to own a home. Farmers without an ownership variable are excluded from the matrix.

Table A5. The migration premium is stable after controlling for pre-migration outcomes

	Log Earnings score				Log wage income, for wage workers			
Sample	Moved between 1920-1940	Moved between 1920-1940	Moved between 1930-1940	Moved between 1930-1940	Moved between 1920-1940	Moved between 1920-1940	Moved between 1930-1940	Moved between 1930-1940
Ever migrant	0.104	0.104	0.088	0.088	0.150	0.149	0.102	0.096
	(0.007)	(0.007)	(0.011)	(0.011)	(0.018)	(0.018)	(0.029)	(0.029)
Education	0.033	0.033	0.029	0.024	0.045	0.044	0.043	0.033
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.005)	(0.005)
Log earnings score in 1920		0.024		0.010		0.114		0.091
		(0.007)		(0.009)		(0.021)		(0.025)
Log earnings score in 1930				0.179				0.376
				(0.013)				(0.031)
HH FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, birth order controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	42,198	42,198	26,563	26,563	19,952	19,952	12,328	12,328
	Percentile Rank				Upward rank mobility			
Ever migrant	5.144	5.112	3.832	3.771	0.108	0.108	0.081	0.079
	(0.367)	(0.367)	(0.589)	(0.576)	(0.008)	(0.008)	(0.013)	(0.012)
Education	1.872	1.801	1.650	1.225	0.029	0.028	0.026	0.019
	(0.078)	(0.079)	(0.103)	(0.101)	(0.002)	(0.002)	(0.002)	(0.002)
Percentile Rank in 1920		0.096		0.061		0.001		0.001
		(0.010)		(0.012)		(0.000)		(0.000)
Percentile Rank in 1930				0.271				0.004
				(0.011)				(0.000)
HH FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, birth order controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	42,198	42,198	26,563	26,563	42,198	42,198	26,563	26,563

Notes: Data are from the 1910, 1920, 1930 and 1940 linked sample. This table shows that controlling for pre-migration earnings scores in 1920 or 1930 do not alter migration premium estimates. The “moved between 1920-1940” sample removes those who moved between 1910 and 1920, and also drops those without an occupational response in 1920. The “moved between 1930-1940” sample removes those who moved between 1910 and 1930, and also drops those without an occupational response in 1930.

Table A6. The interaction between migration and father's status is stronger after accounting for error

	Log earnings score		Log wage income	
Migrant	0.122	0.124	0.199	0.208
	(0.005)	(0.005)	(0.014)	(0.015)
Migrant x Father's log earnings score	-0.134	-0.204	-0.139	-0.226
	(0.012)	(0.019)	(0.033)	(0.054)
Education	0.035	0.035	0.052	0.052
	(0.001)	(0.001)	(0.002)	(0.002)
Education x Father's log earnings score	0.002	0.001	-0.003	-0.014
	(0.002)	(0.003)	(0.005)	(0.008)
Household FE	Yes	Yes	Yes	Yes
2SLS	No	Yes	No	Yes
Age and birth order controls	Yes	Yes	Yes	Yes
Observations	118,802	118,802	59,128	59,128
	Percentile rank		Upward rank mobility	
Migrant	11.406	13.798	0.232	0.281
	(0.572)	(0.775)	(0.015)	(0.019)
Migrant x Rank of Father (div. by 100)	-0.103	-0.150	-0.002	-0.003
	(0.009)	(0.014)	(0.000)	(0.000)
Education	1.959	2.011	0.030	0.033
	(0.106)	(0.144)	(0.003)	(0.004)
Education x Rank of Father (div. by 100)	0.001	-0.000	0.000	-0.000
	(0.002)	(0.002)	(0.000)	(0.000)
Household FE	Yes	Yes	Yes	Yes
2SLS	No	Yes	No	Yes
Age and birth order controls	Yes	Yes	Yes	Yes
Observations	118,802	118,802	118,802	118,802

Notes: Data are from the 1910, 1920, 1930 and 1940 linked sample, with the father linked to either 1900 or 1920. This table shows results when accounting for measurement error by instrumenting the interaction between migration and father's status and the interaction between education and father's status with a second observation. The table shows that the main results hold: the migration premium is high and higher for poorer families. See Figure 6 for predicted premium for the log earnings score. See Appendix D for a further discussion of measurement error.

Table A7. The main results are robust to an alternatively linked sample

	I Triple-linked sample	II Single-linked sample	III Measurement error sample
Intercounty Migrant	0.139	0.134	0.143
	(0.010)	(0.003)	(0.013)
Intercounty Migrant x Father's Earnings Score	-0.159	-0.150	-0.160
	(0.027)	(0.008)	(0.034)
Education (years)	0.035	0.040	0.034
	(0.002)	(0.001)	(0.003)
Education x Father's Earnings Score	-0.001	-0.002	0.002
	(0.004)	(0.001)	(0.006)
HH FE	Yes	Yes	Yes
Birth order and age controls	Yes	Yes	Yes
Observations	949,333	2,691,933	494,031
R-squared	0.953	0.910	0.949

Notes: This table shows that the migration premium is high and higher for children from poorer families when using alternatively linked samples. The difference between this table and the main specification (Equation (3)) is that I use $IntercountyMigrant_{i,h,g}$ as the migration variable instead of $EverMigrant_{i,h,g}$. While $EverMigrant_{i,h,g}$ indicates whether the son was ever in another county in 1920, 1930 or 1935, $IntercountyMigrant_{i,h,g}$ indicates whether the son was in another county in 1935 or 1940. This specification can be run on a single-linked, triple-linked or quadruple-linked sample. Column I is my main sample where sons are linked 1910-1920-1930-1940. Column II is a single linked sample between 1910-1940. Column III is the measurement error sample which takes the main sample in Column I and then finds fathers in the 1900 or 1920 censuses. Estimates are higher than the Ever migration variable because the sample mostly includes permanent migrants.

Table A8. Difference in standardized linking scores across migrants and non-migrants

	Full sample	Above-median links
Ever Migrant	-0.0524	0.0114
	(0.00490)	(0.00111)
Constant	-0.0246	0.654
	(0.00355)	(0.000848)
Observations	211,549	105,773
R-squared	0.001	0.001

Notes: Data are from the 1910, 1920, 1930 and 1940 linked sample. This table shows the difference in linking scores between ever migrants and stayers. Linking scores are first summed from the 1920-1930, 1930-1940 and 1910-1940 links, then standardized. Column I shows the difference for the full sample, and Column II limits the sample to those with above-median scores. See Figure A6 for migration premium estimates when limiting the sample to above-median quality linking scores.

Table A9. Results are robust when accounting for the Great Depression, relief spending, age of child, and early moves

Sample:	Main	Main	Keep Age>5 in 1910	Drop 1910-1920 Movers
Migrant	0.117	0.112	0.111	0.123
	(0.005)	(0.005)	(0.010)	(0.005)
Migrant x Father's log earnings score	0.036	0.036	0.034	0.035
	(0.001)	(0.001)	(0.002)	(0.001)
Education	-0.138	-0.132	-0.138	-0.148
	(0.013)	(0.013)	(0.027)	(0.014)
Education x Father's log earnings score	-0.001	-0.000	0.002	0.000
	(0.002)	(0.002)	(0.004)	(0.002)
Log difference in county retail sales (1929-1933)		0.005		
		(0.022)		
Log relief spending (1933-1939)		0.067		
		(0.008)		
		-0.320		
Household FE	Yes	Yes	Yes	Yes
Age and birth order controls	Yes	Yes	Yes	Yes
Observations	211,558	211,554	125,577	163,824

Notes: This table recreates the results from Figure 5 and Table 4 (log earnings score) from the main text. I additionally control for the log fall in per capita retail sales by county between 1929 and 1933, which controls for the downturn from the Great Depression. I also control for the log per capita relief spending at the county level. These are created after merging data from Fishback et al. (2006) to the 1930 county. In the third column, I only keep older movers in case younger movers migrated with their family. In the last column, I show that the results are robust to dropping 1910-1920 movers, where individuals may have moved with their family.

