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

Yannick Dupraz† Andreas Ferrara‡

October 2019

Abstract

We estimate the causal effect of losing a father in the U.S. Civil War on children's long-run socioeconomic outcomes. Linking military records from the 2.2 million Union Army soldiers with the 1860 U.S. population Census, we track soldiers' sons into adulthood. Sons of soldiers who died had a lower a occupational income score in 1880 and were less likely to have a high- or semi-skilled job as opposed to being low-skilled or farmers. Our results are robust to instrumenting paternal death with the mortality rate of the father's regiment. Effects are largely driven by the increased downward mobility of the sons of semi-skilled fathers, who were more likely to become low-skilled as a result of paternal death. Pre-war family wealth is a strong mitigating factor: there is no effect of losing a father in the top quartile of the wealth distribution.

JEL codes: N11, J13, J62
Keywords: U.S. Civil War; orphans; intergenerational mobility.

*We thank Martha Bailey, Sascha O. Becker, James Feigenbaum, James Fenske, Victor Gay, and Matt Nelson, as well as seminar participants at Queen's University Belfast, University of Michigan, University of Oxford, Paris School of Economics, University of Warwick, 5th Australasian Cliometrics workshop, 78th Economic History Association annual meeting, 30th European Association of Labour Economists conference, 44th Social Science History Association annual meeting, and at the 18th World Economic History Congress for valuable comments and discussion. Vlad Barrow, Bridgid Mogeni, Ziya Springwala, and Aidan Tee provided excellent research assistance. We thank Christian Dippel and Stephan Heblich for joint efforts in digitizing and collecting the military data. We gratefully acknowledge financial support for this project from CAGE (Dupraz), the Royal Economic Society, and the Leverhulme Trust (Ferrara).

†University of Warwick, Department of Economics and Center for Competitive Advantage in the Global Economy (CAGE). Email: y.dupraz@warwick.ac.uk

‡University of Pittsburgh, Department of Economics and CAGE. Email: a.ferrara@pitt.edu

1 Introduction

The persistence of inequalities in income, wealth or education across generations has been well documented (see Black and Devereux, 2011; Chetty et al., 2014). Part of this persistence is explained by parental investments in their children. How do individuals fare in the long-term after having experienced the loss of a parent in childhood, and does this shock affect labor market outcomes?[1] This is a challenging question because modern datasets on orphans tend to be small and have a short time-frame, thus allowing the study of short-term outcomes, such as education or health, but not long-run outcomes relating to income, occupational choice, migration, or marriage.[2]

To answer the above question, we turn to a historical setting and study the U.S. Civil War (1861-65), the deadliest conflict in U.S. history. We link military records of the 2.2 million Union Army soldiers to the 1860 Census to identify soldiers who were fathers. Then we track their sons into adulthood by linking them to the 1880 Census. Our linked data consists of 32,240 men in 1880 who were sons of Union Army soldiers, 13.27% of of whom lost their father during the war. This means that we observe more than 4,200 orphans from age 0 to 20 in 1860 and again when they are aged 20 to 40 in 1880. This allows us to study the effect of paternal deaths on children's occupational income score, geographic mobility, and marriage decisions when they are adults. This goes beyond previous work which tended to focus on education and health outcomes due to restrictions on the size of the sample or the number of years for which the orphaned children could be observed.[3]

Our results show a negative and persistent effect of paternal death on children's later life outcomes. To limit issues of selectivity into the military, we only compare sons of fathers who fought in the war.[4] Controlling for detailed pre-war socioeconomic characteristics of both parents as well as a rich set of military controls, OLS results show that losing a father in the war decreased a son's occupational income score in 1880 by 4% of a

[1]This question is particularly relevant for the more than 140 million children worldwide who lost either or both parents (UNICEF Press Center, 2017).

[2]Beegle et al. (2010) provide the orphan study with the longest panel with 13 years of data, while the largest cross-sectional study was conducted by Gertler et al. (2004) which included 3,119 orphaned children.

[3]See Case et al. (2002), Gertler et al. (2004) Ainsworth and Filmer (2006), Evans and Miguel (2007), or Senne (2014).

[4]Later in the war it was possible for rich men to buy out of service or to furnish a substitute (McPherson, 1988).

standard deviation and raised the probability to be a farmer by 1.6 percentage points relative to being in all other occupational groups. The wage drop corresponds to a lifetime loss in earnings of more than a year of work, which in aggregate means a total loss of $112M dollars in 1860 value for all orphaned sons.[5] In 2019 dollar values, this corresponds to a total loss of $3.2bn. This compares to the $954M in estimated human capital losses due to the deaths and disabilities of soldiers which were computed by Goldin and Lewis (1975). Their estimate did not take into account intergenerational effects. Our result for costs of the war stemming from the intergenerational effect is a lower-bound estimate since we show later that losing brothers, other family members, and men in the local community imposed additional wage penalties. This suggests that the negative effects of the Civil War from lost human capital and earnings may be significantly higher than what was previously estimated.

A major concern with the empirical strategy is the potential endogeneity of paternal death and the difficulty in finding plausibly exogenous variation in it (see Adda et al., 2011). To test the robustness of our OLS estimates, we instrument a father's probability of dying in the war with the mortality rate in his regiment.[6] The argument is that regimental mortality rates tend to be determined by military strategy, not by the socioeconomic characteristics of soldiers, let alone the future socioeconomic characteristics of their children. Using digitized battle maps, we provide evidence that the location of regiments on the battle field was not determined by socioeconomic characteristics of the units, i.e. it is not the case that poor regiments were placed in the front which would invalidate our instrument.

Our IV estimates confirm the OLS findings, though they are substantially larger. We show this is partly explained by heterogeneous effects for the takers of the instrument. We also attribute part of the difference to measurement error in the binary treatment variable because of false links across Census years. Taking into account the regiment's rate of disabled soldiers and death rates of other men in the network of the son (brothers, uncles, neighbors) does not substantially decrease the effect of father death.

[5]The aggregate number sums the compounded lifetime loss of all Union Army orphans. These are around 150,000 affected individuals who lost a father. The compounding assumes a fifty years work-life, an interest rate of 6%, and uses the average wage for male adults in 1860 reported by Long (1960), table 47. The average male income in 1860 was $546 and the compounded loss of 2.2% of these earnings ($12) for the given parameters leads to a lifetime loss of $638.

[6]When computing mortality rates of a father's regiment, we omit the father from this computation in order to avoid a mechanical correlation of this instrument with a father's death indicator.

In terms of mechanisms, we find evidence for the role of an income effect, where the missing father income prevents fatherless sons from accumulating labor market skills through education and training. We find that the negative effect is concentrated in the middle of the skill distribution and largely driven by the downward mobility of the sons of semi-skilled fathers, who are more likely to have a low-skilled occupation. We also find that father wealth in 1860 (before the war) is a strong mitigating factor: we estimate no negative effect in the top quartile of the wealth distribution. We do not find strong evidence for heterogeneity with respect to age at father enlistment.

We contribute to two strands of the literature: first, the literature on the relationship between family structure, parental investments, and the economic mobility of children, and second, the literature focusing on the economic history and long-term consequences of the U.S. Civil war. Today in the United States, one of the strongest determinant of income-mobility for children is whether they live in a single-parent households (Chetty et al., 2014).[7] The long-term consequences of losing a parent as shock to family structure are less well known.[8] The literature on the topic of parental loss and child outcomes mostly includes short-run studies of the effects of parental deaths on education and health. This literature tends to find negative effects (Gertler et al., 2004; Evans and Miguel, 2007; Kovac, 2017), though Ainsworth and Filmer (2006) find that the difference in enrollment rates between orphans and non-orphans is small in many countries. Few papers study labor-market outcomes of orphans in the long run. An exception is Adda et al. (2011) who use linked Swedish administrative data to show that parental loss decreases boys' earnings by 6-7 percent. The U.S. in the 19th century did not have the generous welfare state of Sweden in the 20th century, so our setting might be more suited to understand the effect of parental loss in the absence of a strong safety net.

The second literature we contribute to is the literature on the economic history of the Civil War and its long-term consequences (Goldin and Lewis, 1975; Feigenbaum et al., 2018; Ager et al., 2019). A random sample of Union Army regiments based on Fogel (2000) has been used to estimate the income effect of the Union Army pension program

[7]This correlation holds also when considering only the income mobility of children whose parents are both married, which means family structure correlates with income mobility not only at the family level, but also at the community level.

[8]Note that our work differs from studies that have relied on disruptive events such as divorce (Painter and Levine, 2000; Corak, 2001; Gruber, 2004), or imprisonment (Bhuller et al., 2018; Dobbie et al., 2018) in the sense that incarcerated or divorced parents are still potentially available to invest in their children, the effect of the death of a parent means a permanent loss for the child.

(Costa, 1995), and its effect on living arrangements (Costa, 1997), the impact of combat unit homogeneity on desertion (Costa and Kahn, 2003), the short and long run impact of diversity for black Union soldiers (Costa and Kahn, 2006), and the importance of social networks in survival in POW camps (Costa and Kahn, 2007). Since most soldiers in the Union Army were not fathers and because we want to be able to explore heterogeneity, our study relies on a new and much larger data set of linked Union Army soldiers using newly digitized information from all of the 2.2 million Union Army soldiers.

