ONLINE APPENDICES *for*

"HOW WELL DO AUTOMATED LINKING METHODS PERFORM?

LESSONS FROM U.S. HISTORICAL DATA"

Martha Bailey,[1,2] Connor Cole,[1] Morgan Henderson,[1] Catherine Massey[1]
[1] University of Michigan [2] National Bureau of Economic Research

October 27, 2020

A. *Details Regarding the Implementation of Algorithms and Effects of Race-Blocking*

1. Implementation of Feigenbaum (2016)

We generate links for the Feigenbaum (2016) two ways. The "Iowa" coefficient results use the parameters that Feigenbaum estimated with his training data that links the 1915 Iowa Census to the 1940 Census, which we use directly. Because the performance of this algorithm hinges critically on the quality of the training data and the similarity of the training data to the data to be linked, we also examine performance based on coefficients that we estimate using a random sample of the training data. For the LIFE-M and synthetic data, we use a random sample of 2,000 observations (our simulations show the size of the random sample has little effect on outcomes above a sample size of 2,000 for these samples). Because the number of linked observations is smaller in the Early Indicators data, we select a smaller random sample of 500 genealogically linked observations. For each of these samples, we estimate a probit model on the training dataset using Feigenbaum's suggested covariates, with the outcome being the binary variable denoting a "true" match as indicated in the training data. Then, we apply the estimated parameters to the data (less the training data) to classify links. Our estimates of match rates, Type I and Type II error rates are calculated off the model's performance in the sample, which excludes the training data.

Appendix Table A1 reports the estimation results using each dataset. Column 1 presents the covariates reported by Feigenbaum (2016) and the remaining columns present the coefficients for each sample. The parameter estimates vary somewhat, but the signs and magnitudes are generally similar across datasets. Notably, lower Jaro-Winkler scores, indicating more dissimilar names, decrease the probability that a potential match is a match, and higher age distances also decrease the probability that a potential match is a match. However, the coefficient on an indicator for whether or not first and last names exactly match is positive and significant for the LIFE-M sample, but negative and significant in the other two samples. The missing variables in the probit reflect the fact that the relevant data either does not have variation in the variable or that the excluded variables perfectly predict matches.

Appendix Table A1. Probit Coefficients Estimated Using Different Datasets

	(1)	(2)	(3)	(4)
Variables	Feigenbaum (2016)	LIFE-M	Simulated Data	Early Indicators
First and Last Names Match	0.632***	0.468***	-0.366***	-0.472**
	(0.086)	(0.123)	(0.0651)	(0.186)
First Name Jaro-Winkler Score	-6.071***	-7.458***	-4.407***	-10.57***
	(0.525)	(1.461)	(0.693)	(2.003)
Last Name Jaro-Winkler Score	-10.285***	-13.85***	-13.97***	-13.61***
	(0.487)	(1.276)	(0.635)	(1.619)
Absolute Value of Difference in Year of Birth is 1	-0.708***	-1.184***	-0.907***	-0.992***
	(0.044)	(0.0668)	(0.0540)	(0.109)
Absolute Value of Difference in Year of Birth is 2	-1.562***	-1.649***	-0.901***	-1.363***
	(0.065)	(0.0997)	(0.0554)	(0.132)
Absolute Value of Difference in Year of Birth is 3	-2.316***	-1.689***	-	-1.891***
	(0.102)	(0.105)	-	(0.171)
First Name Soundex Match	0.153***	0.509***	-0.0623	-0.300***
	(0.054)	(0.183)	(0.0966)	(0.275)
Last Name Soundex Match	0.698***	0.982***	0.285***	1.320***
	(0.069)	(0.141)	(0.0758)	(0.201)
Number of Potential Matches	-0.064***	-0.0115***	-0.0146***	-0.0202***
	(0.002)	(0.00127)	(0.0008)	(0.00251)
Number of Potential Matches Squared	0.0003****	0.00003***	0.00004***	0.00004***
	(0.00002)	(0.00001)	(0.000003)	(0.00001)
More than One Exact Match on First and Last Name	-1.690***	-1.738***	-0.938***	-1.219***
	(0.093)	(0.0696)	(0.0444)	(0.119)
First Letter of First Name Matches	0.871***	1.123**	0.598***	-
	(0.130)	(0.496)	(0.105)	-
First Letter of Last Name Matches	0.886***	-	-0.0834	-0.563**
	(0.148)	-	(0.0815)	(0.257)
Last Letter of First Name Matches	0.147***	-0.0661	1.074***	0.0334
	(0.053)	(0.140)	(0.0944)	(0.234)
Last Letter of Last Name Matches	0.649***	0.626***	1.051***	0.380***
	(0.070)	(0.147)	(0.0945)	(0.186)
Middle Initial Matches (0 if Middle Initial Does Not Exist or Does Not Match)	0.537***	0.767***	1.661***	1.385***
	(0.097)	(0.0771)	(0.0609)	(0.142)
State is Ohio	N/A	0.258***	0.112***	N/A
		(0.0548)	(0.0377)	
Constant	-1.479***	-2.478***	-2.115***	0.958**
	(0.225)	(0.577)	(0.206)	(0.470)
Observations	38,091	64,490	58,783	8,587
Log-Likelihood	-2440	-1372	-2847	-470.1
Akaike Inf. Crit.	4916	2778	5728	972.3

Notes: Column 2 drops the estimate for “First Letter on Last Name,” because LIFE-M blocked on this variable to create the list of potential links.

Appendix Table 2 lists the other parameters that are relevant for Feigenbaum's algorithm. The *B1* value is the minimum threshold of estimated probability from the probit that a potential link must meet for it to be judged a match, and the *B2* value is the minimum threshold the ratio between the first best and second best potential links for a given observation must meet for a potential link to be judged as a match. In panel B, we also report the average results from 200 draws of random samples to ensure our findings are not driven by an extreme draw. The *B1* and *B2* values are similar across datasets, although the *B2* value is notably higher and the *B1* value lower for the LIFE-M data compared to the other datasets.

Appendix Table A2. Features of the Feigenbaum Classifier using the Estimated Coefficients

	(1) Reported in Feigenbaum (2016)	(2) LIFE-M	(3) Simulated Data	(4) Early Indicators
A. First random sample				
B1 Value	0.140	0.075	0.125	0.225
B2 Value	1.375	2.625	1.100	1.350
Match Rate	0.57	0.52	0.64	0.57
Type I Error	0.14	0.29	0.26	0.16
Type II Error	0.51	0.63	0.53	0.52
B. Average across 200 random samples				
Average B1 Value		0.07	0.14	0.23
		(0.03)	(0.04)	(0.06)
Average B2 Value		2.42	1.12	1.32
		(0.78)	(0.12)	(0.26)
Average Match Rate		0.54	0.63	0.56
		(0.03)	(0.04)	(0.03)
Average Type I Error		0.31	0.26	0.16
		(0.02)	(0.02)	(0.02)
Average Type II Error		0.63	0.54	0.53
		(0.01)	(0.02)	(0.02)

Notes: Table shows estimated parameters B1 and B2, as well as match rates, Type I and Type II errors for Feigenbaum classifier results reported in the paper, and an exercise where the training and testing data are permuted over 200 samples.

As a final investigation of the performance of this algorithm, we varied the algorithm's γ parameter, chosen by researchers, that determines the relative weight the algorithm places on measures of performance in the training data to choose *B1* and *B2.* Specifically, the algorithm selects values of *B1* and *B2* that result in the highest sum of $\gamma PPV + (1-\gamma)TPR$ where PPV is positive predictive value and TPR is the true positive rate. These values are calculated as follows:

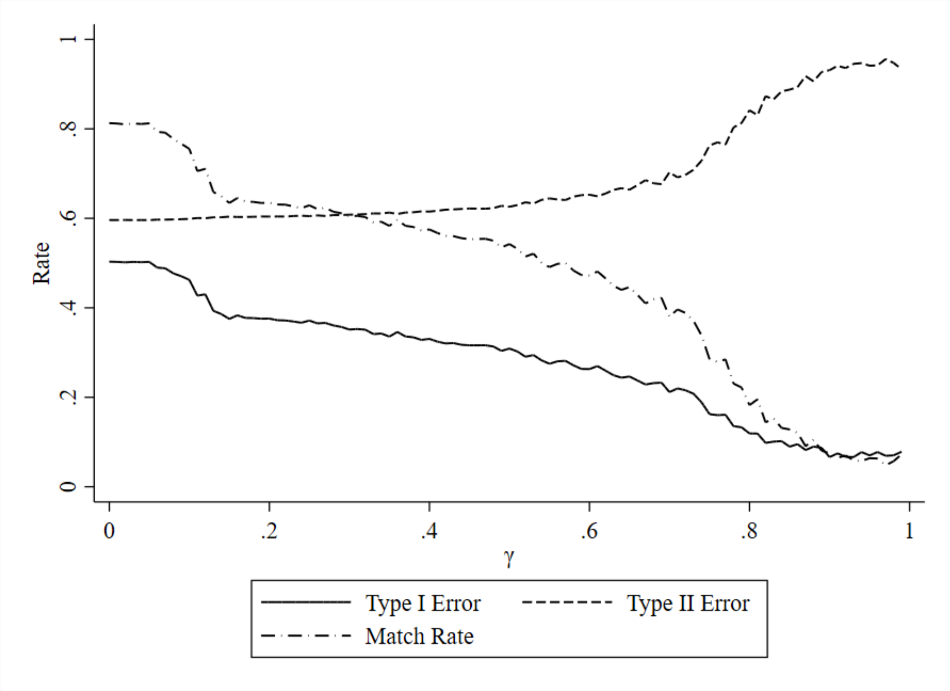
$$PPV = \frac{correct\ matches}{all\ matches\ made}$$

$$TPR = \frac{correct\ matches}{total\ possible\ correct\ matches}$$

Note that *PPV* is the same as one minus the Type I error rate of a matching algorithm, and *TPR* is a scalar multiple of one minus the Type II error rate. Hence, in the context of this algorithm, increasing γ puts a higher weight on achieving a low Type I error rate. For our baseline results in the paper, we choose $\gamma = 0.5$, the same parameter as used in Feigenbaum (2016). Here we vary the choice of γ and look at how algorithm performance changes.

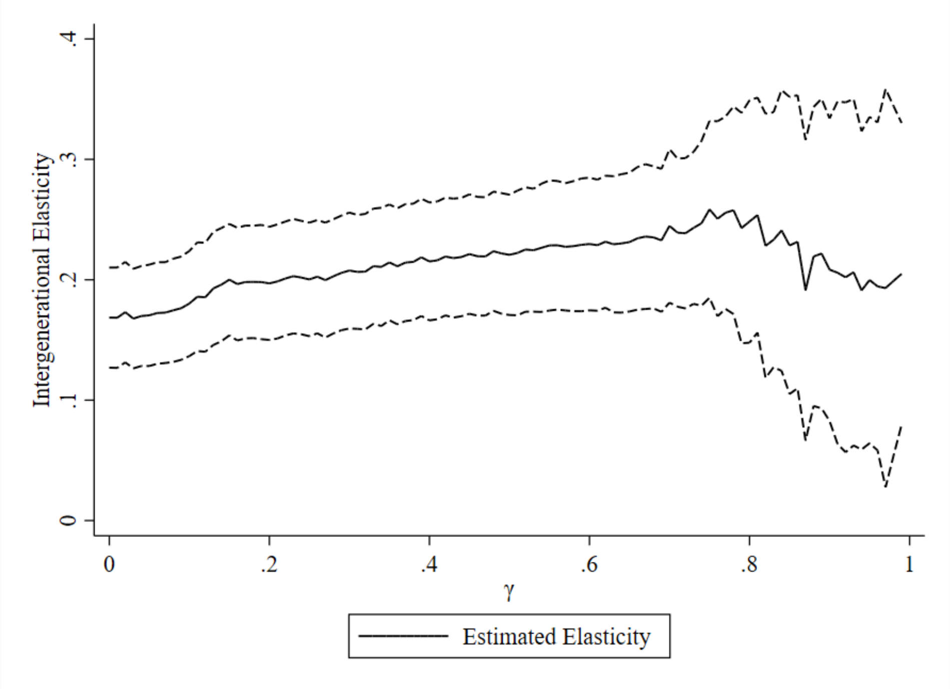
Appendix Figure A1 below shows that, as γ increases, Type I errors fall, and Type II errors rise. This result is unsurprising, as increasing γ puts a higher weight on achieving a lower Type I error rate and a lower weight on Type II errors. Appendix Figure A2 shows that increasing γ tends to increase the estimated intergenerational elasticity, but the intergenerational elasticity estimated with only the records that are correctly linked remains fairly consistent (Appendix Figure A4). This result suggests that incorrect matches attenuate the intergenerational elasticity for low values of γ, and this attenuation does not persist as γ increases. However, the degree of attenuation is slight. As noted in the paper, this result is different than many of the other algorithms we consider in that the estimated intergenerational elasticity with incorrect matches is similar to the estimated elasticity with correctly linked matches. For values of γ larger than 0.3, we nearly always fail to reject the null hypothesis that the estimated elasticity using Feigenbaum's method is the same as the estimated elasticity in the training data. Reweighting slightly increases the probability of failing to reject the null across values of γ (Appendix Figure A3). A selected set of the estimates depicted in these figures are reported in Appendix Table A3.

Appendix Figure A1. Feigenbaum (2016) Performance in LIFE-M Data for Different Choices of γ



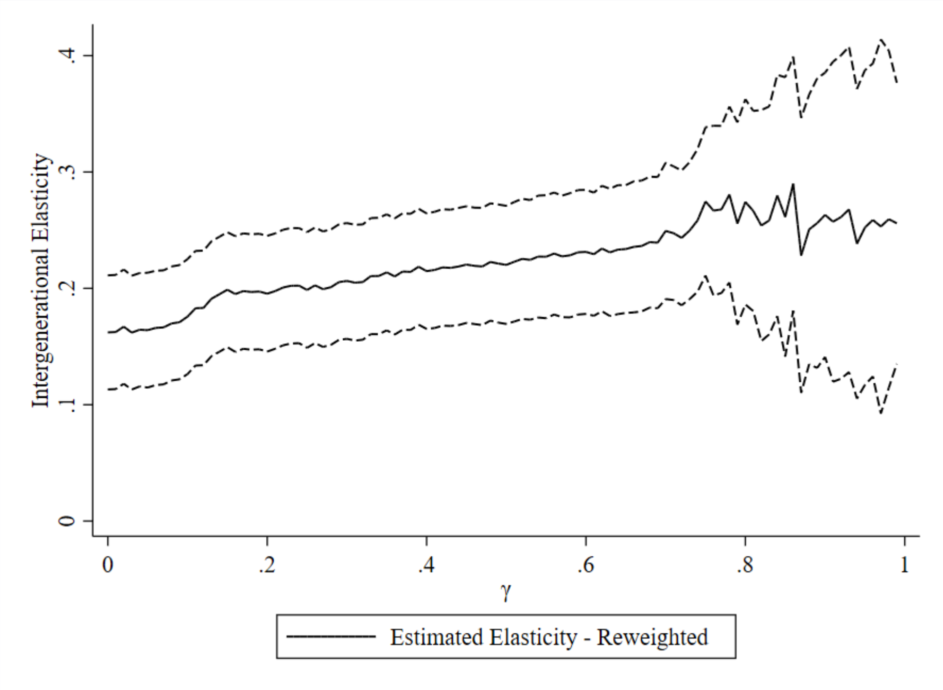
Notes: Graph depicts Type I and Type II errors and match rates on average in LIFE-M data across different choices of γ.