Table A10. The within-brother *nominal* migration premium, alternative moves

	Upward Rank	Percentile Rank	Log Wage Income	Log Earnings score
Intercounty	0.116	5.714	0.174	0.112
	(0.00508)	(0.244)	(0.0138)	(0.00473)
Interstate	0.165	8.026	0.211	0.154
	(0.00861)	(0.413)	(0.0222)	(0.00804)
Interregion	0.191	9.282	0.227	0.184
	(0.0137)	(0.653)	(0.0383)	(0.0132)
Rural to urban	0.297	16.26	0.463	0.306
	(0.0118)	(0.569)	(0.0325)	(0.0114)
Rural to rural	0.108	4.640	0.208	0.104
	(0.0104)	(0.492)	(0.0395)	(0.0105)
Urban to urban	0.0758	4.004	0.169	0.0641
	(0.0130)	(0.601)	(0.0259)	(0.00981)
Urban to rural	0.00697	-2.171	-0.00203	-0.0382
	(0.0199)	(1.018)	(0.0443)	(0.0171)
White	0.112	5.727	0.166	0.107
	(0.005)	(0.248)	(0.013)	(0.005)
Black	0.206	5.118	0.329	0.219
	(0.044)	(1.183)	(0.112)	(0.040)
0-50 miles	0.073	3.320	0.131	0.066
	(0.008)	(0.359)	(0.020)	(0.007)
50-100 miles	0.111	5.147	0.198	0.100
	(0.009)	(0.440)	(0.024)	(0.008)
100-250 miles	0.146	7.154	0.276	0.134
	(0.009)	(0.415)	(0.022)	(0.008)
250-1000 miles	0.204	10.522	0.367	0.197
	(0.008)	(0.401)	(0.021)	(0.008)
1000+ miles	0.262	14.133	0.304	0.264
	(0.010)	(0.508)	(0.027)	(0.009)
Dust Bowl Migrant	0.183	10.02	0.115	0.191
	(0.0150)	(0.783)	(0.048)	(0.014)
Black and from South	0.332	11.08	0.566	0.396
	(0.114)	(3.543)	(0.305)	(0.104)
Permanent	0.147	7.260	0.234	0.140
	(0.00545)	(0.261)	(0.0146)	(0.00505)
Temporary	0.00813	0.228	-0.0504	0.0134
	(0.00793)	(0.378)	(0.0221)	(0.00719)

Notes: Data are from 1910-1940 linked sample. Each cell is a separate regression, always based on within-brother variation. The first row estimates the migration effect for those who are in the North or Midwest census regions in 1940 but were in the South census region in 1910. Dust Bowl migrants are defined based on whether they were in a Dust Bowl county in 1930 and not in one in 1940; non-migrants are those in Dust Bowl counties in both 1930 and 1940. Dust Bowl counties are those where any part of the county had more than 25 percent topsoil erosion (see Hornbeck (2012), Figure 2). Urban is defined as an area with over 2,500 residents. Miles are measured based on straight-line distances between county centroids. Permanent migrants are those who are in a different county in 1940 as they were in 1910. Temporary migrants are those who are in the same county in 1940 as they were in 1910, but are observed in a different county in either 1920, 1930, or according to the 1935 migration question the 1940 Census.

Table A11. The within-brother *real* migration premium, alternative moves

	Upward Rank	Percentile Rank	Log Wage Income	Log Earnings score
Intercounty	0.101	4.658	0.157	0.0859
	(0.00502)	(0.255)	(0.0136)	(0.00457)
Interstate	0.139	6.221	0.182	0.112
	(0.00860)	(0.428)	(0.0219)	(0.00769)
Interregion	0.164	7.157	0.189	0.132
	(0.0135)	(0.674)	(0.0374)	(0.0124)
Rural to urban	0.185	7.731	0.318	0.143
	(0.0124)	(0.593)	(0.0321)	(0.0110)
Rural to rural	0.133	6.283	0.240	0.120
	(0.0102)	(0.534)	(0.0394)	(0.0104)
Urban to urban	0.0399	2.329	0.139	0.0332
	(0.0128)	(0.642)	(0.0259)	(0.00984)
Urban to rural	0.142	7.456	0.159	0.123
	(0.0198)	(1.052)	(0.0443)	(0.0170)
White	0.100	4.731	0.151	0.083
	(0.005)	(0.260)	(0.013)	(0.004)
Black	0.128	2.890	0.265	0.146
	(0.043)	(1.159)	(0.109)	(0.037)
0-50 miles	0.071	3.375	0.135	0.062
	(0.008)	(0.379)	(0.020)	(0.007)
50-100 miles	0.100	4.063	0.181	0.074
	(0.009)	(0.461)	(0.024)	(0.008)
100-250 miles	0.119	5.498	0.247	0.095
	(0.009)	(0.438)	(0.022)	(0.008)
250-1000 miles	0.164	7.922	0.320	0.139
	(0.008)	(0.416)	(0.021)	(0.007)
1000+ miles	0.225	11.333	0.259	0.197
	(0.010)	(0.521)	(0.026)	(0.009)
Dust Bowl Migrant	0.166	8.769	0.0929	0.152
	(0.033)	(0.813)	(0.0472)	(0.0138)
Black and from South	0.290	5.391	0.437	0.243
	(0.119)	(3.660)	(0.302)	(0.0992)
Permanent	0.126	5.870	0.211	0.106
	(0.00542)	(0.272)	(0.0143)	(0.00487)
Temporary	0.0117	0.357	-0.0470	0.0148
	(0.00770)	(0.401)	(0.0221)	(0.00705)

Notes: Data are from 1910-1940 linked sample. Each cell is a separate regression, always based on within-brother variation. The first row estimates the migration effect for those who are in the North or Midwest census regions in 1940 but were in the South census region in 1910. Dust Bowl migrants are defined based on whether they were in a Dust Bowl county in 1930 and not in one in 1940; non-migrants are those in Dust Bowl counties in both 1930 and 1940. Dust Bowl counties are those where any part of the county had more than 25 percent topsoil erosion (see Hornbeck (2012), Figure 2). Urban is defined as an area with over 2,5000 residents. Miles are measured based on straight-line distances between county centroids. Permanent migrants are those who are in a different county in 1940 as they were in 1910. Temporary migrants are those who are in the same county in 1940 as they were in 1910, but are observed in a different county in either 1920, 1930, or according to the 1935 migration question the 1940 Census.

Table A12. The within-brother rural-urban migration premium does not widely vary by population of destination

	Nominal				Real			
	Upward Rank	Rank of Score	Log(Wage Inc.)	Log(Inc. Score)	Upward Rank	Rank of Score	Log(Wage Inc.)	Log(Inc. Score)
Population in 1940:								
2,500-25,000	0.283	15.210	0.423	0.282	0.166	6.882	0.281	0.123
	(0.015)	(0.735)	(0.042)	(0.014)	(0.015)	(0.768)	(0.041)	(0.013)
25,000-50,000	0.280	16.372	0.551	0.308	0.176	8.031	0.408	0.149
	(0.023)	(1.107)	(0.060)	(0.020)	(0.022)	(1.159)	(0.060)	(0.019)
50,000-100,000	0.287	15.328	0.427	0.287	0.153	7.534	0.292	0.134
	(0.024)	(1.149)	(0.058)	(0.021)	(0.026)	(1.196)	(0.057)	(0.021)
100,000-250,000	0.292	16.271	0.498	0.292	0.165	8.354	0.362	0.139
	(0.023)	(1.156)	(0.055)	(0.022)	(0.023)	(1.208)	(0.054)	(0.022)
250,000+	0.327	17.803	0.467	0.346	0.210	8.782	0.313	0.174
	(0.017)	(0.798)	(0.044)	(0.017)	(0.018)	(0.827)	(0.043)	(0.016)
Observations	96,883	96,883	56,669	96,883	96,883	96,883	56,669	96,883
R-squared	0.782	0.812	0.837	0.805	0.762	0.786	0.824	0.779

Notes: Data are from 1910, 1920, 1930 and 1940 linked sample. The sample is only of rural-urban movers between 1910 and 1940, who are compared to stayers who remained in a rural county. Wage income is only for wage workers. Real income is adjusted for across state and rural-urban cost differences. Earnings scores are imputed based on occupation, race and region (see Appendix C).