2 Historical Background

The U.S. Civil War (1861–65) was a defining moment for the United States and the deadliest conflict in U.S. military history with over 650,000 fallen soldiers. This was a substantial shock to a population of only 31 million.[9] While the war was primarily fought over the abolition of slavery and for the preservation of the country's unity, several other political and economic factors played a role. We refer the reader to the work of historians for a comprehensive review of the history of the American Civil War (McPherson, 1988, for example). Instead this historical background section focuses primarily on the military and institutional setting which will help in framing the empirical analysis.

The Civil War started with the Confederate attack on Fort Sumter on April 12, 1861. The regular Army at the time was small with only 16,000 personnel. The Union Army was instead raised as volunteer force and increased significantly in size after Lincoln's call for 75,000 volunteers in the Summer of 1861. With 600,000 initial enlistments, this call was exceeded by a wide margin (Chambers, 1987). Volunteer regiments were raised and organized by the individual states with little centralized intervention from the government. Participation was high. Of those born between 1838-45, up to 98% were examined for service and up to 81% of these cohorts ended up serving in the war (Costa and Kahn, 2008). The only time when participation rates were higher was during World War II. Ultimately, 2.2 million soldiers would serve in the Union and, even though the regular Army also increased in size, 94% of all soldiers were volunteers. The draft lottery, which was introduced in 1863, was largely ineffective and only raised 10,000 additional soldiers. The main purpose of the draft was to act as potential threat that could be

[9] As enumerated by the 1860 Census.

expanded if not enough volunteers could be found (Chambers, 1987).

The structure of the Union Army closely resembled those of modern Armies.[10] The most important unit in terms of both recruitment and fighting was the regiment. A typical infantry regiment had 1,000 soldiers and was composed of ten companies of 100 men each. A Colonel would usually lead the regiment, and companies were commanded by a captain and two lieutenants. Military leaders were often not trained soldiers themselves but prominent men from the community in which a regiment was raised, such as politicians, factory owners, or other prominent figures, and "no company had the ability to pick the best officers and soldiers" (Costa and Kahn, 2008, p. 57). This is true even for the higher unit commanders as half of the Union generals were military amateurs rather than professionally trained soldiers (Chambers, 1987).

The typical contract length for a soldier was 3 years. 61% of the regiments we observe were recruiting on the basis of 3-year contracts. The next common contract lengths were 1- and 2-year contracts, which increased the pressure on recruiters starting in 1863. This led to the passing of the Enrollment Act of 1863 which provided the basis for introducing a national draft in case not enough volunteers could be found. Unlike the South, where a very efficient draft had been established early in the war, the North mainly sought to promote volunteering by the threat of establishing a national draft and with generous enlistment bounties. In general, the early recruits of 1861 and 1862 were of higher quality than later enlistees. The early recruits tended to be positively selected as they were on average taller, richer, less likely to be low-skilled laborers, or married, and they were also more likely to be born in the U.S., Germany, or Ireland (Costa and Kahn, 2008).

The massive organizational scale as well as the size of the conflict generated unexpected problems. In the beginning, politicians of both sides estimated a total duration of no more than six months for the war (Chambers, 1987). Likewise the first battles were unexpectedly bloody as unusually large forces clashed and where old-style Napoleonic military tactics of closed-rank formations were confronted with the devastating power of modern weaponry. In addition, the size of the military posed all sorts of logistic issues. Experience with field hygiene and medical treatment on this a scale was lacking. A little under half of all deaths were therefore due to disease and illness, closely followed by battle deaths.

[10]The organizational structure of the infantry is shown in appendix figure B.1.

3 Data Sources and Record Linking

The main data sources used in this paper are the full-count U.S. Census files for 1860 and 1880, and military records from the Union Army. The Census was provided by IPUMS-USA (Integrated Public Use Microdata Series), while the military data were digitized from various printed volumes published after the War, usually called "Report of the Adjutant General" or "Roster".[11] The Adjutant Generals' reports were compiled after the war to keep a record of veterans and deceased soldiers for accounting purposes, especially for the payment of bounties and pensions. An excerpt for the records of the 22nd Massachusetts volunteer infantry regiment is shown in figure 1.

Figure 1: 22nd MA Volunteer Infantry Regiment Records Example

Twenty-Second Regiment Infantry, M. V.—(Three Years.)—Continued.

NAME AND RANK.	Age.	Bounty.	Residence or Place credited to.	Date of Muster.	Termination of Service and cause thereof.
Company E—Con.					
Murphy, Charles,	25	–	Roxbury,	Sept. 9, '61,	Killed June 27, 1862, Gaines' Mills, Va.
Nayson, William E.,	30	–	Roxbury,	13, '61,	Dec. 7, 1863, disability.
Nickerson, James,	30	–	Roxbury,	9, '61,	Killed July 1, 1862, Malvern, Hill, Va.
Nolan, Henry J.,	22	–	Boston,	9, '61,	Died Oct. 27, 1862, New York Harbor.
Norton, James,	30	–	Roxbury,	9, '61,	Dropped from rolls, Oct. 1, 1861.
Noyes, Joseph P.,	40	–	Lynn,	9, '61,	Oct. 21, 1862, disability.
Pearl, George W.,	18	–	Boston,	28, '61,	4, 1864, expiration of service.
Petterson, Leonard,	29	–	Roxbury,	18, '61,	Killed May 8, 1864, Laurel Hill, Va.
Pierce, Philip R. W.,	39	–	Roxbury,	9, '61,	Nov. 5, 1862, disability.
Quinn, William,	30	–	Roxbury,	13, '61,	Sept. 24, 1862, disability.
Ray, John, J.,	19	–	Boston,	9, '61,	Feb. 1, 1864, to re-enlist.
Raymond, William T.,	19	–	Roxbury,	9, '61,	Sept. 24, 1862, disability.
Richardson, James,	34	–	Roxbury,	9, '61,	Oct. 20, 1864, expiration of service.
Robinson, John,	43	–	Boston,	Aug. 28, '62,	Dec. 15, 1862, disability.

Even though similar data exists for the Confederacy, the Union records are of much higher quality and completeness. This motivated our focus on the Union. Quoting the Adjutant General of Massachusetts in his final report (1866): "*[M]ost of the regiments and batteries are perfect, every man accounted for; of the whole number there are but 1,205 who are not accounted for*" (p. 121). These unaccounted soldiers make up 1.1% of the overall number of enlisted men from Massachusetts which totaled 106,330. Our dataset comprises 2,922 regiments and about 2.7 million military records, covering almost all of the 2.2 million Union soldiers.

The data provide us with less information than the data set constructed by Costa and Kahn (2003, 2007) based on the random 1.3% sample collected by Fogel (2000), for which they added further information on pensions, families, and later life outcomes of

[11] A full list of the different sources is provided in the appendix in table A.1. We are grateful to Christian Dippel and Stephan Heblich for joint efforts in collecting the military data.

soldiers.[12] The advantage of our long but narrower data set is that we can observe the entirety of Union units and almost the entire universe of the 2.2 million Union Army soldiers. The main reason as to why this data still does not comprise the full universe of soldiers is the unprecedented scale of the war. Not everyone could receive a proper burial under those circumstances. Sometimes individuals could not be identified anymore due to the severity of the injuries inflicted by the new weaponry available or because of the weather conditions.

Information on each individual soldier includes their full name, enlistment and discharge date, military rank at enlistment and discharge, regiment and company, duration and terms of enlistment (commissioned, drafted, volunteered), and state of enlistment. The information on soldier's exit and reason for exit from a unit includes information on casualties and the type of casualties. We identify soldiers who died during the war and the reason of death, as well as those who were severely wounded or disabled.

Table 1 provides summary statistics for the 2.7 millions of military records in our dataset. The average age at enlistment is 25, 94% of all observations are enlisted volunteers. 84% entered service in the rank of a Private, and 74% entered service in an Infantry regiment. The death rate in our data set is 12.5%. This is lower than the historians' estimation for the Union Army (16.6%), but it should be noted that the unit of observation in our dataset is the record, not the soldier. The number of records is larger than the number of Union Army soldiers (2.2 millions) because of re-enlistments.[13] Information on whether the soldier died or survived, is missing in about 20% of cases.

Some information is less systematically available across states, like place of residence, place of enlistment, and age at enlistment. We are able to infer the birth year from age and date of enlistment for only 41% of records. We are able to geolocate the county of enlistment for 27% of records, and the county of residence for 34% of records (table 1).[14] These missing values explain the strategy we employ below for linking military records with the 1860 census.

[12]For the final result of this tremendous research effort see `http://uadata.org/`

[13]We tried to identify duplicate soldiers (who reenlisted) based on their first and last name, date of enlistment and age, but we managed to reduce the number of records by about 200,000 only. Soldiers who appear several times in the military records will likely not be linked to the Census because resolving linking ambiguities for them will be difficult.

[14]56% of records have information on either the county of residence or the county of enlistment.