Appendix Figure A2. Feigenbaum (2016) Intergenerational Elasticities and Confidence Intervals for Different Choices of γ in LIFE-M Data



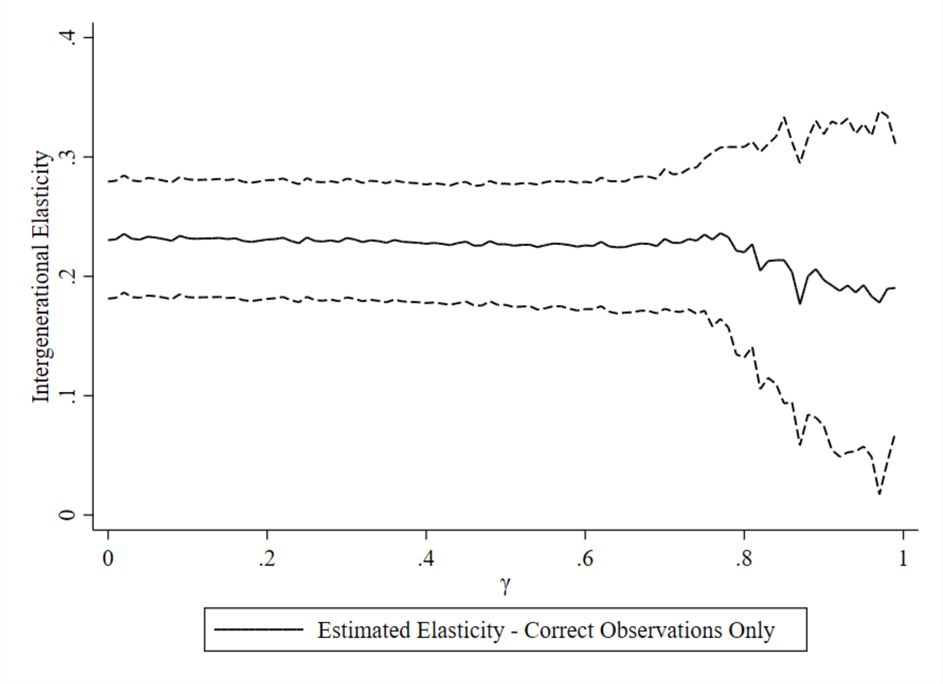
Notes: Graph depicts estimated intergenerational elasticity across different choices of γ in LIFE-M data. Dotted lines depict 95% confidence interval of estimated intergenerational elasticity.

Appendix Figure A3. Feigenbaum (2016) Reweighted Intergenerational Elasticities and Standard Errors for Different Choices of γ



Notes: Graph depicts estimated reweighted intergenerational elasticity across different choices of γ. Dotted lines depict 95% confidence intervals of estimated intergenerational elasticity.

Appendix Figure A4. Feigenbaum (2016) Intergenerational Elasticities and Standard Errors with Correct Observations for Different Choices of γ



Notes: Graph depicts estimated intergenerational elasticity across different choices of γ using only observations correctly matched. Dotted lines depict 95% confidence intervals of intergenerational elasticity.

Appendix Table A3. Select Intergenerational Elasticities by Gamma Value

	(1)	(2)	(3)
Gamma	Estimated Intergenerational Elasticity	Estimated Intergenerational Elasticity - Correct Observations Only	Estimated Intergenerational Elasticity - Reweighted
0.05	0.170	0.233	0.164
	(0.0214)	(0.0251)	(0.0226)
0.10	0.180	0.231	0.175
	(0.0222)	(0.0251)	(0.0236)
0.20	0.197	0.230	0.195
	(0.0239)	(0.0253)	(0.0258)
0.30	0.207	0.232	0.206
	(0.0245)	(0.0253)	(0.0263)
0.40	0.215	0.227	0.214
	(0.0249)	(0.0253)	(0.0269)
0.50	0.220	0.226	0.220
	(0.0254)	(0.0258)	(0.0271)
0.60	0.229	0.225	0.231
	(0.0281)	(0.0271)	(0.0294)
0.70	0.244	0.231	0.249
	(0.0325)	(0.0298)	(0.0351)
0.80	0.248	0.220	0.274
	(0.0513)	(0.0450)	(0.0547)
0.90	0.208	0.196	0.263
	(0.0641)	(0.0623)	(0.0731)
0.95	0.199	0.192	0.252
	(0.0691)	(0.0689)	(0.0789)

Notes: Table reports estimated intergenerational elasticities shown in Figures A2 through A4.

2. Implementation of Abramitzky, Boustan and Eriksson (2012, 2014)

The authors shared the most updated version of code used in Abramitzky et al. (2014), which was updated and posted here: https://people.stanford.edu/ranabr/matching-codes . We implement all algorithms using the code provided by the authors (labeled "Abramitzky et al 2014" in Appendix Table A3).

To implement Abramitzky et al. (2014), one must decide which dataset is the "destfile" and the "sourcefile" (the terms in Abramitzky et al. (2014) code shared with us are "childfile" and "adultfile," respectively). The authors indicated in correspondence that the their 2014 paper uses the full Census data as the "destfile" and the sample file as the "sourcefile," so we report these estimates using the same ordering in the main text of this paper. For the LIFE-M and synthetic data, the 1940 Census is the "destfile" and the birth certificate sample is the "sourcefile." For the Early Indicators data, we make the UA Army Data the “sourcefile” and the 1900 Census data the “destfile.” For the IPUMS data, we make the non-1880 data the “sourcefile” and the 1880 full Census the “destfile.”

An alternative ordering that treats the Census file as the “destfile” and the non-Census year as the “sourcefile.” This second specification is not published elsewhere, and we present it to emphasize the importance of order for Abramitzky et al. (2014) with the existing publicly-available STATA code. Since the code pre-emptively drops all duplicate name-age matches in the “sourcefile,” making a Census file the “sourcefile” may incorrectly link individuals from the “destfile” to the second-best link in the “sourcefile.”

Consider the following example. Suppose that a smaller sample has been made the “destfile” and a Census the “sourcefile”, and there is one John Smith born in Ohio who would have been age 25 in the later year (the “destfile”) but in the “sourcefile” there are three John Smiths born in Ohio age 25 (record 1-3 in the example below) and one John Smith born in Ohio age 27 (record 4) (and no other John Smith observations in the age range of 23-27). Since the Census is the “sourcefile,” Abramitzky et al. (2014) pre-emptively drops all John Smiths who are age 25 in the “sourcefile” and would link in error the sample John Smith age 25 to the Census John Smith age 27 (record 4), even though presumably better matches existed that were ruled out with the first step. On the other hand, if the Census is the “destfile” and the smaller sample the “sourcefile,” the code would not match the John Smith observation in the “sourcefile” to any of the observations in the Census since multiple potential exact links exist.

Example Linking Problem where Order Matters

“destfile”

“destfile” Name	Birthplace	Age
…	…	…
John Smith	OH	25

“sourcefile”

Record number	“sourcefile” Name	Birthplace	Age
~~1~~	~~John Smith~~	~~OH~~	~~25~~
~~2~~	~~John Smith~~	~~OH~~	~~25~~
~~3~~	~~John Smith~~	~~OH~~	~~25~~
4	John Smith		27

Choices about which file is the “sourcefile” and “destfile” are, therefore, consequential for error rates and may be more consequential in settings where the sample files are generally much smaller than a population enumeration. It would be less consequential in settings where researchers are linking full populations to one another, because multiple John Smiths would likely appear in both files and, therefore, be eliminated as matches. Note that the robustness check with the two-year radius reported in our paper eliminates this problem: in our example, it would drop all of the John Smith observations in the “sourcefile.” The table A4.A and A4.B demonstrate how match and error rates change as with the choice of which file is the “sourcefile.”

Appendix Table A4.A. Summary of Match and Error Rates for Linking with Census as "destfile", by Dataset

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed Links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Abramitzky et al. 2014									
Name	0.41	0.41	0.44	0.25	0.29	0.21	0.69	0.71	0.65
NYSIIS	0.42	0.42	0.48	0.32	0.33	0.24	0.72	0.72	0.64
SDX	0.39	0.42	0.50	0.41	0.38	0.28	0.77	0.74	0.64
Abramitzky et al. 2014 (Robustness)									
Name	0.28	0.29	0.35	0.18	0.20	0.17	0.77	0.77	0.71
NYSIIS	0.24	0.26	0.33	0.23	0.23	0.17	0.81	0.80	0.72
SDX	0.18	0.20	0.30	0.32	0.28	0.18	0.88	0.86	0.75

Appendix Table A4.B. Summary of Match and Error Rates for Linking with Census as "sourcefile", by Dataset

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed Links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Abramitzky et al. 2014									
Name	0.47	0.46	0.49	0.34	0.36	0.28	0.69	0.71	0.65
NYSIIS	0.51	0.51	0.56	0.45	0.44	0.35	0.72	0.72	0.64
SDX	0.53	0.55	0.61	0.56	0.52	0.40	0.77	0.74	0.63
Abramitzky et al. 2014 (Robustness)									
Name	0.28	0.29	0.35	0.18	0.20	0.17	0.77	0.77	0.71
NYSIIS	0.24	0.26	0.33	0.23	0.23	0.17	0.81	0.80	0.72
SDX	0.18	0.20	0.30	0.32	0.28	0.18	0.88	0.86	0.75

Notes: In Table A3.A, we treat the Census file as the "childfile" (or "destfile") in the code. In Table A3.B, we treat the Census file as the "adultfile" (or "sourcefile"). EI stands for the "Early Indicators" data. "Baseline" refers to the primary matching algorithm used in Abramitzky et al. 2014. "Robustness" refers to specifications where both the original dataset and the data being linked to are limited to unique name combinations within a five-year age band prior to linkage.

3. Implementation of the Abramitzky, Mill and Perez (2018)

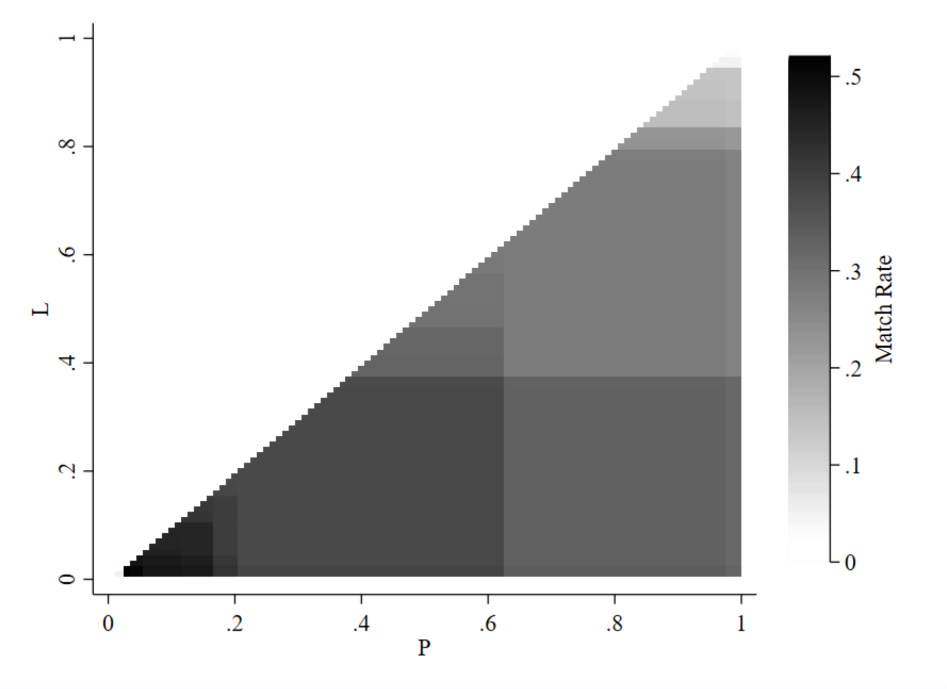
We estimate results for Abramitzky et al. (2018) two ways in the main paper. In order to classify a set of potential matches as matches in their algorithm, a researcher must choose two parameters: a minimum value of acceptable probability for a given match (referred to as *P* in their documentation), and a minimum allowable difference between the maximum probability and the second-highest probability of a match for that particular observation (referred to as *L*). Note that these cut-offs are conceptually similar to Feigenbaum (2016) using *B1* and *B2*, but these cut-offs are chosen by a researcher (not estimated using training data).

Following the cut-offs used by Abramitzky et al. (2018) in a series of replication exercises for other research, we use an *L*-cutoff of 0.1 and a *P*-cutoff of 0.05 for results labeled "Abramitzky et al. 2018 (Less conservative), and use an *L*-cutoff of 0.70 and a *P*-cutoff of 0.65 for results labeled "Abramitzky et al. 2018 (More conservative)." Note that the second set of cutoffs are more restrictive than the first because they require a higher minimum probability and specify that the closest, second-best match be further away in probability. Thus, while the less conservative version would accept a match with an estimated probability of being a match of 0.7 and the second-best match for an observation has an estimated probability of 0.5, the more conservative version would reject that match.

To demonstrate the robustness of our findings, we show how changing these cut-offs would change the match rates, error rates, and intergenerational elasticity inference results in the LIFE-M data. The patterns of match rates and error rates are similar in the synthetic and Early Indicators data for these metrics of algorithm performance, but we omit them here for brevity.

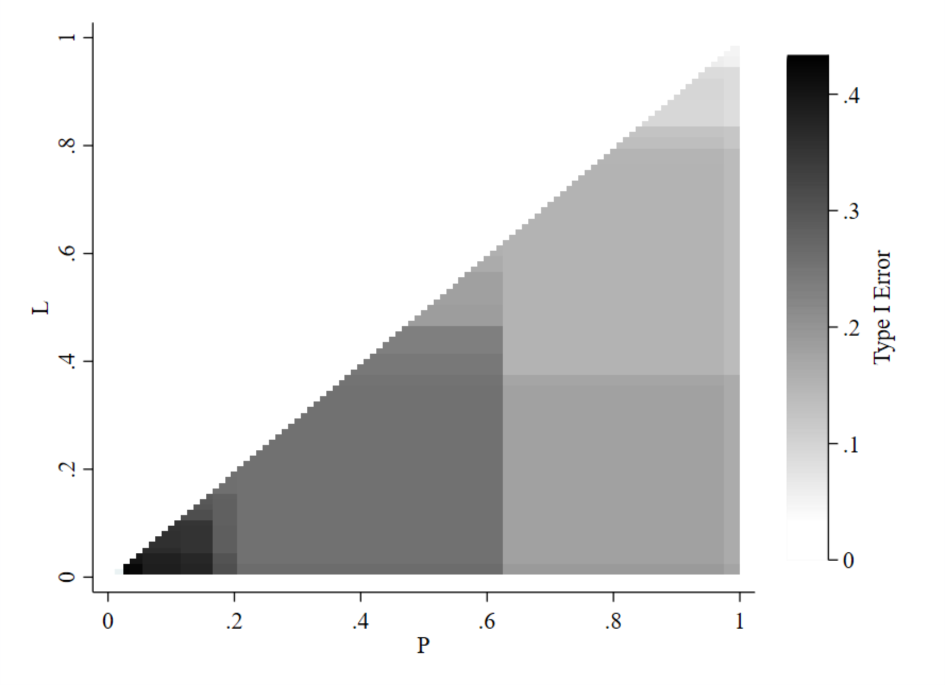
Increasing *P* tends to decrease match rates and decrease the share of observations that are incorrectly matched, as is clear in Appendix Figures A5 and A6 respectively. This pattern makes sense: increasing this cutoff focuses attention on the smaller subset of potential matches that the model predicts are highly likely to be matches. However, for most levels of *P* or *L*, the estimated intergenerational elasticities remain consistently near 0.19 to 0.20, as demonstrated in Appendix Figure A8. Notably for 97 percent of these estimates, we reject the null hypothesis that the estimated elasticity from this data is the same as the estimated elasticity from the hand-linked and reviewed LIFE-M data at the 10-percent significance level. For these same sets of data, however, the estimated elasticity from the correct links is consistent, as can be seen in Appendix Figure A10, with the elasticity among these observations hovering between 0.22 and 0.24. Reweighting increases estimated intergenerational elasticities slightly as is apparent in Appendix Figure A9, and makes the difference in estimated elasticities statistically indistinguishable at the 10% significance level in 34 percent of estimates.

