Fatherless: The Long-Term Effects of Losing a Father in the U.S. Civil War
Online Appendix

Yannick Dupraz
Aix-Marseille University, CNRS, AMSE

Andreas Ferrara
University of Pittsburgh

January 7, 2023

Table of Content

List of Figures

List of Tables

A Appendix Tables

Table A.1: List of Sources for the Union Soldier Data

- **Connecticut**: Barbour, L.A., Camp, F.E., Smith, S.R., and White, G.M. (1889) "Record of Service of Connecticut Men in the Army and Navy of the United States During the War of the Rebellion", Case, Lockwood, & Brainard Company, Hartford, CT
- **Illinois**: Reece, J.N. (1900) "Report of the Adjutant General of the State of Illinois", Vols. 1-9, Philips Bros. State Printers, Springfield, IL
- **Indiana**: Terrell, W.H.H. (1866) "Report of the Adjutant General of the State of Indiana", Vols. 1-5, Samuel M. Douglass State Printers, Indianapolis, IN
- **Iowa**: Thrift, W.H. (1908) "Roster and Record of Iowa Soldiers in the War of Rebellion", Vol. 1-6, Emory H. English State Printers, Des Moines, IA
- **Kansas**: Fox, S.M. (1896) "Report of the Adjutant General of the State of Kansas", The Kansas State Printing Company, Topeka, KS
- **Maine**: Adjutant General (1861-66) "Supplement to the Annual Reports of the Adjutant General of the State of Maine", Stevens & Sayward State Printers, Augusta, ME
- **Massachusetts**: Schouler, W. (1866) "Report of the Adjutant General of the Commonwealth of Massachusetts", Wright & Potter State Printers, Boston, MA
- **Michigan**: Crapo, H.H. (1862-66) "Report of the Adjutant General of the State of Michigan", John A. Kerr & Co. State Printers, Lansing, MI
- **Minnesota**: Marshall, W.R. (1861-66) "Report of the Adjutant General of the State of Minnesota", Pioneer Printing Company, Saint Paul, MN
- **New Hampshire**: Head, N. (1865) "Report of the Adjutant General of the State of New Hampshire", Vols. 1& 2, Amos Hadley State Printers, Concord, NH
- **New Jersey**: Stryker, W.S. (1874) "Report of the Adjutant General of the State of New Jersey", Wm. S. Sharp Steam Power Book and Job Printers, Trenton, NJ
- **New York**: Sprague, J.T. (1864-68) "A Record of the Commissioned Officers, Non-Commissioned Officers and Privates of the Regiments which were Organized in the State of New York into the Service of the United States to Assist in Suppressing the Rebellion", Vols. 1-8, Comstock & Cassidy Printers, Albany, NY
- **Ohio**: Howe, J.C., McKinley, W., and Taylor, S.M. (1893) "Official Rosters of the Soldiers of the State of Ohio in the War of the Rebellion 1861-65", Vols. 1-12, The Werner Company, Akron, OH
- **Pennsylvania**: Russell, A.L. (1866) "Report of the Adjutant General of Pennsylvania", Singerly & Myers State Printers, Harrisburg, PA
- **Rhode Island**: Dyer, E. (1893-95) "Annual report of the Adjutant General of the state of Rhode Island and Providence Plantations", Vols. 1-2, E.L. Freeman Publishing, Providence, RI
- **Vermont**: Peck, T.S. (1892) "Revised Roster of Vermont Volunteers and Lists of Vermonters who Served in the Army and Navy of the United States during the War of the Rebellion 1861-66", Press of the Watchman Publishing Co., Montpelier, VT
- **Wisconsin**: Rusk, J.M. and Chapman, C.P. (1886) "Roster of Wisconsin Volunteers, War of the Rebellion 1861-65", Democrat Printing Company, Madison, WI

Table A.2: Military Records Summary Statistics

	Obs.	Mean	St. Dev.	Min	Max
Age at enlistment	1,129,902	25.425	7.367	11	70
Date of enlistment	2,592,682	Jan 16 1863		Jun 10 1801	Jul 22 1869
Birthyear known	2,739,719	0.412	0.492	0	1
Reason for joining					
Enlisted	2,697,272	0.940	0.238	0	1
Commissioned	2,697,272	0.030	0.171	0	1
Drafted	2,697,272	0.016	0.124	0	1
Substitute	2,697,272	0.014	0.119	0	1
Rank (at enlistment)					
Private	2,739,719	0.840	0.366	0	1
Corporal	2,739,719	0.055	0.228	0	1
Sergeant	2,739,719	0.043	0.202	0	1
Low-ranking officer	2,739,719	0.025	0.156	0	1
High-ranking officer	2,739,719	0.002	0.045	0	1
Musician	2,739,719	0.014	0.116	0	1
Other	2,739,719	0.010	0.101	0	1
Unit type (at enlistment)					
Infantry	2,739,719	0.741	0.438	0	1
Cavalry	2,739,719	0.159	0.366	0	1
Artillery	2,739,719	0.076	0.265	0	1
Special (fighting)	2,739,719	0.003	0.051	0	1
Special (non-fighting)	2,739,719	0.006	0.076	0	1
Casualties					
Died	2,186,785	0.125	0.331	0	1
Died (combat)	2,186,785	0.045	0.207	0	1
Died (disease)	2,186,785	0.049	0.216	0	1
Died (other)	2,186,785	0.031	0.173	0	1
Disabled	2,160,457	0.095	0.293	0	1
Injured	2,739,719	0.060	0.237	0	1

Note: Summary statistics for the 2.7 million Union Army Military Records. The number of soldier is 2.2 million Union Army soldiers but the number of records is larger due to re-enlistments and transfers across units. Substitutes are those who replaced a drafted man for payment. Low-ranking officers are lieutenants and captains, high-ranking officers are majors, lieutenant colonels, and colonels. Other ranks include cooks, wagoners, and other support occupations. Specialized fighting units are sharpshooters and specialized non-fighting units are staff units, for example. Other deaths include accidents, suicides, or natural causes.

Table A.3: OLS Robustness of Results to Alternative Measures of Occupational Income

	Panel a: all sons						
	(1) log IPUMS 1950 occ. score	(2) IPUMS 1950 occ.score ($)	(3) percentile IPUMS 1950 occ. score	(4) log Iowa 1915 occ. score	(5) log P&H 1900 occ. score	(6) log 1870 wealth score (median)	(7) log 1870 wealth score (mean)
Father died	-0.022***	-44.352***	-1.454***	-0.013	-0.015**	-0.016	-0.024*
	(0.007)	(16.407)	(0.486)	(0.009)	(0.007)	(0.022)	(0.014)
Mean dep. var.	2.906	1868.175	44.709	2.543	6.112	1.032	2.500
Observations	27,081	29,269	29,269	27,279	27,438	16,092	27,080
	Panel b: excluding sons of farmers						
	(1) log IPUMS 1950 occ. score	(2) IPUMS 1950 occ.score ($)	(3) percentile IPUMS 1950 occ. score	(4) log Iowa 1915 occ. score	(5) log P&H 1900 occ. score	(6) log 1870 wealth score (median)	(7) log 1870 wealth score (mean)
Father died	-0.022**	-61.471***	-2.025***	-0.021*	-0.016*	-0.014	-0.023
	(0.010)	(22.914)	(0.672)	(0.012)	(0.009)	(0.027)	(0.019)
Mean dep. var.	2.986	2023.448	50.569	2.552	6.193	1.087	2.627
Observations	16,824	18,091	18,091	16,821	17,064	11,409	16,824

Note: Regression of sons' occupational income in 1880 on an indicator for whether their father died in the U.S. Civil War. Column (1) reproduces the result of Table 2, panel a, column (1) as a benchmark. Other columns explore robustness to alternative measures of occupational income. Column (2) considers IPUMS 1950 occupational income in $ rather than in logs. Column (3) considers the percentile in the IPUMS 1950 occupational income score distribution. Column (4) considers the occupational income score built by Feigenbaum (2018) using the 1915 Iowa population census. Column (5) considers the occupational income score built by Olivetti and Paserman (2015) using the 1900 occupational earnings distribution obtained from the tabulations in Preston and Haines (1991) (farmers are assigned the average income of occupations in the 1910 census that were coded as farmers in the 1950 occupational classification). Columns (6) and (7) consider occupational wealth scores based on 1870 census data: we assign each occupation the median (column 6) or average (column 7) wealth (sum of real estate and personal property) of this occupation in the full count 1870 census. Following Olivetti and Paserman (2015), we adjust farmers' personal property downward by the average value of farm equipment and livestock in the 1870 census of agriculture and we adjust real estate property by subtracting the average cash value of farms in the 1870 census of agriculture. Controls are the same as in Table 2. Panel b reproduces the results of panel a excluding from the sample the sons of farmers, who are more likely to be farmers themselves. Standard errors are clustered by the father's last regiment of service and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: OLS Robustness of Results to Different Standard Error Clustering