Table 1: 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 15 1863		Jun 10 1801	Jul 22 1869
Birthyear known	2,739,730	0.412	0.492	0	1
County of enlistment located	2,739,730	0.265	0.441	0	1
County of residence located	2,739,730	0.336	0.472	0	1
Reason for joining					
Enlisted	2,697,273	0.940	0.238	0	1
Commissioned	2,697,273	0.030	0.171	0	1
Drafted	2,697,273	0.016	0.124	0	1
Substitute	2,697,273	0.014	0.119	0	1
Rank (at enlistment)					
Private	2,739,730	0.840	0.366	0	1
Corporal	2,739,730	0.055	0.228	0	1
Sergeant	2,739,730	0.043	0.202	0	1
Low-ranking officer	2,739,730	0.025	0.156	0	1
High-ranking officer	2,739,730	0.002	0.045	0	1
Musician	2,739,730	0.014	0.116	0	1
Low-ranking service personnel	2,739,730	0.006	0.078	0	1
High-ranking service personnel	2,739,730	0.005	0.071	0	1
Other	2,739,730	0.010	0.101	0	1
Unit type (at enlistment)					
Infantry	2,739,730	0.741	0.438	0	1
Cavalry	2,739,730	0.159	0.366	0	1
Artillery	2,739,730	0.076	0.265	0	1
Special (fighting)	2,739,730	0.003	0.051	0	1
Special (non-fighting)	2,739,730	0.006	0.076	0	1
Casualties					
Died	2,186,753	0.125	0.331	0	1
Died (combat)	2,186,753	0.045	0.207	0	1
Died (disease)	2,186,753	0.049	0.216	0	1
Died (other)	2,186,753	0.031	0.173	0	1
Disabled	2,160,456	0.095	0.293	0	1
Injured	2,739,719	0.060	0.237	0	1

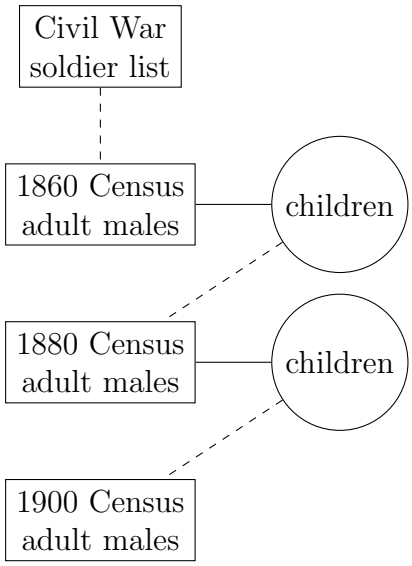
Notes: Summary statistics for the 2.7 millions Union Army Military Records. Records obviously corresponding to the same soldier were collapsed, but the total number of records remains larger to the total number of soldiers in the Union Army because of reenlistment.

3.1 Linking Censuses and Military Records

Linking the military records to the 1860 Census allows us to identify fathers who fought in the Civil War, as well as their children who we then track into later Census years. Tracking children whose fathers fought, comparing those who lost their father to those who did not, limits the problem of selection into the Union Army.

To build an inter-generational dataset on the family members of the Civil War soldiers, we proceed in the following way: 1) we start by linking the Union Army military records with the U.S. population Census of 1860; 2) we then link men younger than 20 in 1860 to the 1880 Census; 3) our final dataset consists of children observed in 1860 and linked to the 1880 Census who had a father linked to the Union Army Records. We also try to link the grandchildren of Union Army soldiers, observed in 1880 in the household of the son of a Union Army soldier, to the 1900 Census, but we end up with a very small sample size. Figure 2 provides a schematic of this linking procedure.

Figure 2: Record Linkage Schematic



Note: Record linkage schematic showing the links created between military and Census records. Dashed lines are generated links, solid lines are available links inside a given data set.

Several algorithmic methods for linking historical records exist and recently a best practice guide has been established on how to proceed with such linkage algorithms in a systematic way (Abramitzky et al., 2019). Bailey et al. (2017) compare the performance of several of these methods with respect to the percentage of links generated and the

associated type-I error (percentage of wrong links). We can think of the choice of these methods as presenting a trade-off between statistical power and accuracy. More restrictive methods produce more accurate, but fewer matches. In this paper, we use the simple algorithm of Ferrie (1996), also used by Abramitzky et al. (2012) and Abramitzky et al. (2014): we link individuals exactly on first name, last name, and state of birth.[15] We then keep links that have an absolute birth year difference of < 2 years. In case of multiple links, we keep the link with the smallest age difference. If this does not resolve ties, we decide we cannot link the record. If there are multiple possible links in a +/- 2 year window, we decide we cannot link the record. We do not use phonetic name cleaning like Soundex and NYSIIS because Bailey et al. (2017) show that this tend to increase false link rates.

Linking Union Army records to the 1860 census is complicated by missing information, notably on birth year, which is missing in 59% of cases. We start with a sample of men aged 10-60 in 1860 and we find all the records in the soldier list that match exactly on first name, last name, and state of residence or enlistment, and we start by excluding all links with a birth year difference larger than 5 years.[16] If birth year is never missing for any of the potential links, we proceed exactly as above — we exclude links with an age difference larger than 2 years, keep the link with the smallest age difference, and decide we cannot keep the record if this does not resolve ties.[17] In cases where the birth year is missing for some of the potential links, we use the distance between counties of residence to infer the correct link.[18] We then exclude all links with a birth year difference ≥ 2 years (when the soldier birth year is known). If a tie is still unresolved, the record is excluded. If a soldier is linked to more than one record in the 1860 Census, we exclude all links using this soldier.

[15]For linking Army records to the 1860 census, we use state of residence, see below.

[16]When we do not know the state of residence of the soldier (63% of cases), we use the state of enlistment, when we do not know state of enlistment (37% of cases), we use the state of service (for example Massachusetts for the 22nd MA Volunteer Infantry Regiment). This is probably innocuous because recruitment in the Union Army was local.

[17]We do not exclude multiple possible links in a +/- 2 year window like in the census to census linking because of re-enlistment and the fact some soldiers appear twice in the Union Army records.

[18]We were able to geolocate the county of residence in the UA records in 33.6% of cases. In 22.6% of cases, the county of residence is missing, but we were able to geolocate the county of enlistment (recruitment in the UA Army was very local, so that soldiers usually enlisted very close to their place of residence). We compute the geodesic distance between the county of residence in the 1860 census and the county of residence/enlistment in the UA records and we keep the links with the smallest county distance. Finally, even if county of residence is missing for some of the potential links, we keep links with a county distance of zero (same county).

Bailey et al. (2017) show that Ferrie-type algorithms perform well in terms of minimizing the type I error when we restrict potential links to rare names. Unfortunately, such a restriction considerably reduces our final sample size and the associated statistical power. However, we show the robustness of our main results to Ferrie-type linking with an uncommon name restriction, considering only individuals whose combination of first and last name appears less than 10 times in the Union and border states in the fighting generation (men aged 13-45 in 1860).[19] We also show the robustness of our main results to varying the birth year difference threshold, excluding a link if there are other possible links in a 5 year age window (instead of 2).

3.2 Final dataset

Linking military records to the 1860 Census produced 482,983 links (17.6% linkage rate).[20] We have information on survival and disability for 435,626 of them. 14.92% died during the war, 10.58% returned home with a major disability.

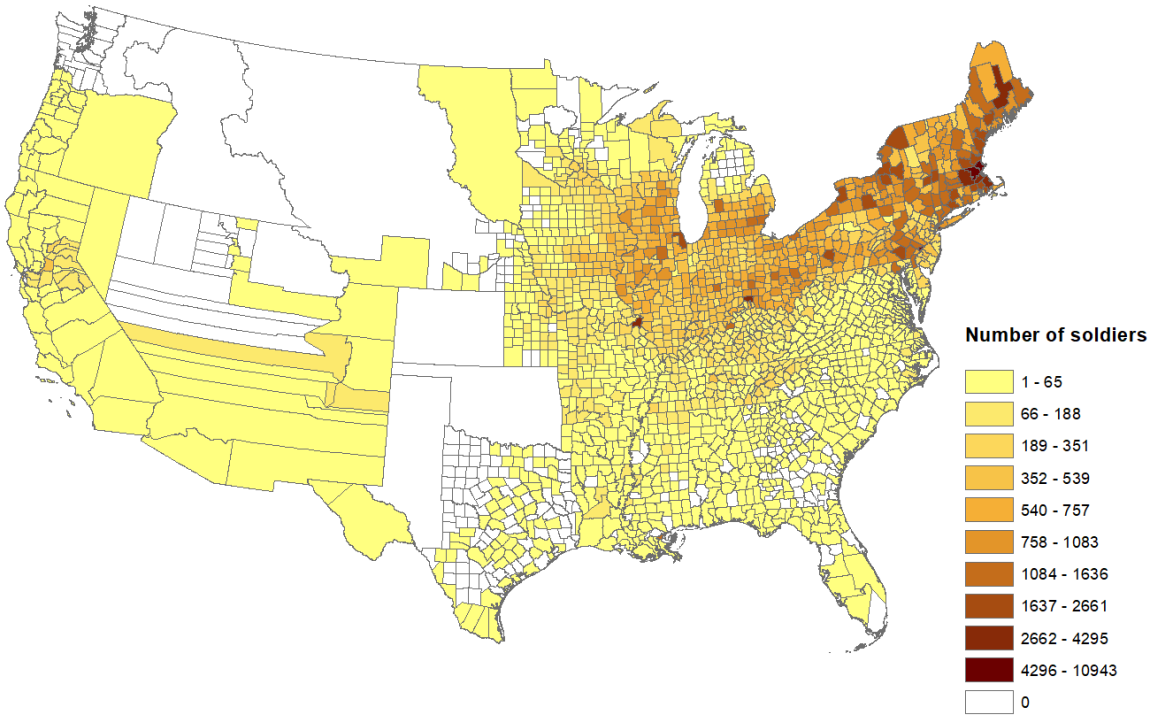
Figure 3 shows the geographical distribution (by 1860 county) of the soldiers in our dataset. This corresponds to the geography of enlistment in the Union Army: the vast majority of Union Army soldiers came from the Northeastern states and the Great Lakes region. Some recruits were residing in the Southern states and came to enlist in the North, but they are a very small minority. Finally, there are very few recruits in the newly settled territories of the West and the Pacific coast. It is worth noting that the Civil War was fought predominantly in the Southern and border states (appendix figure B.2), which allows us to estimate the effect of parental loss separately from other consequences of wars such as spreading disease of passing soldiers or capital destruction (Feigenbaum et al., 2018). Figure 4 displays the geographical distribution of death rates. There is no obvious geographical pattern in the distribution of death rates across counties — counties were few soldiers enlisted naturally display a larger variation in death rates.

