Marriage and the Intergenerational Mobility of Women: Evidence from Marriage Certificates 1850-1910 Supplemental Online Appendix

Jacqueline Craig and Katherine Eriksson and Gregory T. Niemesh

Craig: Research Assistant, Federal Reserve Bank of Chicago. Eriksson: Assistant Professor of Economics, University of California - Davis, and NBER. Niemesh: Assistant Professor of Economics, Miami University, and NBER. We would like to thank Charles Moul, Bhashkar Mazumder, Melissa Thomasson, and participants at Northwestern University's Seminar in Economic History for helpful suggestions. The views expressed in this paper are solely those of the authors and do not reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Appendices

A Matching

In our context, the matching algorithm consists of two matches: the first match from marriage certificates to the post-marriage census, and the second match from the post-marriage census to the pre-marriage census. Figures A1a and A1b provide a visual of the matching process for both cohorts. The first match begins by constructing for each couple i in the marriage certificate sample a pool of potential matches in the census. To be considered a potential link, households in the census must have both spouses present, be within 2 years' difference in year of birth, and have the surname and both given names be similar in terms of string distance to couple i.[1] A random sample of 3,500 marriage certificates for each cohort is chosen for which a potential match is manually chosen as the true match. Because it is conducted by hand, the choice of true match inherently relies on researcher judgment. Only potential matches that are a clear and certain match are marked as a true link. Those without a clear match, either with no good option or multiple potential matches that appear equally likely, are marked as unmatched. The sample of hand-linked matches is split into a set of training data ($N = 2,500$) and sample for cross-validation of the model ($N = 1,000$).

The training data is used to estimate a probit model to predict a true link based on functions of observable characteristics on the marriage certificates and census records. The most important characteristics are the string distances and functions of surnames, given names, and ages of the couple. Predictors in the probit model include: Jaro-Winkler distance for both given names and surname, absolute value of birth year difference for both, polynomial in number of potential matches, and indicators for if the match is exact (Jaro-Winkler distances equal 1 for all 3 names), if the match is exact and birth difference equals 0 for both, if first letter of each name matches, and if last letter of each name matches. The full set of probit coefficients are listed in table A5.

[1] String distance is measured using the Jaro-Winkler method and must be within 0.2 for all names as an initial screen for entry into the pool of potential matches.

With a trained probit model, we estimate predicted values for all potential matches to predict the probability that a potential match is a correct match. To identify a correct match, we are interested in the score, or predicted value, and ratio of the top score to second best score of all potential matches for a specific couple. A correct match must satisfy the following requirements: 1) the match has the highest score of all potential matches, 2) the score is sufficiently high, and 3) the second best score is sufficiently different from the first. To identify which matches satisfy these requirements, we develop thresholds for score and ratio.[2] We calibrate optimal parameters for score and ratio using knowledge of the true matches from the training data and two standard machine learning assessment measures: true positive rate (TPR) and positive predictive value (PPV). TPR is an efficiency metric, which identifies the proportion of true algorithm matches for the total number of true matches. PPV is an accuracy metric, which identifies the proportion of true algorithm matches for the total number of matches by the algorithm.

We estimate separate parameters for couples with one unique potential match vs. couples with multiple potential matches for two reasons: 1) the score thresholds will be significantly different and 2) a ratio will not exist for a unique potential match without the existence of a next best match. We test a grid of values between 0-1 for the score and 1-2 for the ratio to identify the correct combination of parameters that maximize the sum of TPR and PPV.

The potential match with the highest predicted score is taken as the true match. We choose parameters that take into account the match quality of the best potential match (predicted score must be above a cutoff α), as well as the match quality of other potential matches (the true match must have a score significantly better than the next highest score - a distance of β). The choice of α and β captures the trade-offs in increasing the match rate and likelihood of true matches. As α and β increase, the cutoffs for the predicted score become more conservative, decreasing the rate of false positives, but at the cost of decreasing the rate of predicting true positives. As trained and cross-validated on out-of-sample predictions for cohort 1, the algorithm captures 83% of true matches and 87% of the matches are actual true matches. For cohort 2 the rates are respectively 81% and 91%. Appendix Table A7 presents more information about the cross-validation exercise. We successfully link 64,857 couples to the 1880 census in cohort 1 out of 208,026 marriages, and 130,389 couples to the 1910 census for cohort 2 out of 375,195 marriages, giving a 31% and 35% match rate.

The second match applies the same methodology to the successful links from the first match by linking each couple to a pre-marriage census in which we observe the father's economic status. Cohort 1 marriages are linked to the 1850 census, and cohort 2 marriages are linked to the 1880 census. In this match, we seek to link each individual (husband and wife) in the couple separately. The pool of potential matches in this case includes records within two years difference in year of birth, the same birth state, and the given and surname similar to the individual being matched. The algorithm for the second match is independently trained from that used for the first match. For each sex and cohort, 3,500 observations are randomly chosen to hand-code true matches, of which

[2] $Ratio = \frac{BestScore}{SecondBestScore}$

2,500 observations are used to train the model and 1,000 observations used as the cross-validation sample. Probit model predictors include: Jaro-Winkler distances for individual's given name and surname, Jaro-Winkler distances for father and mother's given name, birth year difference, and indicators for if the match is exact (all four Jaro-Winkler distances equal one), if the match is exact and birth difference equals 0, if first letter of each name matches, and if last letter of each name matches. The full set of probit coefficients are listed in Appendix Table A6. For cohort 1, the cross-validated trained model gives a true positive rate of 81% for men and 88% for women, and positive prediction rate of 92% and 88%. For cohort 2 the rates are 91% and 82% for true positives, and 83% and 91% for positive prediction. The match rate for men is 36% in both cohorts and 31% for women in cohort 1 and 30% in cohort 2. In total, we successfully link 20,231 women, and 23,655 men across both matches for cohort 1, and 38,754 women and 47,105 men across both matches for cohort 2. We refer to this as the individually matched sample. The matched couples sample consists of 10,852 links in cohort 1 and 20,413 links in cohort 2 for which both spouses were successfully linked to a pre- and post-marriage census.

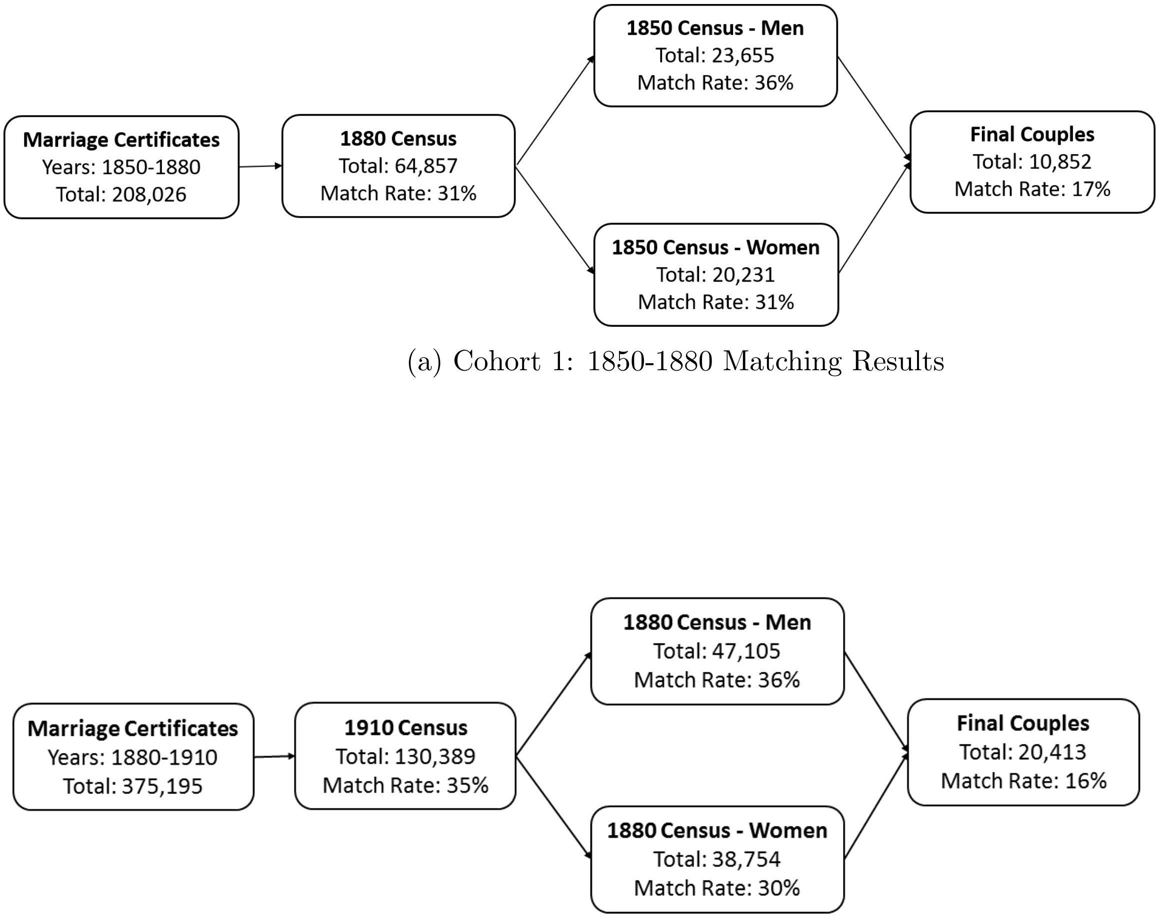
Appendix Table A1 provides a complete breakdown of the causes for match failure. In the first match, we lose 35 percent of cohort 1 and 30 percent of cohort 2 observations because no potential match was found. Note in this step we are looking for complete households with both husband and wife present. Marriages that ended in divorce, separation, or the death of one or both spouses will not be matched by our linking method and will end up in this category. The same for any married couple living apart, or any spouse that migrated internationally. Non-enumeration in the census would also potentially lead to no potential match. Additionally, transcription error in the surname or given names of either spouse such that the string distance is too far away from names listed on the marriage certificate will also lead to no potential matches found.[3] We lose an additional 33-34 percent of marriage observations because of the best potential match not being similar enough, or multiple likely potential matches that cannot be differentiated. Finally, any time a census observation is matched to multiple marriage observations, we remove those marriage observations from our linked sample. We are left with match rates of 32 percent for cohort 1 and 36 percent for cohort 2.