Table A13. The migration premium varies by length of residence

	I Log earnings score	II Log wage income	III Percentile rank	IV Upward Rank Mobility
Last moved 1910-1920	0.080	0.168	3.922	0.098
(>20 years of residence)	(0.007)	(0.017)	(0.381)	(0.008)
Last moved 1920-1930	0.147	0.290	7.676	0.145
(10-20 years of residence)	(0.005)	(0.011)	(0.237)	(0.005)
Last moved 1930-1935	0.114	0.138	5.865	0.118
(5-10 years of residence)	(0.004)	(0.011)	(0.232)	(0.005)
Last moved 1935-1940	0.090	0.127	4.441	0.097
(<5 years of residence)	(0.004)	(0.011)	(0.218)	(0.005)
HH FE	Yes	Yes	Yes	Yes
Birth order, age and education controls	Yes	Yes	Yes	Yes
Observations	209,072	104,920	209,072	209,072
R-squared	0.745	0.677	0.744	0.682

Notes: Data are from 1910, 1920, 1930 and 1940 linked sample. The sample splits ever migrants into when they last moved. The table shows that the migration premium increases between 0 and 20 years of residence before dropping down.

Table A14. Selection into temporary migration based on pre-migration earnings score

	I	II	III	IV
	Log earnings score		Percentile Rank	
Panel A. 1920 census (sample limited to 18+ year olds)				
Eventual permanent migrant	-0.142	0.006	-5.546	0.480
(migrated b/w 1920 and 1940)	(0.009)	(0.027)	(0.296)	(0.889)
Eventual temporary migrant	-0.167	-0.016	-6.314	-0.262
(migrated b/w 1920 and 1940)	(0.020)	(0.044)	(0.593)	(1.619)
HH FE	No	Yes	No	Yes
Age, education, birth order control	Yes	Yes	Yes	Yes
Observations	53,491	53,491	53,491	53,491
R-squared	0.097	0.949	0.133	0.946
Panel B. 1930 census				
Eventual permanent migrant	-0.087	0.014	-3.878	0.747
(migrated b/w 1930 and 1940)	(0.006)	(0.010)	(0.247)	(0.401)
Eventual temporary migrant	-0.093	-0.008	-4.054	0.169
(migrated b/w 1930 and 1940)	(0.020)	(0.027)	(0.738)	(1.123)
HH FE	No	Yes	No	Yes
Age, education birth order control	Yes	Yes	Yes	Yes
Observations	114,860	114,860	114,860	114,860
R-squared	0.183	0.855	0.196	0.860

Notes: Data are from 1910, 1920, 1930 and 1940 linked sample. This recreates the analysis for Table 2, but splits ever migrants into those who returned home by 1940 ("temporary migrant") and those did not ("permanent migrant"). The table shows that there was no within-brother selection into temporary or permanent migration prior to moving.

Table A15. The temporary migration premium is similar for a triple-linked sample and single-linked sample.

	I Triple-linked sample	II Single-linked sample	III Triple-linked sample	IV Single-linked sample
	Log earnings score		Log wage income, wage workers	
Permanent Migrant	0.140	0.139	0.234	0.245
	(0.004)	(0.001)	(0.009)	(0.004)
Temporary (pre-1935 return home)	0.011		-0.044	
	(0.006)		(0.015)	
Temporary (post-1935 return home)	0.021	0.008	-0.071	-0.050
	(0.009)	(0.004)	(0.027)	(0.011)
HH FE	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes
Observations	209,072	995,016	104,920	539,014
R-squared	0.747	0.734	0.681	0.667
	Percentile rank		Upward rank mobility	
Permanent Migrant	7.261	7.236	0.147	0.147
	(0.189)	(0.077)	(0.004)	(0.002)
Temporary (pre-1935 return home)	0.108		0.006	
	(0.309)		(0.007)	
Temporary (post-1935 return home)	0.612	0.162	0.016	0.014
	(0.490)	(0.219)	(0.010)	(0.004)
HH FE	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes
Observations	209,072	995,016	209,072	995,016
R-squared	0.747	0.732	0.684	0.667

Notes: Data are from 1910, 1920, 1930 and 1940 linked sample ("triple-linked sample") and the 1910-1940 linked sample ("single-linked sample"). This table shows that the return to temporary migration is similar in the single-linked sample as the triple-linked sample. One concern is that the triple-linked sample assigns temporary migration status to false links, a concern which is reduced by the similarity of results across triple-linked and single-linked samples.

Table A16. Temporary migrants had higher earning scores in 1930 before returning by 1940

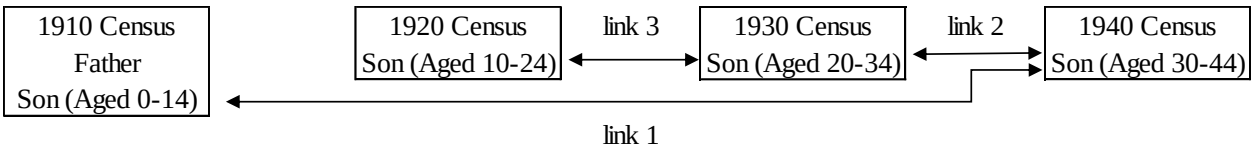
	1930 log earnings score		1930 percentile rank	
Permanent migrant (stayed between 1930 and 1940)	0.074	0.232	2.878	10.557
	(0.003)	(0.009)	(0.158)	(0.381)
Temporary migrant (returned between 1930 and 1940)	0.053	0.175	1.130	6.855
	(0.007)	(0.013)	(0.325)	(0.565)
Education	0.079	0.034	4.182	1.858
	(0.001)	(0.001)	(0.025)	(0.055)
HH FE	No	Yes	No	Yes
Age, Birth Order Controls	Yes	Yes	Yes	Yes
Observations	177,301	177,301	177,301	177,301
R-squared	0.206	0.810	0.218	0.813

Notes: Data are from 1910, 1920, 1930 and 1940 linked sample. This table shows that the return to temporary migration was high prior to returning home. The results suggest that temporary migrants may have experienced a negative shock between 1930 and 1940.

Appendix B. Details on linking data

I combine three different linked datasets in this paper: 1910-1940 (sons from childhood to adulthood); 1940-1930 (sons to another observation); and 1930-1920 (sons to another observation) – see Figure B1. The first link (1910-1940) allows me to observe the adult outcomes of both the father and son because I observe fathers and sons at the same time in the household in 1910, and then the son's adult outcome in 1940. The second link takes sons in 1940 and finds them in the 1930 censuses; the third link finds them in the 1920 census. I will describe each of the three links (1910-1940; 1920-1930; and 1930-1940) links in detail.[1] Note that all links are made in the same basic way (that is, based on the method described in Feigenbaum (2016)).

Figure B1. Linking Process to build the dataset



Building the set of potential matches.

I build new datasets of US-born whites and US-born Blacks by linking the 1910-1940, 1920-1930 and 1930-1940 censues. I use the same broad strategy as in Feigenbaum (2016) where I build a set of potential links, handlink a subset of them, and then train a probit to pick the best link.

I first extract the entire set of US-born white and Black males who are over 10 and under 40 years of age in both 1920 and 1930. For the 1910-1940 link, I extract US-born sons who are 0-14 years old from the 1910 census. After dropping those with the exact same combinations of first name string, last name string, race, state of birth and year of birth, I then search for all possible combinations in the second census that meet the following criteria:

1) First letter of first name match
2) First letter of last name match
3) Jaro-Winkler distance of first name is less than 0.20
4) Jaro-Winkler distance of last name is less than 0.20
5) Year of birth is less than three years in difference

[1] Note that the 1910-1940 link is the same one created in Kosack and Ward (2020, Appendix B).

6) State of birth and race match exactly

The first two criteria differ from Feigenbaum (2016), who does not block on first letters of last or first name; I keep these criteria to reduce computing costs and keep the matching process manageable when matching complete to complete-count censuses. The race match requirement also misses some matches because race identification may change between censuses; therefore, the US-born Black results only apply to fathers and sons listed as Black in all censuses. Finally, I do not block on mother or father's state of birth because there appears to be some error in how these variables are recorded, perhaps because another person of the household was answering the enumerator for the entire household. (Further, mother and father's state of birth available in the 1940 census). However, mother and father's state of birth is useful for choosing the best matches for the 1920-1930 and 1930-1940 links, so I will incorporate it into the probit model.

Based on these linking criteria, not everyone in the starting census has a potential match in the second census. For the white population, about 70 to 80 percent of the starting sample has a possible match in the second census; for the Black population, only about 60 percent of the starting sample has a possible match. The different rates for the Black and white population may reflect differential mortality between the two groups, or that true matches in the Black population are less likely to meet the above criteria. Either way, the results suggest that the maximum linking rate is not near 100 percent even if I could find a true link among the set of potential matches. However, I first need to determine which of the potential matches is the true link.