Linking men younger than 20 in the 1860 Census to the 1880 Census produces 1,118,277 links (15.6% linkage rate). Combining the linked soldier file to the linked

[19] We do not consider Southern states because our sample is composed practically exclusively of individuals residing in a Union or border state in 1860. Men aged 13 in 1860 were aged 18 in 1865, at the end of the war.

[20] We obtain the figure of 17.6% by dividing the number of links by the number of Union Army records (2,739,730). if we divide by the number of soldiers instead (2.2 millions), we obtain a higher linking rate of 22%.

Figure 3: Geographic Distribution of Soldiers Linked to the 1860 Census



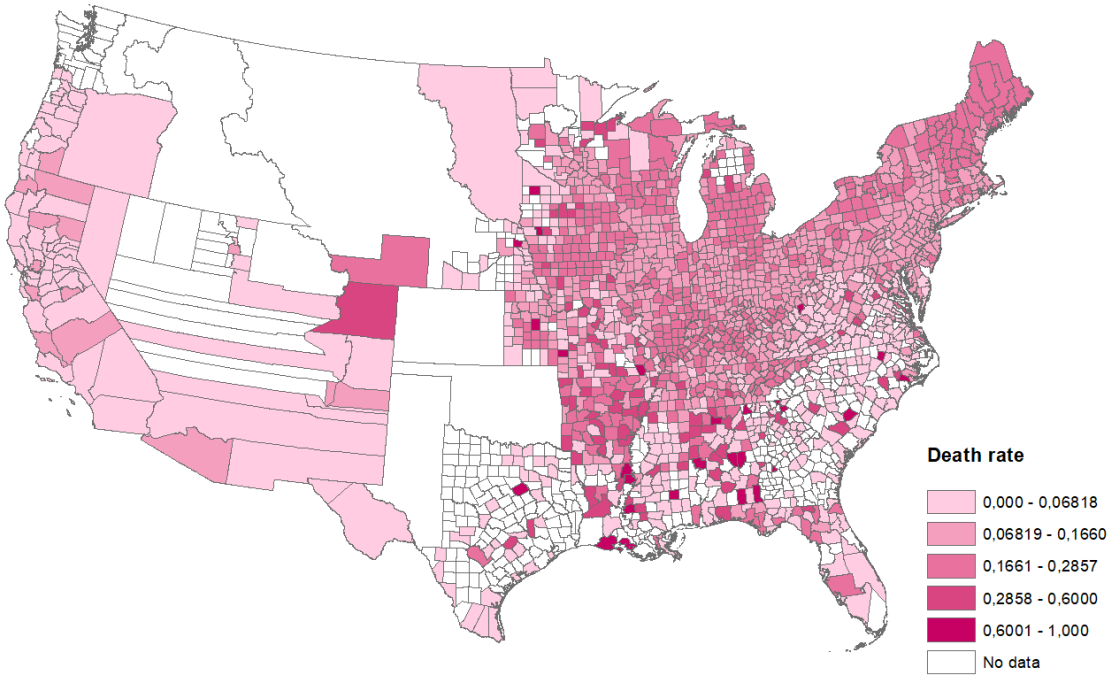
Censuses file, we obtain a sample of 32,240 men observed in 1860 and 1880 whose father fought in the War. 13.27% lost their father during the war, and 11.5% had a father who came back from the war with a severe disability.[21]

To investigate the long-run intergenerational effects of the war, we use the socio-economic variables available in the U.S. Censuses of 1860 and 1880. Variables in the 1860 Census are used as baseline controls, while variables in the 1880 Census are used as outcomes. Our main variable of interest is a measure of income which we derive from an individual's occupation. Occupation is provided as a string variable in the Censuses and we follow the classification of IPUMS USA to build occupation categorical variables and the occupational income score. The occupational income assigns the median income in 1950 of a given occupation to an individual as proxy for income. We normalize it to have mean zero and unit variance.

The occupational indicator variables are: 1) a dummy for having a high skilled occupation (professional, technical, manager, officials, and proprietors), 2) a dummy for having a semi-skilled occupation (sales, craftsmen, operatives), 3) a dummy for having a

[21]The original sample size is 37,560, but we lack information on death and disability for about 11% of fathers who fought.

Figure 4: Death Rates by County



low-skilled occupation (service workers, laborers, including farm laborers), 4) a dummy for being a farmer. In the generation of fathers, farmer is by far the most common category: 40.1% of the soldiers-fathers in our database were farmers in 1860, 15% had a low skilled occupation. 25.5% of their sons were farmers, 28.3% were low-skilled workers. The "farmer" category encompasses many different situations: while the vast majority of farmers were poor, some of them likely had large farms and a high income.[22] This is the main reason why we do not simply rely on occupational income score and present results on the probability to be part of each category. In 1860 only, the Census provides information on real estate and personal wealth measured in dollars. Because wealth tends to be log-normally distributed and because a large number of soldier-fathers (16.9%) had no wealth, we take the inverse hyperbolic sine transform of wealth (Friedline et al., 2015).[23] Literacy is given in 1860 — it is already very high, as 94.10% of soldier-fathers can read. We can also measure cross-county migration in 1880 (whether the individual's county of residence changed between 1860 and 1880), as well as marital status.

[22]The IPUMS USA 1950 occupational income score puts farmers at 14, while the average low-skilled occupational income score is 14.4.

[23]The inverse hyperbolic sine transform of y is $\log(y + \sqrt{y^2 + 1})$. Except for small values of y, the inverse hyperbolic sine is approximately equal to $\log(2y)$, so that it can be interpreted as a log variable.

To what extent is our sample of fathers different from the universe of fathers in the U.S. population in 1860? There are two main reasons why they would be different: 1) selection into the Union Army (for example, our sample obviously massively over-represents the North of the country, even though some men from the South enlisted in the Union Army), 2) selection due to the linking by name. It is much easier to link rare first name-last name combinations, and people with rare names tend to be, on average, more educated, richer, and more often born abroad (Bailey et al., 2017).

Table 2: Final sample of fathers: summary statistics and representativeness

	(1) All free men 10-60 in 18600	(2) UA soldiers linked to 1860 csus	(3) Diff.	(4) All free fathers in 1860	(5) Final sample of father	(6) Diff.
Age	28.26	23.92	-4.54***	39.26	35.50	-3.80***
			(0.02)			(0.06)
Born abroad	0.21	0.15	-0.07***	0.25	0.18	-0.07***
			(0.00)			(0.00)
Non white	0.02	0.01	-0.01***	0.01	0.00	-0.01***
			(0.00)			(0.00)
Illiterate	0.05	0.03	-0.02***	0.08	0.05	-0.03***
			(0.00)			(0.00)
Wealth	1,717	671	-1,097***	3,671	2,696	-984
			(117)			(755)
Occ. score	12.04	11.71	-0.34***	17.75	17.77	0.02
			(0.02)			(0.07)
High-skilled	0.05	0.04	-0.01***	0.08	0.07	-0.01***
			(0.00)			(0.00)
Low-skilled	0.17	0.21	0.04***	0.15	0.16	0.00*
			(0.00)			(0.00)
Semi-skilled	0.18	0.21	0.03***	0.25	0.29	0.04***
			(0.00)			(0.00)
Farmer	0.22	0.15	-0.07***	0.43	0.38	-0.05***
			(0.00)			(0.00)
Observations	9,110,945	434,716	9,110,945	3,098,802	27,762	3,098,802

Notes: the final samples are soldiers/fathers with non-missing survival information. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2 compares the universe of free men aged 10 to 60 in 1860 to our sample of linked soldiers (columns 1 and 2, difference in 3) and the universe of free fathers in 1860 to our sample of soldier-fathers (columns 4 and 5, and difference in 6). Men in our sample are younger, less likely to be born abroad, more likely to be white, more literate, have lower wealth, are less likely to be high-skilled or farmers, and more likely to be semi- or low-skilled. These differences are very comparable if we exclude the Southern states,

where only very few men enlisted in the Union Army (appendix table B.1).

Sample selection is not a problem for our identification strategy, because we never compare the sons of fathers in our linked sample to the son of fathers in the unlinked population. However, it might be a concern for external validity, especially in the presence of very heterogeneous effects. To alleviate this concern, we create customized weights following the method of (Bailey et al., 2019): on the whole sample of fathers in the 1860 census, we create a variable l_j equal to j 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.[24] This gives us, for each father in the 1860 census, a probability $\hat{p}$ to be in the final sample predicted from observable. Appendix figure B.3 displays the kernel density of this predicted probability for fathers in the final sample and not in the final sample. As expected, fathers not in the final sample have, on average, a lower predicted probability to be linked, but the two distribution have a fairly large common support, which means that we can re-weight fathers in the final sample to be more representative of the broader population of fathers (Bailey et al.). When 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. We show below that weighted and unweighted results are very similar.