Appendix Figure A5. Match Rates in LIFE-M with Abramitzky et al. (2018) Under Different Choices of *L* and *P*



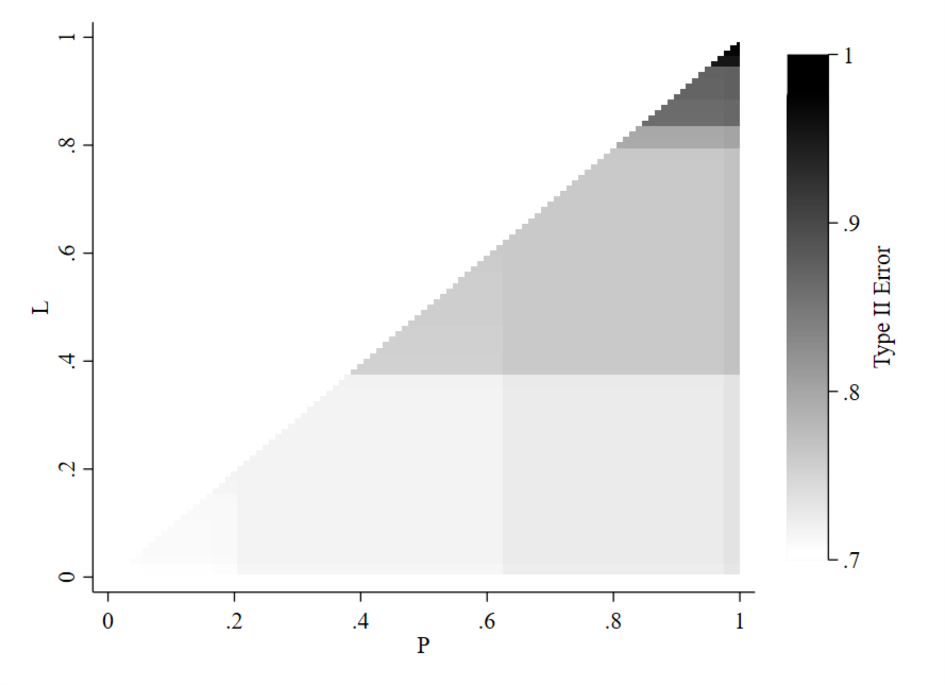
Notes: Figure depicts match rates in LIFE-M data from observations linked using method from Abramitzky et al. (2018) under different choices of *L* and *P*.

Appendix Figure A6. Type I Error Rates in LIFE-M with Abramitzky et al. (2018) Under Different Choices of *L* and *P*



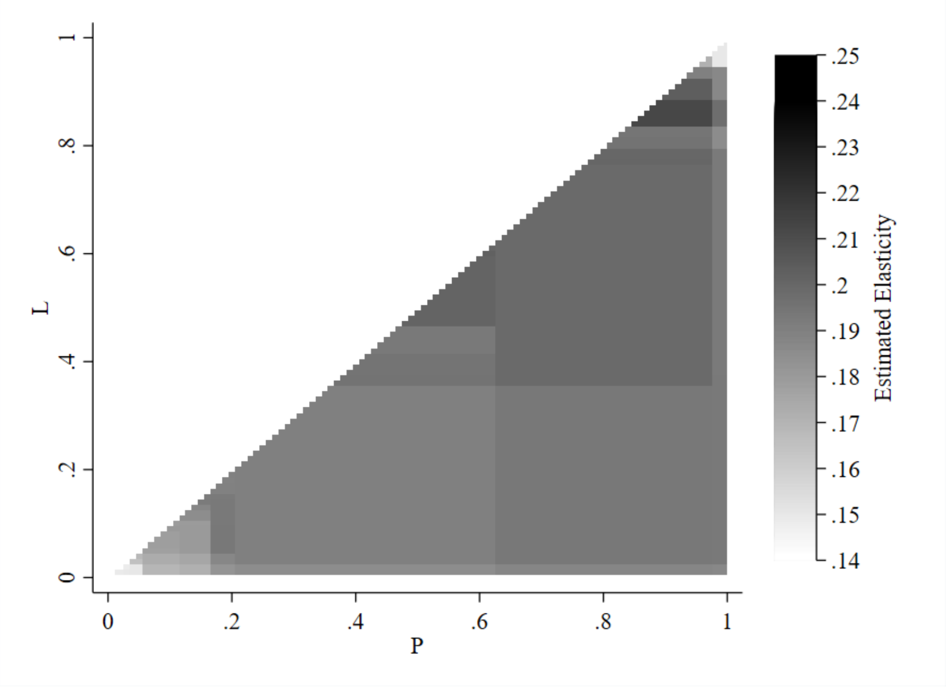
Notes: Figure depicts Type I error rates in LIFE-M data from observations linked using method from Abramitzky et al. (2018) under different choices of *L* and *P*.

Appendix Figure A7. Type II Error Rates in LIFE-M with Abramitzky et al. (2018) Under Different Choices of *L* and *P*



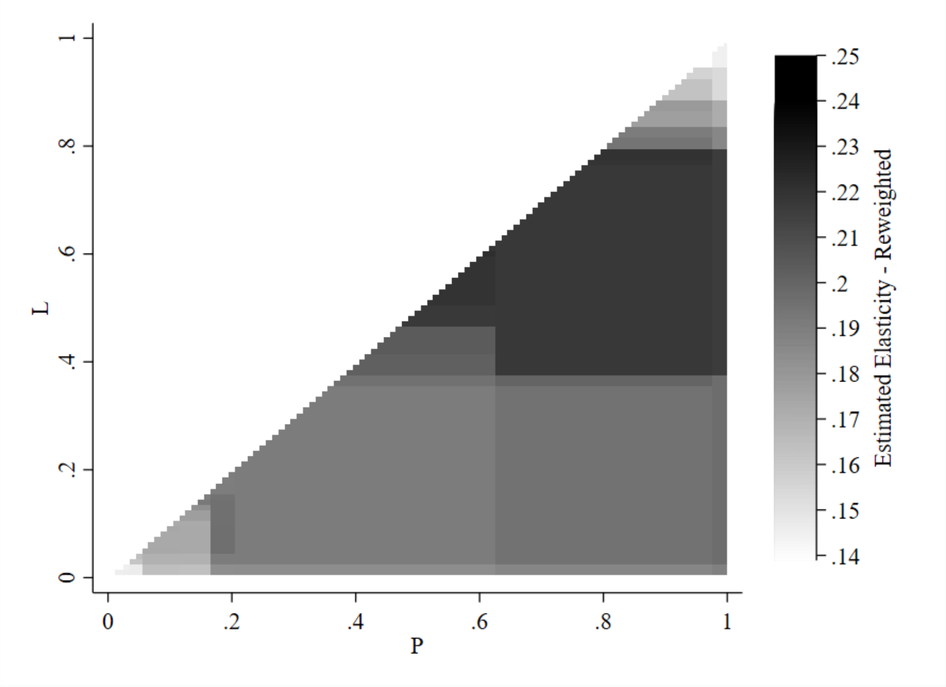
Notes: Figure depicts Type II error rates in LIFE-M data from observations linked using method from Abramitzky et al. (2018) under different choices of *L* and *P*.

Appendix Figure A8. Estimated Elasticity in LIFE-M with Abramitzky et al. (2018) Under Different Choices of *L* and *P*



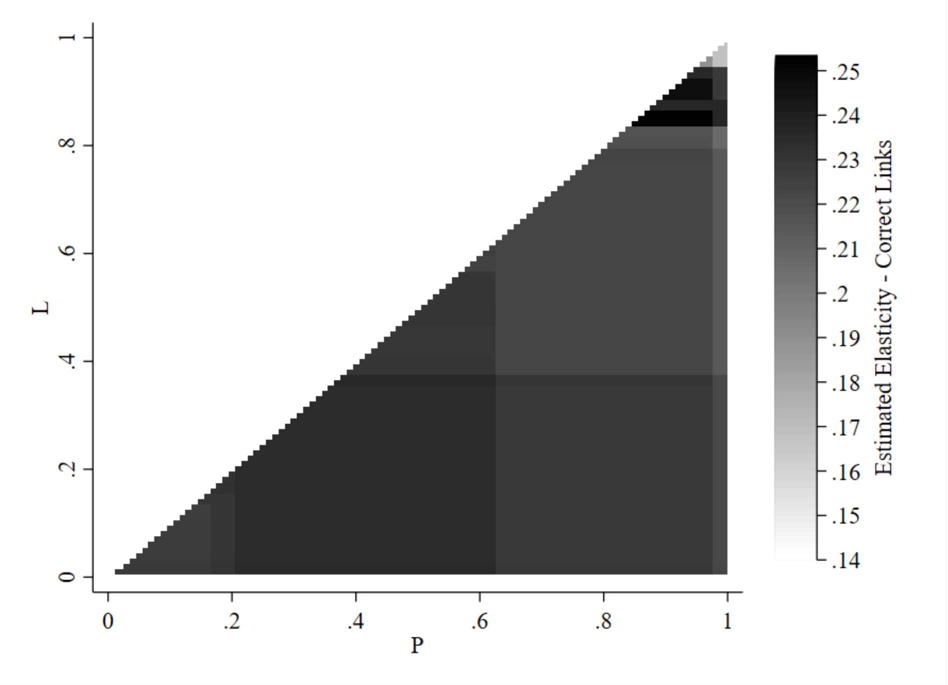
Notes: Figure depicts elasticity estimated in LIFE-M data from observations linked using method from Abramitzky et al. (2018) under different choices of *L* and *P*.

Appendix Figure A9. Estimated Elasticity in LIFE-M with Abramitzky et al. (2018) Under Different Choices of L and P –Reweighted



Notes: Figure depicts elasticity estimated in LIFE-M data from observations linked using method from Abramitzky et al. (2018) under different choices of L and P.

Appendix Figure A10. Estimated Elasticity in LIFE-M with Abramitzky et al. (2018) Under Different Choices of L and P –Correct Observations Only



Notes: Figure depicts elasticity estimated in LIFE-M data from observations linked using method from Abramitzky et al. (2018) using correctly linked observations only under different choices of *L* and *P*.

Lastly, we note that this algorithm's results can be sensitive to the starting value chosen for the maximization procedure. When using the starting values embedded in the code downloaded from the website, the estimated model produces some implausible results for the LIFE-M data. For instance, running the model with no alterations in the starting values produced a set of estimated parameters that assigned potential matches with more dissimilar names a higher probability of being a match. Consequently, using the algorithm without any alterations in starting values resulted in match rates of 33.6 percent and 2.0 percent in the less conservative and more conservative models, respectively, and in Type I error rates in these same matches of 94.0 percent and 78.8 percent. We ran the algorithm with a variety of starting values in each data until we found results that were consistent. Therefore, we strongly recommend that researchers using this method consider a range of potential starting values.

B. Description of LIFE-M Data

The LIFE-M samples used in this paper are based on the first two years of hand-linking in this project. This section supplements the description in the paper.

1. Determining Sex

Because it was the norm for girls to change their name at marriage and around 90 percent of surviving girls married in the middle of the 20th century, we only attempt to link male infants from birth certificates to their records in the 1940 Census. In some cases, sex is missing or apparently incorrect. To determine which of these records were very likely to be boys, we generated an empirical distribution of sex and first names using all vital records in the LIFE-M collection. We use this distribution to classify names as "male" if there were at least 50 records with the name and at least 99 percent of the records with that name were male. If there was ambiguity about whether the infant was a boy, we did not attempt to link the record.

2. Constructing Samples of Male Birth Certificates

To construct data for Ohio, we drew a random sample of 13,270 birth certificates for individuals born in Ohio from 1909 to 1920: a 2 percent sample of 1909 and a 1-percent sample each year between 1910 and 1920. Next, sets of parents for these infants were linked to parents of other infants using the first and last name at birth (for women, this is often called "maiden name") for both fathers and mothers (four distinct fields), thus recovering siblings for each sampled birth certificate. The final sample of reconstructed families consists of 53,721 children, 19,090 of which we determined to be boys (see discussion above). After cleaning potential duplicates among birth certificates (e.g., due to data entry duplication or delayed birth certificates), we were left with 18,461 boys. Appendix Figure C1 plots the distribution of age in 1940 for the 18,461 Ohio boys.

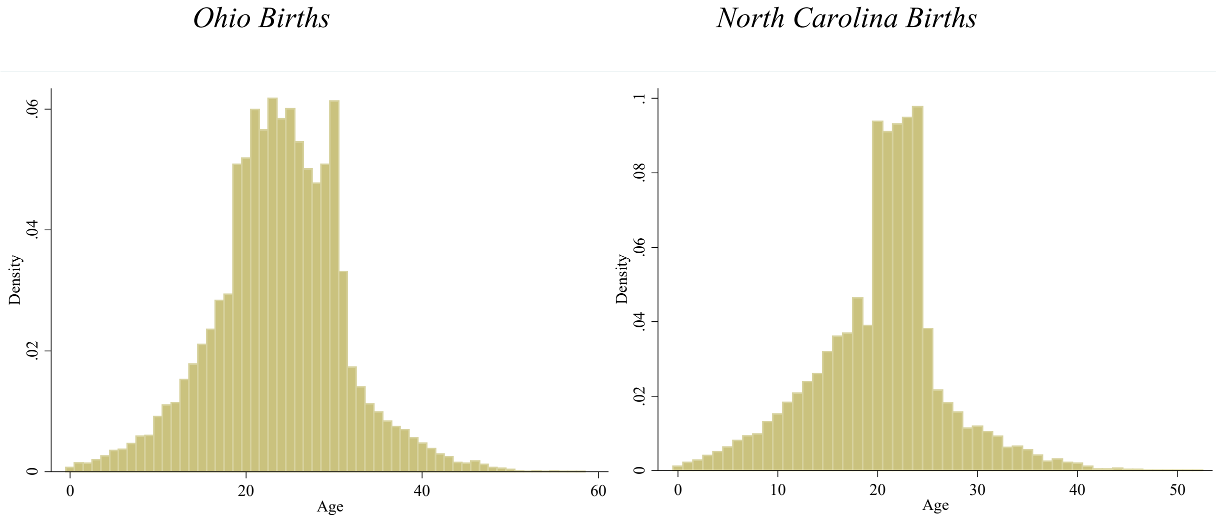
To construct the North Carolina data, we drew a random sample of 23,073 birth certificates for individuals born in North Carolina from 1915 to 1919. These sampling years differ from Ohio due to the incompleteness of the North Carolina birth data before 1915. We used the same process as in Ohio to link parents' names to identify siblings. This resulted in a total sample of 86,209 infants, 26,352 of whom we determined to be boys (see discussion above). After cleaning potential duplicates among birth certificates, we were left with 24,481 boys. Appendix Figure C1 shows the distribution of age in 1940 for these 24,481 boys.

3. Linking Birth Certificate Samples to the 1940 Census

A final step linked the Ohio and North Carolina samples to the 1940 Census using the process described in the paper. During the hand-linkage process, each potential link was initially reviewed by two trainers in a double-blind review process. If these trainers agreed on whether to link or not to link a case, we take their decision as the truth. If they disagreed, an additional three additional trainers reviewed the record. We treated records for which 4/5 of the trainers agreed as a match, the presumption being that one of the original trainers made an error. The first two data trainers disagreed 10 percent of the time for male infants in Ohio and 8.3 percent of the time for male infants in North Carolina. This matching process resulted in 11,751 matches in North Carolina and 9,658 matches in Ohio. As a final step to determine LIFE-M links, we dropped all matches where more than one birth certificate linked to the same 1940 Census observation. This last step dropped 297 matches in Ohio and 2,011 matches in North Carolina for a grand total of 19,100 matches for the LIFE-M data.

As described in the paper, we engage in a final layer of review to determine correct matches where we send all matches from automated methods that differ from LIFE-M through the "police line-up" process, where between 2 to 5 trainers saw both the automated method's link and the existing LIFE-M link, if one was present, along with a set of other potential matches. Trainers were not aware what the previous LIFE-M link was, or what the automated method's link was.