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.022***	-0.003	-0.020***	0.007	0.016**	-0.007	0.016**
	(0.007)	(0.005)	(0.008)	(0.008)	(0.007)	(0.009)	(0.007)
s.e. clustered by:							
Father id	0.00818	0.00529	0.00827	0.00838	0.00755	0.00970	0.00768
1860 county	0.00809	0.00503	0.00798	0.00800	0.00751	0.00938	0.00692
Conley s.e. (50km)	0.00791	0.00514	0.00777	0.00785	0.00745	0.00917	0.00676
Conley s.e. (100km)	0.00784	0.00554	0.00788	0.00794	0.00755	0.00925	0.00675

Note: Regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors in the main specification are clustered by the father's last regiment of service and are reported in parentheses. The lower panel reports standard errors using alternative clustering variables and methods. The spatial autocorrelation robust standard errors by Conley (1999) were estimated with a 50 and 100km distance cutoff. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: OLS Results Robustness to Various Fixed Effects

	(1) log income score	(2) log income score	(3) log income score	(4) log income score	(5) log income score	(6) log income score	(7) log income score
Father died	-0.022***	-0.022**	-0.023**	-0.019**	-0.017*	-0.021**	-0.019**
	(0.007)	(0.010)	(0.010)	(0.008)	(0.010)	(0.010)	(0.008)
County F.E.	✓			✓	✓	✓	✓
Town F.E.		✓					
Post office F.E.			✓				
Regiment F.E.				✓			
Company F.E.					✓		
Last name F.E.						✓	
First name F.E.							✓
Observations	27,081	27,081	26,614	26,842	22,618	23,314	26,721
R^2	0.17	0.45	0.40	0.24	0.46	0.34	0.19
PPS test							
χ^2		0.01	0.05	0.68	0.28	0.01	1.27
p-value		0.92	0.82	0.41	0.60	0.94	0.26

Note: Regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War. Columns (2)-(7) estimate the model jointly with the baseline model reported in column (1) in a seemingly unrelated regression framework and test for differences in the effect of father death across each pair of models. The two bottom lines of the table report the χ^2 statistic and associated p-value of the coefficient comparison test developed by Pei, Pischke and Schwandt (2018). The effect of father death in the augmented models (with additional dimensions of fixed effects) is never statistically different from the effect in the baseline model. The number of the respective fixed effects is as follows: there are 9,534 different townships/wards, 8,051 different post office areas, 2,413 regiments and 12,992 companies, 8,925 different last names and 1,019 different first names. All models include characteristics of sons in 1880 and 1860 baseline characteristics of fathers and mothers. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: OLS Robustness to Double ML Covariate Selection

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.022***	-0.004	-0.019**	0.008	0.018**	-0.007	0.020***
	(0.007)	(0.005)	(0.008)	(0.008)	(0.007)	(0.009)	(0.007)
Son controls	✓	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓	✓
Father other controls	✓	✓	✓	✓	✓	✓	✓
Mothercontrols	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓
Observations	27,081	29,269	29,269	29,269	29,269	29,269	28,590

Note: Regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using the post-double selection (PDS) machine learning algorithm by Belloni, Chernozhukov and Hansen (2014). The PDS algorithm takes all controls, their squares, and cross-term interactions and selects the union of significant predictors of the treatment and the outcome and then runs the original regression with the set of selected controls in either step. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: OLS Robustness to Different Linking Techniques

	(1) baseline results	(2) excluding multiple links in 5 year window	(3) Ferrie rare names	(4) Only nonmissing birthyear	(5) Large sample size linking
	Dep. var.: log income score				
Father died	-0.022***	-0.024***	-0.028**	-0.038***	-0.015**
	(0.007)	(0.009)	(0.011)	(0.010)	(0.006)
Observations	27,081	21,042	13,637	13,166	45,547

Note: Regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using different linkage methods: Column (2): we exclude all links that are not unique in a 5 year window instead of 2. Column (3): we consider only individuals whose combination of first and last names appear less than 10 times in the Union and border states in the fighting generation (men aged 13-45 in 1860) and we keep the link closest in age in a 5 year window. Column (4): we drop all links with missing birth year in the Union Army records. Column (5): we consider all links closest in age in a 5 year window (instead of 2) and we do not exclude links not unique in a 2 or 5-year window. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Instrument Balance Test with pre-war variables on the left hand side

	(1)	(2)	(3)
	Coefficient of death rate	Coefficient of death rate	Observations
Age	-2.4677**	-2.3567*	28,911
	(1.2159)	(1.3864)	
Foreign born	0.003	0.030	28,911
	(0.048)	(0.052)	
Occupational income (ihs)	0.010	-0.016	28,911
	(0.146)	(0.159)	
High-skilled	0.014	0.031	28,911
	(0.031)	(0.034)	
Semi-skilled	-0.075	-0.034	28,911
	(0.059)	(0.064)	
Low-skilled	0.058	0.017	28,911
	(0.048)	(0.053)	
Farmer	0.018	-0.024	28,911
	(0.061)	(0.066)	
Illiterate	-0.004	-0.003	28,911
	(0.028)	(0.030)	
Wealth (ihs)	0.115	0.546	28,911
	(0.381)	(0.424)	
Wife age	-1.837	-1.266	24,337
	(1.230)	(1.356)	
Wife wealth (ihs)	0.132	0.162	24,337
	(0.131)	(0.147)	
Wife illiterate	-0.024	-0.026	24,337
	(0.037)	(0.041)	
Wife occupational income (ihs)	-0.020	-0.024	24,337
	(0.065)	(0.072)	
Wife not in household	0.003	0.000	28,911
	(0.034)	(0.038)	
County fixed effects	✓	✓	
Enlistment date polynomial	✓	✓	
Ex ante service duration polynomial	✓	✓	
Additional military controls		✓	
Regiment socioeconomic controls		✓	

Note: Each cell gives the outcome of a different regression, where the pre-war father characteristic is regressed on the instrument (the "leave-one-out" regiment death rate) and the controls, exactly like in the first stage (but without the father, mother and son controls). For occupational income and wealth, we consider the inverse hyperbolic sine transform, which allows to interpret coefficient as percentage changes without excluding zero values. Enlistment date polynomial: father enlistment date in days and enlistment date squared. Ex ante service duration polynomial: ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Additional military controls: fixed effects for father rank at enlistment and characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Regiment socioeconomic controls: socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment, that is weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29 (see Appendix D for details). Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Accounting for Nonlinearities in the First Stage Regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Death rate	1.345***	1.365***	1.318***	1.077***	0.961***	0.990***	0.977***
	(0.093)	(0.092)	(0.091)	(0.115)	(0.127)	(0.127)	(0.127)
Death rate^2	-1.184***	-1.175***	-1.297***	-0.713	-0.407	-0.433	-0.384
	(0.417)	(0.398)	(0.395)	(0.442)	(0.472)	(0.469)	(0.469)
County F.E.		✓	✓	✓	✓	✓	✓
Enl. date poly			✓	✓	✓	✓	✓
Days of service poly				✓	✓	✓	✓
Other military controls					✓	✓	✓
Rgmt socioecon. controls						✓	✓
Father controls							✓
Mother controls							✓
Son controls							✓
Observations	28,911	28,911	28,911	28,911	28,911	28,911	28,911
F-stat	955.26	785.49	632.79	261.85	195.26	204.41	204.66

Note: Regressions of an indicator for whether a father from our linked sample died in the U.S. Civil War on the mortality rate in their last regiment and its squared term. Enlistment date polynomial: father enlistment date in days and enlistment date squared. Ex ante service duration polynomial: ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Other military controls are fixed effects for father rank at enlistment and characteristics of his last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Regmt socioeconomic ctrls: socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment, that is weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. Son controls: age and age squared in 1880. The son's controls are included since they are also conditioned on in the second stage. Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10: First Stage Regression Robustness to Different Standard Error Clustering

	Dependent variable: Pr(Father died)=1				
Death rate	0.865***	0.865***	0.865***	0.865***	0.865***
	(0.047)	(0.048)	(0.050)	(0.050)	(0.054)
Observations	28911	28911	28911	28911	28911
s.e. clustered by:	regiment id	father id	1860 county	Conley (50km)	Conley (100km)