Compared to the marriage certificate to census match, the census to census match is less likely to fail from no potential matches and more likely to fail from no potential match being similar enough to the adult we are trying to find as a child. Many of the no potential match observations are foreign-born immigrants that we would not expect to find in a childhood home in any case. The cutoffs to enter the pool of potential matches is less stringent for the census-census link. We block on state of birth and year of birth within 2 years of that reported on the marriage certificate. Both the given name and surname of the child must have a Jaro-Winkler string distance below 0.20 relative to the adult observation. The key point is that we have additional information on the given names of both parents of the adult observation that helps us decide the best potential match.

[3] Recall that we require a Jaro-Winkler string distance of 0.20 or less as an initial screen to enter the pool of potential matches.

Hence, we get a bigger pool of potential matches for each adult observation, but it is more likely that none of them will be similar enough to coded as the true match.

Additionally, we find that our parameters as well as assessment measures are relatively comparable toFeigenbaum (2016), which provides reassurance that this method is correctly applied to the data. Table A7 reports the TPV and PPV for trained model applied to the cross-validation sample. The rate of false positives ranges from 9 percent to 13 percent. The false negative rate ranges from 9 to 19 percent.



(a) Cohort 1: 1850-1880 Matching Results

(b) Cohort 2: 1880-1910 Matching Results

Figure A1: Illustration of double match procedure and corresponding match rates

Table A1: Summary of Matching Results and Reasons for Non-match

Panel A: Marriage Certificate to Census Link

Category	Cohort 1: 1850-1880			Cohort 2: 1880-1910		
Total couples	**208,026**		**100%**	**375,195**		**100%**
No potential matches found	**72,185**	**100%**	**35%**	**114,193**	**100%**	**30%**
Causes of Match Failure	**69,338**	**100%**	**33%**	**129,201**	**100%**	**34%**
Top potential match score too low (Hits = 1)	14,316	21		28,567	22	
Top potential match score too low (Hits >1)	34,315	49		71,213	55	
Top potential match ratio too low	2,547	4		3,230	3	
Top potential match score and ratio too low	15,640	22		19,828	15	
Census record double matched	2,520	4		6,363	5	
Matches	**66,503**	**100%**	**32%**	**131,801**	**100%**	**36%**
Unique matches (Hits = 1)	37,969	57		66,836	51	
Non-unique matches (Hits >1)	28,534	43		64,965	49	

Panel B: Census to Census Link (Cohort 1: 1850-1880)

Category	Men			Women		
Total	**66,503**		**100%**	**66,503**		**100%**
No potential matches found	**7,263**	**100%**	**11%**	**7,529**	**100%**	**11%**
Causes of Match Failure	**35,450**	**100%**	**53%**	**38,416**	**100%**	**58%**
Top potential match score too low (Hits = 1)	4,535	13		4,519	12	
Top potential match score too low (Hits >1)	23,085	65		26,622	79	
Top potential match ratio too low	788	2		789	2	
Top potential match score and ratio too low	6,981	20		6,424	17	
Census record double matched	61	0		62	0	
Matches	**23,790**	**100%**	**36%**	**20,588**	**100%**	**31%**
Unique matches (Hits = 1)	17,167	72		16,975	83	
Non-unique matches (Hits >1)	6,623	28		3,583	17	

Panel C: Census to Census Link (Cohort 2: 1880-1910)

Category	Men			Women		
Total	**131,801**		**100%**	**131,801**		**100%**
No potential matches found	**21,117**	**100%**	**16%**	**27,977**	**100%**	**21%**
Causes of Match Failure	**62,671**	**100%**	**48%**	**65,057**	**100%**	**49%**
Top potential match score too low (Hits = 1)	9,528	15		11,303	17	
Top potential match score too low (Hits >1)	47,137	75		48,995	75	
Top potential match ratio too low	797	1		577	1	
Top potential match score and ratio too low	5,056	8		4,000	6	
Census record double matched	153	0		182	0	
Matches	**48,013**	**100%**	**36%**	**38,767**	**100%**	**30%**
Unique matches (Hits = 1)	37,393	78		32,104	83	
Non-unique matches (Hits >1)	10,620	22		6,663	17	

Table A2: Comparison of Linked and Full Sample Characteristics: Marriage Certificate to Census

	Cohort 1: 1850-1880			Cohort 2: 1880-1910		
	Matched Sample	All Marriages	P-value of difference	Matched Sample	All Marriages	P-value of difference
Panel A: Personal Characteristics						
Year of marriage	1865.33	1865.03	0.000	1896.60	1895.83	0.000
Husband age at marriage (mean)	26.16	26.06	0.000	27.06	26.76	0.000
Husband age at marriage (Std. Dev.)	5.03	5.00		5.28	5.27	
Wife age at marriage (mean)	23.38	23.40	0.432	24.69	24.41	0.000
Wife age at marriage (Std.Dev.)	4.62	4.69		4.95	5.03	
Age difference at marriage	2.78	2.66	0.000	2.37	2.35	0.081
Panel B: String Characteristics						
Last name commonness	0.07	0.07	0.013	0.06	0.06	0.007
Husband first name commonness	2.92	3.08	0.000	2.39	2.47	0.000
Wife first name commonness	3.48	3.68	0.000	2.22	2.38	0.000
Last name length	6.27	6.30	0.000	6.43	6.49	0.000
Husband first name length	5.95	5.88	0.000	6.06	6.00	0.000
Wife first name length	5.97	5.91	0.000	5.94	5.90	0.000
N	131,801	375,195		66,501	208,026	

Notes: The table displays sample means for the sample of all potential marriage certificates that could be matched compared to the marriage certificates that were matched to the first census. Commonness index was calculated as the share of 100 individuals in the census with the same name (Feigenbaum, 2016). Observations with missing information cause the slight differences in number of observations relative to table A1.

Sources: Marriage certificates from *FamilySearch.org*.