Appendix Figure B1. Distribution of Boys Age in 1940



Ohio Births

North Carolina Births

C. *Representativeness Regression Results*

Appendix Table C1. Representativeness Results for LIFE-M Data

	LIFE-M	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Day of Birth	0.00010***	0.00006***	0.00005**	0.00003**	0.00008***	0.00008***	0.00006***	0.00005**	0.00002	0.00000	0.00009***	0.00009***	0.00007***
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Number of Siblings	-0.00212***	0.00075	0.00010	0.00018	-0.00312***	-0.00380***	-0.00376***	0.00432***	0.00243***	0.00030	-0.00216***	-0.00263***	-0.00385***
	(0.00081)	(0.00076)	(0.00073)	(0.00066)	(0.00082)	(0.00082)	(0.00081)	(0.00077)	(0.00068)	(0.00059)	(0.00081)	(0.00081)	(0.00080)
Length of Child's Name	-0.00052	0.00550***	0.00394**	0.00481***	0.00505***	0.00437**	0.00806***	0.00507***	0.00707***	0.00815***	-0.00340*	0.00983***	0.00975***
	(0.00194)	(0.00182)	(0.00174)	(0.00158)	(0.00195)	(0.00195)	(0.00193)	(0.00188)	(0.00171)	(0.00154)	(0.00191)	(0.00192)	(0.00192)
Length of Father's Name	0.00329***	0.00170**	0.00655***	0.00430***	0.00071	0.00367***	0.00419***	-0.00333***	-0.00494***	-0.00197***	0.00108	0.00362***	0.00449***
	(0.00074)	(0.00070)	(0.00067)	(0.00060)	(0.00074)	(0.00074)	(0.00074)	(0.00069)	(0.00061)	(0.00053)	(0.00073)	(0.00073)	(0.00073)
Length of Mother's Name	0.00429***	0.00195**	0.00399***	0.00192***	0.00132	0.00240***	0.00217***	-0.00037	-0.00140**	-0.00017	0.00196**	0.00288***	0.00206***
	(0.00081)	(0.00076)	(0.00073)	(0.00066)	(0.00081)	(0.00081)	(0.00080)	(0.00076)	(0.00066)	(0.00057)	(0.00080)	(0.00080)	(0.00080)
Share of Observations with Mispelled Mother's Name	0.00168	0.00629	0.00907	0.01441*	-0.00901	-0.00191	0.00557	-0.02438***	-0.01379*	-0.00645	-0.01893**	0.00240	0.00136
	(0.00917)	(0.00856)	(0.00823)	(0.00754)	(0.00922)	(0.00922)	(0.00915)	(0.00868)	(0.00762)	(0.00658)	(0.00906)	(0.00911)	(0.00905)
Share of Observations with Mispelled Fatherr's Name	0.04389***	0.01920**	0.03758***	0.03542***	0.02018**	0.02946***	0.03757***	-0.01823**	-0.02840***	-0.01176*	0.02253**	0.03590***	0.04819***
	(0.00916)	(0.00852)	(0.00824)	(0.00759)	(0.00920)	(0.00921)	(0.00914)	(0.00857)	(0.00759)	(0.00655)	(0.00907)	(0.00911)	(0.00907)
Link in Ohio	0.10124***	0.11289***	0.05580***	-0.00718*	0.08943***	0.07651***	0.03819***	-0.04444***	-0.01095**	0.02388***	0.09201***	0.07924***	0.04220***
	(0.00521)	(0.00494)	(0.00469)	(0.00418)	(0.00523)	(0.00523)	(0.00517)	(0.00488)	(0.00426)	(0.00363)	(0.00518)	(0.00518)	(0.00513)
Constant	0.27803***	0.17442***	0.05946***	0.06952***	0.36366***	0.31065***	0.25751***	0.71103***	0.84150***	0.83569***	0.34330***	0.22351***	0.21894***
	(0.01728)	(0.01618)	(0.01543)	(0.01395)	(0.01735)	(0.01731)	(0.01714)	(0.01645)	(0.01479)	(0.01284)	(0.01707)	(0.01707)	(0.01702)
Observations	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869
R-squared	0.015	0.016	0.011	0.003	0.009	0.009	0.005	0.004	0.004	0.002	0.010	0.010	0.006
Wald	634.74	688.29	445.93	130.61	412.25	402.59	208.82	178.84	148.16	104.66	454.58	456.99	255.68
Prob > W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors are in parentheses. Regressions are a test of representativeness in the LIFE-M data by regressing for all observations in the birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether being linked is systematically and jointly related to covariates.

Appendix Table C1. Representativeness Results for LIFE-M Data - Continued

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Day of Birth	0.00007***	0.00009***	0.00010***	0.00012***	0.00013***
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Number of Siblings	0.00017	0.00536***	0.00549***	-0.00410***	-0.00005
	(0.00070)	(0.00083)	(0.00085)	(0.00081)	(0.00072)
Length of Child's Name	0.01152***	0.00508***	0.00008	-0.00216	0.00765***
	(0.00165)	(0.00196)	(0.00201)	(0.00194)	(0.00173)
Length of Father's Name	0.00567***	0.00399***	0.00378***	0.00258***	0.00339***
	(0.00064)	(0.00074)	(0.00076)	(0.00074)	(0.00065)
Length of Mother's Name	0.00415***	0.00298***	0.00355***	0.00209***	0.00278***
	(0.00070)	(0.00082)	(0.00083)	(0.00081)	(0.00072)
Share of Observations with Mispelled Mother's Name	0.00354 (0.00781)	0.02068** (0.00927)	0.00270 (0.00949)	-0.00059 (0.00916)	-0.00193 (0.00803)
Share of Observations with Mispelled Fatherr's Name	0.04025*** (0.00789)	0.03336*** (0.00924)	0.01106 (0.00947)	0.03141*** (0.00916)	0.01635** (0.00807)
Link in Ohio	0.06382***	0.02450***	0.06073***	0.12096***	0.14314***
	(0.00450)	(0.00523)	(0.00536)	(0.00521)	(0.00474)
Constant	-0.01445	0.33334***	0.34507***	0.34083***	0.06039***
	(0.01471)	(0.01743)	(0.01783)	(0.01726)	(0.01536)
Observations	42,869	42,869	40,869	42,869	42,869
R-squared	0.014	0.005	0.008	0.018	0.031
Wald	568.64	195.75	334.92	788.30	1350
Prob > W	0.00	0.00	0.00	0.00	0

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors are in parentheses. Regressions are a test of representativeness in the LIFE-M data by regressing for all observations in the birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether or not matched status systematically relates to covariates.

Appendix Table C2. Representativeness Results for Synthetic Data

	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Day of Birth	0.00005**	0.00004**	0.00003	0.00005**	0.00006***	0.00003	-0.00004**	-0.00003*	-0.00003*	0.00002	0.00006***	0.00003
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Number of Siblings	-0.00361***	-0.00315***	-0.00132*	-0.00653***	-0.00615***	-0.00579***	0.00218***	0.00241***	0.00026	-0.00650***	-0.00598***	-0.00621***
	(0.00077)	(0.00074)	(0.00068)	(0.00082)	(0.00082)	(0.00081)	(0.00077)	(0.00068)	(0.00058)	(0.00081)	(0.00081)	(0.00081)
Length of Child's Name	0.00655***	0.00387**	0.00209	0.00570***	0.00595***	0.00369*	0.00913***	0.00997***	0.00911***	0.00075	0.01293***	0.00870***
	(0.00185)	(0.00179)	(0.00168)	(0.00195)	(0.00195)	(0.00195)	(0.00186)	(0.00169)	(0.00146)	(0.00193)	(0.00194)	(0.00194)
Length of Father's Name	0.00108	0.00417***	0.00367***	-0.00059	0.00124*	0.00290***	-0.00381***	-0.00694***	-0.00348***	-0.00047	0.00185**	0.00321***
	(0.00070)	(0.00067)	(0.00062)	(0.00074)	(0.00074)	(0.00074)	(0.00070)	(0.00063)	(0.00053)	(0.00073)	(0.00073)	(0.00073)
Length of Mother's Name	-0.00031	0.00219***	0.00048	-0.00055	0.00036	0.00106	-0.00200***	-0.00333***	-0.00122**	-0.00047	-0.00001	0.00048
	(0.00076)	(0.00073)	(0.00067)	(0.00081)	(0.00081)	(0.00081)	(0.00077)	(0.00069)	(0.00058)	(0.00080)	(0.00081)	(0.00081)
Share of Observations with Mispelled Mother's Name	0.00116	-0.00107	-0.00277	0.01324	0.00366	-0.00165	0.02261***	0.01895**	0.01048	0.00765	-0.00535	-0.00062
	(0.00864)	(0.00830)	(0.00765)	(0.00921)	(0.00923)	(0.00918)	(0.00877)	(0.00785)	(0.00665)	(0.00909)	(0.00915)	(0.00912)
Share of Observations with Mispelled Fatherr's Name	-0.01509*	-0.03099***	-0.03265***	0.00457	-0.01169	-0.03471***	0.03516***	0.04723***	0.03024***	0.00603	-0.00593	-0.03130***
	(0.00858)	(0.00830)	(0.00780)	(0.00916)	(0.00918)	(0.00917)	(0.00872)	(0.00787)	(0.00668)	(0.00905)	(0.00909)	(0.00912)
Link in Ohio	0.08319***	0.05076***	-0.00654	0.07865***	0.08789***	0.06783***	-0.08308***	-0.02802***	0.02339***	0.06483***	0.07287***	0.05281***
	(0.00495)	(0.00474)	(0.00431)	(0.00522)	(0.00522)	(0.00521)	(0.00498)	(0.00439)	(0.00366)	(0.00517)	(0.00519)	(0.00518)
Constant	0.26150***	0.17933***	0.17725***	0.40611***	0.38932***	0.37827***	0.68149***	0.81725***	0.83420***	0.40307***	0.31594***	0.33769***
	(0.01867)	(0.01803)	(0.01674)	(0.01985)	(0.01987)	(0.01980)	(0.01905)	(0.01717)	(0.01460)	(0.01959)	(0.01966)	(0.01972)
Observations	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869	42,869
R-squared	0.009	0.007	0.002	0.009	0.010	0.007	0.011	0.009	0.004	0.006	0.009	0.006
Wald	390.66	277.50	71.10	378.12	446.95	310.61	452.14	363.92	174.67	271.91	387.18	257.77
Prob > W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a test of representativeness in the synthetic data by regressing for all observations in the synthetic birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether being linked is systematically and jointly related to covariates.

Appendix Table C2. Representativeness Results for Synthetic Data - Continued

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Day of Birth	0.00004*	0.00003	0.00002	0.00007***	0.00006***
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Number of Siblings	-0.00349***	-0.00095	0.00228***	-0.00773***	-0.00476***
	(0.00071)	(0.00083)	(0.00082)	(0.00082)	(0.00076)
Length of Child's Name	0.01133***	0.00721***	0.01217***	0.00243	0.01127***
	(0.00175)	(0.00195)	(0.00193)	(0.00196)	(0.00181)
Length of Father's Name	0.00416***	0.00108	0.00172**	0.00001	0.00236***
	(0.00065)	(0.00074)	(0.00073)	(0.00074)	(0.00069)
Length of Mother's Name	0.00185***	0.00127	0.00033	-0.00044	0.00030
	(0.00071)	(0.00081)	(0.00081)	(0.00081)	(0.00076)
Share of Observations with Mispelled Mother's Name	-0.00719	-0.00838	-0.00293	-0.00438	-0.01025
	(0.00803)	(0.00924)	(0.00920)	(0.00927)	(0.00857)
Share of Observations with Mispelled Fatherr's Name	-0.01671**	-0.01520*	-0.00947	-0.02313**	-0.02510***
	(0.00797)	(0.00918)	(0.00910)	(0.00923)	(0.00854)
Link in Ohio	0.05929***	0.01045**	-0.01101**	0.09319***	0.10438***
	(0.00459)	(0.00521)	(0.00518)	(0.00522)	(0.00490)
Constant	0.10722***	0.49889***	0.53884***	0.50876***	0.20977***
	(0.01734)	(0.01993)	(0.01988)	(0.01996)	(0.01843)
Observations	42,869	42,869	40,869	42,869	42,869
R-squared	0.009	0.001	0.002	0.011	0.016
Wald	397.75	34.86	62.15	472.55	673.05
Prob > W	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a test of representativeness in the synthetic data by regressing for all observations in the synthetic birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether being linked is systematically and jointly related to covariates.

Appendix Table C3. Representativeness Results for Early Indicators Data

	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Age	0.01013**	0.00267	0.00627	0.00705	0.00280	0.00359	0.01447***	0.01713***	0.01277**	0.00987*	0.00435	0.00582
	(0.00493)	(0.00522)	(0.00504)	(0.00519)	(0.00540)	(0.00544)	(0.00542)	(0.00616)	(0.00547)	(0.00533)	(0.00571)	(0.00565)
Is Currently Married	0.09295***	0.03236	0.00883	0.06755*	0.02805	0.00019	0.05187	0.03705	0.03197	0.06574*	0.02339	0.00700
	(0.03603)	(0.03621)	(0.03582)	(0.03695)	(0.03701)	(0.03682)	(0.03609)	(0.03412)	(0.03236)	(0.03585)	(0.03703)	(0.03694)
Foreign Born	0.00327	-0.03386	-0.02838	-0.03353	-0.00542	-0.05508	-0.08665*	-0.07468	-0.08622*	-0.05197	0.02664	-0.06653
	(0.05197)	(0.05185)	(0.04888)	(0.05278)	(0.05349)	(0.05338)	(0.05170)	(0.04822)	(0.04477)	(0.05181)	(0.05307)	(0.05353)
Day of Year Birth	0.00013	-0.00000	0.00012	0.00004	-0.00002	0.00012	0.00011	0.00012	0.00017*	0.00012	-0.00004	0.00017
	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00010)	(0.00009)	(0.00011)	(0.00011)	(0.00011)
Literate	0.15583**	0.11815*	0.11231*	0.08594	0.02120	0.05211	0.09121	0.09661	0.13983**	0.07334	0.06526	0.05431
	(0.06212)	(0.06209)	(0.05931)	(0.06666)	(0.06634)	(0.06625)	(0.06687)	(0.06356)	(0.06403)	(0.06498)	(0.06625)	(0.06682)
Length of First Name	0.01989**	0.02785***	0.03538***	0.01358	0.01867**	0.02464***	0.00006	-0.00214	-0.00379	-0.01049	0.00434	0.01472*
	(0.00848)	(0.00842)	(0.00826)	(0.00863)	(0.00866)	(0.00859)	(0.00859)	(0.00836)	(0.00818)	(0.00850)	(0.00869)	(0.00863)
Length of Second Name	-0.02518***	0.02089***	0.02163***	-0.03036***	0.00226	0.01086	-0.04926***	-0.04592***	-0.02650***	-0.02779***	-0.00272	0.00949
	(0.00707)	(0.00724)	(0.00715)	(0.00707)	(0.00727)	(0.00726)	(0.00678)	(0.00660)	(0.00630)	(0.00702)	(0.00731)	(0.00731)
Mother is Foreign Born	-0.09971*	-0.07775	-0.08133	-0.06753	-0.10125*	-0.02324	-0.07276	-0.05270	-0.01759	-0.04376	-0.11211**	-0.02917
	(0.05138)	(0.05193)	(0.05092)	(0.05191)	(0.05225)	(0.05279)	(0.04834)	(0.04469)	(0.04132)	(0.05230)	(0.05247)	(0.05329)
Father is Foreign Born	0.03299	0.01159	-0.02760	0.04602	0.02970	-0.03241	0.12288***	0.11829***	0.10936***	0.03311	0.01859	-0.00111
	(0.04990)	(0.04931)	(0.04872)	(0.04979)	(0.04896)	(0.04976)	(0.04542)	(0.04138)	(0.03676)	(0.05051)	(0.04996)	(0.05044)
Year of Birth	0.00916*	0.00479	0.00644	0.00418	-0.00063	-0.00067	0.01537***	0.01945***	0.01376**	0.00743	0.00151	0.00269
	(0.00521)	(0.00547)	(0.00530)	(0.00552)	(0.00572)	(0.00578)	(0.00575)	(0.00639)	(0.00571)	(0.00567)	(0.00601)	(0.00598)
Constant	-17.19199*	-8.93633	-12.27542	-7.62326	1.38067	1.29755	-28.35389***	-35.94292***	-25.31993**	-13.71172	-2.62138	-5.00706
	(9.85815)	(10.35117)	(10.02231)	(10.44422)	(10.82675)	(10.93410)	(10.88880)	(12.10543)	(10.81505)	(10.73442)	(11.36783)	(11.30958)
Observations	1,886	1,886	1,886	1,886	1,886	1,886	1,886	1,886	1,886	1,886	1,886	1,886
R-squared	0.024	0.020	0.028	0.018	0.009	0.014	0.039	0.042	0.025	0.017	0.007	0.009
Wald	47.45	38.20	56.96	35.53	15.97	25.57	75.59	69.92	43.56	32.25	12.65	17.13
Prob > W	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.24	0.07

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a test of representativeness in the Early Indicators data by regressing for all observations in the Early Indicators birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether being linked is systematically and jointly related to covariates. Note that the representativeness of the data are measured against the universe of ‘high quality’ links stated by the Early Indicators project for the Oldest Old sample.