Note: Regressions of an indicator for whether a father from our linked sample died in the U.S. Civil War on the mortality rate in their last regiment. The table replicates the specification in column 7 of the first stage regression in Table 5 with different types of standard error clustering methods. Column 1 is the baseline result with standard errors clustered by father's last regiment of service. The spatial autocorrelation robust standard errors by Conley (1999) were estimated with a 50 and 100km distance cutoff. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11: IV Results for Father's Death and Son's Socioeconomic Outcomes in 1900

	Panel a: Parsimonious specification						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.189***	-0.053	-0.050	0.060	0.091	-0.122*	-0.001
	(0.066)	(0.052)	(0.061)	(0.047)	(0.058)	(0.063)	(0.043)
Mean dep. var.	2.995	0.165	0.323	0.159	0.292	0.679	0.895
Observations	23,198	24,698	24,698	24,698	24,698	24,698	24,679
K-P F-stat	322.92	330.26	330.26	330.26	330.26	330.26	330.40
	Panel b: Full set of controls						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.170**	-0.028	-0.064	0.056	0.084	-0.114*	-0.018
	(0.067)	(0.052)	(0.063)	(0.049)	(0.059)	(0.067)	(0.046)
Mean dep. var.	2.995	0.165	0.323	0.159	0.292	0.679	0.895
Observations	23,198	24,698	24,698	24,698	24,698	24,698	24,679
K-P F-stat	287.65	292.35	292.35	292.35	292.35	292.35	292.43

Note: Instrumental variables regressions of sons' socioeconomic outcomes in 1900 on an indicator for whether their father died in the U.S. Civil War. The indicator for a father's death in the war is instrumented with the mortality rate in their last regiment. When computing the regimental mortality rate the father himself was excluded to not create a mechanical correlation. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1900, and ever married is an indicator for having been married in 1900 including those who became widowers or divorcees before the enumeration date. Panel a (parsimonious specification) controls only for 1860 county of residence fixed effects, enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared. Ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Panel b (full set of controls) also controls for fixed effects for father rank at enlistment, and characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Panel b also controls for socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment: weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29 (see Appendix D for details). Panel b also controls for father characteristics in 1860 (age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth), mother characteristics in 1860 (the same variables as for the father and an indicator for whether there was a mother present in the household) and son characteristics (age and age squared in 1880). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12: IV Results with Increasingly Stringent Geographic Fixed Effects

	Panel a: Parsimonious specification			
	(1) log income score	(2) log income score	(3) log income score	(4) log income score
Father died	-0.134**	-0.160**	-0.141**	-0.143*
	(0.059)	(0.067)	(0.069)	(0.073)
County F.E.	✓			
Town F.E.		✓		
Post office F.E.			✓	
Neighborhood F.E.				✓
Observations	26,753	24,129	23,825	22,342
K-P F-stat	399.83	279.05	296.60	253.20
	Panel b: Full set of controls			
	(1) log income score	(2) log income score	(3) log income score	(4) log income score
Father died	-0.123*	-0.179**	-0.153**	-0.166**
	(0.064)	(0.072)	(0.073)	(0.077)
County F.E.	✓			
Town F.E.		✓		
Post office F.E.			✓	
Neighborhood F.E.				✓
Observations	26,753	24,129	23,825	22,342
K-P F-stat	346.99	235.80	248.50	212.20

Note: This table replicates the results of Table 6, column (1), adding increasingly stringent geographic fixed effects. In the sample, there are 688 different counties, 6,985 different towns/wards, 6,998 different post office districts. We call neighborhood the smallest geographical unit that can be identified in the 1860 census using the post office district and the town/ward (some post office districts contain several towns/wards, some town/wards contain several post office districts. There are 10,044 neighborhoods in our data. Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.13: Placebo IV regressions

	(1) income score (ihs)	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) wealth (ihs)	(7) foreign born	(8) illiterate
Father died	-0.012	0.037	-0.044	0.015	-0.017	0.740	0.033	-0.004
	(0.184)	(0.039)	(0.074)	(0.061)	(0.075)	(0.484)	(0.061)	(0.035)
Mean dep. var.	3.22	.0677	.294	.154	.382	6.01	.185	.0449
Observations	28,911	28,911	28,911	28,911	28,911	28,911	28,911	28,911
K-P F-stat	341.49	341.49	341.49	341.49	341.49	341.49	341.49	341.49
County F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Enlistment date polynomial	✓	✓	✓	✓	✓	✓	✓	✓
Ex ante service duration polynomial	✓	✓	✓	✓	✓	✓	✓	✓
Additional military controls	✓	✓	✓	✓	✓	✓	✓	✓
Regmt socioeconomics controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: In this placebo exercise, pre-war characteristics of fathers are regressed on an indicator for whether they died in the U.S. Civil War instrumented with the mortality rate in their last regiment. For occupational income and wealth, we consider the inverse hyperbolic sine transform which allows to interpret the coefficient as a percentage change without excluding zero value (10% of fathers had no income in 1860, 18% had no wealth). Enlistment date polynomial: father enlistment date in days and enlistment date squared. Ex ante service duration polynomial: ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Additional military controls: fixed effects for father rank at enlistment and characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Regiment socioeconomic controls: socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment, that is weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29 (see D for details). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.14: IV Sensitivity to Father Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep. var.: log income score							
Father died	-0.118*	-0.120*	-0.120*	-0.120*	-0.117*	-0.124*	-0.117*	-0.123*
	(0.065)	(0.065)	(0.065)	(0.064)	(0.065)	(0.064)	(0.065)	(0.064)
Fth age		-0.001**						-0.009***
		(0.001)						(0.003)
Fth age squared			-0.000*					0.000***
			(0.000)					(0.000)
Fth foreign born				0.055***				0.056***
				(0.008)				(0.008)
Fth cannot read					-0.046***			-0.040***
					(0.013)			(0.013)
Fth occ. score						0.006***		0.006***
						(0.000)		(0.000)
Fth wealth							-0.001	-0.003***
							(0.001)	(0.001)
Observations	26,753	26,753	26,753	26,753	26,753	26,753	26,753	26,753
K-P F-stat	346.24	346.61	346.63	347.50	346.11	347.38	346.34	346.99

Note: Instrumental variables regressions of sons' socioeconomics outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War. The indicator for a father's death in the war is instrumented with the mortality rate in their last regiment. When computing the regimental mortality rate the father himself was excluded to not create a mechanical correlation. This table investigates the sensitivity of results to the inclusion of observable father characteristics in the model. All regressions control for father military variables, son variables and mother variables like in Table 6, panel b (full set of controls). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.15: IV Robustness to Different Linking Techniques

	(1) baseline results	(2) excluding multiple links in 5 year window	(3) Ferrie rare names	(4) Only nonmissing birthyear	(5) Large sample size linking
	Dep. var.: log income score				
Father died	-0.123*	-0.159**	0.081	-0.203**	-0.090*
	(0.064)	(0.071)	(0.090)	(0.082)	(0.048)
Observations	26,753	20,782	13,429	13,021	44,990

Note: Instrumental variables regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using different linkage methods: Column (2): we exclude all links that are not unique in a 5 year window instead of 2. Column (3): we consider only individuals whose combination of first and last names appear less than 10 times in the Union and border states in the fighting generation (men aged 13-45 in 1860) and we keep the link closest in age in a 5 year window. Column (4): we drop all links with missing birth year in the Union Army records. Column (5): we consider all links closest in age in a 5 year window (instead of 2) and we do not exclude links not unique in a 2 or 5-year window. The indicator for a father's death in the war is instrumented using the "leave-one-out" mortality rate in their last regiment. We control for the same variables as in Table 6, panel b (full control set). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.16: IV Results excluding disease deaths from the instrument

	Panel a: Parsimonious specification						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.160*	-0.067	-0.165*	0.070	0.074	-0.035	-0.105
	(0.088)	(0.058)	(0.086)	(0.085)	(0.076)	(0.100)	(0.097)
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	136.62	145.62	145.62	145.62	145.62	145.62	139.63
	Panel b: Full set of controls						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.147	-0.020	-0.198*	0.047	0.039	-0.008	-0.028
	(0.100)	(0.069)	(0.106)	(0.103)	(0.091)	(0.120)	(0.090)
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	93.30	97.69	97.69	97.69	97.69	97.69	94.34

Note: Instrumental variables regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War. The instrument for father death is the regimental death rate excluding deaths from disease (the percentage of soldiers who died minus the percentage of soldiers who died of disease). When computing the regimental mortality rate the father himself was excluded to not create a mechanical correlation. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Panel a (parsimonious specification) controls only for 1860 county of residence fixed effects, enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared. Ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Panel b (full set of controls) also controls for fixed effects for father rank at enlistment, and characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Panel b also controls for socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment: weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29 (see Appendix D for details). Panel b also controls for father characteristics in 1860 (age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth), mother characteristics in 1860 (the same variables as for the father and an indicator for whether there was a mother present in the household) and son characteristics (age and age squared in 1880). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.17: IV Robustness to Double ML Covariate Selection