Choosing the best link.

After creating the set of potential matches, I draw a sample of 2,000 Black and 2,000 white individuals and all of their potential matches in the later census. I do this each for the 1920-1930 match and the 1930-1940 match, where the 2,000 are drawn from the starting census. These three datasets will form the basis of the training data, but first I need to handlink the people in the dataset.

From the dataset of potential matches, I handpick which is the best match. If there are two close potential links that look similar to the original link, then I do not pick a match because I am not confident which one is the true link. The matching rates for the training data are given in Table B1. After going through this handlinking process, I am able to find a true link for about 60 percent of the white population with at least one potential match, and about 50 percent of the Black

population with at least one potential match. Part of the reason why I fail to find a link for all of the training data is that none of the potential links are close in names or year of birth; part of the reason is that there are multiple good matches.

Table B1. Details for the handlinked dataset

	1910-1940		1920-1930		1930-1940	
	White	Black	White	Black	White	Black
Random sample in base year	2,000	2,000	2,000	2,000	2,000	2,000
Potential links in second census	16,248	9,695	15,547	9,302	14,550	9,137
Successfully linked	1,121	862	1,224	992	1,198	958
Handlinking Rate for training data (given 1 potential match)	56.1	43.1	61.2	49.6	59.9	47.9

Notes: Data are from the handlinked sets from 1910-1940, 1920-1930 or 1930-1940.

With the training dataset of potential links and actual links in hand, I model the true link as a function of observable differences between matches. I include the Jaro-Winkler distance in the first and last name; absolute difference in year of birth; number of potential links and its square; mother's place of birth and father's place of birth. I also include information on whether there are unique and exact matches for either the first or last name in terms of NYSIIS codes or exact string match; this is based on the handlinking process where having the same last name that was unique (that is, no other potential links have the same last name) was a strong predictor of a link. The probit models for each of the 1920-1930 and 1930-1940 matches, separately by Black and white, are shown in Table B2, and the one for the 1910-1940 link is in Table B3.

The probit models estimate a predicted match score for each potential link in the training dataset. From this information, I set two tuning parameters to determine who will be included in my linked dataset. The first parameter is the cut off for predicted probability, where a potential link needs to have a predicted probability above this level to be included in the linked dataset. The second parameter is the ratio of the 1st best probability to the 2nd best probability; this ensures that I do not keep a match that has a close alternative. I set these parameters to maximize the efficiency of the algorithm in terms of true positive rate (TPR, or the percentage of true links that I keep), as long as the positive predictive value (PPV) is at least 0.9. The positive predictive value is the ratio of true positives to total matches; viewed from the opposite direction, it sets the false-positive rate

to 10 percent. This false positive rate is slightly lower than Feigenbaum's (2016) training data in Iowa and thus is on the conservative end; however, one could easily change this parameter to be more or less restrictive. A consequence of the decision to limit false positives is that it reduces the matching rate for the full sample. See Table B4 for the tuning parameters and the resulting PPV and TPR.

Table B2. Predicting the handlinked match using a probit model

	1920-1930 White	1920-1930 Black	1930-1940 White	1930-1940 Black
Jaro-Winkler Distance, First name	-6.073***	-6.992***	-7.622***	-6.012***
	(0.553)	(0.530)	(0.669)	(0.542)
Jaro-Winkler Distance, Last name	-13.07***	-13.11***	-14.38***	-13.17***
	(0.869)	(1.012)	(0.969)	(0.977)
Year of Birth Difference = 1	-0.180	-0.172	-0.678***	-0.247
	(0.128)	(0.178)	(0.138)	(0.161)
Year of Birth Difference = 2	-0.544***	-0.254	-1.102***	-0.442***
	(0.144)	(0.173)	(0.165)	(0.158)
Year of Birth Difference = 3	-0.920***	-0.634***	-1.546***	-0.630***
	(0.156)	(0.179)	(0.191)	(0.164)
No. of potential links	-0.0257	-0.100***	-0.0932***	-0.126***
	(0.0192)	(0.0199)	(0.0212)	(0.0219)
No. of potential links squared	-9.29e-05	0.00253***	0.00211***	0.00301***
	(0.000668)	(0.000803)	(0.000739)	(0.000901)
Unique and Exact NYSIIS First name match	0.315**	0.266**	-0.0212	0.121
	(0.156)	(0.132)	(0.164)	(0.119)
Unique and Exact NYSIIS Last name match	-0.00698	0.797**	0.0484	0.157
	(0.304)	(0.340)	(0.307)	(0.509)
Unique and Exact NYSIIS First AND Last name match	0.616***	0.961***	1.062***	0.994***
	(0.128)	(0.119)	(0.134)	(0.115)
Unique Exact Last name String match	0.491***	0.137	0.769***	-0.242
	(0.180)	(0.219)	(0.215)	(0.226)
Middle initial match, if have one	0.718***	1.557***	1.121***	1.163***
	(0.113)	(0.387)	(0.139)	(0.271)
NYSIIS last name match AND Year of Birth Diff=0	1.539***	0.850***	0.889***	1.343***
	(0.245)	(0.285)	(0.235)	(0.453)
NYSIIS last name match AND Year of Birth Diff=1	1.172***	0.790***	0.915***	1.147**
	(0.241)	(0.274)	(0.251)	(0.452)
NYSIIS last name match AND Year of Birth Diff=2	0.914***	0.169	0.470*	0.792*
	(0.255)	(0.270)	(0.275)	(0.445)
2 Potential links with NYSIIS last name match	-0.582***	-0.498**	-0.622***	-0.811***
	(0.178)	(0.228)	(0.193)	(0.260)
>2 potential links with NYSIIS last name match	-0.952***	-0.286	-0.684***	-1.014**
	(0.227)	(0.263)	(0.228)	(0.439)
2 Potential links with last name string match	-0.798***	-0.450**	-0.935***	-0.321*
	(0.151)	(0.182)	(0.184)	(0.193)
>2 Potential links with last name string match	-1.415***	-1.168***	-1.433***	-1.501***
	(0.133)	(0.144)	(0.142)	(0.144)
One potential link	0.636***	0.645***	0.986***	0.980***
	(0.170)	(0.133)	(0.180)	(0.123)
Difference in length of last name strings	-0.383***	-0.479***	-0.619***	-0.552***
	(0.0586)	(0.0711)	(0.0738)	(0.0683)
Mother place of birth match	0.653***	0.327***		
	(0.0784)	(0.119)		
Father place of birth match	0.546***	0.0173		
	(0.0775)	(0.110)		
Constant	0.269	0.648***	1.845***	1.322***
	(0.182)	(0.229)	(0.190)	(0.200)
Observations	15,547	9,302	14,205	9,096

Notes: Data are from the handlinked sample between 1920-1930 or 1930-1940. The coefficients are from a probit model that predicts the correct link.

Table B3. Modeling the linking process with a probit, 1910-1940

	White	Black
Jaro-Winkler Distance, First name	-4.885***	-4.203***
	(0.576)	(0.635)
Jaro-Winkler Distance, Surname	-13.64***	-12.35***
	(0.853)	(1.113)
Year of Birth Difference = 1	-0.557***	0.0740
	(0.114)	(0.198)
Year of Birth Difference = 2	-0.906***	-0.199
	(0.131)	(0.202)
Year of Birth Difference = 3	-1.426***	-0.355*
	(0.157)	(0.203)
Number of Potential links	-0.114***	-0.129***
	(0.0187)	(0.0235)
Number of Potential links squared	0.00217***	0.00283***
	(0.000639)	(0.000922)
Exact surname match AND unique surname	0.619***	0.228
	(0.235)	(0.303)
Exact first and surname string match AND unique first and surname string	0.382**	0.763***
	(0.159)	(0.154)
Exact first name match AND unique first name	-0.391*	-0.0317
	(0.205)	(0.178)
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***
	(0.192)	(0.168)
Exact Soundex first and surname match AND unique soundex first and surname	0.829***	0.622***
	(0.188)	(0.158)
Exact NYSIIS first name match AND unique NYSIIS first name	0.204	0.489**
	(0.298)	(0.215)
Exact NYSIIS surname match AND unique NYSIIS surname	0.445	-2.197
	(0.320)	(126.1)
Exact NYSIIS first and surname match AND unique NYSIIS first and surname	0.0126	0.253
	(0.196)	(0.171)
Middle initial match, if have one	1.212***	1.068***
	(0.103)	(0.201)
NYSIIS last name match AND YOB Diff=0	1.131***	4.537
	(0.234)	(126.1)
NYSIIS last name match AND YOB Diff=1	1.066***	4.175
	(0.243)	(126.1)
NYSIIS last name match AND YOB Diff=2	0.795***	3.757
	(0.255)	(126.1)
2 Potential links with NYSIIS last name match	-0.308**	-0.951**
	(0.147)	(0.407)
>2 potential links with NYSIIS last name match	-0.637***	-3.921
	(0.224)	(126.1)
2 Potential links with last name string match	-1.090***	-0.682**
	(0.181)	(0.273)
>2 Potential links with last name string match	-1.575***	-1.194***
	(0.123)	(0.162)
One potential link	0.398**	0.407***
	(0.173)	(0.143)
Constant	1.370***	0.318
	(0.167)	(0.233)
Observations	16,248	9,695

Notes: This paper 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.