Appendix Table C3. Representativeness Results for Early Indicators Data – Continued

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Age	#REF!	0.01763***	0.01892***	0.02032***	0.01995***
	(0.00479)	(0.00573)	(0.00690)	(0.00618)	(0.00579)
Is Currently Married	0.01894	0.06344*	0.02031	0.07041*	0.05539
	(0.03480)	(0.03618)	(0.04345)	(0.03646)	(0.03530)
Foreign Born	-0.00633	-0.10089*	-0.15828***	-0.07842	-0.04085
	(0.04776)	(0.05261)	(0.06083)	(0.05426)	(0.05040)
Day of Year Birth	0.00010	0.00002	0.00007	0.00006	0.00006
	(0.00011)	(0.00011)	(0.00013)	(0.00011)	(0.00011)
Literate	0.15867***	0.13228**	0.14818**	0.18051***	0.23114***
	(0.05065)	(0.06445)	(0.07289)	(0.06480)	(0.04995)
Length of First Name	0.01139	0.02472***	0.02469**	0.00743	0.01101
	(0.00788)	(0.00850)	(0.00974)	(0.00847)	(0.00816)
Length of Second Name	0.01623**	-0.00598	-0.02338***	0.00374	-0.00585
	(0.00692)	(0.00707)	(0.00819)	(0.00719)	(0.00707)
Mother is Foreign Born	-0.11261**	-0.06226	-0.04290	-0.04392	-0.05287
	(0.04934)	(0.05069)	(0.05888)	(0.05206)	(0.05026)
Father is Foreign Born	0.05106	0.02935	0.07655	-0.00974	0.02654
	(0.04720)	(0.04729)	(0.05553)	(0.04864)	(0.04816)
Year of Birth	0.00155	0.01193**	0.01509**	0.01347**	0.01336**
	(0.00503)	(0.00599)	(0.00719)	(0.00638)	(0.00598)
Constant	-3.12545	-22.68073**	-28.46547**	-25.70009**	-25.68240**
	(9.51725)	(11.33626)	(13.62244)	(12.08197)	(11.31869)
Observations	1,866	1,866	1,866	1,866	1,866
R-squared	0.014	0.028	0.035	0.028	0.026
Wald	31.15	50.03	43.98	46.61	51.39
Prob > W	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a test of representativeness in the Early Indicators data by regressing for all observations in the Early Indicators birth certificates data a dummy variable indicating whether or not an observation was matched on the variables included in this table. The Wald statistic tests whether being linked is systematically and jointly related to covariates. Note that the representativeness of the data are measured against the universe of ‘high quality’ links stated by the Early Indicators project for the Oldest Old sample.

D. *Correlation of Incorrect Links with Baseline Sample Characteristics*

Appendix Table D1. Correlation of Incorrect Links with Baseline Sample Characteristics in LIFE-M Data

	LIFE-M	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Day of Birth	-0.00004***	-0.00002	-0.00007**	-0.00000	-0.00005	-0.00010***	-0.00005	-0.00009***	-0.00012***	-0.00011***	-0.00005*	-0.00012***	-0.00008**
	(0.00001)	(0.00003)	(0.00004)	(0.00005)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00003)	(0.00003)	(0.00003)
Number of Siblings	-0.00028	-0.00035	-0.00330**	-0.00826***	-0.00173	-0.00420***	-0.00585***	0.00412***	0.00201**	0.00006	-0.00273**	-0.00478***	-0.00643***
	(0.00048)	(0.00121)	(0.00141)	(0.00175)	(0.00110)	(0.00119)	(0.00130)	(0.00097)	(0.00090)	(0.00083)	(0.00112)	(0.00122)	(0.00133)
Length of Child's Name	-0.00013	0.00409	0.00075	0.00332	0.00290	0.00211	0.00351	0.00265	0.00124	0.00060	0.00536**	-0.00557*	-0.01047***
	(0.00120)	(0.00286)	(0.00329)	(0.00412)	(0.00263)	(0.00281)	(0.00304)	(0.00238)	(0.00221)	(0.00205)	(0.00273)	(0.00293)	(0.00309)
Length of Father's Name	0.00015	-0.00084	-0.00421***	-0.00312**	-0.00023	-0.00468***	-0.00280**	-0.00308***	-0.00741***	-0.00495***	-0.00124	-0.00533***	-0.00371***
	(0.00045)	(0.00107)	(0.00123)	(0.00154)	(0.00098)	(0.00104)	(0.00113)	(0.00088)	(0.00082)	(0.00076)	(0.00100)	(0.00107)	(0.00116)
Length of Mother's Name	-0.00054	-0.00245**	-0.00554***	-0.00602***	-0.00275***	-0.00739***	-0.00656***	-0.00409***	-0.00695***	-0.00458***	-0.00204*	-0.00771***	-0.00697***
	(0.00048)	(0.00114)	(0.00132)	(0.00167)	(0.00106)	(0.00114)	(0.00125)	(0.00096)	(0.00089)	(0.00083)	(0.00108)	(0.00117)	(0.00128)
Share of Observations with Mispelled Mother's Name	-0.00015	0.01555	0.01267	0.04566**	0.00766	0.00402	0.02270	-0.00464	-0.00917	0.00419	0.00085	0.01623	0.01133
	(0.00573)	(0.01384)	(0.01582)	(0.01945)	(0.01256)	(0.01325)	(0.01433)	(0.01098)	(0.01015)	(0.00930)	(0.01296)	(0.01375)	(0.01472)
Share of Observations with Mispelled Fatherr's Name	0.00692	-0.00991	-0.01719	-0.02270	-0.00817	-0.03018**	-0.02070	-0.03766***	-0.05208***	-0.03385***	0.00183	-0.03432**	-0.01398
	(0.00577)	(0.01391)	(0.01568)	(0.01891)	(0.01254)	(0.01309)	(0.01410)	(0.01093)	(0.01014)	(0.00938)	(0.01293)	(0.01349)	(0.01439)
Link in Ohio	-0.01667***	-0.14169***	-0.08939***	-0.03818***	-0.16671***	-0.09832***	-0.05245***	-0.23326***	-0.14229***	-0.07718***	-0.16613***	-0.10710***	-0.06334***
	(0.00300)	(0.00734)	(0.00864)	(0.01093)	(0.00670)	(0.00729)	(0.00802)	(0.00618)	(0.00577)	(0.00534)	(0.00685)	(0.00749)	(0.00816)
Constant	0.06050***	0.30640***	0.46256***	0.48435***	0.39765***	0.59571***	0.59963***	0.68943***	0.86785***	0.86091***	0.36669***	0.64347***	0.69707***
	(0.01033)	(0.02509)	(0.02926)	(0.03683)	(0.02304)	(0.02468)	(0.02697)	(0.02084)	(0.01934)	(0.01785)	(0.02361)	(0.02558)	(0.02738)
Observations	19,100	13,952	11,834	8,559	19,820	19,656	17,757	29,394	33,873	36,840	17,760	17,971	16,839
R-squared	0.003	0.033	0.019	0.009	0.037	0.020	0.009	0.058	0.032	0.013	0.040	0.026	0.013
Wald Statistic	54.97	468.3	242.9	81.49	772.1	429	157.7	1859	1148	471.2	744.4	500.9	223.2
Prob > W	4.47e-09	0	0	0	0	0	0	0	0	0	0	0	0

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in the LIFE-M data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the LIFE-M 'police line-up' review process described in the paper) on the variables included in this table from the LIFE-M birth certificates data. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

Appendix Table D1. Correlation of Incorrect Links with Baseline Sample Characteristics in LIFE-M Data – Continued

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Day of Birth	-0.00008**	-0.00006**	-0.00007***	-0.00016***	-0.00004
	(0.00004)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
Number of Siblings	-0.00490***	0.00869***	0.00901***	-0.00469***	-0.00228**
	(0.00145)	(0.00106)	(0.00105)	(0.00119)	(0.00115)
Length of Child's Name	-0.00663*	-0.00506**	-0.00149	0.00068	0.00622**
	(0.00349)	(0.00256)	(0.00253)	(0.00281)	(0.00275)
Length of Father's Name	-0.00448***	-0.00098	0.00052	-0.00333***	-0.00176*
	(0.00128)	(0.00096)	(0.00095)	(0.00106)	(0.00104)
Length of Mother's Name	-0.00674***	-0.00411***	-0.00296***	-0.00503***	-0.00117
	(0.00138)	(0.00104)	(0.00103)	(0.00115)	(0.00112)
Share of Observations with	0.02171	0.00673	-0.00280	0.02001	-0.00745
Mispelled Mother's Name	(0.01687)	(0.01205)	(0.01207)	(0.01351)	(0.01342)
Share of Observations with	-0.01007	-0.01694	-0.02431**	-0.00464	0.00488
Mispelled Fatherr's Name	(0.01635)	(0.01203)	(0.01207)	(0.01336)	(0.01383)
Link in Ohio	-0.07921***	-0.23678***	-0.22273***	-0.13214***	-0.07256***
	(0.00905)	(0.00652)	(0.00643)	(0.00734)	(0.00730)
Constant	0.51234***	0.52367***	0.41680***	0.59496***	0.21869***
	(0.03058)	(0.02262)	(0.02230)	(0.02475)	(0.02433)
Observations	10,373	22,370	21,320	19,603	12,008
R-squared	0.021	0.071	0.066	0.027	0.012
Wald Statistic	239	1806	1559	559.3	139.8
Prob > W	0	0	0	0	0

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in the LIFE-M data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the LIFE-M 'police line-up' review process described in the paper) on the variables included in this table from the LIFE-M birth certificates data. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

Appendix Table D2. Correlation of Incorrect Links with Baseline Sample Characteristics in Simulated Data

	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Day of Birth	-0.00002	-0.00007*	-0.00001	-0.00001	-0.00004	-0.00003	-0.00007**	-0.00006**	-0.00004*	-0.00003	-0.00004	-0.00005
	(0.00003)	(0.00004)	(0.00004)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00003)	(0.00003)	(0.00003)
Number of Siblings	0.00202	-0.00019	-0.00055	0.00036	-0.00105	-0.00205	0.00288***	0.00163*	0.00007	-0.00014	-0.00072	-0.00185
	(0.00124)	(0.00141)	(0.00167)	(0.00119)	(0.00122)	(0.00127)	(0.00100)	(0.00093)	(0.00087)	(0.00119)	(0.00122)	(0.00127)
Length of Child's Name	-0.00286	-0.00265	0.00028	-0.00677**	-0.00538*	-0.00457	-0.00286	-0.00441*	-0.00438**	-0.00465*	-0.00593**	-0.00722**
	(0.00277)	(0.00313)	(0.00367)	(0.00273)	(0.00280)	(0.00293)	(0.00240)	(0.00226)	(0.00211)	(0.00270)	(0.00276)	(0.00287)
Length of Father's Name	-0.00142	-0.00374***	0.00019	-0.00208**	-0.00512***	-0.00144	-0.00332***	-0.00642***	-0.00372***	-0.00227**	-0.00467***	-0.00266**
	(0.00111)	(0.00125)	(0.00148)	(0.00105)	(0.00108)	(0.00112)	(0.00091)	(0.00085)	(0.00079)	(0.00106)	(0.00108)	(0.00111)
Length of Mother's Name	-0.00048	-0.00055	0.00058	-0.00100	-0.00246**	-0.00013	-0.00218**	-0.00374***	-0.00179**	-0.00139	-0.00336***	-0.00070
	(0.00121)	(0.00139)	(0.00163)	(0.00116)	(0.00119)	(0.00124)	(0.00100)	(0.00093)	(0.00086)	(0.00116)	(0.00119)	(0.00123)
Share of Observations with Mispelled Mother's Name	-0.00337	0.00107	0.00739	0.00040	-0.00442	-0.01360	-0.00329	0.00390	-0.00039	-0.00182	-0.01314	-0.01926
	(0.01400)	(0.01581)	(0.01846)	(0.01341)	(0.01367)	(0.01419)	(0.01138)	(0.01054)	(0.00978)	(0.01348)	(0.01368)	(0.01416)
Share of Observations with Mispelled Fatherr's Name	0.00770	0.01949	-0.00764	0.01413	0.02043	0.00085	0.02389**	0.03289***	0.02479**	0.02178	0.02698**	0.00794
	(0.01408)	(0.01558)	(0.01806)	(0.01336)	(0.01349)	(0.01384)	(0.01132)	(0.01054)	(0.00978)	(0.01337)	(0.01351)	(0.01381)
Link in Ohio	-0.05742***	-0.02780***	0.00398	-0.06504***	-0.01606**	0.02525***	-0.12005***	-0.04389***	0.01399**	-0.05939***	-0.01000	0.03821***
	(0.00768)	(0.00887)	(0.01057)	(0.00732)	(0.00753)	(0.00787)	(0.00642)	(0.00594)	(0.00548)	(0.00736)	(0.00759)	(0.00786)
Constant	0.29414***	0.36260***	0.30140***	0.44120***	0.54717***	0.51349***	0.58758***	0.70985***	0.71066***	0.38517***	0.48643***	0.47841***
	(0.02953)	(0.03336)	(0.03931)	(0.02848)	(0.02903)	(0.03024)	(0.02457)	(0.02289)	(0.02126)	(0.02828)	(0.02908)	(0.02996)
Observations	14,058	12,162	9,282	19,383	19,817	18,740	28,411	33,084	36,767	17,534	18,144	17,837
R-squared	0.006	0.003	0.000	0.006	0.003	0.001	0.017	0.007	0.002	0.006	0.003	0.002
Wald Statistic	79.29	35.07	0.89	115.33	64.39	17.41	504.97	231.96	58.36	100.19	64.30	41.66
Prob > W	0.00	0.00	1.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in in the synthetic data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the process used to create the synthetic data) on the variables included in this table from the synthetic birth certificates. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

Appendix Table D2. Correlation of Incorrect Links with Baseline Sample Characteristics in Simulated Data - Continued