4 Result of fixed effect estimation

Two main concerns arise when trying to estimate the effect of military death on the socioeconomic outcomes of the son: selection into the army (men who fought might be different from men who did not fight), and selection into death given enlistment in the army (men who died might be different from men who returned).

Tables 2 and B.1 show that men in the Union Army differed from the general population. To solve the problem of selection into the army, we restrict the analysis to children of men who fought in the Union Army, comparing the children of those who fought and died to the children of those who fought and returned home. This means our effect is estimated on a selected sample of families that are not necessarily representative of the general population, which is why we also present results of estimations using customized weights to make the sample more representative of the 1860 US population (Bailey et al.,

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

2019).

Given participation in the conflict, there is of course no reason to think that the probability of dying was as good as random. The two main causes of death in the war were death in battle (45% of all death in the military records) and death of disease (49%). In military camps with poor sanitation, epidemics of dysentery, typhoid, pneumonia and malaria spread and killed a large number of soldiers. Both types of death can be correlated with a soldier's socioeconomic status, if only because military ranks and roles in the army are partly determined by education and occupation.

Fortunately, our linked dataset allows us to control for a large set of father characteristics, both military characteristics coming from the Union Army records (rank, type of military unit, date of enlistment) and socioeconomic characteristics coming from the 1860 Census. In a first empirical strategy, we regress a son's socioeconomic status on whether his father died, controlling for a rich set of father socioeconomic controls, father military controls, and county fixed effects. To show that this strategy captures potential selection into death, we show the invariability of our main estimates to adding additional geographic military unit, and first and last name fixed effects capturing unobserved socioeconomic status.

We estimate the following equation by OLS,

$$y_{ijc,1880} = \beta_0 + \beta_1 died_j + x'_{j,1860}\theta_1 + m'_j\theta_2 + s'_{i,1860}\theta_3 + \alpha_{c,1860} + \varepsilon_{ij} \quad (1)$$

where $y_{ijc,1880}$ is a socioeconomic outcome (e.g. occupational income score) observed in the 1880 census for individual i, whose father is j. County of residence of father and son in the 1860 census is indexed by c. The treatment $died_j$ is an indicator equal to 1 if j (the father of i) died in the war. The vector $x_{j,1860}$ includes baseline controls for father j measured in the 1860 census: age and age squared, a nonwhite indicator, occupational income score, the inverse hyperbolic sine of wealth, literacy, and being foreign born. The vector also contains the same variable for the mother observed in 1860.[25] The vector $m'_{j,1860}$ contains military variables for father j observed in the Union Army Records: enlistment date, enlistment rank fixed effects, and regiment type fixed effects.[26] The

[25]When the mother is missing (in 15% of cases), we set these controls to zero and we also control for a dummy equal to 1 if the mother is missing.

[26]The ranks are: private, musician, low-ranking service personnel, corporal, sergeant, high-ranking service personnel, low-ranking officer, high-ranking officer. The regiment types are: infantry, cavalry, artillery, specialized (fighting) and specialized (non-fighting). 5.6% of fathers in our sample finished the

vector $s'_{i,1860}$ contains 1860 baseline controls for son i: age, age squared and literacy. County of residence fixed effects are absorbed by the vector α_c. Therefore, identification does not come from the comparison of counties with different death rates but from the comparison, within the same county of residence in 1860, of the sons of fathers who fought and died to the sons of fathers who fought and returned home.[27] Some fathers have several sons, so that standard errors should at least be clustered by father, but we cluster by regiment to have the same level of clustering as in the IV estimation below.[28]

Table 3 displays the effect of losing a father on socioeconomic characteristics of sons estimated with equation (1). Losing a father during the Civil War decreases a son's occupational income score in 1880 by 4% of a standard deviation on average. Men who lost their father during the war are 2.1 percentage points less likely to be semi skilled, 1 percentage points more likely to be low skilled, and 1.6 percentage points more likely to be farmers. We find no effect on the probability to have migrated between 1860 and 1880 (that is, to live in a different county in 1880 than in 1860). Sons who lost their father are 1.6 percentage points more likely to be married. The average age of the sons of Civil War soldiers in 1880 is 25, so that this increase in the probability to be married should not be interpreted as an overall increase in the probability of ever marrying, but rather as a decrease in age at marriage for sons who lost their father during the war.

The estimates are robust to the inclusion of additional dimensions of fixed effects to capture unobserved characteristics of fathers. Table 4 shows that the baseline result barely changes when we replace the county fixed effects with more demanding geographical fixed effects like township/ward fixed effects (9,534 different townships/wards in the sample) or post office fixed effects (8,051 different post office areas in the sample) — columns (2) and (3). Likewise, the baseline estimate is barely affected when we include, along with county fixed effects, military unit fixed effects for the 2,413 regiments and 12,992 companies — columns 4 and 5. Finally, in columns (6) and (7) we also control for last name fixed effects (8,925 different last names in the sample) and first name fixed effects (1,019 different first names). Last names and first names capture unobserved

war in a different regiment that the one they enlisted in. Regiment type is the regiment type of the last regiment of service (the one that mattered for whether the soldier survived or not). Controlling for regiment types for the first regiment does not affect results (not reported).

[27] There are 1,166 counties in the original dataset. Including county fixed effects results in the exclusion of 165 counties with a single observation.

[28] We observe more than one son for 16.7% of fathers in our dataset.

Table 3: Effect of father death on socioeconomic characteristics of sons in 1880

	(1) occupational score (normalized)	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.040***	-0.003	-0.021***	0.010	0.016**	-0.002	0.016**
	(0.015)	(0.005)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Mean dep. var.		0.091	0.309	0.283	0.242	0.561	0.468
Observations	32,121	32,121	32,121	32,121	32,121	32,121	31,378
Own controls	✓	✓	✓	✓	✓	✓	✓
Father mil. controls	✓	✓	✓	✓	✓	✓	✓
Father other controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

socioeconomic status (if only because they give information on country of origin and assimilation) — on last name and socioeconomic status, see for example Clark and Cummins (2015), on first names and social status, see for example Lieberson (2000). In the bottom two line of table 4, we follow the recommendation of Pei et al. (2018) and test the equality of the coefficients of the baseline and augmented models by estimating them jointly in a seemingly unrelated regression system. None of the coefficients in columns (2)-(7) are statistically different from the baseline coefficient in table (1).

In table D.3, we estimate equation (1) weighting observations by customized weights built to make the sample of fathers representative of the population of fathers in 1860 (Bailey et al., 2019). Weighted estimates are very similar to unweighted ones. Table D.4 displays the effect of father death on occupational score estimated on samples obtained using different linking techniques. Finally, to provide a robustness check with respect to issues of selection on observables and functional form, we apply the post-double machine learning selection algorithm by Belloni et al. (2014). The algorithm takes all controls, their squares, and cross-term interactions including all fixed effects, and uses the LASSO to select significant predictors of the treatment and of the outcome. The original regression is then run again including the union of selected controls in the previous two LASSO selection steps. The algorithm potentially improves inference by excluding irrelevant controls while reducing potential biases from misspecified functional forms in the set

of controls. Results are reported in the appendix in table D.5 and show that the baseline results remain unchanged.

Table 4: Robustness of results to various dimensions of fixed effects

	(1) occ. score	(2) occ. score	(3) occ. score	(4) occ. score	(5) occ. score	(6) occ. score	(7) occ. score
Father died	-0.040***	-0.033*	-0.040**	-0.032**	-0.042**	-0.045**	-0.037**
	(0.015)	(0.020)	(0.020)	(0.016)	(0.020)	(0.019)	(0.015)
Observations	32,121	32,121	31,552	31,816	27,018	28,064	31,715
R^2	0.14	0.40	0.34	0.21	0.44	0.31	0.16
County F.E.	✓			✓	✓	✓	✓
Town F.E.		✓					
Post office F.E.			✓				
Regiment F.E.				✓			
Company F.E.					✓		
Last name F.E.						✓	
First name F.E.							✓
PPS test							
χ^2		0.34	0.00	1.92	0.36	0.14	0.52
p-value		0.56	1.00	0.17	0.55	0.71	0.47

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns (2)-(7): we estimate the model jointly with the model of column (1) and we test the equality of the effect of father death. The two bottom lines of the table give the χ^2 statistic and associated p-value. The effects of father death in the augmented models (with additional dimensions of fixed effects) is never statistically different from the effect in the baseline model.

4.1 Economic Significance

The dimension of the associated wage drop of orphans can be compared to the estimated human capital costs of the war provided by Goldin and Lewis (1975). In our sample, 25% of soldiers had children in 1860. Given the mortality rate in the Union Army, assuming that father and non-father soldiers died at the same rate, and that fathers had on average two children, the aggregate loss in orphans' life-time incomes can be computed as \$112M in 1860 value.[29] This is equivalent to removing a little more than an entire year worth of wages from each orphan's life-time earnings. The corresponding value in

[29]Using the log of the occupational income score on the left-hand-side of equation (1), we find that father death decreases occupational income by 2.2%. The average male income in 1860 was \$546 (Long, 1960, table 47) and the compounded loss of 2.2% of these earnings (\$12) leads to a lifetime loss of \$638 per orphan, assuming a fifty year work-life and an interest rate of 6%.