Table B4. Tuning parameters for determining who to keep in the linked sample

Census Years	Race	Cutoff for predicted probability	Score Ratio of 1st best link to 2nd best	PPV	TPR
1910-1940	White	0.335	2.6	0.901	0.790
	Black	0.784	5.8	0.901	0.580
1920-1930	White	0.412	2.5	0.900	0.786
	Black	0.587	2.8	0.901	0.596
1930-1940	White	0.33	4.4	0.900	0.836
	Black	0.639	1.7	0.901	0.631

Notes: PPV stands for positive predictive value and gives the ratio of true positives to all links. TPR stands for true positive rate and gives the proportion of true links that would appear in the final linked dataset.

Provided these estimates from the probit model, I then predict the linking scores for the full to full count match with the probit model; afterward, I keep only those who meet the parameters set in Table B4. See Table B5 and B6 for the linking rates when applying this process to the full-count data. I link of 32 to 36 percent of the white population, and 15 to 17 percent of the Black population for the decadal links. The 1910-1940 link is 29.8 for the white population and 11.9 for the Black population. These linking rates are lower than Feigenbaum's link from the 1915 Iowa Census to the 1940 Federal Census of near 60 percent. This may be due to several reasons: because Iowa is a smaller state and thus has fewer other potential matches, because the data quality is higher from Iowa, because modeling the hand-linking process is easier for Iowans versus the rest of the country, or because there are lower mortality rates for Iowans relative to the rest of the country. I also have more restrictive requirements for a potential link (requiring the first letter of the first name and first letter of the last name to match exactly). While the linking rate is somewhat low, I still have millions of individuals linked across censuses.

Table B5. Applying the probit model to the full 1920-1930 and 1930-1940 link, details

	1920-1930 Census		1930-1940 Census	
	White	Black	White	Black
Starting group in base year	21,234,490	2,819,891	25,676,888	3,193,903
Starting group in base year with a potential link ten years later	16,320,377	1,489,797	19,144,294	1,979,364
Potential links ten years later	159,810,633	7,071,600	258,313,387	9,217,794
Unique match amongst links	7,559,217	418,958	8,197,666	539,359
Overall Linking Rate	**35.6**	**14.9**	**31.9**	**16.9**
Linking Rate given Potential Match	46.3	28.1	42.8	27.2

Table B6. Applying handlinking results to full 1910-1940 link, details

	Anglo American	African Americans
Starting group in 1910	12,567,861	1,851,076
Starting group in 1910 with a potential link in 1940 based on criteria	10,180,244	1,094,394
Potential links in 1940	136,372,727	6,085,262
Unique match in 1940 amongst links	3,748,917	220,145
Overall Linking Rate	**29.8**	**11.9**
Linking Rate given Potential Match	36.8	20.1

Getting into the sample used in the main analysis

To be included in the final linked sample used in this paper, an individual must be in the 1910-1940, 1930-1940 and 1920-1930 link. That is, the son must survive being triple linked. The resulting sample is of 949,333 sons. This number is only 9.1 percent of the 1910 children that I could have possibly linked to the 1940 census. Given that the general linking rate of two censuses is around 33 percent, it would be expected that about (0.33)(0.33)(0.33) = 3.6 percent of individuals would be linked three times if linking is independent across censuses. The actual linking rate is higher than 3.6 percent, indicating that being successfully linked is not independent,

perhaps because individuals have unique names. The linking rate for Black sons is even lower at 2.0 percent of the original sons; this reflects that the Black linking rate is lower at around 15 percent.

Weighting

Only a select group (9.1 percent) of the original population shows up in the triple-linked sample. Therefore, this group may be unrepresentative of the original population and provide misleading information on the convergence of economic gaps. I address this problem by reweighting the data to be representative of the population. To weight the data, I use the inverse probability weight approach as suggested by Bailey et al. (2019). The process is as follows: pool the linked and linkable sample together (that is, the children in 1910), run a probit to determine which observables predict being in the linked sample, and then weight each observation in the linked sample using the inverse probability weight.[2]

The representativeness of the sample (as found in the probit model) is shown in Table B7. I run representativeness checks separately by the white population and Black population so that I tailor the weights to be race-specific. Table B7 shows that there is selection into the linked sample, where fathers with white-collar jobs and farmers are more likely to be in the sample than unskilled or semi-skilled fathers. Further, those in the Midwest and West are more likely to be in the sample than those in the South or in the Northeast. Therefore, estimating the mobilities using the unweighted data will erroneously reflect Midwestern rural states like Iowa, rather than the full population. The weighted representative characteristics are also shown in Table B7.

The final step to the weighting process is to upweight African Americans. Because the regressions in Table B7 are done separately by Black and white, the average weight for the white sample is about one and the average weight in the Black sample is about one. However, because the linking rate is lower for African Americans, they only end up being 2.22 percent of the linked sample, in contrast to 9.25 percent of population (of 30-44 year old adults in 1940). Therefore, I reweight the Black sample up by multiplying the Black weights by 9.25/2.22; I also reweight the white sample down by multiplying their weights by 90.75/97.78. Now the Black weighted proportion of the sample reflects the population proportion.

[2] Let q be the share of linked records and p be the predicted probability. The weight is $[(1-p)/p] \times [q/(1-q)]$

Table B7. Representativeness of the linked sample on a probit

	White, unweighted	White, weighted	Black, unweighted	Black, weighted
Length of First name	0.0267***	0.00387***	0.0247***	0.00340**
	(0.000310)	(0.000316)	(0.00157)	(0.00153)
Length of last name	0.0203***	0.000425	0.00220	0.000912
	(0.000308)	(0.000322)	(0.00172)	(0.00180)
Urban	-0.0390***	0.00777***	0.0169**	0.00815
	(0.00141)	(0.00150)	(0.00818)	(0.00853)
Father has white-collar job	0.0879***	-0.00281	0.150***	0.0104
	(0.00177)	(0.00186)	(0.0201)	(0.0213)
Father is farmer	0.124***	-0.00610***	0.156***	0.00625
	(0.00177)	(0.00185)	(0.0145)	(0.0155)
Father has unskilled job	-0.0145***	0.000245	0.150***	-0.00289
	(0.00165)	(0.00175)	(0.0141)	(0.0152)
Age in 1910	-0.0231***	-0.00122**	-0.0185***	-0.00308
	(0.000461)	(0.000474)	(0.00229)	(0.00230)
Age in 1910 squared	0.000818***	7.73e-05**	0.000302*	0.000220
	(3.21e-05)	(3.34e-05)	(0.000160)	(0.000165)
Midwest	0.141***	0.00703***	0.0378**	6.14e-05
	(0.00140)	(0.00145)	(0.0151)	(0.0139)
South	-0.0976***	0.0177***	-0.326***	0.00157
	(0.00160)	(0.00174)	(0.0124)	(0.0118)
West	0.203***	0.0100***	0.0604*	-0.0139
	(0.00213)	(0.00215)	(0.0361)	(0.0325)
Constant	-1.543***	-1.217***	-1.874***	-1.969***
	(0.00374)	(0.00386)	(0.0247)	(0.0248)
Observations	10,462,321	10,462,321	1,070,349	1,070,349

Notes: The regression is the pooled linkable and linked sample between 1910 and 1940. The dependent variable is an indicator for being in the linked sample. The likelihood of being in the linked sample is modelled in a probit. The probit coefficients are reported in this table.