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.114*	-0.062	-0.150**	0.114*	0.057	0.016	-0.012
	(0.060)	(0.040)	(0.064)	(0.059)	(0.051)	(0.072)	(0.054)
Son controls	✓	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓	✓
Father controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓
Mean dep. var.	2.91	.0924	.318	.28	.236	.55	.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	348.01	350.68	352.01	350.90	352.29	350.79	342.63

Note: Instrumental variables regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using the post-double selection (PDS) machine learning algorithm by Belloni et al. (2014). The PDS algorithm takes all controls, their squares, and cross-term interactions and selects the union of significant predictors of the treatment and the outcome and then runs the original regression with the set of selected controls in either step. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. The set of controls for the PDS algorithm to select from is the full set of controls in Table 6, panel b. We always include as controls the quadratic polynomial in enlistment date and the quadratic polynomial in ex ante service duration (the difference between enlistment date and the date of disbandment of the regiment) because they are important predictors of regimental death rate that could be correlated with soldier characteristics (see Table 4). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B External Validity and Weighting

Sample selection introduced by linking is not a concern for our identification strategy, because we never compare the sons of fathers in our linked sample to the sons of fathers in the unlinked population. However, it might be a concern for external validity, especially in the presence of effect heterogeneity. To alleviate this concern, we create customized weights following the method of Bailey, Cole and Massey (2020a). We create two types of weights: 1) weights to make our sample representative of northern fathers in 1860, 2) weights to make our sample representative of fathers in 1860 linked to Union Army records. We cannot create weights to make our sample representative of all Union Army fathers (including those we could not link), because record linking by name between the census and the Union Army records is the only way for us to infer whether a man observed in 1860 later enrolled into the Union Army.

To create the first set of weights, we start with the population of all fathers residing in core Union states in the 1860 census. We create a variable l_j equal to 1 if father j is in the final sample of soldier-fathers. We then use a probit model to regress l_j on covariates measured in the 1860 census.[1] This gives us, for each father of sons in the 1860 census, a probability $\hat{p}$ to be in the final sample predicted from observables. The top panel of Appendix Figure B.1 displays the kernel density of this predicted probability for fathers in the final sample and absent from the final sample. As expected, fathers absent from the final sample have, on average, a lower predicted probability to be linked, but the two distributions have a fairly large common support, which means that we can re-weight fathers in the final sample to be more representative of fathers in 1860 (Bailey et al., 2020a). We then create weights as $((1-\hat{p})/\hat{p}) \times q/(1-q)$ where q is the share of fathers who end up in the final sample.

To create the second set of weights, we use the exact same method, but we start with the sample of fathers in 1860 who we could link to Union Army records. These weights only take care of the selection problem due to linking sons of soldiers between 1860 and 1880. The bottom panel of Appendix Figure B.1 shows that the predicted probabilities to end up in the final sample for fathers present in our sample and absent from our sample have common support.

Appendix Tables B.1 and B.2 show that weighted results are very similar to baseline results, whatever the type of weights used. In the IV specification, effect sizes are somewhat lower

[1] Age, whether born abroad, White, illiterate, the inverse hyperbolic sine of wealth, occupational income score and occupational skill dummies.

when using the first type of weights (about 25% lower for the log income score), but given relatively large standard errors, it is hard to conclude that these effects are statistically different from our baseline results.

Figure B.1: The predicted probabilities to be in the final sample have a broad common support

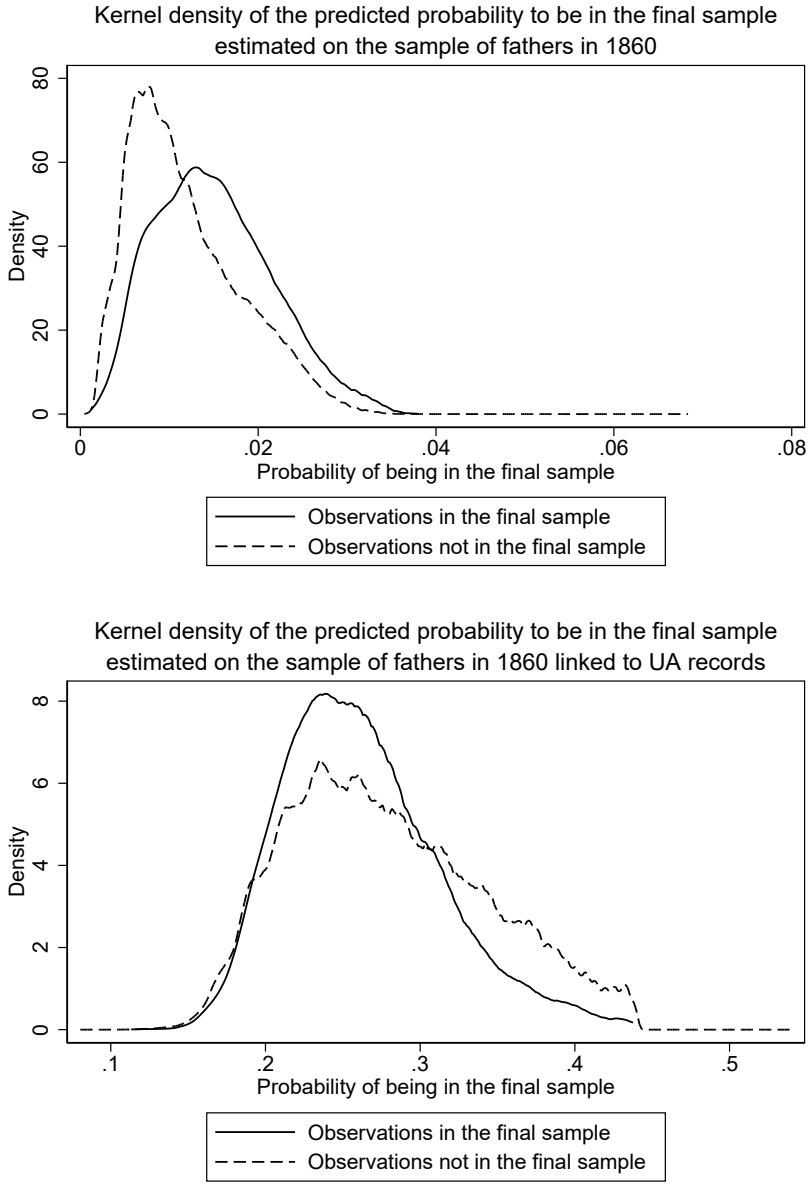


Table B.1: Effect of Father Death on Socioeconomic Characteristics of Sons in 1880 with Customized Weights

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
	Panel a: to make the sample representative of fathers in 1860						
Father died	-0.022***	0.002	-0.022**	0.005	0.015*	-0.010	0.012
	(0.009)	(0.006)	(0.009)	(0.009)	(0.008)	(0.010)	(0.008)
	Panel b: to make the sample representative of fathers in 1860 linked to Union Army records						
Father died	-0.021***	-0.002	-0.022***	0.007	0.015**	-0.005	0.017**
	(0.008)	(0.005)	(0.008)	(0.009)	(0.007)	(0.009)	(0.007)
Son controls	✓	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓	✓
Father other controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓
Observations	27,081	29,269	29,269	29,269	29,269	29,269	28,590

Note: Regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using the re-weighting scheme by Bailey et al. (2020a) to increase sample representativeness. In panel a, the weights make the sample representative of fathers in 1860. In panelb, the weigths make the sample representative of father in 1860 linked to Union Army records by name. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors clustered by the father's last regiment of service and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: IV Estimation with Customized Weights

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
	Panel a: to make the sample representative of fathers in 1860						
	Parsimonious specification						
Father died	-0.115	-0.128**	-0.041	0.039	0.071	0.044	-0.087
	(0.074)	(0.053)	(0.078)	(0.071)	(0.065)	(0.088)	(0.085)
K-P F-stat	254.20	264.92	264.92	264.92	264.92	264.92	262.52
	Full set of controls						
Father died	-0.081	-0.070	-0.090	0.032	0.024	0.075	-0.068
	(0.080)	(0.057)	(0.086)	(0.077)	(0.068)	(0.099)	(0.073)
K-P F-stat	212.05	215.26	215.26	215.26	215.26	215.26	212.63
	Panel b: to make the sample representative of fathers in 1860 linked to Union Army records						
	Parsimonious specification						
Father died	-0.120*	-0.102**	-0.141**	0.128**	0.055	0.028	-0.046
	(0.062)	(0.041)	(0.064)	(0.061)	(0.049)	(0.071)	(0.065)
K-P F-stat	409.98	411.95	411.95	411.95	411.95	411.95	396.15
	Full set of controls						
Father died	-0.127*	-0.085*	-0.154**	0.137**	0.015	0.063	-0.023
	(0.066)	(0.045)	(0.072)	(0.067)	(0.052)	(0.081)	(0.060)
K-P F-stat	355.04	348.67	348.67	348.67	348.67	348.67	335.79
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244