2019 dollars is $3.2bn.[30]

This compares to the $954M in estimated human capital losses due to the deaths of soldiers which were computed by Goldin and Lewis (1975). Their estimate did not take into account intergenerational effects but instead is based on the lost life-time earnings of deceased soldiers. Adding our estimate of lost orphan life-time earnings, this would increase the Goldin and Lewis estimates by almost 12%. However, our result for costs of the war stemming from the intergenerational effect is a lower-bound estimate since we show later that losing brothers, other family members, and men in the local community imposed additional wage penalties. This suggests that the negative effects of the Civil War from lost human capital and earnings may be significantly higher than what was previously estimated when intergenerational spillover effects are taken into account.

5 IV estimation

One concern regarding estimation of 1 via fixed effects regression is that unobserved confounders remain in the error term that both affect the probability of a father dying as well as their sons' later life outcomes. For example, if fathers of poor health both transmit this condition to their sons, thus reducing their future income, but also are more likely to die because of it during the war, we would underestimate the effect of paternal deaths on sons' outcomes in 1880.

To estimate the causal effect of paternal death on a son's socioeconomic status, we use the death rate of the father's regiment as an instrument. Our identification strategy takes advantage of the high within regiment correlation of death rates. The high within-regiment correlation of death rate is explained by the fact that regiments fought and camped together. The fact they fought together explains the high within regiment correlation of combat death rate. The fact they camped together explains the high within regiment correlation of disease death rates, as epidemics such as dysentery, typhoid, pneumonia and malaria spread from soldier to soldier in military camps.

Because regiments were recruited locally, there was also within regiment correlation in socioeconomic characteristics of soldiers (Costa and Kahn, 2008). For this reason, it is important to keep controlling for county fixed-effects: our estimation strategy consists in

[30]The inflation adjustment was based on the estimated consumer price index by the Minneapolis Fed for 1860 which is available at: `https://www.minneapolisfed.org/community/financial-and-economic-education/cpi-calculator-information/consumer-price-index-1800`

comparing, within the same county of residence in 1860, the sons of soldiers who fought in a regiment with a high death rate to the sons of soldiers who fought in a regiment with a lower death rate. We also control for a number of characteristics of regiments, including the average age, occupational income score and wealth in 1860 of the regiment's soldiers we were able to link to the 1860 census.

Our identifying assumption is that, conditional on regiment controls, the level of risk a father's regiment was exposed to, as measured by the regimental death rate, is orthogonal to son's unobservable characteristics that affect their outcomes in 1880 and that relate to a father's probability of dying in the war. We will provide some evidence for this below, after showing results for the first stage.

The regiment death rate we use as an instrument is always the "leave-one-out" regimental death rate. For each father-soldier j, the instrument is computed as:

$$deathrate_{jr} = \frac{\sum_{s \neq j} died_s}{N_r - 1} \tag{2}$$

where N_r is the number of men who served in regiment r.[31] These death rates are computed using the universe of Union Army records and not only the sample of fathers we were able to link to the census. In practice, because regiments are 1,000 men on average, the contribution of a single soldier to the regimental death rate is small. Figure 5 displays the distribution of regimental death rates in the sample of father-soldiers. The average regimental death rate is 12.9% and the standard deviation is 7.9 percentage points.[32]

5.1 First stage

In a first stage, we estimate the following equation:

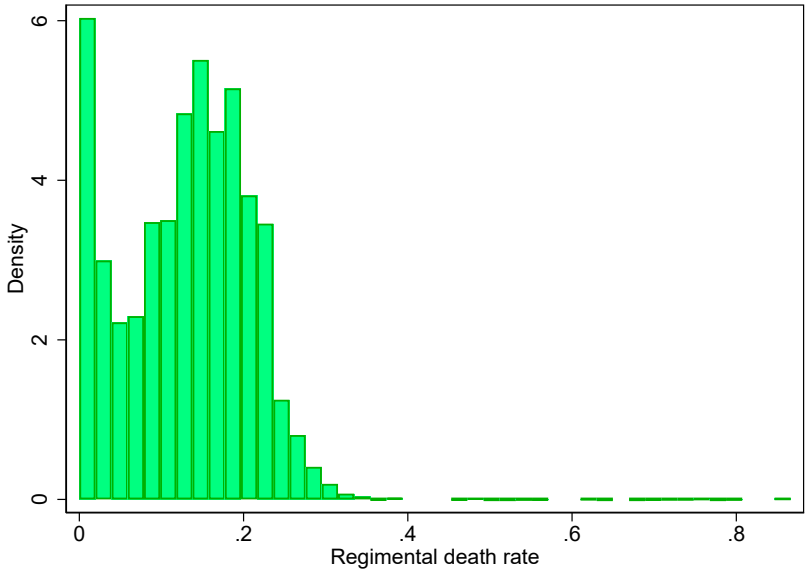
$$died_{ijr} = \gamma + \delta deathrate_{ijr} + z'_r\lambda_1 + x'_{j,1860}\lambda_2 + m'_j\lambda_3 + s'_{i,1860}\lambda_4 + \alpha_{c,1860} + \nu_{ijr} \tag{3}$$

where $died_{ijr}$ is an indicator equal to 1 if j, the father of i died during the war; $deathrate_{ijr}$ is the death rate of father j's last regiment of service r (indexed by j as

[31] It is actually the number of men who served in regiment r for whom we have information on survival.

[32] We can obtain two additional instruments by computing the regiment combat and disease death rate, though these do not sum to the overall death rate because the cause of death is missing for some soldiers. Though these instruments capture exposure to two different kinds of risk (epidemiological risk and battle risk), we show below that they give broadly similar results.

Figure 5: Distribution of regimental death rate in the sample of father-soldiers



well as r because it is the "leave-one-out" rate). z'_r is a vector of regiment controls computed from the Union Army record data: regiment type dummy, share of privates in the regiment,[33] average enlistment date, socio-economic characteristics of the regiment computed from information on counties of enlistment of soldiers.[34] Regiment level controls computed from our sample of linked soldiers include the average age, occupational income score, wealth and share of foreign-born soldiers.[35] x, m and s are the same family and son controls as in equation 1. $\alpha_{c,1860}$ are county of residence fixed effects, which means that we are comparing soldiers from the same county who joined regiments that were exposed to different levels of risk during the war. Though Civil War recruitment was local, there is within-county variation in regiments, with an average of 31 different regiments per county.

Table 5 presents the result of the first stage for the overall death rate. A one-percentage point increase in the "leave-one-out" death rate of a father's regiment increases the probability to die by a bit more than one percentage point, 1.06 without any controls. In columns (2) to (7), controls are added progressively. The only control that

[33]This controls for administrative units which were primarily staffed with officers and that did not participate in combat.

[34]We link soldiers' residence counties to economic and population data from the 1860 county-level census. For each county level variable x_c, we compute the regimental average weighted by the number of soldiers belonging to each county. See appendix C for the list of county-level variables. We also control for the number of different counties of residence in the regiment.

[35]All these averages are "leave-one-out": they do not include the characteristic of father j.

reduces the relationship between death rate and the probability of dying is the polynomial in enlistment date. For mechanical reason, the date of enlistment is a very good predictor of both the probability to die and the average death rate in the regiment — the longer a regiment was active in the war, the higher the death rate. In our preferred specification (column 7), a one percentage point increase in regimental death rate increase the probability to die by 1.03 percentage points.[36] Appendix tables D.7 and D.8 present the result of the first stage using the regimental combat and disease death rates.

Table 5: First stage: regression of probability to die on regiment "leave-one-out" death rate

	Dependent variable: probability of dying						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Regimental death rate	1.061***	1.083***	1.081***	1.036***	1.037***	1.034***	1.032***
	(0.028)	(0.030)	(0.034)	(0.035)	(0.035)	(0.034)	(0.034)
State F.E.		✓	✓	✓	✓		
Regiment controls			✓	✓	✓	✓	✓
Enl. date poly				✓	✓	✓	✓
Enl. rank F.E.					✓	✓	✓
County F.E.						✓	✓
Father controls							✓
Mother controls							✓
Own controls							✓
Observations	31,460	31,460	31,460	31,460	31,460	31,460	31,460
F-stat	1,460.96	1,331.09	1,016.33	895.70	899.32	939.43	933.94

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The regimental death rate is the "leave-one-out" death rate of the father's last regiment.

Our identifying assumption is that, conditional on regiment controls, the level of risk a father's regiment was exposed to, as measured by the regimental death rate, is orthogonal to the son's unobservable characteristics that affect their outcomes in 1880 and that relate to a father's probability of dying in the war. In appendix C, we show that Union regiments that were on the front line during important battles were not different in their socioeconomic composition. We collected and digitized 128 battle maps from the Civil War Preservation Trust and we show that there is no correlation between a Union regiment's socioeconomic characteristics and its distance from the nearest enemy unit at various stages of the battle. This implies that regiments composed of poorer fathers

[36]This is a linear estimation of a non-linear relationship. Non-parametric estimations (not reported) reveal that the slope of the relationship between regiment death rate and probability to die is steeper for lower values of the death rate.

were not systematically placed in the front line, which would have induced a correlation between fathers current and therefore son's future socioeconomic status that would have posed a problem in terms of the exclusion restriction.