Appendix C. Earnings score

The main earnings score used for estimating rank-rank associations follows the process of Collins and Wanamaker (2017), who provide more detail on assumptions behind the adjustments. I provide the general strategy here for the interested reader. The benefit of the score over the traditional occupational score based on *occscore* from IPUMS is that this earnings score further adjusts income by state of residence and race. These adjustments are key for estimating the economic benefit from internal migration, because income gaps were large across geography and Black/white families.

I first use the 1940 census and limit the sample to 25 and 55-year-olds to measure occupational-based earnings at prime ages in the lifecycle. I then separate the sample into cells based on 3-digit occupational code (occ1950), race/ethnicity (Black/white/Mexican), and state of residence. After splitting the 1940 full-count census into cells, I use the average income in the cell as the earnings score. While this may seem straightforward, a few further corrections need to be made before taking the average income to address that the 1940 census only includes wage income but not self-employed earnings. That is, I need to estimate how much self-employed workers earn. To do so, I take the ratio of total income to wage income by occupational code in the 1960 census. I then multiply this ratio by the average wage income in the 1940 census in the occupation/race/state cell for self-employed workers.

I also make adjustments for farmers and farm laborers to address that some compensation may be in-kind. To do so, in the 1960 census I increase farmer income by 35 percent and farm laborer income by 19 percent (Collins and Wanamaker 2017). Then, I take the ratio of farmer to farm laborer income in the 1960 census in order to proxy for how much farmers earn over farm laborers in the 1940 census. Before multiplying this ratio by farm laborer wages in the 1940 census, I increase farm laborer wages by 26 percent in 1940 to reflect perquisites in this earlier period. Note that I use the same ratios for land-holding farmers (who are assumed to be farm owners) and non-land-holding farmers (who are assumed to be farm tenants). According to my data, this results in increasing farmer income by about 43 percent to account for perquisites.

After these adjustments, the average earnings in the occupation/race/state cell is the earnings score. If there are fewer than 30 people in the regional cell, then the earnings score is the

national average in the occupation/race cell. Finally, if there are less than 30 people in the national cell, then the earnings score is the average income at the 1-digit level by race.

Appendix D. The importance of measurement error

One issue with estimating the relationship between migration and intergenerational mobility is bias from measurement error. It is difficult to measure the father's and son's economic status with a single observation, whether due to issues recording occupations or because permanent status is not observed (Solon 1992, Ward 2020). For example, consider the following regression:

$$y^*_{son} = \beta_0 + \beta_1 y^*_{father} + \beta_2 EverMigrant + \varepsilon$$

Where the son's true outcome is regressed on the father's true outcome and whether one ever migrated. However, typically one observes the father's and son's status with error:

$$y_{father} = y^*_{father} + u_{father}$$

$$y_{son} = y^*_{son} + v_{son}$$

The observed values are y_{father} and y_{son}. The error terms u_{father} and v_{son} are usually assumed to be independent of the true values and of the main error term ε (i.e., classical measurement error).

It is well known that classical measurement error attenuates the father's coefficient, but a less recognized bias is that on the coefficients for correctly measured variables, like for β_2 on $EverMigrant$. Given classical measurement error, the migration coefficient can be biased in unclear ways. The measurement error formula is:[3]

$$plim\ \widehat{\beta_2} = \beta_2 + \left(\frac{\sigma^2_{u_{father}}}{\sigma^2_{y^*_{father}} + \sigma^2_{u_{father}}} \right) \beta_1 \alpha_2$$

Where σ^2_{father} is the variance of the auxiliary regression of the true y^*_{father} on $EverMigrant$ and α_2 is the coefficient on $EverMigrant$ in the auxiliary regression. The overall bias depends on selection into internal migration on father's status (α_2), the true father-son association (β_1), and the amount of error in the father's status $(\sigma^2_{u_{father}})$. Because it is clear that the father-son association β_1 and error $\sigma^2_{u_{father}}$ are positive, the direction of the bias comes down to selection into internal migration on the father's status. The pattern in my data suggests that selection is negative, and therefore that migration premium is negatively biased. A negative bias to the

[3] See Garber and Klepper (1980).

migration premium reinforces my main argument that the association between migration and the son's outcome is large.

My preferred method of estimating the return to migration includes household fixed effects. A household fixed effect model addresses measurement error in the father's status because the father's outcome (and error) are absorbed and therefore the migration premium is estimated without bias from measurement error. However, a main question in this paper is whether the migration premium is larger or smaller for children raised in poorer households. That is, how the migration premium varies with the father's economic status:

$$y_{son} = b_1 EverMigrant + b_2(EverMigrant \times y_{father}) + \eta_h + \varepsilon + u_{son}$$

For this specification, classical error in the father's outcome attenuates b_2 and therefore will understate the relative premium for poorer households relative to richer households. The bias to b_1 depends on the excluded group ($y_{father} = 0$), but in general the migration premium is understated for poorer households and overstated for richer households. If one instead uses the percentile rank of the father's and son's outcomes, which can be written as the location in the cumulative distribution $F(y_{father})$ and $F(y_{son})$, then the bias is less clear because the error is now non-classical (Nybom and Stuhler 2017). However, because measurement error also attenuates rank-rank associations, then the bias to b_1 and b_2 are likely similar for percentile ranked and non-ranked outcomes. In the next section, I will simulate the effect of measurement error to clarify the direction and magnitude of error.

One way to address measurement error is to instrument one observation of the father's outcome with a second. Under classical measurement error, the error in the second observation is uncorrelated with the first observation, which leads to consistent estimates of b_1 and b_2. However, instrumenting may not be as successful for ranked outcomes due to non-classical error, because the errors may be correlated across observations for the top or bottom of the distribution. I will also simulate how 2SLS method performs when estimating the migration premium in the next section. I will also show the importance of the more traditional method of averaging two father and son observations together, because my main dataset also contains two father and two son observations.

A final complication arises when using binary dependent variables, such as whether the son improved on the father's outcome. Error in a binary dependent variable leads to a special situation of non-classical error, partially because the error always goes in the opposite direction of the true value (i.e., negatively correlated). For example, let $Upward^* = 1[y^*_{son} > y^*_{father}]$. Let π_{10} be $P[Upward = 1|Upward^* = 0]$ and π_{01} be $P[Upward = 0|Upward^* = 1]$. If one simply estimates the association between migration and upward mobility:

$$Upward = \delta_0 + \delta_1 EverMigrant + \varepsilon$$

Then the effect of migration on upward mobility can be written as:[4]

$$\widehat{\delta_1} = [1 - (\pi_{10} + \pi_{01})]\delta_1$$

The upshot is that false positives and false negatives attenuate the effect of migration on upward mobility. A similar result would hold for the effect of migration on upward rank mobility.

Monte Carlo Simulation

To clarify the effect of measurement error on the estimated effect of migration, I run a simulation. I specify the data generating process (DGP) so that the son's true outcome is a function of the father's status, migration and the interaction between migration and the father's status:

$$y^*_{son} = \gamma_0 + \gamma_1 EverMigrant_i + \gamma_2 y^*_{father} + \gamma_3 (EverMigrant_i \times y^*_{father}) + \eta_h + \varepsilon_i$$

My goal is to test how the coefficients are biased when using values of y^*_{son} and y^*_{father} that are measured with error. A household-specific constant is also included to reflect fixed factors that influence brothers' outcomes. For convenience, I will describe the results as if y^*_{son} and y^*_{father} are the log earnings scores for the father and son.

I create the data to follow the patterns from the early 20th century data. Specifically, I assume that the return to migration is positive for children from average-status households ($\gamma_1 = 0.15$) and the return to migration falls across the economic distribution ($\gamma_3 = -0.10$). I also assume that the reliability ratio $\left(\sigma^2_{y^*} \Big/ \sigma^2_{y^*} + \sigma^2_u\right)$ is about two-thirds, which fits the results in Ward

[4] See Bound et al. (2001).

(2020). I assume that the errors are classical such that they are independent of the true values or the main error term. While I specify the above DGP, I estimate household fixed effects specifications such that η_h drops from the model.