Table A3: Comparison of Linked and Full Sample Characteristics: Census to Census

	Cohort 1: 1850-1880			Cohort 2: 1880-1910		
	Matched Sample	All Observations	P-value of difference	Matched Sample	All Observations	P-value of difference
Panel A: Husbands						
Internal Migrant	0.24	0.23	0.067	0.30	0.54	0.000
Immigrant Parent	0.05	0.33	0.000	0.26	0.54	0.000
Urban residence	0.36	0.44	0.000	0.80	0.81	0.000
Literate	1.00	0.99	0.000	1.00	0.98	0.000
Farmer	0.16	0.17	0.008	0.08	0.09	0.540
Occscore	27.22	25.66	0.000	29.89	27.86	0.000
Wealthscore	4255.95	3334.55	0.000	4219.14	3215.55	0.000
Age (mean)	40.67	40.82	0.008	40.33	40.44	0.010
Age (std. dev.)	5.20	5.55		5.30	5.51	
Last name commonness	0.07	0.07	0.016	0.05	0.06	0.000
First name commonness	2.30	2.92	0.000	2.10	2.40	0.000
Last name length	6.28	6.27	0.648	6.34	6.43	0.000
First name length	6.01	5.95	0.000	6.13	6.06	0.000
Father first name commonness	2.22	2.67	0.000	2.48	2.51	0.255
Father first name length	5.83	5.82	0.567	5.87	5.89	0.158
Mother first name commonness	3.03	3.37	0.000	2.66	2.74	0.001
Mother first name length	5.94	5.91	0.216	5.85	5.97	0.000
Panel B: Wives						
Internal Migrant	0.25	0.24	0.014	0.30	0.54	0.000
Immigrant Parent	0.05	0.34	0.000	0.27	0.56	0.000
Urban residence	0.36	0.44	0.000	0.80	0.81	0.000
Literate	1.00	0.99	0.000	1.00	0.98	0.000
Age (mean)	37.65	38.04	0.000	37.92	38.08	0.000
Age (std. dev.)	4.92	5.38		5.09	5.36	
Last name commonness	0.08	0.07	0.000	0.07	0.06	0.000
First name commonness	2.70	3.48	0.000	1.75	2.21	0.000
Last name length	6.26	6.35	0.000	6.40	6.69	0.000
First name length	5.97	5.97	0.708	5.87	5.94	0.000
Father first name commonness	2.34	2.72	0.000	2.59	2.58	0.660
Father first name length	5.82	5.78	0.001	5.87	5.87	0.745
Mother first name commonness	3.01	3.29	0.000	2.67	2.78	0.000
Mother first name length	5.96	5.85	0.000	5.85	5.95	0.000
N	10,741	65,132		19,718	125,624	

Notes: The table displays sample means for the sample of marriage certificates matched to the first census compared to the sample that also matched to the second census. Internal migrant is defined by living in a different state than the individual's birth state. Immigrant parent is defined as the parent's birth place falling outside of the United States measured in the adult census (1880 and 1910). Commonness index was calculated as the share of 100 individuals in the census with the same name (Feigenbaum, 2016). A slight difference in the total observations in this table compared to matched observations in table A2 can be attributed to missing information in the census for variables of interest.

Table A4: Inverse Propensity Score Probits

	Cohort 1 (1850-1880)			Cohort 2 (1880-1910)		
	Men	Women	Couple	Men	Women	Couple
Hus given name length	0.002***	0.002***	0.001*	0.007***	0.006***	0.003***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Wife given name length	-0.003***	-0.001**	-0.001***	-0.002***	-0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus surname length	-0.001*	-0.001**	-0.000	-0.006***	-0.004***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Wife surname length	-0.001***	-0.002***	-0.001***	-0.006***	-0.008***	-0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus fthr given name length	-0.000	-0.001	0.000	-0.002***	-0.002***	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Wife fthr given name length	0.003***	0.005***	0.003***	-0.001	0.001	0.001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus mthr given name length	0.002***	0.001*	0.001***	-0.004***	-0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Wife mthr given name length	0.003***	0.004***	0.003***	-0.002***	-0.002***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus age of marriage	-0.001	0.004**	0.001	-0.001	0.002*	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Wife age of marriage	-0.002	-0.001	0.002	0.010***	0.009***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Hus age of marriage squared	0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Wife age of marriage squared	-0.000	-0.000**	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of marriage	0.002***	-0.000	0.000***	0.003***	0.002***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus surname similarity score	-0.032***	-0.025***	-0.017***	-0.022***	-0.019***	-0.014***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Wife surname similarity score	-0.030***	-0.030***	-0.017***	-0.021***	-0.025***	-0.017***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Hus given name similarity score	0.008***	0.003	0.004**	-0.021***	-0.017***	-0.011***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Wife given name similarity score	-0.004*	-0.005**	-0.004**	0.002	0.002	0.001
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Hus surname commonness index	-0.024***	-0.028***	-0.015***	-0.030***	-0.050***	-0.023***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)
Wife given name commonness index	-0.005***	-0.005***	-0.003***	-0.006***	-0.007***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hus given name commonness index	-0.007***	-0.004***	-0.004***	-0.004***	-0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	208,026	208,026	208,026	375,195	375,195	375,195
Pseudo-Rsquared	0.023	0.024	0.031	0.023	0.024	0.024

Notes: Marginal effects from probit displayed with standard errors in parentheses. The sample size in each model is the total number of potential marriage certificates that were eligible to be matched. For each cohort, probit estimates the probability that the male or female of the couple will be matched individually, as well as, the probability that both will be matched. Length is measured as string length of each name. Similarity scores were computed by Simpson et al. (2013) as a measure of visual similarity. Commonness index was calculated as the share of 100 individuals in the census with the same name. For cohort 1, a pool of names from the 1850 and 1910 censuses were used. For cohort 2, a pool of names from the 1880 and 1910 censuses were used.

Table A5: Marriage Certificate to Census Match Probit Weights

Predictors	Cohort 1850-1880	Cohort 1880-1910
All names exact match	1.172***	1.109***
	(0.118)	(0.119)
All names exact match and birth difference = 0	0.948***	0.637**
	(0.213)	(0.195)
Husband first name distance	-0.022	-1.885
	(1.567)	(1.489)
Wife first name distance	-4.274**	-2.871*
	(1.337)	(1.234)
Last name distance	-11.292***	-12.113***
	(0.696)	(0.794)
Absolute value difference in wife birth year = 1	-0.211**	-0.224**
	(0.072)	(0.074)
Absolute value difference in wife birth year = 2	-0.545***	-0.691***
	(0.081)	(0.085)
Absolute value difference in husband birth year = 1	-0.058	-0.226**
	(0.072)	(0.073)
Absolute value difference in husband birth year = 2	-0.431***	-0.701***
	(0.082)	(0.086)
Husband first name Soundex match	0.368	0.393
	(0.263)	(0.241)
Wife first name Soundex match	-0.182	0.371*
	(0.190)	(0.165)
Last name Soundex match	0.393***	0.593***
	(0.087)	(0.094)
Hits	-0.133***	-0.095***
	(0.007)	(0.006)
Hits-squared	0.001***	0.001***
	(0.000)	(0.000)
Multiple exact matches on all names	-3.545***	-3.344***
	(0.148)	(0.127)
First letter of last name matches	0.309*	0.365*
	(0.125)	(0.146)
Last letter of last name matches	0.339***	0.357***
	(0.079)	(0.089)
First letter of husband first name matches	0.585	0.748
	(0.356)	(0.418)
Last letter of husband first name matches	0.389*	0.091
	(0.183)	(0.199)
First letter of wife first name matches	0.867***	0.601**
	(0.182)	(0.190)
Last letter of wife first name matches	0.442**	0.219
	(0.149)	(0.145)
Constant	-1.908***	-2.021***
	(0.543)	(0.598)
Observations	14,163	14,656

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.1$ Results are the coefficients from a probit model of an indicator for correct match on matching variables for the set of potential matches in the training data. The first column is on the set of marriages between 1850-1879 to the 1880 census (Cohort 1). The second column is on the set of marriages from 1880-1909 to the 1910 census (Cohort 2). Name distance refers to one minus the Jaro-Winkler score between a name listed on the marriage certificate and in the census. Hits signifies total number of potential matches found for an individual marriage certificate that satisfied the initial screen that given names of both spouses and last name of couple all have a Jaro-Winkler score above 0.80, and the census age is within two years of the year of birth listed on the marriage certificate for both spouses.