One way to test the identification assumption is to use fathers' observable characteristics in 1860, before the war, and to put them on the left hand side of equation 3. It consists in estimating the following equation:

$$y_{ijr} = \gamma + \delta deathrate_{ijr} + z_r'\lambda_1 + m_j'\lambda_3 + \alpha_{c,1860} + \nu_{ijr} \tag{4}$$

where y is a socioeconomic characteristic of soldier j, father of i, observed in 1860, before the war. If the identifying assumption holds, then there should be no statistically significant correlation between y_{ijr} and the instrument. Table 6 shows the results of estimating this equation for 10 variables measured in 1860 and for the three instruments (death rate, combat death rate, and disease death rate). The only statistically significant coefficient is the correlation between the combat death rate and the probability of being foreign born (significant at the 10% level). All our regressions control for whether the father is foreign born (and for all characteristics in table 6), and results are robust to simply excluding all foreign born father from the analysis (see appendix table D.11).

5.2 Results

Table 7 displays the effect of losing a father on socioeconomic outcomes of the son, instrumenting father death by regimental death rate. Losing a father decreases occupational income in adulthood by 28% of a standard deviation, decreases the probability to be high-skilled by 5 percentage points and the probability to be semi-skilled by 12 percentage points. The probability of being low-skilled increases by 10 percentage points and the probability to be a farmer by 3 percentage points. Compared to the OLS estimation, results are qualitatively similar (except for the probability of migrating between 1860 and 1880, which increases with father death in the IV estimation), but effect sizes are larger (7 times larger for the occupational income score). We'll explore below the reasons that might explain this discrepancy.

Appendix table D.6 shows that the 2sls results are insensitive to the inclusion of father socioeconomic controls. Column (1) estimates the effect of father death on the

Table 6: Balance test of the three instruments

	(1)	(2)	(3)
	Instrument regimental death rate	Instrument regimental combat death rate	Instrument regimental disease death rate
Foreign born	0.044	0.135*	-0.044
	(0.033)	(0.073)	(0.056)
Non white	0.009	0.005	0.001
	(0.009)	(0.015)	(0.011)
Age	0.097	0.003	0.052
	(0.079)	(0.177)	(0.159)
Occupational score	-0.111	-0.163	-0.046
	(0.098)	(0.194)	(0.171)
High-skilled	-0.039	-0.057	-0.024
	(0.026)	(0.051)	(0.042)
Semi-skilled	-0.034	-0.036	-0.029
	(0.045)	(0.093)	(0.077)
Low-skilled	0.030	0.067	0.081
	(0.038)	(0.079)	(0.063)
Farmer	0.051	0.001	-0.020
	(0.046)	(0.087)	(0.078)
Cannot read	-0.002	0.021	0.031
	(0.025)	(0.048)	(0.039)
Wealth (IHS)	-0.203	-0.726	-0.224
	(0.298)	(0.603)	(0.495)
Observations	31,460	31,460	31,460

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

occupational income score of the son without controlling for father socioeconomic characteristics; columns (1) to (8) add the controls one by one, and column (9) corresponds to our preferred specification with the full set of father controls. Though father controls are economically and statistically significant predictors of a son's occupational income score, they barely affect the estimated effect of father death. This, along with the balance test previously presented, reassures us on the validity of our instrumental strategy.

Our results are robust to alterations of the instrument when we consider using the combat death rate (table D.9) or the disease death rate in a regiment (table D.10) as instrument. Results are also not affected by the exclusion of foreign born soldiers (table D.11) or issues related to selection on observables and functional form (table D.12). We obtain similar results when weighting observations by customized weights to make the sample of fathers more representative of the general population of fathers (table D.13). We also obtain broadly similar results on samples constructed using different linking

techniques (table D.14).

Table 7: Effect of losing a father: results of IV estimation

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.283***	-0.050*	-0.117***	0.098**	0.028	0.091*	0.040
	(0.089)	(0.027)	(0.040)	(0.041)	(0.035)	(0.047)	(0.034)
Mean dep. var.		0.091	0.310	0.283	0.241	0.561	0.468
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,731
K-P F-stat	933.53	933.53	933.53	933.53	933.53	933.53	897.14

Notes: Standard errors clustered by last regiment of service of the father in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" death rate of the father's last regiment of service.

There are three potential explanations for the difference between the OLS and the IV estimation: heterogeneous treatment effects for the "compliers" of the instrument, measurement error in the treatment variable, and violation of the exclusion restriction.

5.3 Who are the compliers?

The "compliers" of our instrument are the children of fathers who died because they were exposed to a higher level of risk, and who would not have died otherwise. In this context, the "always takers" are the sons of fathers who would have died anyway, maybe because of very poor health, and the "never-takers" are the sons of fathers who were protected, for example high-ranking officers serving in an administrative unit. Defining who the never-takers, always-takers and compliers are is complicated by the fact that the instrument is not a binary variable. Appendix table D.15 presents descriptive statistics for three types of soldiers: those whose probability of dying as predicted by the first stage is higher than 14% (the median) and who died as well as those whose probability of dying is lower than 14% and who survived — the complier fathers; those whose probability of dying is higher than 14% but who survived — the never-taker fathers; and those whose probability of dying is lower than 14% but who died — the always-taker fathers.

Compared to the "non-compliers", the "compliers" are more likely to be private soldiers and volunteers, but less likely to be in an infantry regiment. They are more likely to be semi-skilled and less likely to be farmer and more likely to live in an urban county. We

will show below that the skill group of the father and the urbanization rate of the county of residence in 1860 are important dimensions of heterogeneity. Appendix table D.16 shows the result of estimating the OLS model of equation (1) separately on the sample of "compliers" and "non-compliers". The effect of father death is not statistically different from zero in the sample of "non-compliers". In the sample of "compliers", it is 1.6 times higher than in the full sample, at 6.3% of a standard deviation in occupational income score. Even though the IV estimate remains higher, this is evidence that treatment effect heterogeneity plays a role.

5.4 Measurement error in the treatment variable

A second explanation for why IV estimates are larger than OLS estimates is measurement error in the treatment variable (whether the father died). Because we are linking multiple sources, measurement error due to wrong links across data sets is a possibility. We should underline that father regimental death rate is probably measured with less error than father death: this is because errors are averaged out in the regimental death rate, and because, if we make an error and link a 1860 individual to the wrong soldier, it is possible we actually link him to the right regiment, since recruitment was local and we use county of residence as matching variable in the linking. In the classical case, measurement error in the treatment variable can be corrected by an instrument. However, measurement error in our case is non-classical since the treatment is binary. Non-classical measurement error in the treatment variable leads to an attenuation bias in OLS and an inflation bias in IV (Bingley and Martinello, 2017). We discuss this in more detail in appendix E, where we attempt to quantify the biases introduced by linkage errors which reverse the treatment status of sons. Denoting ν the share of false positives (children who did not lose their father but who were marked as fatherless), and η the share of false negatives (children who lost their father but were not marked as fatherless), the biases of OLS and IV are:

$$\text{plim}\widehat{\beta}_{OLS} = \beta(1 - \nu - \eta) \tag{5}$$

$$\text{plim}\widehat{\beta}_{IV} = \beta\frac{1}{1 - \nu - \eta} \tag{6}$$

which, for $\nu + \eta < 1$ leads to an attenuation bias for OLS and an inflation bias for IV. Bailey et al. (2017) find that the rate of linking error when using Ferrie's algorithm with common names is 30%. In the extreme case where a linking error always flips the treatment status of an individual, we have $\nu + \eta = 0.3$. In this case, the attenuation bias of OLS is 70% and the inflation bias of IV is 143%. This could explain 42% of the difference in the OLS and IV coefficients for the occupational income score.[37]

5.5 Disability and death of other men

Are IV estimates larger than OLS estimates because of a violation of the exclusion restriction? More precisely, is the death rate of the father's regiment affecting sons through a channel other than the death of the father? There are two main concerns here: the first is that the death rate of a father's regiment also affects fathers who survived through psychological trauma and/or physical disability — 10% of men in our dataset returned home with a disability. The second concern is that, because recruitment was local, the death rate of a father's regiment is also a predictor of the death of other family members and neighbors in the son's network.

In table 8, we estimate the effect of father disability along with father death. We use the death rate of the father's regiment to instrument for father death and the *rate of disabled men* in the father's regiment to instrument for father disability.[38] The effect of higher exposure to risk during the war does not seem to affect children through disability.

The effect of regimental death rates might also affect a son through the death of other men in his social network. Our linked dataset actually allows us to investigate this channel. We should first underline the fact that all our effects are estimated within counties, so that we are largely abstracting from the broader labor market effects of the demographic shock of the war. In our sample of 33,240 men who had a father in the Union Army, 2,769 also had a brother who fought (471 died), and 895 an other family member who fought (159 died).[39] All men in our sample had Union Army soldiers in what we call the neighborhood: the smallest geographical unit that can be identified in the 1860 census, combining information on the name of the post office and the name

[37] This comes from computing $1 - \frac{(1-0.3)\widehat{\beta}_{\text{IV}} - \frac{1}{1-0.3}\widehat{\beta}_{\text{OLS}}}{\widehat{\beta}_{\text{IV}} - \widehat{\beta}_{\text{OLS}}}$.

[38] It is also the "leave-one-out" disability rate.

[39] We define "other family member" as someone living in the same household but having a different last name than the child.