Table D1 shows that classical measurement error does not strongly bias the migrant coefficient γ_1 but attenuates the interaction γ_3. Based on the simulated data, the within-brother estimate of γ_1 is 0.151 and the interaction is estimated at -0.101 (the specified values of 0.15 and -0.10 are within the standard errors). After adding error to the father and son's outcome and using one father observation, the estimate of γ_1 remains at 0.148 while the interaction falls to -0.068. Therefore, an attenuation of about 30 percent. The attenuation of the interaction is expected, but the lack of bias to γ_1 may not be. Note that the data is centered at the mean of the father's outcome, so the lack of bias to γ_1 is because the estimate is unbiased for the mean father's outcome. This relationship can be seen in Figure D1 panel A.

One way to fix measurement error is to average father's and son's observations. However, average the son's outcome does not matter because it is the dependent variable and error is classical. When averaging the father's observation, the interaction is 0.083, or 17 percent lower than the true value.

Instrumental variables help to solve the attenuation bias problem when error is classical, as seen in Column 3 of Table D1. After instrumenting the interaction between migration and the father's outcome with an interaction with the second observation, then γ_3 is estimated at -0.111. This estimate is slightly higher than the true estimate. I am unable to reject that this estimate is statistically different from the -0.101 estimate in the actual data or -0.10 in the DGP. Figure D1, Panel B also plots the estimated migration premium from 2SLS, which shows that the predicted migration premium is slightly overstated for children from poorer families but understated for children from richer families.

Figure D2 shows that the measurement error leads to different biases for the percentile ranked version (see point estimates in Panel B of Table B1). Now, the migration premium is attenuated across the entire distribution. However, the size of the bias is more severe for children from poorer households than for children from richer households. Further, unlike the log-log specification, 2SLS does not fix the biases for the percentile ranked specification (see Panel B of

Figure D2). Rather, it tends to overstate the interaction between the father's rank and percentile rank. That is, the true relationship is between the FE and FE-IV specification. Either way, the simulated data suggest that the percentile rank migration premium estimates from the main section understate the true premium.

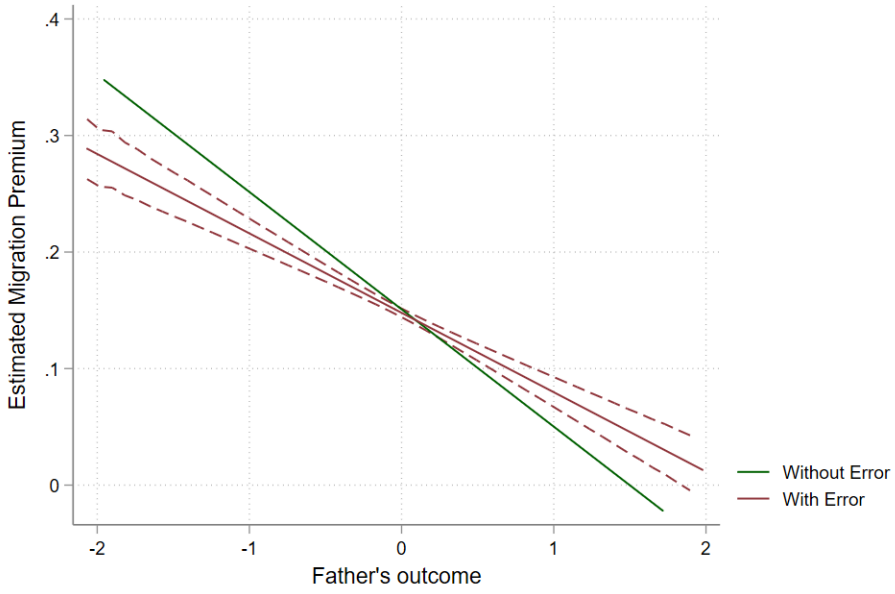
Finally, Figure D3 shows that migration's effect on upward rank mobility is severely attenuated by measurement error across the father's distribution. While the error-free data predict that children from the poorest families are 18 percentage points more likely to be upwardly mobile, the fixed effect estimate is 12.5 percentage points, an attenuation of nearly a third. Using 2SLS does not fix the attenuation bias problem, but again slightly overstates the interaction between the father's status and migration (albeit the difference is not statistically significant).

Measurement error leads to many false positives and negatives in upward rank mobility. According to the simulated data, false positives and negatives jointly make up about 25 percent of the data when the one father and son observation (See Table D2). If one uses the average of father's and son's observations, then the false positives/negatives drop to 19 percent, which is still a large number despite the decrease. It would be helpful to have more father observations in the historical data, but each link loses more data and makes the sample more selective (Bailey et al. 2019).

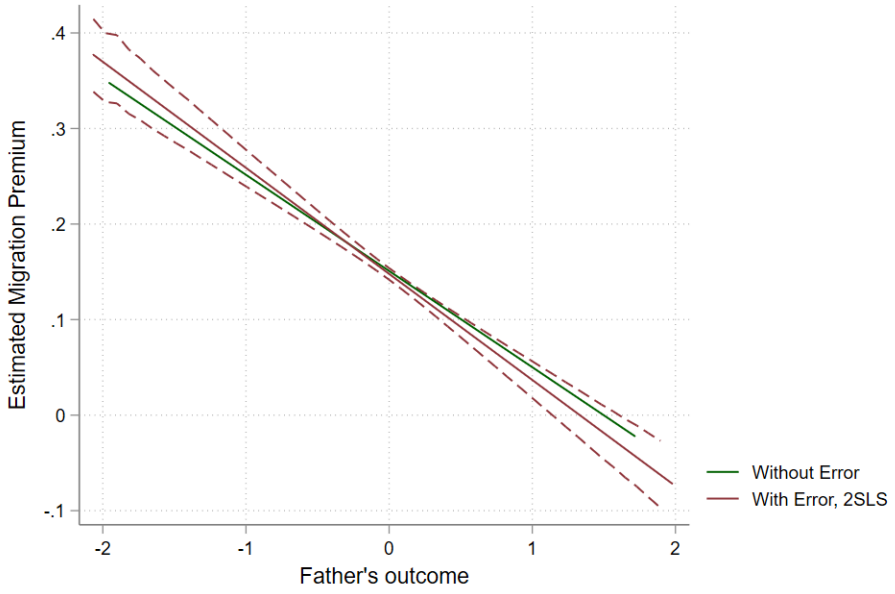
For my actual linked dataset, I can compare the disagreement rate in the simulated data and actual data when using one observation of an average of two observations. Table D3 shows that about 16.5 percent of the simulated data's upward rank mobility measures disagree. In comparison, the actual data (when using the average of 1910 and 1900/1920 for the father, and 1930 and 1940 for the son) has a 17.1 percent disagreement. This similarity suggests that the attenuation bias in the simulated data captures the degree of bias in the actual data. Attenuation bias in upward rank mobility is the main reason why I use the log earnings score as the main dependent variable rather than zero-one dependent variables.

Figure D1. Measurement error when using log earnings score

Panel A. No error versus one father observation measured with error



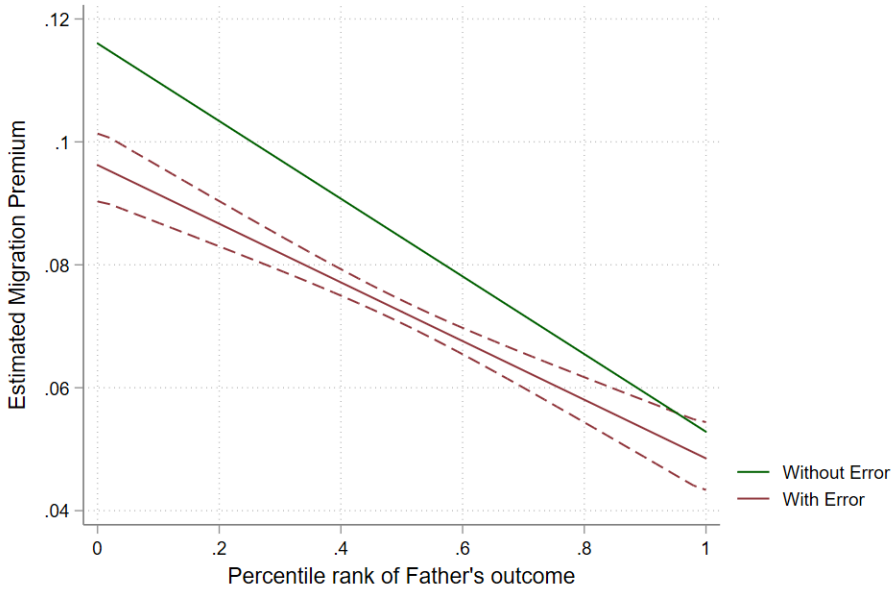
Panel B. No error versus instrumental variables