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Day of Birth	-0.00002	-0.00007***	-0.00006**	-0.00004	-0.00003
	(0.00004)	(0.00003)	(0.00002)	(0.00003)	(0.00003)
Number of Siblings	-0.00151	0.00400***	0.00412***	-0.00138	-0.00065
	(0.00138)	(0.00096)	(0.00092)	(0.00105)	(0.00093)
Length of Child's Name	-0.00090	-0.00787***	-0.00724***	-0.00390	-0.00131
	(0.00301)	(0.00220)	(0.00217)	(0.00248)	(0.00219)
Length of Father's Name	-0.00235*	-0.00165*	-0.00322***	-0.00261***	-0.00157*
	(0.00123)	(0.00086)	(0.00083)	(0.00094)	(0.00083)
Length of Mother's Name	-0.00346**	-0.00051	-0.00045	-0.00103	-0.00050
	(0.00136)	(0.00094)	(0.00092)	(0.00102)	(0.00092)
Share of Observations with	-0.01349	-0.00607	-0.00228	-0.00590	-0.00518
Mispelled Mother's Name	(0.01573)	(0.01077)	(0.01061)	(0.01182)	(0.01062)
Share of Observations with	0.02799*	0.01820*	0.01617	0.02176*	0.00771
Mispelled Fatherr's Name	(0.01542)	(0.01070)	(0.01056)	(0.01170)	(0.01054)
Link in Ohio	-0.00250	-0.10511***	-0.13810***	-0.05398***	-0.01349**
	(0.00881)	(0.00589)	(0.00575)	(0.00654)	(0.00591)
Constant	0.31719***	0.35215***	0.39818***	0.38260***	0.16472***
	(0.03262)	(0.02289)	(0.02267)	(0.02533)	(0.02316)
Observations	10,982	24,007	26,019	21,965	13,664
R-squared	0.002	0.018	0.029	0.005	0.001
Wald Statistic	24.28	448.00	801.98	114.31	17.44
Prob > W	0.00	0.00	0.00	0.00	0.03

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in in the synthetic data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the process used to create the synthetic data) on the variables included in this table from the synthetic birth certificates. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

Appendix Table D3. Correlation of Incorrect Links with Baseline Sample Characteristics in Early Indicators Data

	Ferrie 1996 (Name)	Ferrie 1996 (NYSIIS)	Ferrie 1996 (SDX)	Ferrie 1996 (Name) + common names	Ferrie 1996 (NYSIIS) + common names	Ferrie 1996 (SDX) + common names	Ferrie 1996 (Name) + common names + ties	Ferrie 1996 (NYSIIS) + common names + ties	Ferrie 1996 (SDX) + common names + ties	Abramitzky et al. 2014 (Name)	Abramitzky et al. 2014 (NYSIIS)	Abramitzky et al. 2014 (SDX)
Age	-0.02406**	-0.01888**	-0.01261	-0.03000***	-0.02286***	-0.02345***	-0.02589***	-0.01566**	-0.01972***	-0.03666***	-0.03021***	-0.03126***
	(0.00986)	(0.00937)	(0.01020)	(0.00768)	(0.00774)	(0.00849)	(0.00743)	(0.00651)	(0.00684)	(0.00899)	(0.00896)	(0.01015)
Is Currently Married	0.04963	0.04375	0.01276	0.02051	0.02601	-0.02884	-0.01566	0.03026	0.02061	0.03270	0.02962	-0.03012
	(0.04316)	(0.04352)	(0.04933)	(0.04395)	(0.04303)	(0.04737)	(0.04487)	(0.04137)	(0.04165)	(0.04389)	(0.04308)	(0.04672)
Foreign Born	0.06032	0.10728	-0.03451	-0.01553	0.04480	-0.08238	-0.02012	0.06143	0.00214	-0.06980	0.07432	-0.05057
	(0.07543)	(0.07464)	(0.08775)	(0.07263)	(0.07199)	(0.07669)	(0.06793)	(0.06163)	(0.05987)	(0.07214)	(0.07130)	(0.07679)
Day of Year Birth	-0.00016	-0.00017	-0.00013	-0.00026**	-0.00020	-0.00018	-0.00021	-0.00014	-0.00013	-0.00025*	-0.00026*	-0.00017
	(0.00013)	(0.00014)	(0.00015)	(0.00013)	(0.00013)	(0.00014)	(0.00013)	(0.00012)	(0.00012)	(0.00013)	(0.00013)	(0.00013)
Literate	-0.13760	-0.15172	-0.13014	-0.23181**	-0.26148***	-0.18478**	-0.27362***	-0.19258**	-0.20365***	-0.35098***	-0.16620*	-0.12943
	(0.10451)	(0.10520)	(0.11495)	(0.09685)	(0.08780)	(0.09060)	(0.08499)	(0.07605)	(0.07436)	(0.10413)	(0.09406)	(0.09568)
Length of First Name	-0.02273*	-0.01840	-0.01954	-0.01543	-0.02037*	-0.02160*	-0.01908*	-0.03169***	-0.03565***	-0.01014	-0.02026*	-0.02636**
	(0.01160)	(0.01200)	(0.01275)	(0.01117)	(0.01117)	(0.01153)	(0.01070)	(0.01014)	(0.01013)	(0.01126)	(0.01161)	(0.01073)
Length of Second Name	-0.00265	-0.01274	0.00464	-0.01101	-0.02055**	-0.00355	-0.01893**	-0.04999***	-0.01897**	-0.01421*	-0.02670***	0.00100
	(0.00858)	(0.00920)	(0.00998)	(0.00838)	(0.00871)	(0.00937)	(0.00848)	(0.00836)	(0.00843)	(0.00835)	(0.00918)	(0.00941)
Mother is Foreign Born	0.07438	0.00982	0.10534	0.09580	0.03357	0.16047**	0.06755	0.00330	0.08255	0.12747**	-0.01508	0.11756
	(0.06659)	(0.06667)	(0.07761)	(0.06417)	(0.06749)	(0.06921)	(0.06271)	(0.05973)	(0.05801)	(0.06171)	(0.06839)	(0.07197)
Father is Foreign Born	-0.04429	-0.03511	-0.04827	0.00217	0.00654	-0.01345	0.05509	0.09257*	0.04206	0.01360	0.01269	-0.00599
	(0.05147)	(0.05440)	(0.06286)	(0.05303)	(0.05678)	(0.06036)	(0.05559)	(0.05360)	(0.05348)	(0.04934)	(0.05700)	(0.06297)
Year of Birth	-0.01402	-0.01179	-0.00441	-0.02182***	-0.01742**	-0.01672*	-0.01595**	-0.00653	-0.01120	-0.03049***	-0.02567***	-0.02574**
	(0.01008)	(0.00984)	(0.01063)	(0.00785)	(0.00806)	(0.00875)	(0.00762)	(0.00679)	(0.00713)	(0.00907)	(0.00930)	(0.01031)
Constant	27.69950	23.36568	9.32643	42.56651***	34.16765**	32.84323**	31.71714**	13.96893	22.72117*	58.96482***	49.70969***	49.80274**
	(19.10638)	(18.61725)	(20.13100)	(14.87561)	(15.25372)	(16.57689)	(14.43792)	(12.85489)	(13.50516)	(17.19549)	(17.60716)	(19.54841)
Observations	830	795	723	933	965	959	1,104	1,270	1,369	793	861	907
R-squared	0.049	0.038	0.026	0.055	0.046	0.038	0.053	0.064	0.045	0.084	0.057	0.047
Wald Statistic	45.83	26.21	17.51	48.61	39.16	32.36	65.68	93.17	66.98	54.28	39.37	28.56
Prob > W	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in in the Early Indicators Union Army data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the Early Indicators links to the 1900 Census) on the variables included in this table from the synthetic birth certificates. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

Appendix Table D3. Correlation of Incorrect Links with Baseline Sample Characteristics in Early Indicators Data

	Abramitzky et al. 2014 (NYSIIS, Robustness)	Feigenbaum 2016 (Iowa coef.)	Feigenbaum 2016 (estimated coef.)	Abramitzky et al. 2018 (Less conservative)	Abramitzky et al. 2018 (More conservative)
Age	-0.02982*	-0.02935***	-0.02631**	-0.03306**	-0.01782
	(0.01568)	(0.01125)	(0.01229)	(0.01282)	(0.01260)
Is Currently Married	0.04494	0.02588	-0.01816	0.00543	-0.04059
	(0.04031)	(0.03611)	(0.04482)	(0.03908)	(0.04299)
Foreign Born	0.21013***	-0.04328	-0.00744	-0.05275	-0.13589**
	(0.07168)	(0.06065)	(0.07221)	(0.06392)	(0.06617)
Day of Year Birth	-0.00012	-0.00033***	-0.00036***	-0.00026**	-0.00025**
	(0.00015)	(0.00011)	(0.00013)	(0.00012)	(0.00012)
Literate	-0.07274	-0.19962**	-0.06650	-0.19242**	-0.10671
	(0.11734)	(0.08840)	(0.08615)	(0.09753)	(0.11271)
Length of First Name	-0.01651	-0.01383	-0.01109	-0.01058	-0.00030
	(0.01208)	(0.00906)	(0.01006)	(0.00901)	(0.00828)
Length of Second Name	-0.01143	-0.01580**	-0.00884	-0.00183	-0.00784
	(0.00960)	(0.00680)	(0.00785)	(0.00767)	(0.00601)
Mother is Foreign Born	-0.04362	0.11321**	0.10422*	0.03021	0.11996*
	(0.06510)	(0.05463)	(0.06186)	(0.05652)	(0.06149)
Father is Foreign Born	-0.03773	-0.03280	-0.02407	0.06053	0.03995
	(0.05359)	(0.04556)	(0.05053)	(0.05315)	(0.04816)
Year of Birth	-0.02279	-0.02401**	-0.02273*	-0.02481*	-0.01391
	(0.01598)	(0.01150)	(0.01251)	(0.01306)	(0.01274)
Constant	44.06990	46.50573**	43.78841*	48.11923*	26.96099
	(30.32669)	(21.80944)	(23.73681)	(24.77940)	(24.18086)
Observations	601	1,063	757	1,024	686
R-squared	0.070	0.050	0.042	0.054	0.048
Wald Statistic	32.39	38.88	19.37	43.02	18.27
Prob > W	0.00	0.00	0.04	0.00	0.05

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Robust standard errors in parentheses. Regressions are a comparison of correctly and incorrectly matched observations in in the Early Indicators Union Army data by regressing a dummy variable indicating whether or not an observation was correctly matched (as determined by the Early Indicators links to the 1900 Census) on the variables included in this table from the synthetic birth certificates. The Wald statistic tests whether being linked incorrectly is systematically and jointly related to covariates.

E. Using Propensity Score Matching to Reweight Estimates to Resemble the 1940 Census Samples

We reweight the IGE estimates two ways. First, we reweight the linked sample to look like the random sample of male birth certificates. Second, we reweight the linked sample to look like a stratified sample of the 1940 Census. We chose a *stratified* sample of the 1940 Census to account for the distribution of ages in the birth certificate sample (see Appendix Figure B1) in order to represent the potentially linkable cases from the 1940 Census. To construct the stratified random sample of the 1940 Census, we sampled Ohio- and North Carolina-born cohorts in the 1940 Census such that the year of birth distribution matched that of the birth certificates.

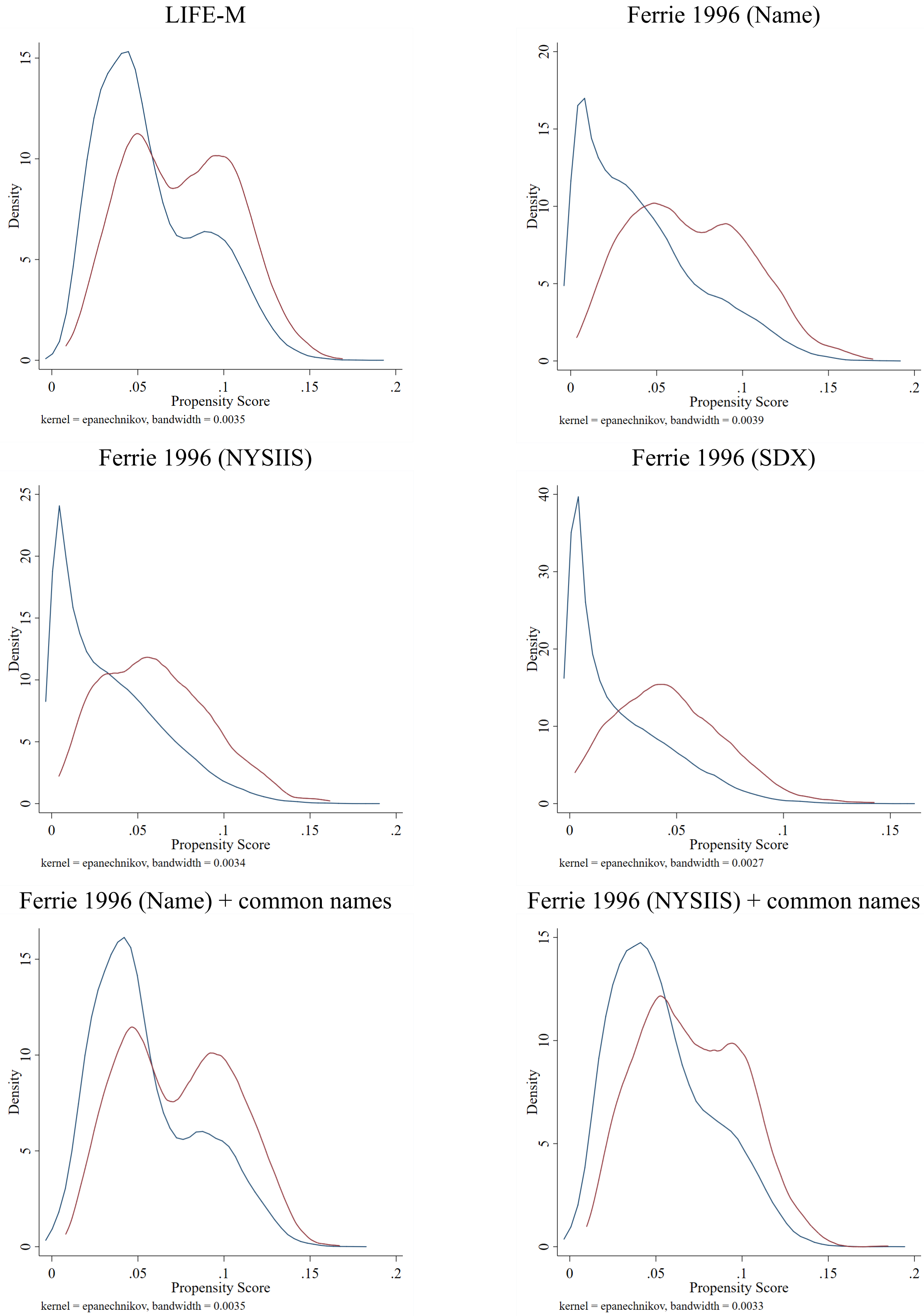
We construct weights using inverse propensity score reweighting (DiNardo et al. 1996). First, we append on data that we use for reweighting and then construct a binary indicator that takes on value one if the data come from our linked sample. We construct weights using probit regressions of whether an observation was linked on characteristics in either the full sample of male birth certificates or the stratified 1940 Census sample. When reweighting for the birth certificates, these characteristics include day of year, age, first and last name commonness indexes, a dummy variable for presence of siblings, polynomials in the number of siblings, polynomials in the length of child, mother, and father name, and state fixed effects. When reweighting to match the 1940 Census, these characteristics include polynomials in a first and last name commonness index, the interaction of the commonness index for first and last name, state fixed effects, cohort fixed effects, polynomials in age, and race-cohort fixed effects. Specifically, we predict the propensity of being linked ($P_i(L_i = 1|\boldsymbol{X_i})$), which we then use to reweight the matches cases by $W_i = (1 - P_i(L_i = 1|\boldsymbol{X_i}))/P_i(L_i = 1|\boldsymbol{X_i}) * q/(1-q)$, where q is the share of records that are linked. We use this linking strategy as our stratified random sample of the 1940 Census does not have a proper subset of all the birth certificates that we linked. In the LIFE-M data, as the birth certificates we linked are necessarily a subset of the birth certificates that we attempted to link, we could also reweight the data using simple inverse-propensity score weights, $1/P(L_i = 1|\boldsymbol{X_i})$**.**

We present the distribution of estimated propensity scores in Appendix Figure E1 (next pages).