Note: Instrumental variable regressions of sons' socioeconomic outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War using the re-weighting scheme by Bailey et al. (2020a) to increase sample representativeness. The indicator for a father's death in the war is instrumented with the mortality rate in their last regiment. When computing the regimental mortality rate the father himself was excluded to not create a mechanical correlation. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Panel a (parsimonious specification) controls only for 1860 county of residence fixed effects, enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared. Ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment (actual, ex post days of service are mechanically correlated with the death variable). Panel b also controls for socioeconomic characteristics of the regiment computed from information on the soldiers' counties of enlistment: weighted averages for railway and water access, urbanization rates, share of improved acres per farm, farm value, farm equipment value, value of livestock, the labor share in manufacturing, value of manufacturing capital, manufacturing output, personal family estate value, churches per capita, church value, foreign-to-native inhabitant ratio, and the share of men aged 14 to 29 (see Appendix D for details). Panel b also controls for father characteristics in 1860 (age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth), mother characteristics in 1860 (the same variables as for the father and an indicator for whether there was a mother present in the household) and son characteristics (age and age squared in 1880). Standard errors clustered by last regiment of service in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

C Estimating the Aggregate Costs of Losing a Father in the Civil War

In this Appendix, we seek to complement the work by Goldin and Lewis (1975) on the cost of the Civil War by estimating the aggregate cost of father loss. For simplicity, we abstract from potential general equilibrium effects. We simply multiply the implied lifetime income loss of father death by the number of paternal orphans, without considering that non-orphaned men could have benefited from opportunities left vacant by orphaned men. Computing these general equilibrium effect would require a completely different empirical and theoretical framework.

The lifetime income loss of these paternal orphans implied by our results is substantial. Assuming a real wage growth of 1.5% per year, as suggested by data by Long (1960), a 50 years working life and a discount rate of 6%, like Goldin and Lewis (1975), our estimate suggests a loss of lifetime income (discounted to 1861) of $172 per child ($5,200 in 2021 terms). We assume that children start working in 1870 (at an average age of 16) for 50 years, that the average wage for male adults in 1860 is $546 and that it grows at a 1.5% per year in real terms (Long, 1960, table 47). Using a discount rate of 6%, like Goldin and Lewis (1975), we find that the 1861 present value of lifetime income is $7,825 for non-orphaned sons and $7,653 for paternal orphans. Multiplying the difference of $172 by an estimated number of orphans of 363,000, we find a total cost of $62.5 million in 1861 present value ($1.9 billion in 2021 terms). This compares to the $954.9 million in costs from killed soldiers computed by Goldin and Lewis (1975) for the Union ($28.8 billion in 2021 terms). Adding our estimates for the intergenerational effects of these deaths implies that the costs from lost human lives to the North are 6.5% larger than what was previously known. This is likely a lower bound, as we probably underestimate the number of paternal orphans, and because measurement error due do linkage likely biases the OLS estimates towards zero.

D Front Line Service and Socioeconomic Regiment Composition

A potential threat to our identification strategy is a correlation between military strategy and the socioeconomic composition of regiments. Suppose leaders place regiments from the poorest areas in the front lines where they have a higher probability of dying. Regression analyses might then attribute too much of the change in children's later-life outcomes to losing a father which absorbs the effect of the lower socioeconomic status. However, the opposite argument is also plausible when leaders want to occupy the front rows with the most able-bodied soldiers. In this case, we would underestimate the effect of losing a father when children come from the upper classes of society which have the means to alleviate such a loss with more wealth and household resources.

To test for such potential selection, we collected and digitized 128 battle maps from the Civil War Preservation Trust.[2] The idea is to compute the distance of Union regiments to the nearest enemy regiment in order to then regress these distances on the economic composition of Union units and their military characteristics. The maps provide information on the location of Union and Confederate regiments and maintain the same color codes and symbols throughout. Regiments are represented by rectangles and artillery units are marked with a canon symbol. Using pattern recognition techniques, we digitized the location of these symbols on each map. The color schemes were used to differentiate between Union and Confederate units, as well as different battle stages.[3]

For each Union unit, the distance to the nearest Confederate unit was computed for a given battle and battle stage as the point-to-point distance on the Cartesian plane. The distance measure therefore does not have an interpretation in geographic units. Generating a geographic distance variable is complicated by the fact that maps are on different scales. For this reason regressions will use log distances and battle fixed effects. Figure D.1 provides an example.

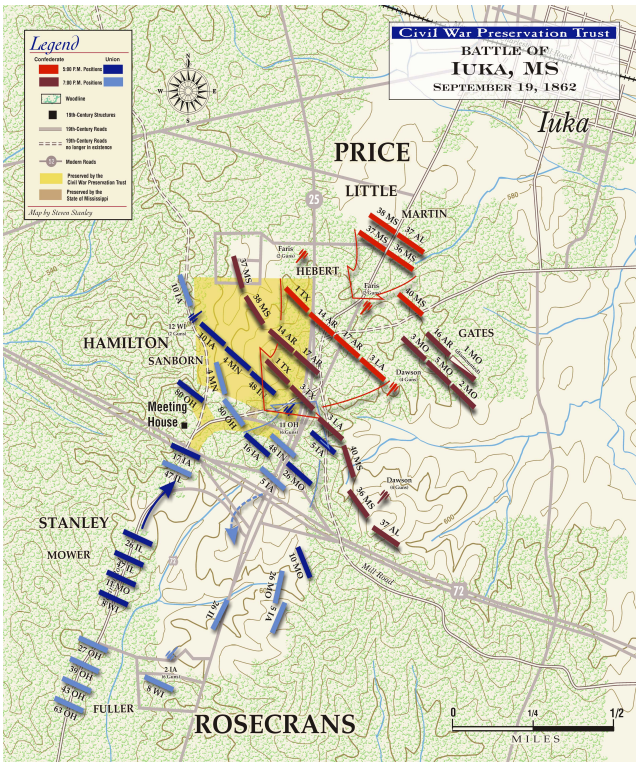
This resulted in 4,147 unit-battle-stage locations for a total of 128 battles and 799 unique Union units. Battles tend to be large with an average number of 20.5 Union units where a typical infantry regiment consists of 1,000 men. To compute the economic composition of each regiment, we used the individual-level soldier data to link soldiers' residence county to economic and population data from the 1860 county-level census. A given census variable x_c

[2]The maps were retrieved from: https://www.battlefields.org/learn/maps on August 27th, 2020.

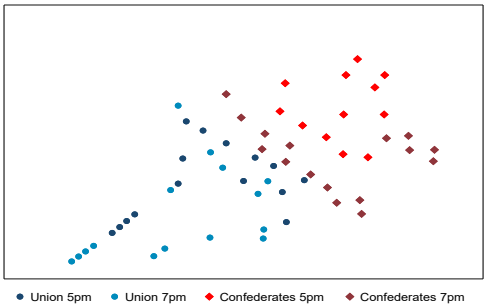
[3]88 of the 128 maps show unit positions for different stages of a battle. This means that there is within-battle variation in the location of regiments. The average battle has 1.45 stages with a maximum of 5.