Table A6: Census to Census Match Probit Weights

	Cohort 1850-1880		Cohort 1880-1910	
	(1)	(2)	(3)	(4)
Predictors	Husband Match	Wife Match	Husband Match	Wife Match
All names exact match	1.083***	0.741***	0.722***	0.380*
	(0.181)	(0.146)	(0.187)	(0.183)
First name distance	-1.148	-3.950*	0.366	0.412
	(1.494)	(1.607)	(2.108)	(1.968)
Last name distance	-7.027***	-7.018***	-6.680***	-6.139***
	(0.885)	(0.872)	(0.926)	(0.946)
Father first name distance	0.879***	0.393	0.404	-1.191**
	(0.221)	(0.235)	(0.255)	(0.390)
Mother first name distance	1.524***	2.032***	0.090	0.048
	(0.184)	(0.208)	(0.214)	(0.228)
Absolute Value Difference in Birth Year = 1	-0.295***	-0.399***	-0.530***	-0.508***
	(0.074)	(0.074)	(0.080)	(0.087)
Absolute Value Difference in Birth Year = 2	-0.769***	-0.723***	-1.042***	-0.949***
	(0.093)	(0.088)	(0.103)	(0.112)
First name Soundex match	0.074	-0.052	0.275	0.312
	(0.245)	(0.227)	(0.333)	(0.261)
Last name Soundex match	0.472***	0.504***	0.884***	0.757***
	(0.137)	(0.127)	(0.143)	(0.145)
Hits	-0.056***	-0.034***	-0.047***	-0.031***
	(0.003)	(0.002)	(0.003)	(0.002)
Hits-squared	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
More than one match for first and last name	-3.103***	-2.672***	–	-2.732***
	(0.665)	(0.607)	–	(0.406)
First letter of first name matches	0.622	0.206	0.186	0.547
	(0.415)	(0.246)	(0.404)	(0.309)
First letter of last name matches	0.607**	0.537*	-0.021	0.426
	(0.232)	(0.223)	(0.221)	(0.235)
Last letter of first name matches	-0.026	0.017	0.084	0.166
	(0.188)	(0.186)	(0.234)	(0.240)
Last letter of last name matches	0.336*	0.054	0.125	0.112
	(0.133)	(0.114)	(0.133)	(0.137)
First letter of father's first name matches	1.853***	1.867***	1.864***	1.819***
	(0.139)	(0.142)	(0.161)	(0.201)
First letter of mother's first name matches	1.769***	2.094***	1.779***	1.754***
	(0.130)	(0.149)	(0.133)	(0.143)
Last letter of father's first name matches	1.268***	1.103***	1.125***	0.840***
	(0.111)	(0.112)	(0.127)	(0.155)
Last letter of mother's first name matches	0.863***	0.935***	0.566***	0.491***
	(0.105)	(0.114)	(0.095)	(0.102)
Constant	-5.669***	-5.266***	-4.266***	-4.715***
	(0.625)	(0.519)	(0.657)	(0.633)
Observations	30,677	48,660	34,408	39,263

Notes: *** p<0.01, ** p<0.05, * p<0.1 Results are the coefficients from a probit model of an indicator for correct match on matching variables for the set of potential matches in the training data. The first two columns are on the set of marriages between 1850-1879 to the 1880 census matched to the 1850 census (Cohort 1). The second two columns are on the set of marriages from 1880-1909 to the 1910 census matched to the 1880 census (Cohort 2). Name distance refers to one minus the Jaro-Winkler score between a name listed on the marriage certificate and in the census. Hits signifies total number of potential matches found for an individual that satisfied the initial screen that given name and last name both have a Jaro-Winkler score above 0.80, the age difference between observations on the two censuses is within two years. We block on the state of birth.

Table A7: Optimized Parameters and Algorithm Quality

	Parameters			**Training Data**		**Testing Data**	
	Score (Hits = 1)	Score (Hits >1)	Ratio	Efficiency (TPR)	Accuracy (PPV)	Efficiency (TPR)	Accuracy (PPV)
Cohort 1:							
1880 - Couple	0.128	0.370	1.536	0.85	0.93	0.83	0.87
1850 - Men	0.100	0.439	1.919	0.87	0.93	0.81	0.92
1850 - Women	0.141	0.400	1.720	0.85	0.91	0.88	0.88
Cohort 2:							
1910 - Couple	0.143	0.363	1.3800	0.85	0.94	0.81	0.91
1880 - Men	0.119	0.377	1.300	0.89	0.91	0.91	0.83
1880 - Women	0.100	0.371	1.278	0.92	0.91	0.82	0.91

Notes: For both cohorts matched, the optimal parameters for score and ratio are reported. Optimal score is the minimum score, or predicted probability of being a match, required to be considered an accurate match. Optimal ratio is the minimum ratio of best score to next best score, also required to be considered an accurate match. Optimized True Positive Rate (TPR) and Positive Prediction Value (PPV) are also listed for applying the specific score and ratio thresholds to the training data. The TPR and PPV are also displayed for an additional testing data set, which was manually linked and compared with matches predicted by the model. Training data consists of 2,500 manually linked matches for each step. Testing data consists of 1,000 manually linked matches for each step.

B Occupational Transition Tables

Table A8: Occupational Transition Tables

	Father's Occupation				
	White collar	Farmer	Skilled/semiskilled	Unskilled	Row total
Men (1850-1880)					
White Collar	2,514	2,957	4,195	589	10,255
	(64.2)	(25.5)	(31.6)	(18.5)	
Farmer	218	3,630	789	297	4,934
	(5.6)	(31.4)	(5.9)	(9.3)	
Skilled /semiskilled	1,040	4,158	7,465	1,661	14,324
	(26.5)	(35.9)	(56.1)	(52.0)	
Unskilled	146	835	849	645	2,475
	(3.7)	(7.2)	(6.4)	(20.2)	
Column total	3,918	11,580	13,298	3,192	31,988
Men (1880-1910)					
White Collar	7,858	3,779	10,703	2,629	24,969
	(63.2)	(32.30)	(37.1)	(30.1)	
Farmer	459	2,550	989	408	4,406
	(3.7)	(21.8)	(3.4)	(4.7)	
Skilled/semiskilled	3,115	3,966	13,796	3,878	24,755
	(25.0)	(33.9)	(47.8)	(44.4)	
Unskilled	1,008	1,406	3,373	1,812	7,599
	(8.1)	(12.0)	(11.7)	(20.8)	
Column total	12,440	11,701	28,861	8,727	61,729
Women (1850-1880)					
White Collar	2,217	2,354	4,307	610	9,488
	(58.4)	(26.7)	(32.8)	(19.3)	
Farmer	280	2,425	1,262	357	4,324
	(7.4)	(27.5)	(9.6)	(11.3)	
Skilled/semiskilled	1,145	3,477	6,664	1,653	12,939
	(30.2)	(39.4)	(50.7)	52.4)	
Unskilled	154	572	909	536	2,171
	(4.1)	(6.5)	(6.9)	(17.0)	
Column total	3,796	8,828	13,142	3,156	28,922
Women (1880-1910)					
White Collar	6,863	2,750	10,603	2,374	22,590
	(61.7)	(34.5)	(38.7)	(29.7)	
Farmer	493	1,477	1,386	419	3,775
	(4.4)	(18.5)	(5.0)	(5.2)	
Skilled/semiskilled	2,961	2,838	12,120	3,820	21,739
	(26.6)	(35.6)	(44.2)	(47.7)	
Unskilled	815	902	3,325	1,394	6,436
	(7.3)	(11.3)	(12.1)	(17.4)	
Column total	11,132	7,967	27,434	8,007	54,540

C Assortative Mating: Subgroup Analysis

Table A9: Assortative Mating: Subgroup Analysis (Rank-Rank regressions)

	Men		Women	
	Cohort 1	Cohort 2	Cohort 1	Cohort 2
	1850-1880	1880-1910	1850-1880	1880-1910
Panel A: Urban vs. Rural Childhood				
Father's total property wealth score	0.140***	0.191***	0.133***	0.202***
	(0.018)	(0.010)	(0.018)	(0.011)
Rural childhood	-3.423**	-1.935**	-4.062**	-1.346
	(1.322)	(0.978)	(1.246)	(0.915)
Fthr's wealth * Rural	0.036	0.047**	0.037	0.029
	(0.024)	(0.017)	(0.023)	(0.018)
Panel B: Immigrant vs. Native-born Parents				
Father's total property wealth score	0.151***	0.172***	0.143***	0.181***
	(0.012)	(0.010)	(0.012)	(0.011)
Immigrant parent	-9.323***	-7.944***	-10.460***	-7.283***
	(2.303)	(0.947)	(2.048)	(0.921)
Fthr's wealth * Immigrant parent	0.118**	0.001	0.092*	-0.034*
	(0.053)	(0.018)	(0.051)	(0.019)
Panel C: Internal Migration				
Father's total property wealth score	0.167***	0.219***	0.159***	0.211***
	(0.014)	(0.009)	(0.013)	(0.010)
Internal migrant	0.684	4.998***	2.636*	4.386***
	(1.677)	(1.142)	(1.506)	(1.108)
Fthr's wealth * Internal migrant	-0.003	-0.055**	-0.001	-0.009
	(0.029)	(0.019)	(0.028)	(0.019)