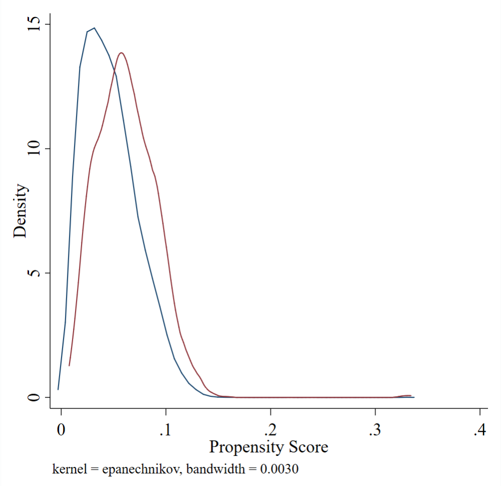
Moreover, we present the second set of IGE estimates reweighted to the stratified sample of the 1940 Census in Appendix Figure E2 (following Appendix Figure E1) and separate regressions for the correct versus incorrect links in Appendix Figure E3.

Appendix Figure E1. Distributions of Inverse Propensity Score Weights

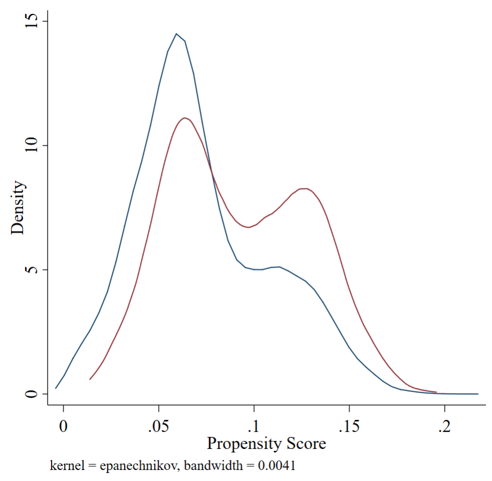
Panel A: Inverse Propensity Score Weights to Birth Certificates



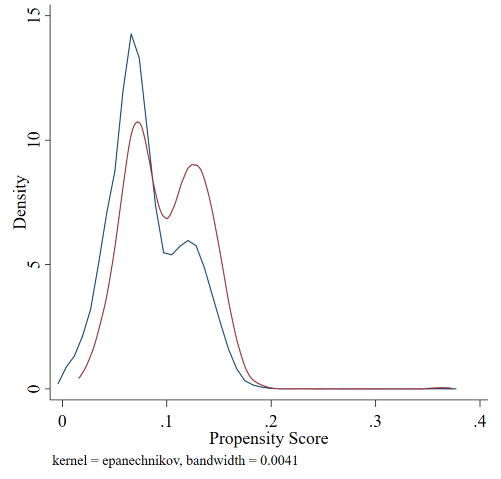
Ferrie 1996 (SDX) + common names



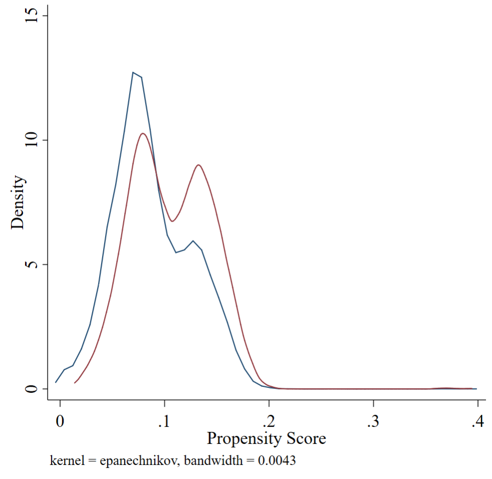
Ferrie 1996 (Name) + common names + ties



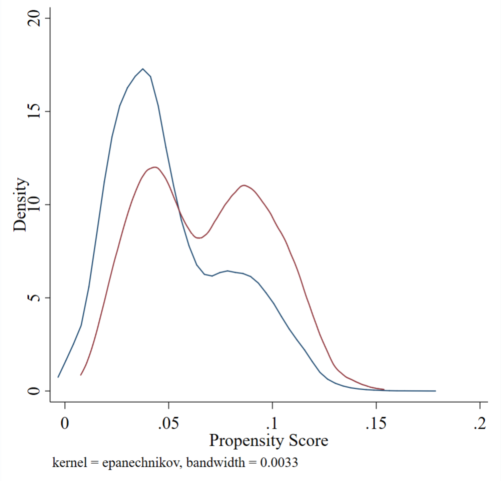
Ferrie 1996 (NYSIIS) + common names + ties



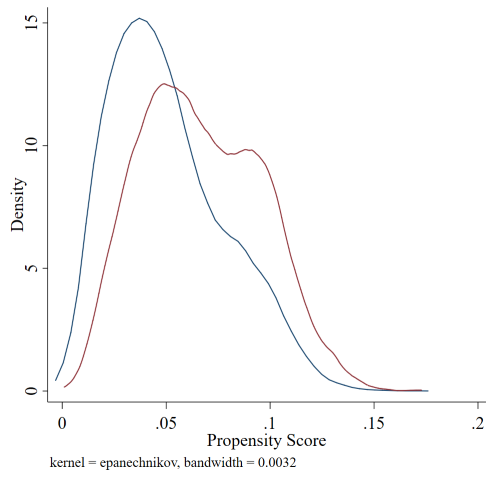
Ferrie 1996 (SDX) + common names + ties

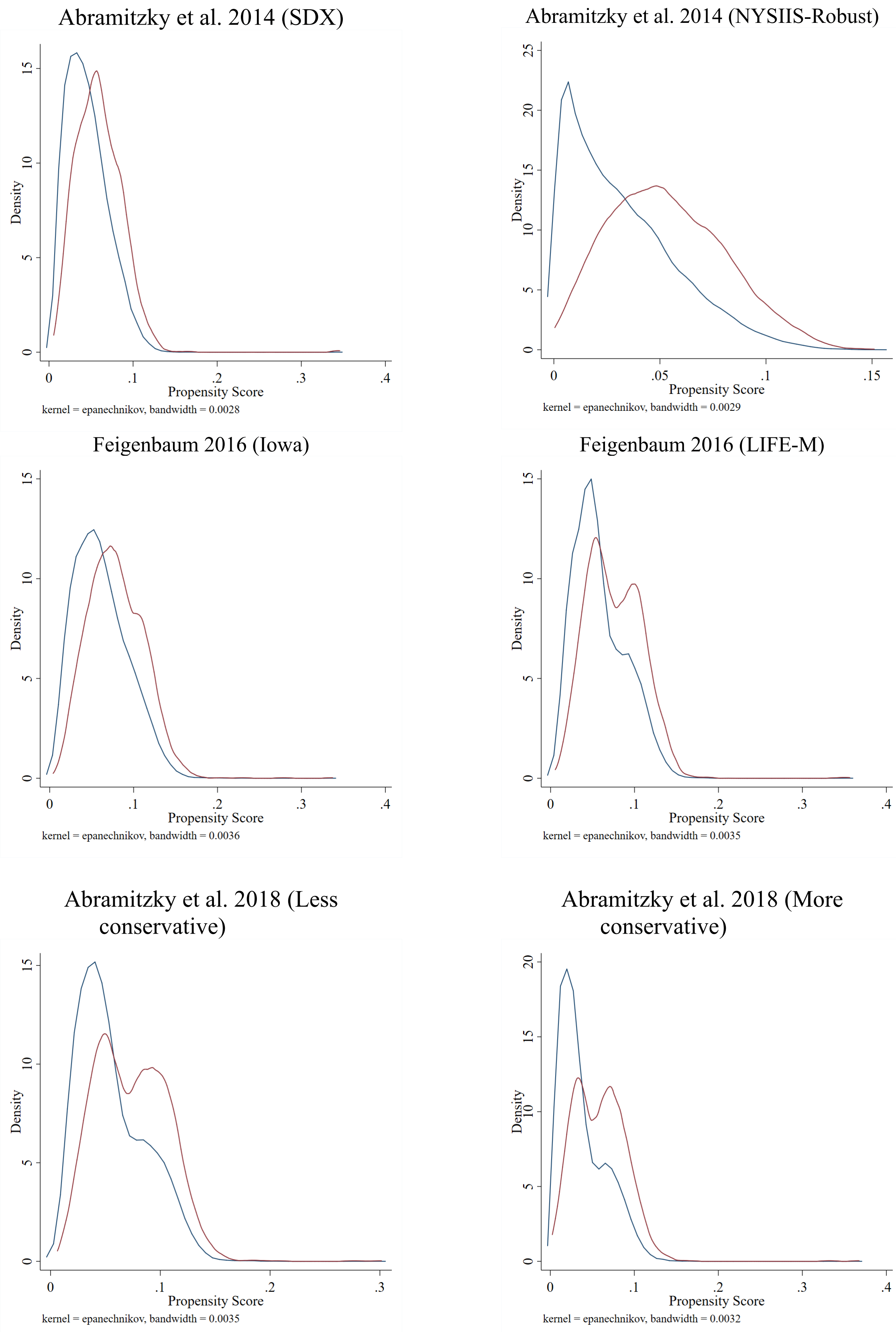


Abramitzky et al. 2014 (Name)



Abramitzky et al. 2014 (NYSIIS)





Abramitzky et al. 2014 (SDX)

Abramitzky et al. 2014 (NYSIIS-Robust)

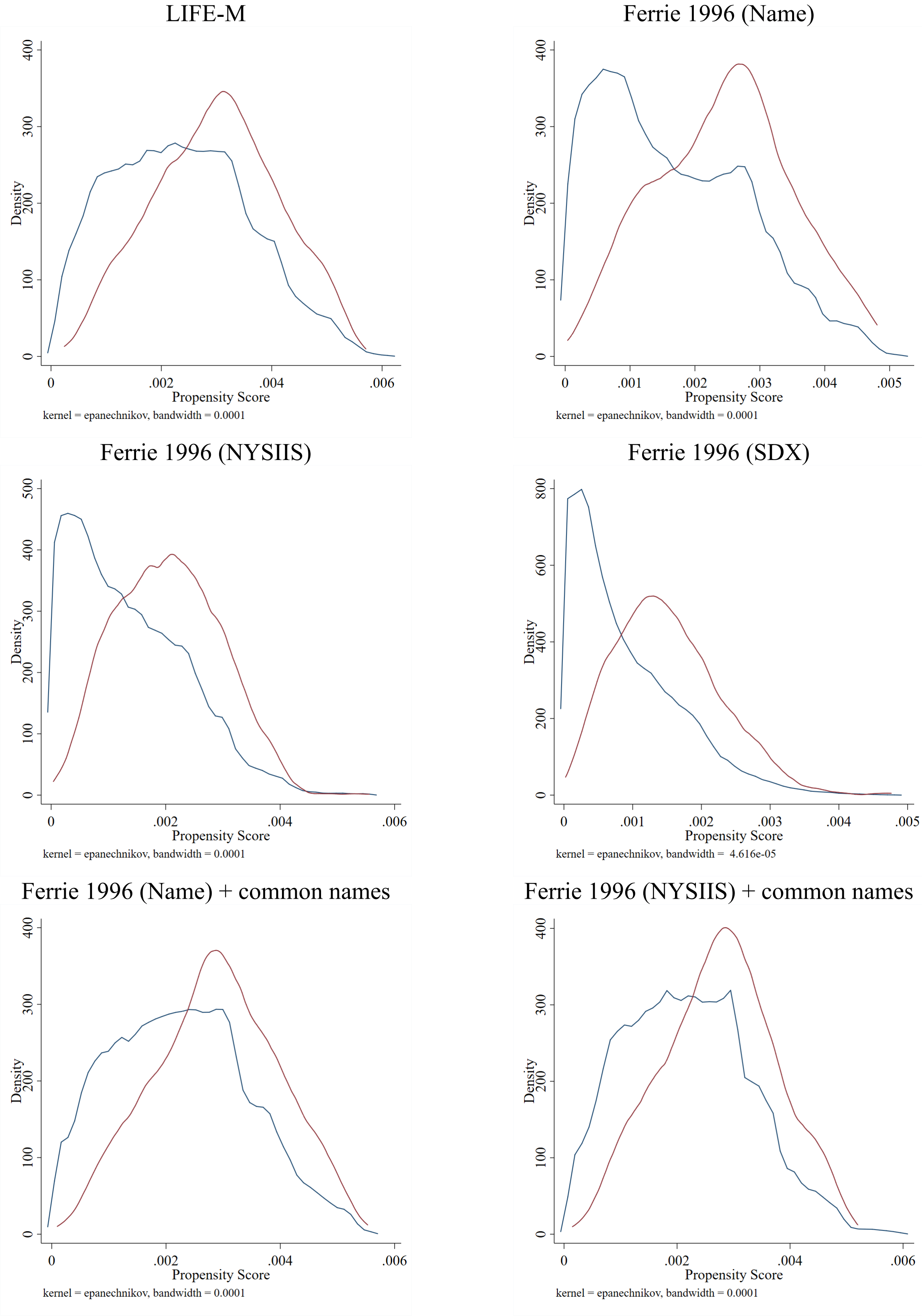
Feigenbaum 2016 (Iowa)

Feigenbaum 2016 (LIFE-M)

Abramitzky et al. 2018 (Less conservative)

Abramitzky et al. 2018 (More conservative)

Panel B: Inverse Propensity Score Weights to the 1940 Census



LIFE-M

kernel = epanechnikov, bandwidth = 0.0001

Ferrie 1996 (Name)

kernel = epanechnikov, bandwidth = 0.0001

Ferrie 1996 (NYSIIS)

kernel = epanechnikov, bandwidth = 0.0001

Ferrie 1996 (SDX)

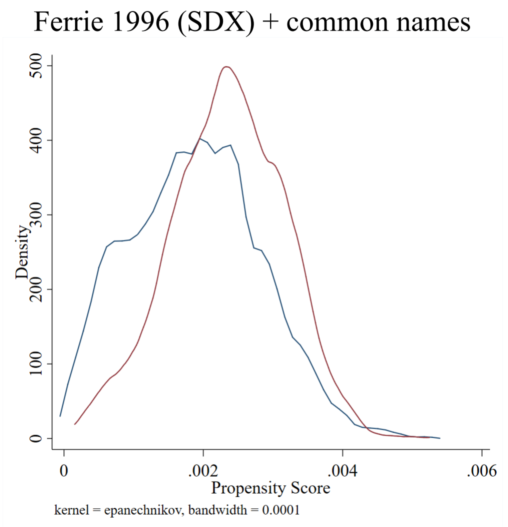
kernel = epanechnikov, bandwidth = 4.616e-05

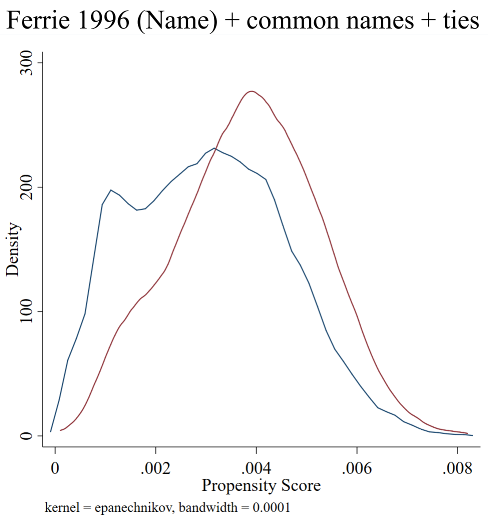
Ferrie 1996 (Name) + common names

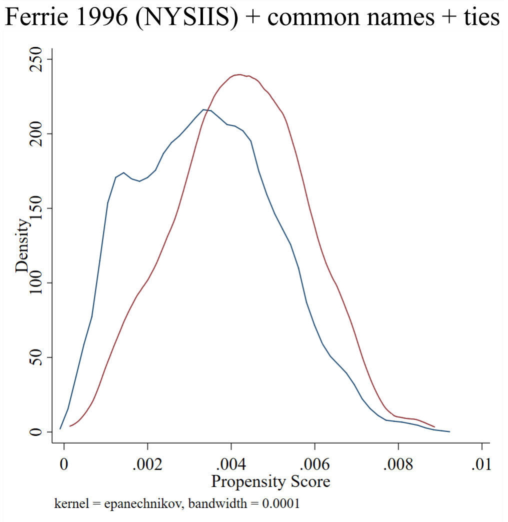
kernel = epanechnikov, bandwidth = 0.0001