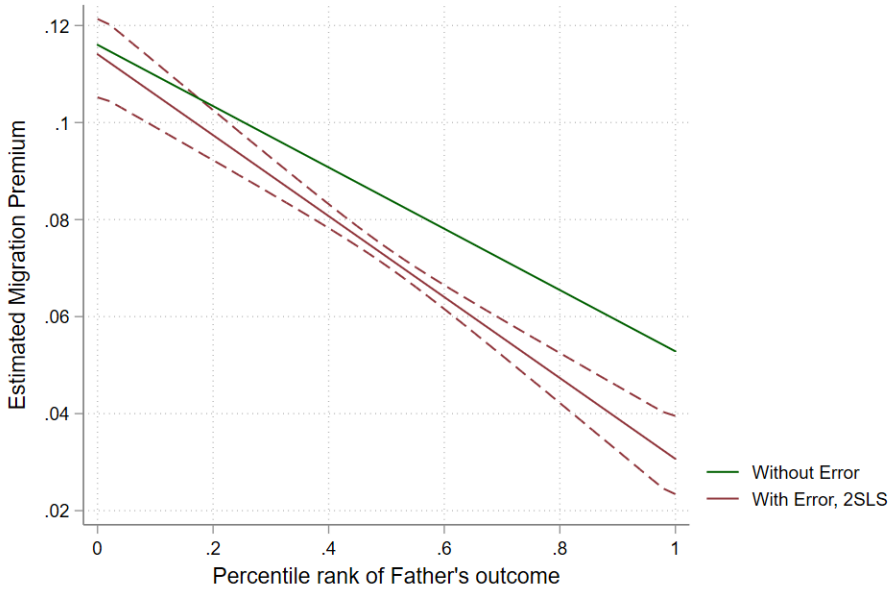
Notes: The above figures show with simulated data that measurement error in the father's status attenuated the migration premium across the distribution.

Figure D2. Measurement error when using percentile ranks

Panel A. No error versus one father observation measured with error



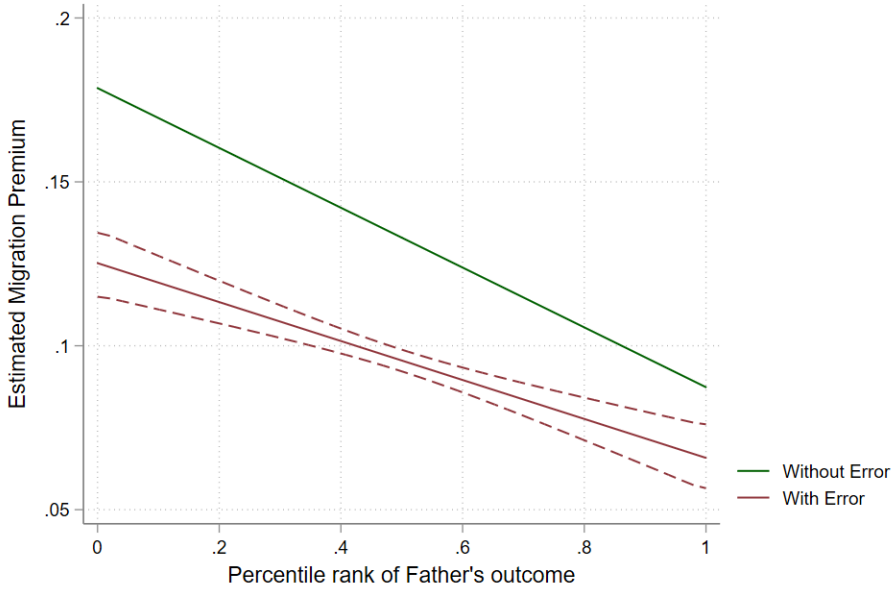
Panel B. No error versus instrumental variables



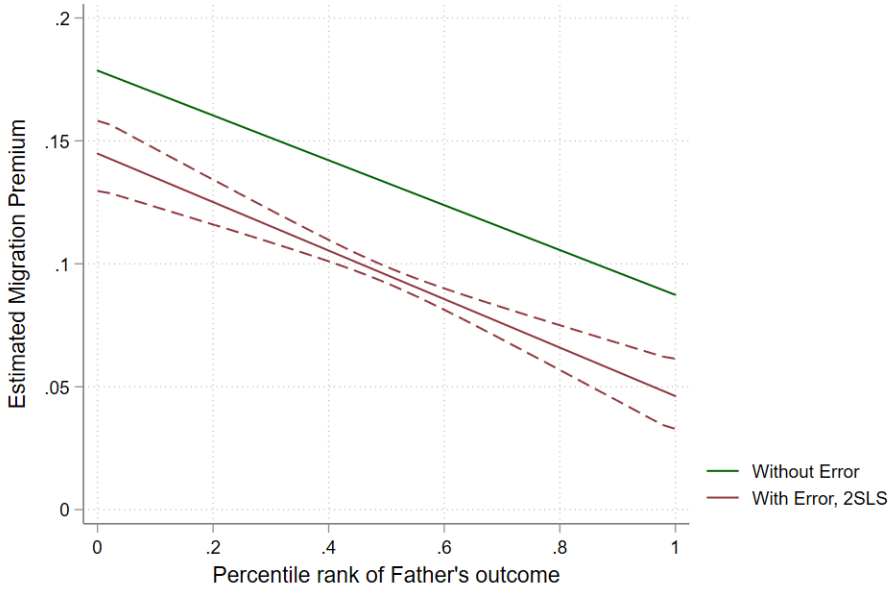
Notes: The above figures show with simulated data that measurement error in the father's status attenuated the migration premium across the distribution.

Figure D3. Measurement error when using upward rank mobility

Panel A. No error versus one father observation measured with error



Panel B. No error versus instrumental variables



Notes: The above figures show with simulated data that measurement error in the father's status attenuated the migration premium across the distribution.

Table D1. Estimates of the migration premium across the father's distribution

	No Error	With Error		
		1 father and son observation	Mean of 2 father and son observation	2SLS
Panel A. Income score				
Migrant	0.151	0.148	0.149	0.148
	(0.003)	(0.003)	(0.003)	(0.003)
Migrant x Father's score	-0.101	-0.068	-0.083	-0.111
	(0.006)	(0.006)	(0.006)	(0.009)
Panel B. Percentile rank				
Migrant	0.116	0.096	0.105	0.114
	(0.003)	(0.003)	(0.003)	(0.004)
Migrant x Father's rank	-0.063	-0.048	-0.055	-0.083
	(0.005)	(0.005)	(0.005)	(0.008)
Panel C. Upward rank mobility				
Migrant	0.179	0.125	0.145	0.145
	(0.006)	(0.006)	(0.006)	(0.008)
Migrant x Father's rank	-0.091	-0.059	-0.067	-0.099
	(0.010)	(0.010)	(0.010)	(0.014)
Observations	200,000	200,000	200,000	200,000
Number of hh	100,000	100,000	100,000	100,000

Notes: Regression estimates underlying Figures D1-D3.

Table D2. False positives and negatives in upward rank mobility in simulated data

Panel A. No error versus one father observation

		Error, one observation	
		0	1
No error	0	75.6	24.8
	1	24.4	75.2

Panel B. No error versus average of two observations

		Error, average of two observations	
		0	1
No error	0	80.9	19.4
	1	19.1	80.6

Notes: False positives and false negatives in simulated data. No error is the "true" father and son's status rank. Panel A compares the true levels to data based on one father and son observation. Panel B compares the true level to data based on two father and son observations.

Table D3. Disagreements in upward rank mobility when using one or a mean of two observations

Panel A. Simulated data

		Error, one observation	
		0	1
Error, mean of 2 observations	0	83.6	16.6
	1	16.4	83.4

Panel B. Actual linked data

		Error, one observation	
		0	1
Error, average of 2 observations	0	89.7	11.4
	1	24.3	75.7

Notes: False positives and false negatives in simulated data (Panel A) and linked data between 1910, 1920, 1930 and 1940 (Panel B). Panel A compares upward rank mobility measures when using data based on one father observation and one son observation in the simulated data. Panel B males the same comparison but for the actual linked data. Average for father is from 1910 and either the 1900 or 1920 census; average for son is the 1930 and 1940 censuses.