Notes: Urban is an indicator equal to one if the observation resided in an urban area as a childhood. Urban is defined following the IPUMS definition of towns and incorporated places of at least 2,500 in population. Immigrant is an indicator equal to one if at least one of the parents was born outside the United States according to FBPL and MBPL IPUMS variables in the adult census (1880 and 1910). Internal migrant is defined as observing an observation in two different states in the childhood census and adult census. First-born is an indicator equal to one if the observation is the oldest child listed in the household. Number of siblings is the total number of children of the head of household listed in the census minus one for the observation at hand.
Sources: 1870 1% sample and complete count 1850, 1880, and 1910 Federal Census data from Ruggles et al. (2017). Marriage certificates from *FamilySearch.org.*

D Upward and Downward Rank Mobility

In addition to the descriptive IGE and rank-rank regressions, we also characterize mobility as upward or downward movement in rank relative to the rank of the father - named "upward rank mobility" (URM) and "downward rank mobility" (DRM) by Bhattacharya and Mazumder (2011). Conditional on the rank of the father, the likelihood of upward or downward movement provides a measure of mobility for children with similar initial positions in the distribution of economic status. Decile-to-decile transition matrices are commonly used to display rank mobility. Instead, we illustrate the differences between men and women – within and across cohorts – graphically using the URM measure. Let R_{son} and R_{father} represent the rank of the son and the rank of the father. Following Bhattacharya and Mazumder (2011) and Corak et al. (2014), the cumulative URM is constructed as:

$$URM_{\tau,s} = Pr(R_{son} - R_{father} > \tau | R_{father} \leq s) \tag{1}$$

, where the father's rank in the interval between 0 and s. We use deciles as the interval cutoffs. The URM can be constructed for varying degrees of upward mobility by changing the value of τ. Downward rank mobility is similarly defined. We also consider a more continuous measure that does not depend on the choice of value for τ. Conditional on an upward movement in rank between generations, the mean gain is defined as:

$$\text{Mean Gain}_s = \frac{1}{N} \sum (R_{son} - R_{father} | R_{son} > R_{father}, \quad s_{lower} \leq R_{father} < s_{upper}) \tag{2}$$

, where now the measure is not cumulative but specific to a given interval. We estimate mean gain for decile intervals. Mean loss is defined similarly.

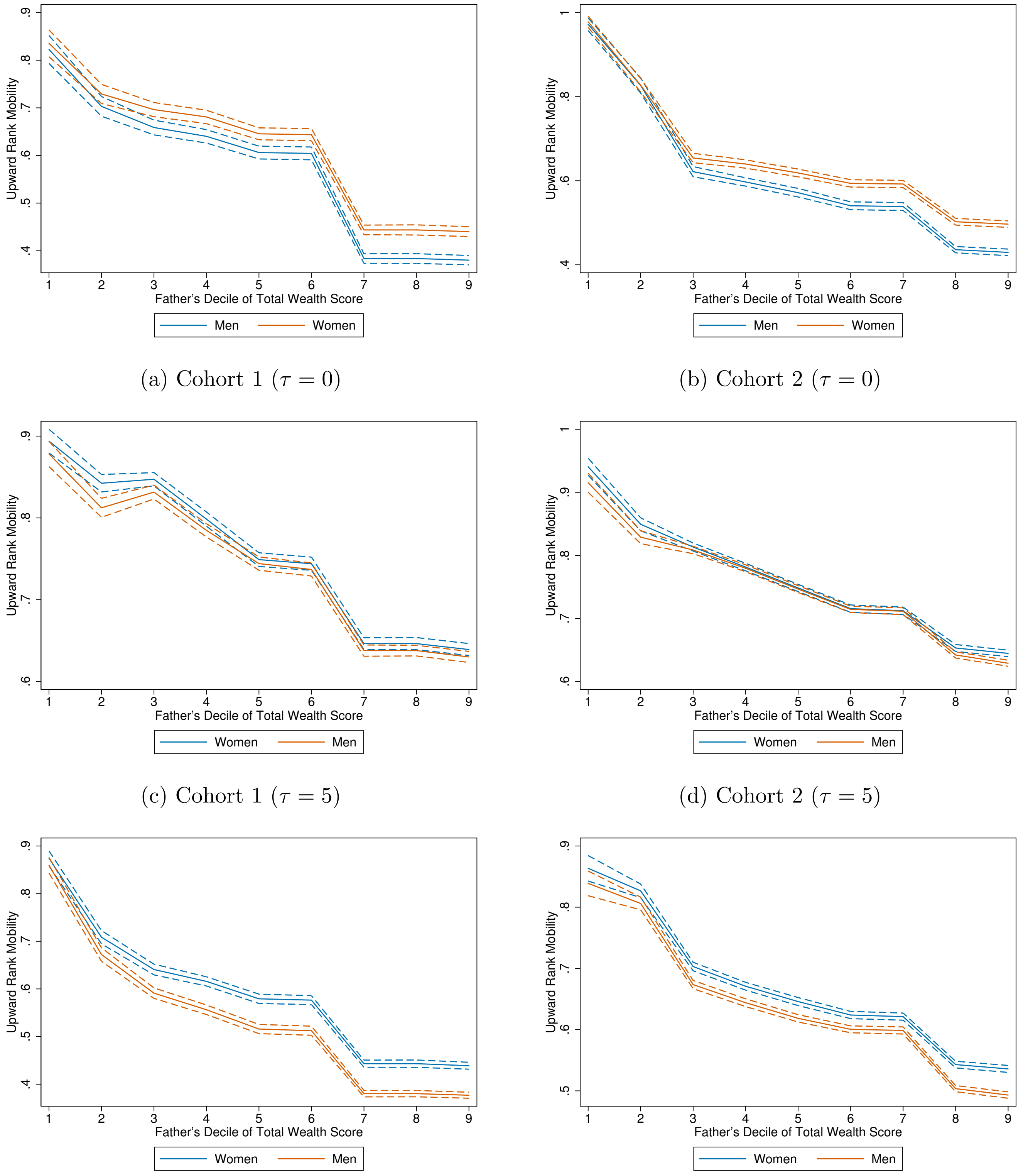
D.1 Results

To explore mobility differences in more detail, we turn to transition probabilities with ranks constructed from the total wealth score. In figure A2 we show how the upward transition probabilities differ between men and women for varying levels of τ. Changing τ allows us to see how the probabilities change for increasingly large upward movements. Panels (a) and (b) suggest high levels of upward mobility for the bottom quintile - roughly a 95 percent chance in both cohorts. We find no meaningful difference between men and women in the probability of an upward rank movement in either cohort. The remaining panels shed light on how this result is consistent with our finding of lower persistence for women in both the rank-rank and IGE specifications. The second set of panels present the probability of an upward rank movement from father to child of at least 5 percentiles, in which we see small differences emerge between men and women. The differences are stark in the final set of panels where the upward movement is of 10 percentiles or greater. While similarly likely to have *any* upward rank movement, we can clearly see in the 1850-1880 cohort that women are significantly more likely to have *large* upward movements. Women maintained their increased likelihood of large upward movements in the 1880-1910 cohort, but the differences between the sexes is smaller than in the 1850-1880 cohort.

To further illustrate the underlying transitions, in figure A3 we present mean gains conditional on an upward transition and mean losses conditional on a downward transition. Note that these figures plot probabilities for each decile individually; they are not cumulative. Conditional on an upward transition, the average size of the gain is quite similar in the tails for the 1850-1880. However, the higher mobility of women emerges for daughters of fathers in the third to seventh deciles of the distribution of occupational total wealth scores. Panel (c) shows that conditional on a downward transition, the size of the loss does not meaningfully differ between the sexes on average. Thus, the large mobility advantage of daughters in the 1850-1880 cohort stems not from a higher probability of upward transition, but from *larger* movements. By the 1880-1910 cohort, the size of average gains had increased for women in the first quintile, mean gains had declined slightly in the 4th through 6th deciles. The average upward movement increased more rapidly for men across the middle of the distribution, causing the convergence in mobility between the sexes that we saw earlier in the IGE and rank-rank estimates.

Finally, to make clear comparisons over time we present the cumulative upward rank mobility results across cohorts for women in figure A4.[4] Panel (a) shows that the probability of an upward transition for women was similar for both cohorts, without a clearly dominant time period despite IGE estimates that are lower in the later cohort. The first quintiles are almost identical. The earlier cohort is more likely to experience upward movement in the second quintile, but rates converged in the third quintile. The largest difference is in the two upper quintiles where women of the later period have an advantage. Similar to the previous discussion of differences between men and women, the differences over time come not from the probability of an upward transition, but from its *size*. Panel (c) plots the cumulative upward rank mobility probability of a movement of at least 10 percentiles. We can clearly see that women in the 1880-1910 cohort are much more likely to have large upward transitions from their father's economic standing to their husband's.

[4] The figures for men tell a similar story.