Figure D.1: Digitizing Civil War Battle Maps

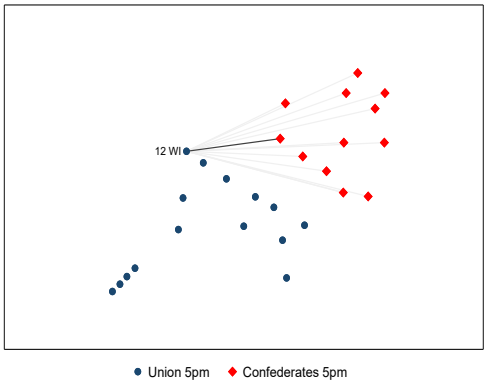
(a) Raw Battle Map



(b) Digitized Battle Map



(c) Minimum Distance to Enemy



Note: Panel a) shows the raw battle map for the Battle of Iuka, Mississippi on September 19, 1862. Union and Confederate regiment positions are shown for two phases of the battle. These are at 5pm (dark blue Union, light red Confederacy) and at 7pm (light blue Union, dark red Confederacy). Panel b) shows the digitized version of the map. Panel c) plots Union and Confederate regiments in their 5pm location, computes the distances to the closes enemy units from the 12th Wisconsin, and marks the minimum distance with a black rather than a gray line. The digitized maps look different due to the way in which they are displayed here, however, relative positions of the regiments to each other are not affected. Battle maps were obtained from the Civil War Preservation Trust (https://www.battlefields.org/learn/maps) and digitized by the authors via pattern recognition algorithms in Python.

for county $c = 1, 2, ..., C$ was then averaged to the regiment level,

$$\overline{x}_r = \frac{\sum_{c=1}^{C} x_c n_{rc}}{\sum_{c=1}^{C} n_{rc}}$$

where the weights $n_{rc} = \sum_{i=1}^{I} n_{irc}$ are the total number of soldiers in regiment r from county c. Variables taken from the 1860 census are the average cash value, number of improved acres, machinery, and livestock value per farm, the share of men aged 14 to 29, the share of employment in manufacturing, the average value of capital, and output per manufacturing establishment, the value of personal real estate per family, the number of churches per 1,000 inhabitants, the average value of church property, and the ratio of foreign- to native-born men.

The military regiment characteristics are the regiment type (infantry, cavalry, artillery), in-

dicators for whether a unit belongs to the regular Army or the U.S. Colored Troops, the average enlistment age of soldiers in the unit, the share of fighting soldiers (to distinguish support units on the field), and measures for unit cohesion such as the total number of counties from which soldiers in the unit joined, and the shares of voluntarily enlisted, soldiers transferred into the unit, and the share of deserted soldiers. Note that most of these measures are only available at the end of the war. This means they should be thought of as totals. For instance, the number of counties in a regiment looks surprisingly large with an average of 30.5. This is mainly due to re-enlistments where soldiers stated a different county and transfers. Hence the average Union regiment had soldiers from about 31 different counties during the entire duration of the war. Summary statistics are reported in Table D.1.

The test for selection into front line service amounts to regressing,

$$\ln(\text{distance})_{rbs} = \delta_b + \phi_s + \overline{x}_r'\gamma + m_r'\beta + \eta_{rbs} \tag{D.1}$$

where the outcome is the natural logarithm of a Union unit's distance to the nearest enemy unit in a given battle b and battle stage s. The vectors $\overline{x}_r$ and m_r contain the economic composition information and military characteristics of the unit, respectively. Battle fixed effects δ_b account for the different geographic scaling of maps while phase fixed effect ϕ_p absorb systematic location differences between earlier and later stages of a battle. Standard errors are clustered at the battle level.

Results are reported in Table D.2. Columns 1 and 2 show the fixed effects only regressions for battles with more than one stage. When adding regiment fixed effects, the adjusted R^2 increase from 47.2 to 49.5 which implies that unobserved time-invariant regiment characteristics are not a major determinant of their distance to the nearest enemy unit. Columns 3 and 4 add military and economic characteristics separately, and jointly in column 5. Again, the adjusted R^2 barely changes and none of the coefficients is a significant correlate with the distance measure in any specification. For most variables these coefficients are tightly estimated zeroes and are not just insignificant due to measurement error in the outcome. The only coefficients with an economically sizable magnitude are those for the artillery and U.S. Colored Troop dummies, however, they are imprecisely estimated. It should also be noted that there are only 16 Black regiments among our 799 units because there were very few Black combat units. Overall there seems to be little evidence for military, economic, and time-invariant regiment specific characteristics to play an important role in the determination of units' front line proximity.

Table D.1: Battle Distance Summary Statistics

	Observations = 4,147			
	Mean	St. Dev.	Min.	Max.
Military Information				
Distance	254.240	278.327	5.099	2,206.181
ln(Distance)	5.152	0.867	1.629	7.699
Number of Union units per battle	20.514	18.318	1	94
Number of battle stages	1.450	0.720	1	5
Infantry	0.948	0.221	0	1
Cavalry	0.030	0.170	0	1
Artillery	0.022	0.146	0	1
Regular Army	0.038	0.192	0	1
US Colored Troops	0.004	0.062	0	1
Mean enlistment age	25.267	2.426	16	39
Share fighting soldiers	98.544	4.062	70.461	100
Share enlisted enlisted	90.456	12.070	17.670	100
Share transferred-in	3.859	8.713	0	82.260
Share deserted	6.645	6.911	0	40.970
Counties present in unit	30.572	24.467	1	161
County Information				
Share men aged 14-29	69.225	3.166	52.285	77.579
Ratio of foreign to native men	0.317	0.230	0.004	1.474
Mean improved acres per farm	63.788	22.149	12.053	195.992
Mean farm value	10,630.411	17,488.969	803.022	80,026.117
Mean machinery value per farm	148.403	83.505	50.444	425.238
Mean value of livestock per farm	472.014	132.702	173.590	1,639.027
Share employed in manufacturing	4.523	3.457	0.241	20.084
Mean capital value per firm	8,064.809	4,530.886	1,512.564	46,688.063
Mean value of output per firm	15,764.820	9,320.380	3,229.907	65,403.676
Value of real estate per family	935.332	527.008	360.179	13,141.862
No. churches per 1,000 population	1.569	0.675	0	5.120
Mean value of church property	9,641.684	11,427.625	0	45,486.945

Note: Summary statistics for the 4,147 unit-battle observations for 799 Union regiments in 128 Civil War battles. Distance to the nearest enemy unit is measured as point-to-point distance on the Cartesian plane. County characteristics are weighted averages at the regiment level. These were computed as the mean characteristic from all counties represented in a regiment weighted by the number of soldiers in the regiment from each county.

Table D.2: Determinants of Distance to Nearest Enemy on the Battlefield

	Outcome: log distance to nearest enemy unit				
	(1)	(2)	(3)	(4)	(5)
Cavalry			0.002		0.005
			(0.060)		(0.061)
Artillery			-0.090		-0.087
			(0.060)		(0.062)
Regular Army			0.034		0.082
			(0.085)		(0.091)
USCT			-0.045		-0.033
			(0.100)		(0.109)
Enlistment age			-0.004		-0.004
			(0.004)		(0.004)
% combat soldiers			0.001		0.001
			(0.003)		(0.004)
% enlisted			0.001		0.002
			(0.001)		(0.002)
% transferred			0.002		0.003
			(0.002)		(0.002)
% deserted			-0.002		0.000
			(0.002)		(0.003)
Improved acres per farm				0.000	0.000
				(0.001)	(0.001)
Mean farm value				0.000	0.000
				(0.000)	(0.000)
Mean farm machinery value				-0.001	-0.001
				(0.000)	(0.000)
% employed in manufact.				0.001	0.003
				(0.007)	(0.007)
Manufact. output value				-0.000	-0.000
				(0.000)	(0.000)
Mean real estate value				-0.000	-0.000
				(0.000)	(0.000)
Ratio foreign to native men				0.065	0.072
				(0.079)	(0.080)
Share men aged 14-29				0.000	0.000
				(0.007)	(0.006)
Observations	3,065	3,065	4,147	4,147	4,147
Battles	88	88	128	128	128
Adj. R^2	0.472	0.495	0.499	0.499	0.498
Regiment FE		Yes			

Note: Regressions of the log point-to-point distance of Union regiments to the nearest Confederate unit on military characteristics and measures of the socioeconomic composition of Union units. Columns (1) and (2) report fixed effects regressions for battles with multiple stages only (88 out of 128 battles). County characteristics are weighted averages at the regiment level. These were computed as the mean characteristic from all counties represented in a regiment weighted by the number of soldiers in the regiment from each county. All regressions include battle and battle stage fixed effects. Standard errors clustered at the battle level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E The Bias of OLS and IV Resulting from Linkage Errors

The linking of census or other historical records without individual identifiers has become a very active research area. Since the first rare-name matching algorithm introduced by Ferrie (1996), more recent papers have introduced supervised (Feigenbaum, 2016) and unsupervised (Abramitzky, Mill and Perez, 2020) machine learning techniques for automated record linkage, as well as evaluations of the performance of such algorithms (Bailey, Cole, Henderson and Massey, 2020b). While a lot of effort is currently devoted to producing more accurate and faster linkage techniques and best practice guides to establish a unified approach (Abramitzky, Boustan, Eriksson, Feigenbaum and Perez, 2021), we know relatively little about what happens to our OLS and IV estimates when we get those links wrong. Abramitzky et al. (2020) state that a promising direction for future research, "is how to adjust regression coefficients when dealing with imperfectly linked data." (p. 11).