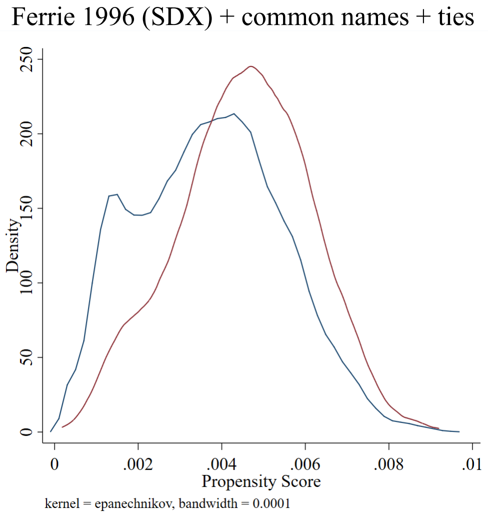
Ferrie 1996 (NYSIIS) + common names

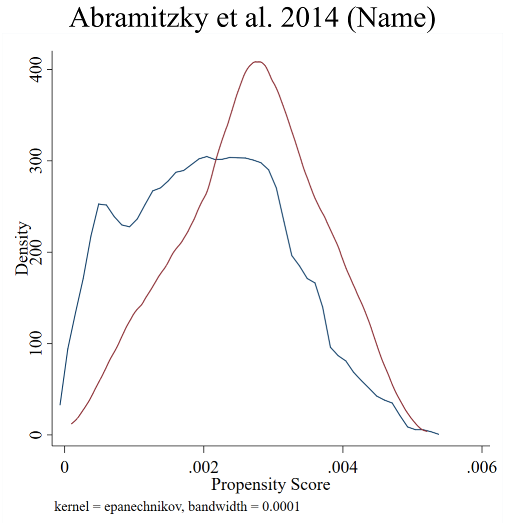
kernel = epanechnikov, bandwidth = 0.0001

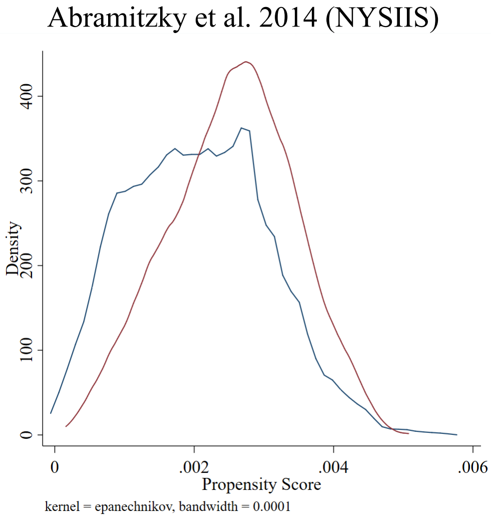

Ferrie 1996 (SDX) + common names
Density
0
100
200
300
400
500
0
.002
.004
.006
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001


Ferrie 1996 (Name) + common names + ties
Density
0
100
200
300
0
.002
.004
.006
.008
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001

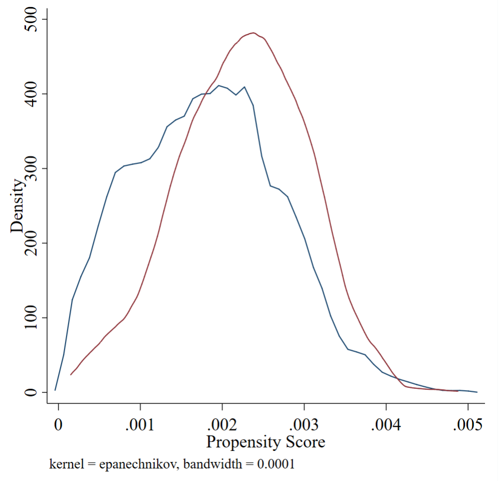

Ferrie 1996 (NYSIIS) + common names + ties
Density
0
50
100
150
200
250
0
.002
.004
.006
.008
.01
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001


Ferrie 1996 (SDX) + common names + ties
Density
0
50
100
150
200
250
0
.002
.004
.006
.008
.01
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001

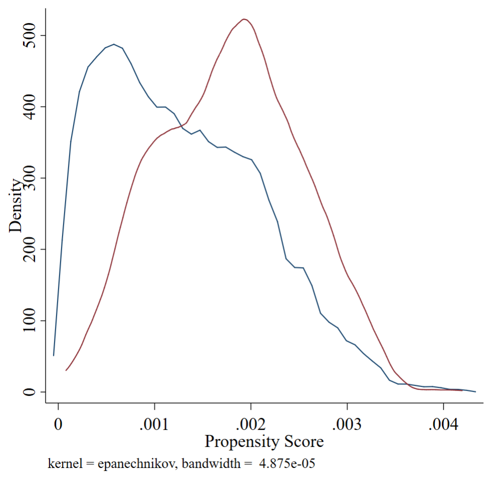

Abramitzky et al. 2014 (Name)
Density
0
100
200
300
400
0
.002
.004
.006
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001


Abramitzky et al. 2014 (NYSIIS)
Density
0
100
200
300
400
0
.002
.004
.006
Propensity Score
kernel = epanechnikov, bandwidth = 0.0001

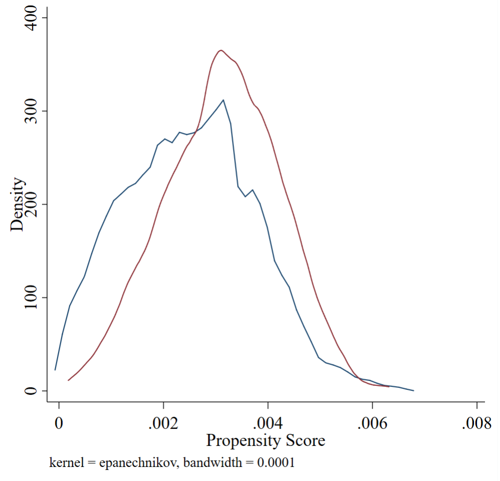
Abramitzky et al. 2014 (SDX)



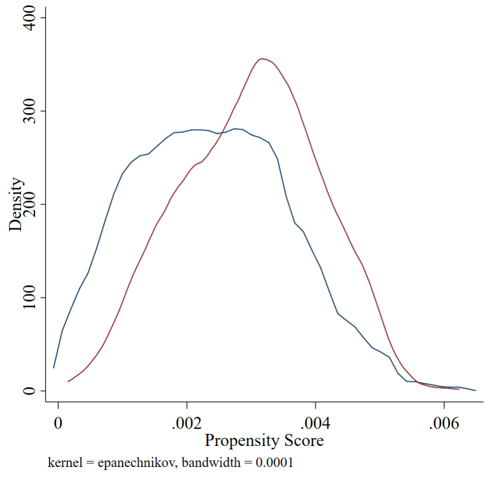
Abramitzky et al. 2014 (NYSIIS-Robust)



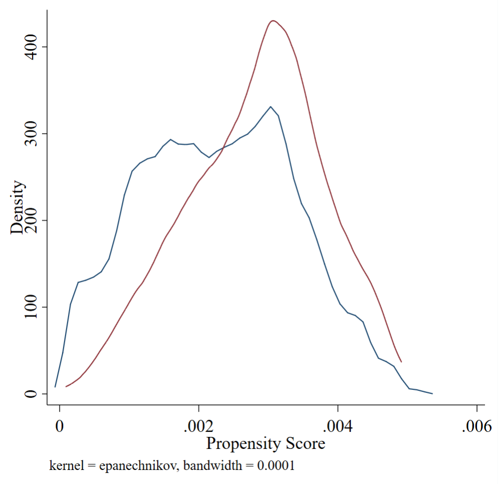
Feigenbaum 2016 (Iowa)



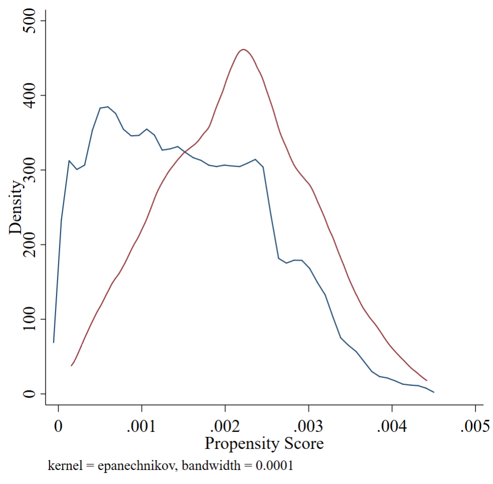
Feigenbaum 2016 (LIFE-M)



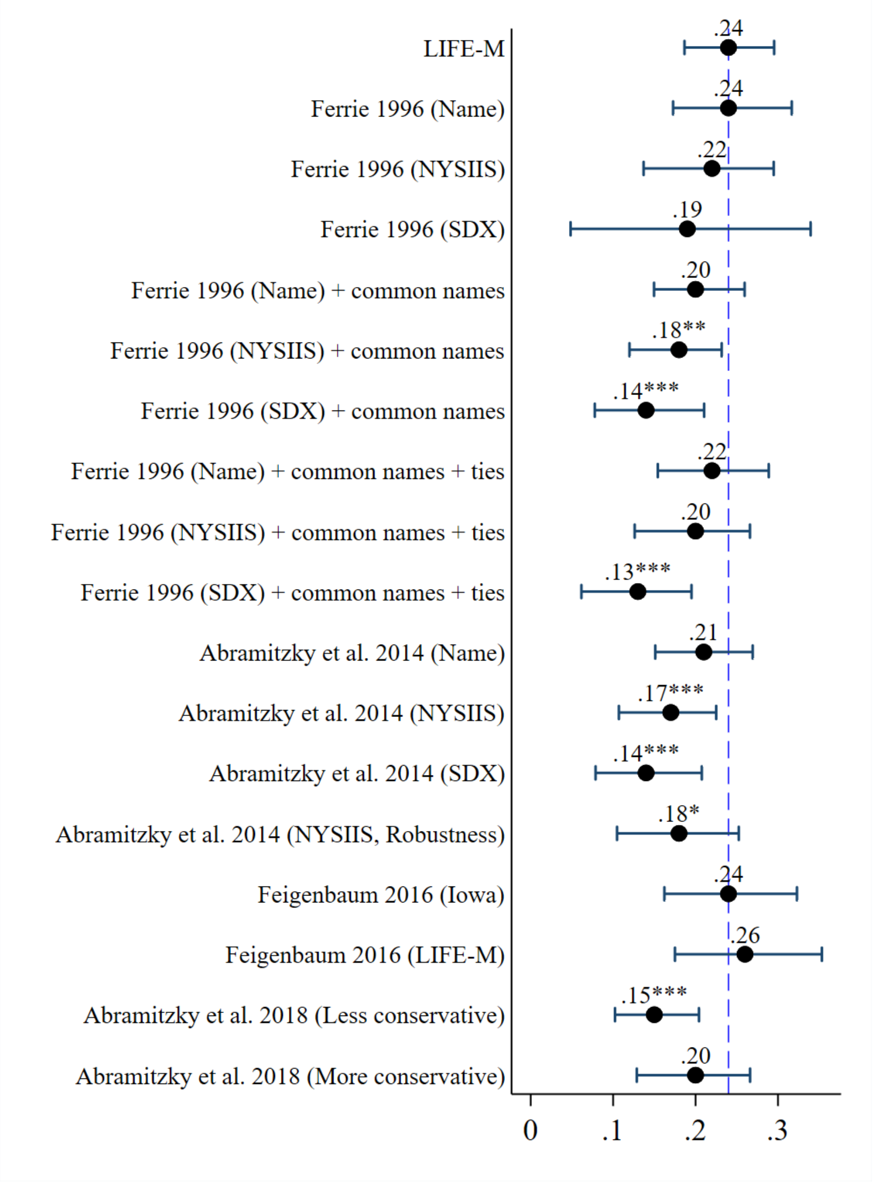
Abramitzky et al. 2018 (Less conservative)



Abramitzky et al. 2018 (More conservative)

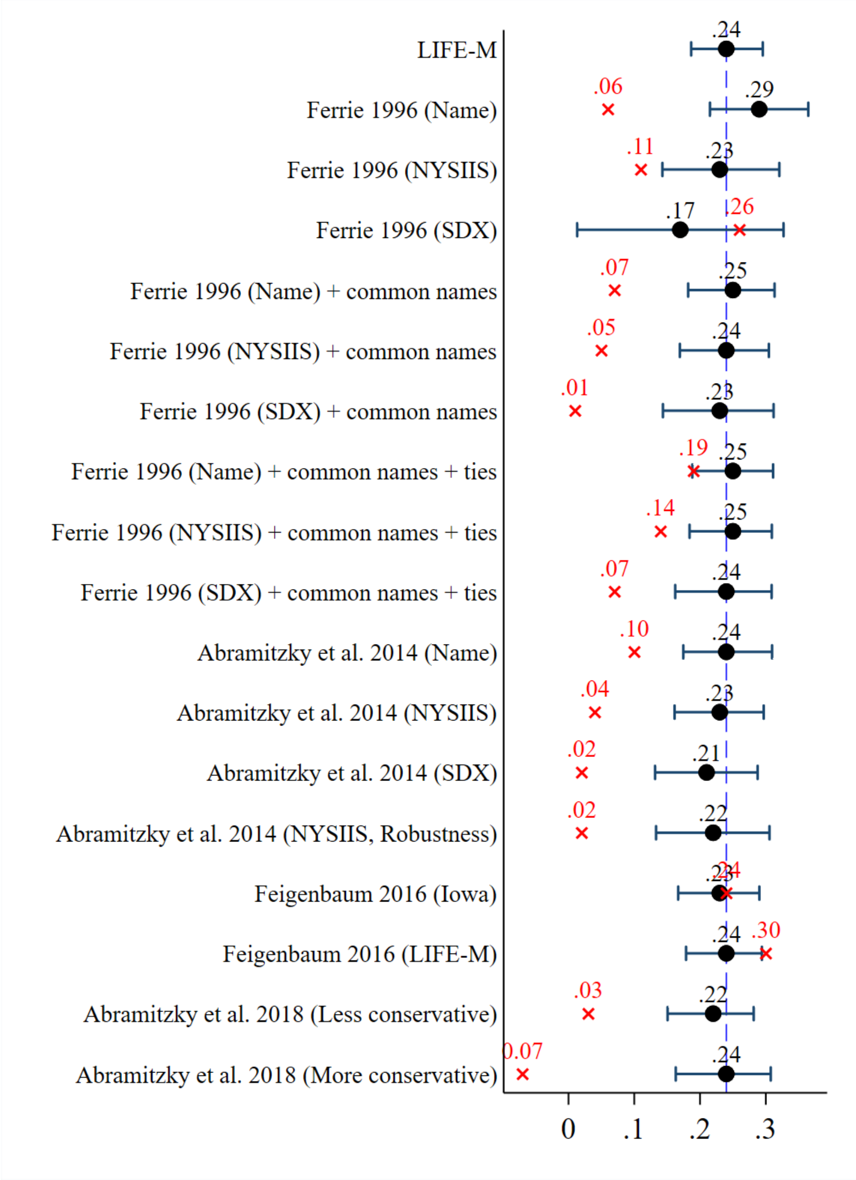


Appendix Figure E2. Intergenerational Income Elasticities Reweighted to Resemble the Characteristics a Cohort- Stratified Sample of the 1940 Census



Notes: See figure notes in the paper for sample sizes. These estimates were reweighted to represent a stratified random sample of the 1940 Census. Weighting variables include a first and last name commonness index and polynomials in this index, the interaction of the commonness index for first and last name, birth state fixed effects, cohort fixed effects, polynomials in age, and race-cohort fixed effects.

Appendix Figure E3. Separate Regressions for Imputed and Correct Links Reweighted to Resemble the Characteristics a Cohort- Stratified Sample of the 1940 Census



Notes: See figure notes in the paper for sample sizes. These estimates were reweighted to represent a stratified random sample of the 1940 Census. For variables used in reweighting, see notes to Appendix Figure E2.