(a) Cohort 1 ($\tau = 0$)

(b) Cohort 2 ($\tau = 0$)

(c) Cohort 1 ($\tau = 5$)

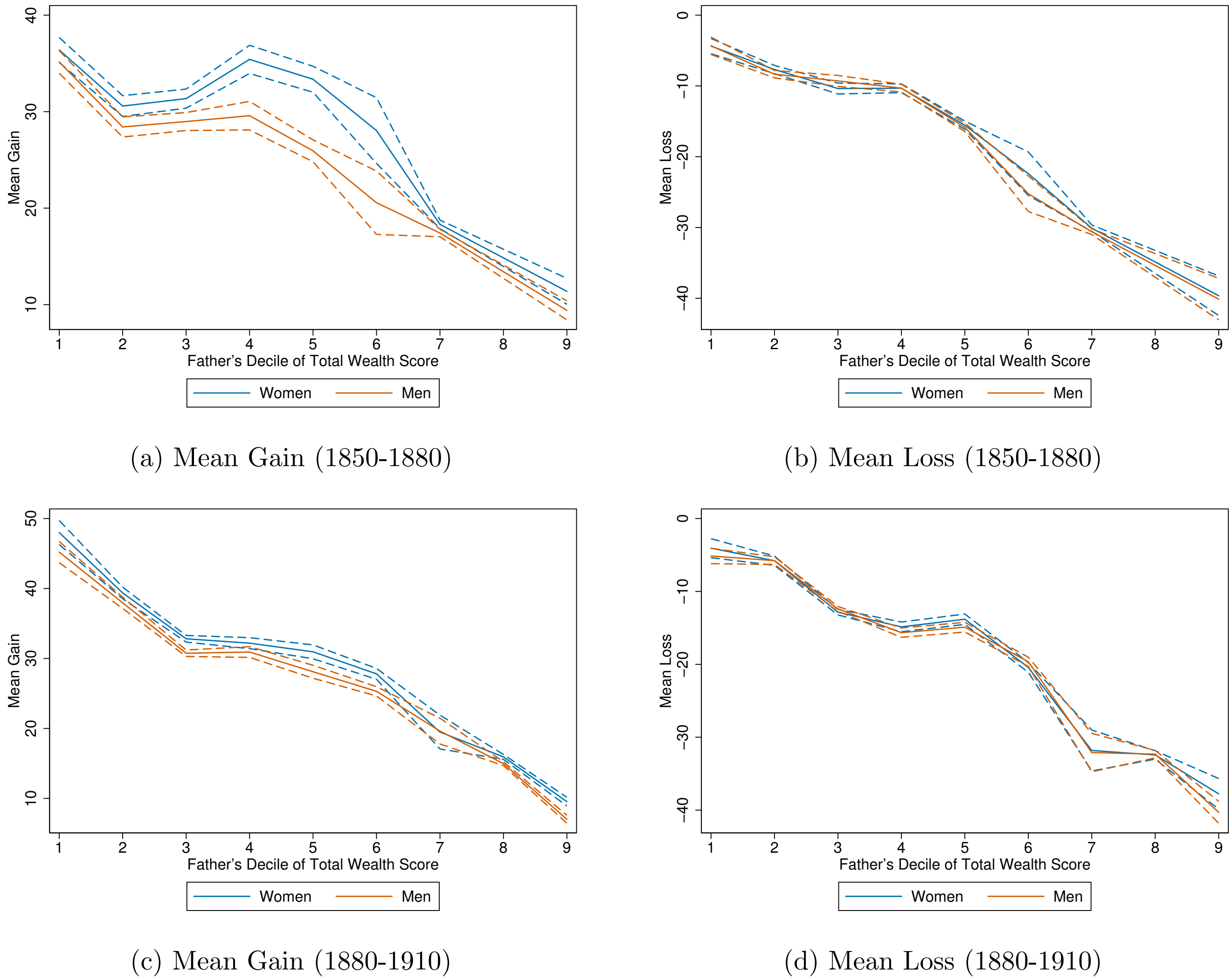
(d) Cohort 2 ($\tau = 5$)

(e) Cohort 1 ($\tau = 10$)

(f) Cohort 2 ($\tau = 10$)

Figure A2: Women vs. Men Cumulative Upward Rank Mobility

Notes: Cumulative probability of an upward rank move greater than τ for all percentiles $\leq r$. Standard errors calculated with the bootstrap procedure of Corak et al. (2014).



(a) Mean Gain (1850-1880)

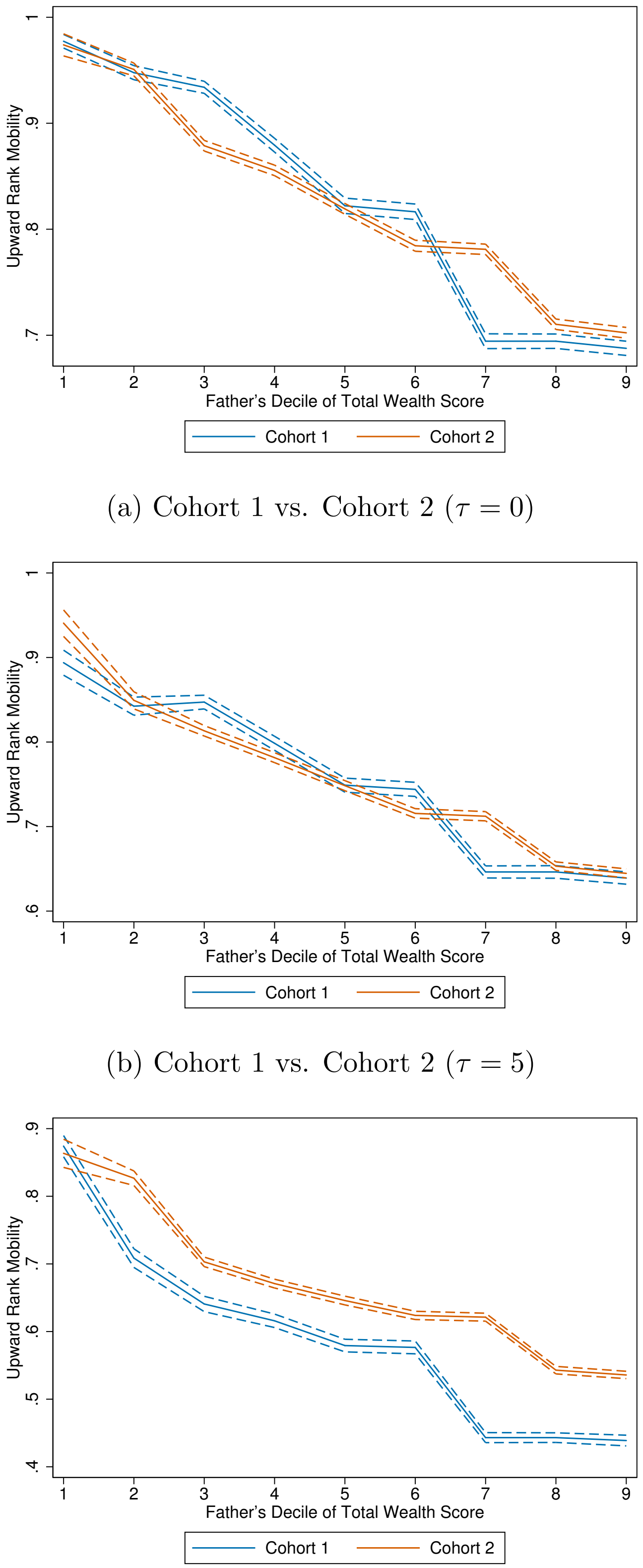
(b) Mean Loss (1850-1880)

(c) Mean Gain (1880-1910)

(d) Mean Loss (1880-1910)

Figure A3: Women vs. Men Mean Gain and Mean Loss

Notes: Mean gain is the average rank gain conditional on an upward rank move in a specific decile of father's total wealth score. Mean loss is the average rank loss conditional on a downward rank move. Standard errors calculated with the bootstrap procedure of Corak et al. (2014).