Thinking about the impact of record linkage errors on different types of estimators is conceptually challenging because this depends on the nature of the right-hand side variable of interest, whether linkage errors are systematically related to individuals' characteristics,[4] and on the number of data sets that need to be linked, e.g. if an instrument comes from an additional data set.

In the following, we provide a first attempt at quantifying a highly simplified worst-case scenario. Assume that we linked two data sets such as the 1860 and 1880 U.S. census. In the case of this paper, let the true share of orphans be denoted by $T^* = \Pr(x^* = 1)$, where a child with $x^* = 1$ is truly an orphan. Variables with a superscript asterisk denote true values, individual subscripts i are omitted for clarity. In the linked sample, we observe a share of $\tilde{T} = \frac{1}{N}\sum x$ individuals marked as orphans, and a share of $\tilde{C} = (1 - \tilde{T})$ individuals who are marked as non-orphans.[5] Among the children marked as orphans, ν are actually non-orphans and among the children marked as non-orphans, η have lost a father but this error is not observed by the econometrician.

Assume the extreme case wherein every linkage error also results in a flip in treatment status. The mis-measured orphan status can be thought of as measurement error and this error is non-classical. Whenever a child is wrongly marked as orphan, the only other value that the

[4] For instance, individuals with longer names can be linked more accurately because they contain more information and are usually rarer than shorter names. However, longer names have been shown to correlate with higher incomes and levels of education (Bailey et al., 2020b).

[5] T and C denote the treatment and control group, respectively.

true orphan status can take is the exact opposite ($x = 1, x^* = 0$). This induces a negative correlation between the true and observed treatment status. This is the framework considered by Aigner (1973) who shows that measurement error in a binary treatment attenuates OLS estimates. The true share of orphans relates to the observed quantities as,

$$T^* = (1 - \nu)\tilde{T} + \eta\tilde{C} \tag{E.2}$$

and the mis-measured orphan status can be expressed as

$$x = x^* + u \tag{E.3}$$

where u is the error induced by wrong record linkages, and $x^* \sim \text{Ber}(T)$ and $x \sim \text{Ber}(\tilde{T})$. To derive the bias of the OLS estimator, Aigner (1973) states the following quantities:

$$\begin{aligned}
\mathbb{E}(u) &= \nu(\tilde{T}) - \eta\tilde{C} \\
Var(u) &= \nu\tilde{T} + \eta\tilde{C} - (\nu\tilde{T} - \eta\tilde{C})^2 \\
Cov(x, u) &= (\nu + \eta)\tilde{T}\tilde{C}.
\end{aligned}$$

Then for the model $y = \alpha + \beta x^* + \epsilon$, the OLS estimator is,

$$\begin{aligned}
\widehat{\beta}_{\text{OLS}} &= \frac{Cov(\alpha + \beta x^* + \epsilon, x^* + u)}{Var(x)} \\
&= \beta \left[\frac{Var(x^*) + Cov(x^*, u)}{Var(x)} \right] \\
&= \beta \left[\frac{T(1 - T) + Cov(x, u) - Var(u)}{\tilde{T}(1 - \tilde{T})} \right]
\end{aligned} \tag{E.4}$$

Now substitute the following quantities into (E.4),

$$\begin{aligned}
Var(x^*) &= T(1 - T) \\
&= \left[(1 - \nu)\tilde{T} + \eta\tilde{C} \right] \left[1 - (1 - \nu)\tilde{T} - \eta\tilde{C} \right] \\
&= (1 - \nu)\tilde{T} - \left[(1 - \nu)\tilde{T} \right]^2 - 2\eta\tilde{T}\tilde{C}(1 - \nu) + \eta\tilde{C} - \left[\eta\tilde{C} \right]^2 \\
Cov(x, u) &= \nu\tilde{T}\tilde{C} + \eta\tilde{T}\tilde{C} \\
Var(u) &= -\nu\tilde{T} - \eta\tilde{C} + \left[\nu\tilde{T} \right]^2 - 2\eta\nu\tilde{T}\tilde{C} + \left[\eta\tilde{C} \right]^2
\end{aligned}$$

to derive the OLS bias as,

$$
\begin{aligned}
\widehat{\beta}_{\text{OLS}} &= \beta \left[\frac{T(1-T) + Cov(x,u) - Var(u)}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[\frac{\left[(1-\nu)\tilde{T} + \eta\tilde{C}\right]\left[1 - (1-\nu)\tilde{T} - \eta\tilde{C}\right] + (\nu\tilde{T}\tilde{C} + \eta\tilde{T}\tilde{C})}{\tilde{T}(1-\tilde{T})} \right] \\
&+ \beta \left[\frac{-\nu\tilde{T} - \eta\tilde{C} + \left[\nu\tilde{T}\right]^2 - 2\eta\nu\tilde{T}\tilde{C} + \left[\eta\tilde{C}\right]^2}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[\frac{\tilde{T} - \nu\tilde{T} - \tilde{T}^2 + 2\nu\tilde{T} - \left[\nu\tilde{T}\right]^2 + 2\eta\nu\tilde{T}\tilde{C} - 2\eta\tilde{T}\tilde{C} + \eta\tilde{C} - \left[\eta\tilde{C}\right]^2 + \nu\tilde{T}\tilde{C} + \eta\tilde{T}\tilde{C}}{\tilde{T}(1-\tilde{T})} \right] \\
&+ \beta \left[\frac{-\nu\tilde{T} - \eta\tilde{C} + \left[\nu\tilde{T}\right]^2 - 2\nu\eta\tilde{T}\tilde{C} + \left[\eta\tilde{C}\right]^2}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[\frac{\tilde{T} - \tilde{T}^2 - 2\nu\tilde{T} + 2\nu\tilde{T}^2 - \eta\tilde{T}\tilde{C} + \nu\tilde{T}\tilde{C}}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[\frac{\tilde{T} - \tilde{T}^2 - 2\nu\tilde{T} + 2\nu\tilde{T}^2 - \eta\tilde{T}(1-\tilde{T}) + \nu\tilde{T}(1-\tilde{T})}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[\frac{\tilde{T}(1-\tilde{T}) - \nu\tilde{T}(1-\tilde{T}) - \eta\tilde{T}(1-\tilde{T})}{\tilde{T}(1-\tilde{T})} \right] \\
&= \beta \left[1 - \nu - \eta\right] \qquad \text{(E.5)}
\end{aligned}
$$

It follows from (E.5) that OLS is biased towards zero for a type I error rate of $\nu + \eta < 1$. For very high error rates that are $\nu + \eta > 1$, the OLS estimate will reverse in sign. Note that if all true orphans are wrongly classified as non-orphans ($\eta = 1$) and if all true non-orphans are classified as orphans ($\nu = 1$), then OLS will recover the true coefficient but with the opposite sign.

For the IV estimator, assume that we have an instrumental variable z which relates to the true orphan status via the first stage regression,

$$x^* = \pi_0 + \pi_{x^*z} z + \xi \qquad \text{(E.6)}$$

and that satisfies the exclusion restriction. Let $\delta_{yz} = \frac{Cov(y,z)}{Var(z)}$ denote the reduced form coeffi-

cient from the regression of y on z. An IV estimate can then be constructed as,

$$\widehat{\beta}_{\text{IV}} = \frac{\delta_{yz}}{\pi_{x^*z}} \tag{E.7}$$

however, while the reduced form is unbiased, the first stage is not. This is because instead of x^* we observe the mis-measured x. Meyer and Mittag (2017) show that the OLS estimate of the first stage with the mis-measured binary dependent variable will be

$$\pi_{xz} = (1 - \nu - \eta)\pi_{x^*z}$$

and therefore the bias of the IV estimator is,

$$\begin{aligned}
\widehat{\beta}_{\text{IV}} &= \frac{\delta_{yz}}{\pi_{xz}} \\
&= \frac{\delta_{yz}}{(1 - \nu - \eta)\pi_{x^*z}} \\
&= \frac{1}{1 - \nu - \eta}\beta_{\text{IV}}
\end{aligned} \tag{E.8}$$

The IV bias is the inverse of the OLS bias. For the case where $\nu + \eta = 1$ exactly, the IV estimator does not exist. And again, if treatment and control group are switched around with $\nu + \eta = 2$, also the IV estimator recovers the true parameter with the opposite sign.