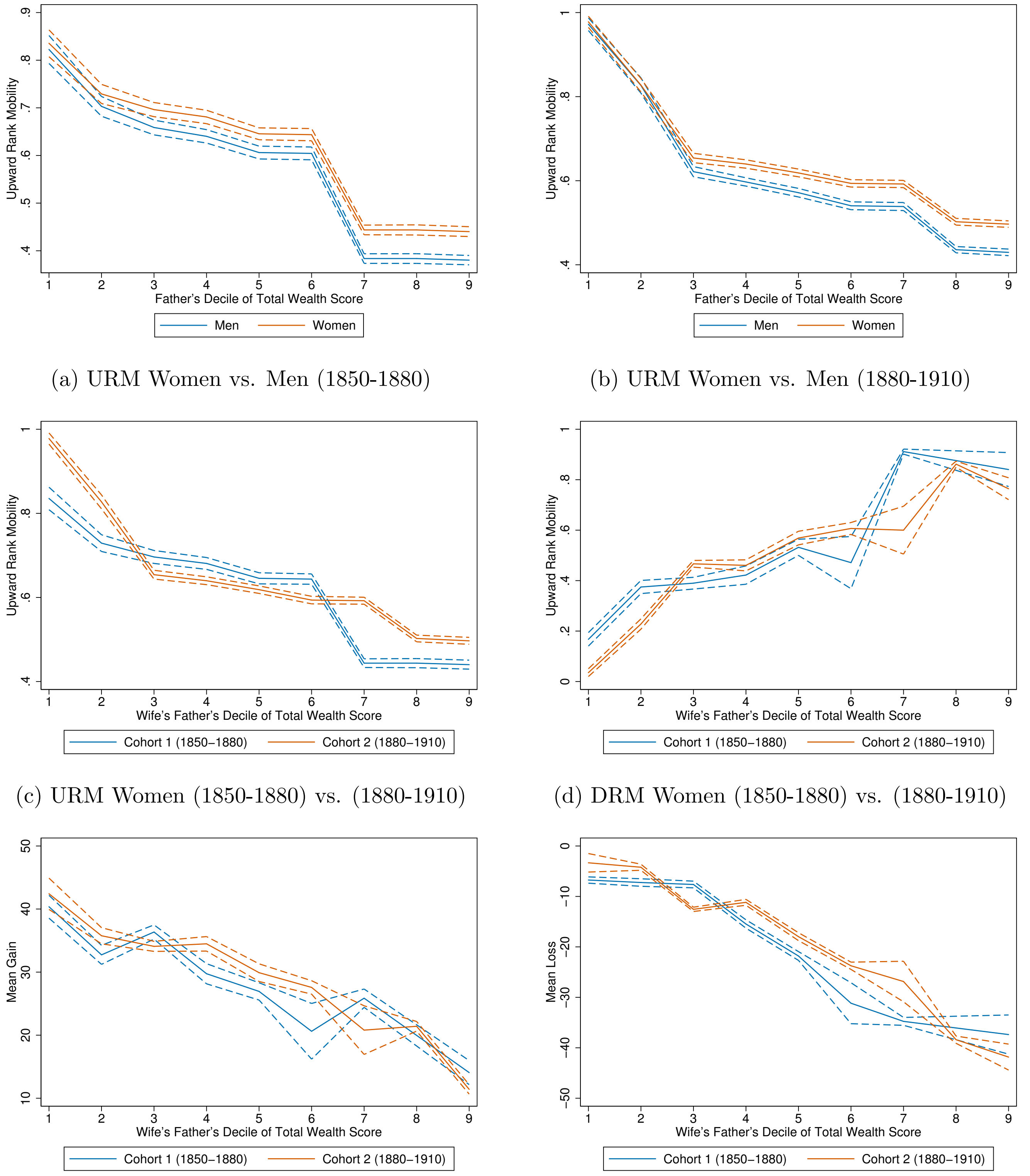
(a) Cohort 1 vs. Cohort 2 ($\tau = 0$)

(b) Cohort 1 vs. Cohort 2 ($\tau = 5$)

(c) Cohort 1 vs. Cohort 2 ($\tau = 10$)

Figure A4: Women between cohort comparison of Cumulative Upward Rank Mobility

Notes: Cumulative probability of an upward rank move greater than τ for all percentiles $\leq r$. Standard errors calculated with the bootstrap procedure of Corak et al. (2014).



(a) URM Women vs. Men (1850-1880)

(b) URM Women vs. Men (1880-1910)

(c) URM Women (1850-1880) vs. (1880-1910)

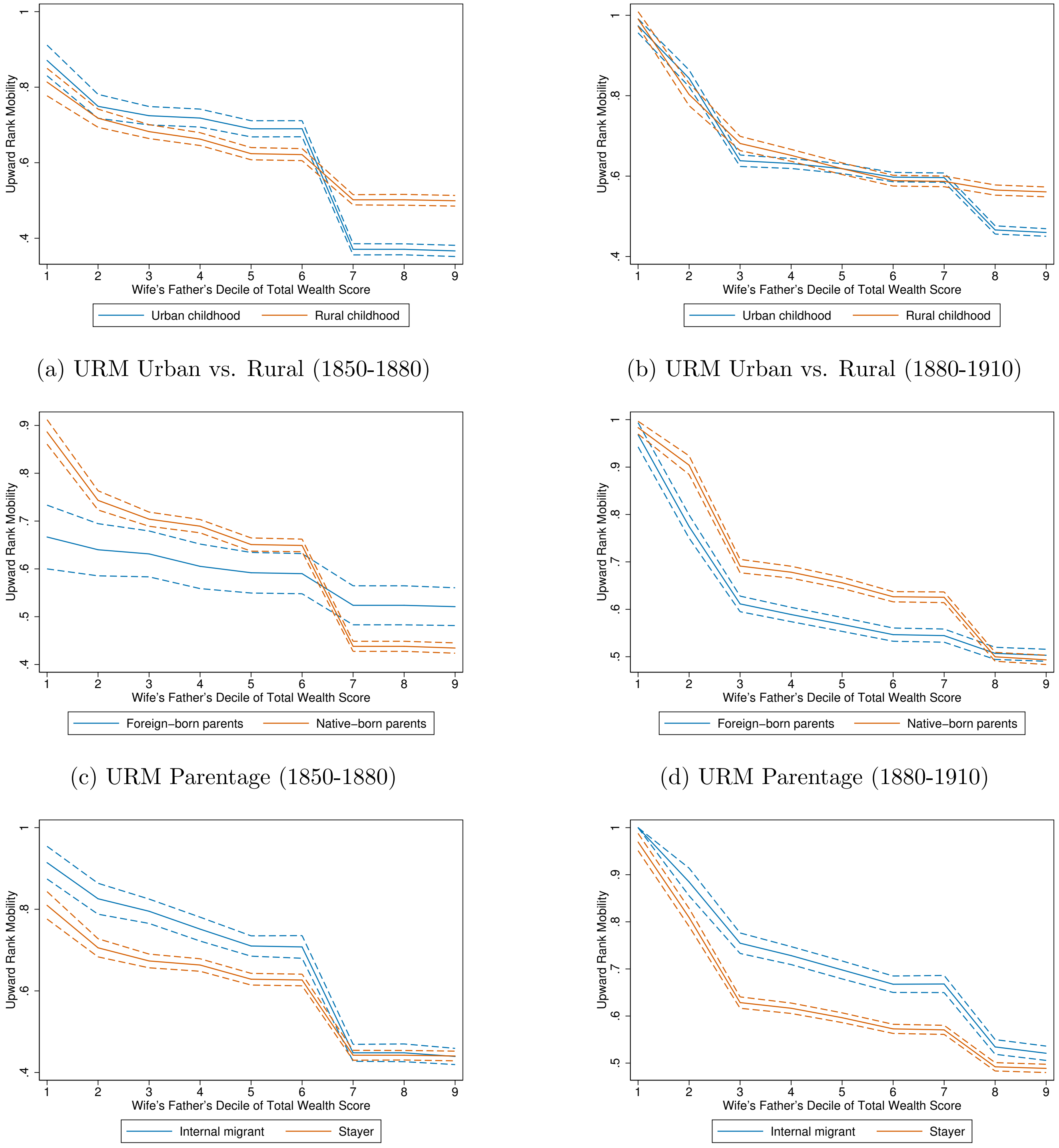
(d) DRM Women (1850-1880) vs. (1880-1910)

(e) Mean Gain Women (1850-1880) vs. (1880-1910)

(f) Mean Loss Women (1850-1880) vs. (1880-1910)

Figure A5: Probability and Size of Marriage Transitions in Father's Occupational Standing

Notes: Rank moves are own father's total wealth score rank vs. spouse's total wealth score rank. URM is the cumulative probability of an upward rank marriage move greater than $\tau = 0$ for all percentiles $\leq r$. DRM is the cumulative probability of a downward rank marriage move greater than $\tau = 0$. Standard errors calculated with the bootstrap procedure of Corak et al. (2014).