How does this result relate to practice? The typical type I error rate of automated linkage methods in Bailey et al. (2020b) ranges between 0.22 and 0.69. For the lowest error rate, OLS will be attenuated to 78% and IV will be inflated to 128% of the true coefficient value. For the highest error rate instead, OLS will only be 31% and IV will be 323% of the true coefficient. Even though the scenario described here is highly simplified and a worst-case situation in which each wrong link leads to a treatment status change, the example shows how linkage errors can potentially lead to large differences between OLS and IV estimates which cannot be motivated with the typical LATE explanation.

Also note that, in the absence of other endogeneity problems, OLS and IV will set identify the true parameter value by providing lower and upper bounds $\widehat{\beta}_{\text{OLS}} < \beta < \widehat{\beta}_{\text{IV}}$. Without further assumptions, these bounds are sharp. This means that even in the presence of linkage errors the OLS and IV estimates can be informative.

E.1 Evidence from a Simulation Exercise

To test the theoretical framework above, we simulate a data set of 10,000 individuals, half of whom are in the treatment and control group respectively, $T = C = 0.5$. For 10% of individuals on both groups we then assume a linkage error that reverses their treatment status, such that $x = 1 - x^*$, implying a total error rate of $\nu + \eta = 0.1 + 0.1 = 0.2$, which is roughly the type I error rate found for the Ferrie (1996) rare-name linkage algorithm in Bailey et al. (2020b). The observed treatment status x is then generated as described above with $x = x^* + u$.

The true estimating equation is,

$$y_i = 1x_i^* + \epsilon_i \tag{E.9}$$

where $\epsilon_i \sim N(0, 1)$ is an *iid* error term, and the coefficient of the true treatment effect is $\beta = 1$. Suppose we have a valid instrument z which relates to x^* via the first stage regression,

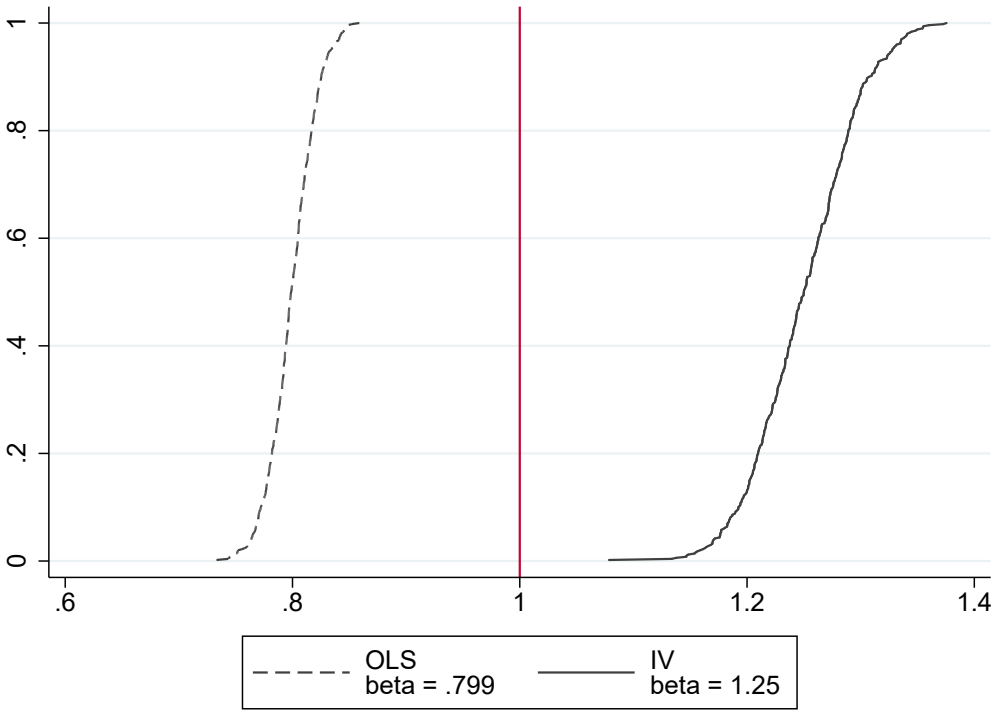
$$x_i^* = \frac{2}{3} z_i + \xi_i \tag{E.10}$$

with $\xi_i \sim N(0, 1)$ *iid* errors, a first stage coefficient $\pi = \frac{2}{3}$, and $Corr(\epsilon, \xi) = 0$.[6] We simulate (E.9) by substituting x^* with x and we do this 500 times to observe the behavior of the OLS and IV estimates. The CDFs of the OLS and IV estimates obtained from these 500 simulations are graphically reported in Figure E.1 and numerically in Table E.1.

As predicted by the theory outlined in the previous section, OLS recovers 80% of the true parameter value while IV is inflated to 125% of the true coefficient. Note that IV has more than twice the dispersion of OLS, yet none of the two estimators includes the true value in their 95% confidence interval. In practice, however, this will depend on the strength of the first stage and whether any other endogeneity concerns are present. The true first stage coefficient is estimated when using the treatment variable without linkage error which yields $\widehat{\pi}_{x^*z} = 0.6669$, while the first stage with the mis-measured treatment produces the predicted coefficient of $(1 - \nu - \eta)\pi_{x^*z} = (1 - 0.2)\frac{2}{3} = 0.5338$. Also the simulation confirms that $\widehat{\beta}_{\text{OLS}} < \beta < \widehat{\beta}_{\text{IV}}$, given that no other endogeneity problem was simulated.

[6] The distinction of whether z is binary or continuous does not matter in this context.

Figure E.1: Simulated OLS and IV Bias with Mis-Measured Binary Treatment due to Linkage Errors



Note: OLS and IV CDFs from 500 simulations of a data set with 10,000 individuals, half of whom are in the treatment group. Misclassification rates for both treatment and control are set to 0.1 each (i.e. a total misclassification error of 20%) and a true treatment effect of 1 which is marked by the red line. The figure reports the median bias of OLS and IV below the graph.

Table E.1: Summary Statistics for Simulated OLS and IV Estimations with a Mis-Measured Binary Treatment due to Linkage Errors

	obs.	mean	st. dev.	min	max
$\widehat{\beta}_{\text{OLS}}$	500	0.7994	0.0207	0.7331	0.8588
$\widehat{\beta}_{\text{IV}}$	500	1.2504	0.0458	1.0785	1.3756
$\widehat{\pi}_{x^*z}$	500	0.6669	0.0031	0.6556	0.6751
$\widehat{\pi}_{xz}$	500	0.5338	0.0072	0.5081	0.5554

Note: Summary statistics for OLS, IV and first stage estimates from 500 simulations of a data set with 10,000 individuals, half of whom are in the treatment group. Misclassification rates for both treatment and control are set to 0.1 each (i.e. a total misclassification error of 20%). Rows from top to bottom are for the OLS estimator $\widehat{\beta}_{\text{OLS}}$, the IV estimator $\widehat{\beta}_{\text{IV}}$, the first stage using the true treatment variable as outcome $\widehat{\pi}_{x^*z}$, and the first stage using the mis-measured treatment as outcome $\widehat{\pi}_{xz}$.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

_ , Roy Mill, and Santiago Perez, "Linking individuals across historical sources: A fully automated approach," *Historical Methods*, 2020, *53* (2), 94–111.

Aigner, Dennis J., "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1973, *1* (1), 49–59.

Bailey, Martha, Connor Cole, and Catherine Massey, "Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples," *Historical Methods*, 2020, *53* (2), 80–93.

_ , _ , Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," *Journal of Economic Literature*, 2020, *58* (4), 997–1044.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *Review of Economic Studies*, 2014, *81* (2), 608–650.

Conley, Timothy G., "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Feigenbaum, James J., "A Machine Learning Approach to Census Record Linking," *mimeo*, 2016.

_ , "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940," *The Economic Journal*, 2018, *128*, 446–481.

Ferrie, Joseph P., "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Goldin, Claudia D. and Frank D. Lewis, "The Economic Cost of the American Civil War: Estimates and Implications," *Journal of Economic History*, 1975, *35* (2), 299–326.

Long, Clarence D., *Wages and Earnings in the United States, 1860-1890*, Princeton University Press, Princeton, NJ, 1960.

Meyer, Bruce D. and Nikolas Mittag, "Misclassification in Binary Choice Models," *Journal of Econometrics*, 2017, *200* (2), 295–311.

Olivetti, Claudia and M. Daniele Paserman, "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1940," *American Economic Review*, 2015, *105* (8), 2695–2724.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt, "Poorly Measured Confounders are More Useful on the Left than on the Right," *Journal of Business and Economic Statistics*, 2018, *37* (2), 205–216.

Preston, Samuel H. and Michael R. Haines, *Fatal Years: Child Mortality in Late Nineteenth-Century America*, Princeton, NJ: Princeton University Press, 1991.