(a) URM Urban vs. Rural (1850-1880)

(b) URM Urban vs. Rural (1880-1910)

(c) URM Parentage (1850-1880)

(d) URM Parentage (1880-1910)

(e) Internal migration (1850-1880)

(f) Internal migraiton (1880-1910)

Figure A6: Women's Probability of Marrying Up

Notes: Rank moves are own father's total wealth score rank vs. spouse's total wealth score rank. URM is the cumulative probability of an upward rank marriage move greater than $\tau = 0$ for all percentiles $\leq r$. Standard errors calculated with the bootstrap procedure of Corak et al. (2014). Urban is an indicator equal to one if the observation resided in an urban area as a childhood. Urban is defined following the IPUMS definition of towns and incorporated places of at least 2,500 in population. Immigrant is an indicator equal to one if at least one of the parents was born outside the United States according to FBPL and MBPL IPUMS variables in the adult census (1880 and 1910). Internal migrant is defined as observing an observation in two different states in the childhood census and adult census.

E Comparison to Pseudo-Linking Methodology

Olivetti and Paserman (2015) make the first attempt at estimating female mobility in the United States separately from males. Without the ability to link a female to her childhood economic status due to name change at the time of marriage, they create a pseudo-link using the assumption that names convey socioeconomic status. They identify the occupational income score (*occscore*) of an individual in a specific census and calculate the average occupational income score for all fathers in the previous census who have a child with that individual's name. For example, for an individual named "John", the income level of his father is calculated as the average income of all fathers in the previous census with a son named "John". They use these pseudo-links to calculate estimates for intergenerational elasticity of income for both men and women from 1850-1940. As a majority of women did not work during this time period, intergenerational elasticity of income for women was calculated as the elasticity between a female's husband and a female's father. Their method does not capture a "true" estimate for elasticity of income because of the absence of direct links between generations. Rather, it calculates a measure that can be compared over time, assuming equal bias over time and across genders. While this method does provide an avenue for uncovering female mobility, it may face attenuation bias from averaged income values and the premium or penalty associated with first names on the labor market. They have since extended this method to include three generations and found that grandparents do matter in mobility level of their children and grandchildren (Olivetti et al., 2016). Due to the methodological differences, our estimate of intergenerational mobility will not be comparable in magnitude to Olivetti and Paserman (2015), but will be useful to compare the mobility of women relative to men and the trends over time.

Our results consistently demonstrate that female mobility is higher than male mobility, when applying a direct-linking methodology between fathers and their children.Olivetti and Paserman (2015) find mixed results during this same time period, when applying a pseudo-linking methodology. To identify how our estimates compare with Olivetti and Paserman (2015) pseudo-linked estimates, we apply their pseudo-linking methodology to the sample of direct links from Massachusetts marriages. We compare our direct and pseudo-linked estimates to Olivetti and Paserman (2015)'s northeastern region elasticities, as our dataset consists mainly of individuals from Massachusetts. To do so, we take all male and female first names from the full sample of direct matches and impute the pseudo-linked father's occscore based on all fathers in the United States.[5] On average, both the pseudo-linked and direct-linked father's occscore in 1880 are less than the husband's occscore in 1910. Significantly less variation in occscore occurs for pseudo-linked fathers, which is not surprising given that this variable is already an average of occscores for fathers with similarly named children.

We perform identical regressions as before, except replacing the direct-linked father's occcore with the imputed pseudo-linked father's occscore. Results show an elasticity of income for females ranging from 0.217 to 0.303 and for males ranging from 0.178-0.374. We compare these estimates

[5] About 100 first names did not exist as a child in the 1880 census, therefore preventing the imputation of the father's occscore. These names can be assumed to have some sort of small error that prevented exact matching to a child in the 1880 census, but allowed for the probabilistic matching process to identify a direct match.

to the Northeast region results reported by Olivetti and Paserman (2015). They find an elasticity of income from 1880-1900 for females of 0.3111 and for males of 0.1677, an increase in mobility for men and a decrease in mobility for women. Pseudo-linked results on the Massachusetts sample also find an increase in mobility for men, but of a smaller magnitude, but, instead, a decrease in mobility for women as well.

Finally, while little has been done on the intergenerational mobility of women in the 19th-century U.S., the exception of note is Olivetti and Paserman (2015), which estimates historical mobility for women by overcoming female linkage problem by using what they call a pseudo-linking procedure. They argue that given names contain economic content. Daughters with certain names tend to come from higher status fathers, where daughters with other names tend to come more often from lower status families. As such, the childhood economic status of a married women with the given name "Sue" in an adult census is proxied by, essentially, the mean income of all fathers in the childhood census with a daughter in the household named "Sue". Olivetti and Paserman (2015) find that father-daughter elasticities were flat in the during the 19th-century and increased during the early part of the 20th-century. Additionally, the elasticity for daughters is higher than for sons for the 1860-1880 cohort, but converge by 1920.

We compare estimates using our direct-linking procedure with those from Olivetti and Paserman (2015)'s pseudo-linking procedure. The levels of the IGE estimates using pseudo-links are not directly comparable to those using a direct-linking procedure, but the relative mobility between men and women and the trends over time are. We find some striking differences between the methods. The direct-linking procedure on the sample of couples finds the IGE for men to be greater than women in both cohorts, with decreases from the 1850-1880 to the 1880-1910 cohort. Depending on whether we use *all fathers* in the census, *fathers residing in New England*, or *fathers born in New England*, we find that the relative difference in the IGE between men and women changes. In the all fathers case, we get a reversal in the trend, an increase in persistence, whereas direct-linking finds a decrease in persistence. Limiting the sample pool of fathers to more closely align with the underlying population for which estimates are being made tends to lead estimates more in line with the direct-linking procedure. We suggest that researchers leverage the complete count census microdata recently made available for the historical U.S. censuses and construct a pool of fathers that mimics the underlying population of daughters in the sample when using the pseudo-linking procedure.[6]

[6] At the time, the complete count censuses were not available to (Olivetti and Paserman, 2015).

Table A1: Direct-linking comparison to pseudo-linking

	Direct-linking		Pseudo-linking					
			All fathers		Fathers reside in NE		Fathers born in NE	
	Men	Women	Men	Women	Men	Women	Men	Women
Panel A: Massachusetts Data								
Occscore IGE (1850-1880)	0.278	0.222	0.374	0.217	0.334	0.303	0.372	0.278
	(0.008)	(0.012)	(0.056)	(0.047)	(0.055)	(0.054)	(0.055)	(0.051)
Obs	9,568	9,568	9,568	9,568	9,568	9,568	9,568	9,568
Occscore IGE (1880-1910)	0.209	0.172	0.178	0.249	0.258	0.250	0.268	0.249
	(0.006)	(0.009)	(0.047)	(0.044)	(0.049)	(0.044)	(0.049)	(0.044)
Obs	18,021	18,021	18,021	18,021	18,021	18,021	18,021	18,021
Panel B: Olivetti & Paserman (2015) Estimates								
Occscore IGE (1850-1870)			0.35	0.34	0.295	0.201		
			(0.02)	(0.02)	(0.04)	(0.04)		
Occscore IGE (1880-1900)			0.344	0.399	0.168	0.311		
			(0.02)	(0.02)	(0.03)	(0.04)		

Heteroskedasticity robust standard errors reported in parentheses. All estimates are statistically significant at the 1 percent level. Direct-linking reprints our main estimates of mobility from Table 1. Pseudo-linking refers to creating father-child pairs by imputing father's occupational income score using the mean occscore of father's in a given pool that have a child listed in the household with the same given name as the child observed as an adult in a later census. The pool of fathers over which mean occscore is computed are: all fathers in childhood census, fathers residing in New England at time of childhood census, and fathers born in New England at time of childhood census. Panel B reprints IGE estimates for the Northeast region from table 8 of Olivetti and Paserman (2015).

1 Bibliography

Bhattacharya, D. and B. Mazumder (2011, November). A nonparametric analysis of black–white differences in intergenerational income mobility in the United States. *Quantitative Economics 2*(3), 335–379.

Corak, M., M. J. Lindquist, and B. Mazumder (2014). A comparison of upward and downward intergenerational mobility in Canada, Sweden and the United States. *Labour Economics 30*(C), 185–200.

Feigenbaum, J. (2016). A machine learning approach to census record linking. Working Paper.

Olivetti, C. and M. D. Paserman (2015, August). In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940. *American Economic Review 105*(8), 2695–2724.

Olivetti, C., M. D. Paserman, and L. Salisbury (2016). Three-generation mobility in the united states, 1850-1940: The role of maternal and paternal grandparents. NBER Working Paper No. 22094.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (2017). Integrated public use microdata series. Version 7.0 [Machine-readable database].