Table 8: Effect of losing a father: instrumenting for father death and father disability

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.310***	-0.060**	-0.121***	0.124***	0.008	0.060	0.032
	(0.100)	(0.030)	(0.046)	(0.047)	(0.040)	(0.052)	(0.038)
Father disabled	0.065	0.025	0.008	-0.063	0.047	0.074	0.019
	(0.110)	(0.033)	(0.053)	(0.047)	(0.040)	(0.058)	(0.044)
Mean dep. var.		0.091	0.310	0.283	0.241	0.561	0.468
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,731
K-P F-stat	171.94	171.94	171.94	171.94	171.94	171.94	172.06

Notes: Standard errors clustered by last regiment of service of the father in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The two instruments are the "leave-one-out" death rate and the "leave-one-out" disability rate of the father's last regiment of service.

of the township or ward. There are 11,705 neighborhoods in our data, with an average population in 1860 of 2,719, an average enlistment rate of 10% and an average death rate of 1.6%.[40]

In order to assess to what extend our instrument, the death rate in the father's regiment, also picks up the effect of the death of other men in an individual's network, we estimate the following equation:

$$\begin{aligned} y_{ijnc,1880} &= \beta_0 + \beta_1 died_j + \beta_2 died_i^{brother} + \beta_3 died_i^{other} + \beta_4 deathrate_i^{neighbors} \\ &+ \gamma_1 fought_i^{brother} + \gamma_2 fought_i^{other} + \gamma_3 enlrate_i^{neighbors} \\ &+ x_{ij}^{father\prime}\theta_1 + x_i^{brother\prime}\theta_2 + x_i^{other\prime}\theta_3 + z_n'\theta_4 + \alpha_c + u_{ijnc} \end{aligned} \tag{7}$$

where $y_{ijnc,188}$ is a socioeconomic outcome in 1880 for individual i, whose father is j, living in county c and neighborhood n in 1860; $died_j$ is an indicator equal to 1 if j died; $died_i^{brother}$ is the number of brothers of i who died in the war, $died_i^{other}$ the number of other family members, and $deathrate_i^{neighbors}$ the death rate among the neighbors of i.[41]

[40]Enlistment rates are the number of soldiers linked to the 1860 census in the neighborhood divided by the total adult male population, death rates are the number of soldiers who died divided by adult male population. These underestimate the actual enlistment and death rates because we could not link all soldiers.

[41]Only 10 individuals had two brothers who died, and none had more than one other family members who died, so these variables can be mostly interpreted as binary variables. $deathrate_i^{neighbors}$ is the number of deceased neighbors divided by the adult male population of the neighborhood (minus the

When we were considering fathers only, we did not need to control for selection in the Union Army because we were considering only the children of Union Army soldiers. But it is in practice impossible to compare only individuals who had the exact same number of brothers, other family members and neighbors in the Union Army, so that we need to control for the number of brothers, other family members and neighbors who fought.[42] s_i contains individual controls, like in equations (1) and (3); x_{ij}^{father} contains all father controls of equation (3): the individual and military father controls, as well as all the regiment controls; $x_i^{brother}$ and x_i^{other} contain all these controls for brothers and other family members who fought.[43] z_n is a vector of neighborhood controls containing socioeconomic and military controls averaged over all men who fought in the neighborhood, as well as total population of the neighborhood in 1860. Finally, α_c is a vector of county fixed effects, like in all previous specifications.

Table 9 displays the results of estimating equation (7). Though the effect of the loss of a father is reduced by about 8% once we take into account the death of other men, it remains large at 26% of a standard deviation in the occupational income score. The effect of losing a brother is noisily estimated, but very large, at 48% of a standard deviations. However, contrary to the loss of a father, the effect of losing a brother on occupational income score seems to come primarily from a large increase in the probability to be a farmer (20 percentage points). A likely explanation is that individuals who lose an older brother are more likely to take over the family farm.

The loss of other men in the son's neighborhood seems to matter as well, though results are again a bit noisy. A 1% increase in the death rate of neighbors decreases the probability to be semi-skilled by 3 percentage points and increases the probability to be low-skilled by 3.7 percentage points (significant at the 5% level) — the magnitudes are in reality lower since we underestimate the number of neighbors who died because we cannot link each and every soldier to the census.[44]

---

father, brothers and other family members).

[42] $enlrate_i^{neighbours}$ is the enlistment rate of the neighbors of i: the number of neighbors who fought divided by the adult male population in the neighborhood (minus the fathers, brothers and other family members.

[43] When several brothers or family members fought, these controls are averaged. When an individual had no brother or other family members who fought, we set the controls to zero and we take care of the selection by controlling for the number of brothers and family members who fought.

[44] We manage to link about 20% of all Union Army soldiers, so the magnitudes are in reality 5 times lower.

Table 9: Effect of losing other men

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.261***	-0.045*	-0.105***	0.085**	0.024	0.074	0.036
	(0.089)	(0.027)	(0.041)	(0.042)	(0.036)	(0.048)	(0.035)
# brothers died	-0.484*	-0.137	-0.100	0.030	0.202	-0.135	0.100
	(0.270)	(0.091)	(0.126)	(0.121)	(0.126)	(0.154)	(0.120)
# other family died	0.152	-0.046	0.028	0.246	-0.120	-0.144	-0.067
	(0.520)	(0.158)	(0.248)	(0.233)	(0.210)	(0.273)	(0.201)
% neighbors died	-0.029	-0.005	-0.029**	0.037**	-0.001	0.018	0.010
	(0.033)	(0.009)	(0.014)	(0.015)	(0.014)	(0.017)	(0.014)
# brothers fought	-0.008	0.008	-0.011	-0.003	0.005	-0.008	-0.045
	(0.060)	(0.019)	(0.029)	(0.031)	(0.032)	(0.036)	(0.030)
# other family fought	-0.067	-0.009	-0.055	0.040	-0.000	0.119**	0.035
	(0.143)	(0.039)	(0.053)	(0.055)	(0.045)	(0.056)	(0.047)
% neighbors fought	0.006	0.001	0.004*	-0.006**	0.000	-0.005*	-0.001
	(0.006)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
Observations	31,442	31,442	31,442	31,442	31,442	31,442	30,713
K-P f-stat	105.96	105.96	105.96	105.96	105.96	105.96	99.43
Own controls	✓	✓	✓	✓	✓	✓	✓
Father controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
Brother controls	✓	✓	✓	✓	✓	✓	✓
Other family controls	✓	✓	✓	✓	✓	✓	✓
Neighbor controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered by last regiment of service of the father in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The four instruments are the "leave-one-out" death rate of the father's last regiment of service, of the brother's last regiment of service (average if more than one brother fought, zero if no brother fought), of the other family member's last regiment of service (average if more than one brother fought, zero if no brother fought), and the average regimental death rate of neighbors who fought.

6 Discussion of mechanisms

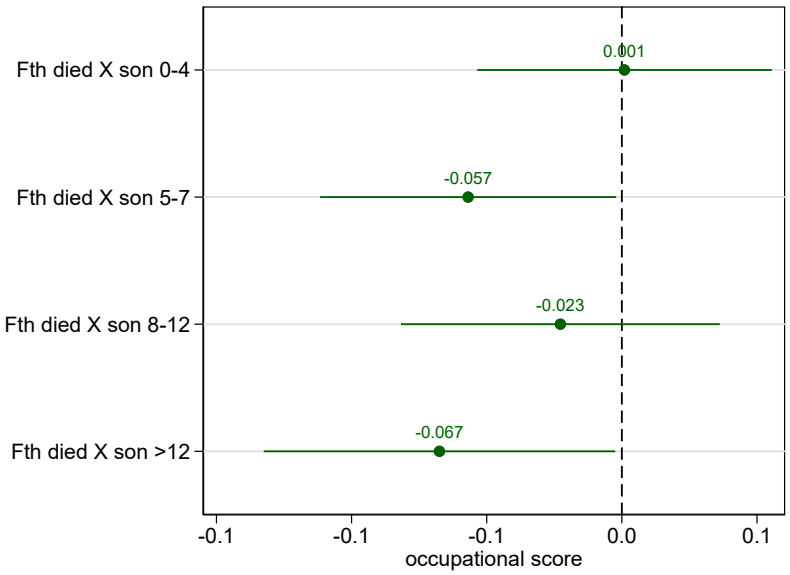
What are the mechanisms linking father absence to later life socioeconomic outcomes? Psychological trauma in childhood could play a role, as the first few years of life are crucial for skill development (Heckman et al., 2016).[45] The lack of paternal exposure and direct transmission of cultural and human capital from the father to the son could also play a role as shown by Kalil et al. (2016). Finally, the negative effect of paternal death on socioeconomic status could reflect an income effect: the missing father's income could

[45]The "fetal origin" literature, documenting the negative effects of in-utero nutrition and mother exposure to stress (Almond and Currie, 2011), is not directly relevant in our setting, since all individuals in our sample were born before the Civil War started.

prevent orphaned sons from accumulating human capital and training in school by forcing a sub-optimally early entry into the labor market.[46]

If psychological trauma in childhood or lack of paternal exposure are important channels explaining our results, we expect to find that the effect is predominantly concentrated at young ages (for paternal exposure, this is because children who lost their father later were exposed to the father longer). If the income effect is important, we expect effects to be heterogeneous with respect to baseline wealth, as family wealth at baseline provides financial protection against the negative income shock.

Figure 6: Heterogeneity by age at father enlistment, occupational income score



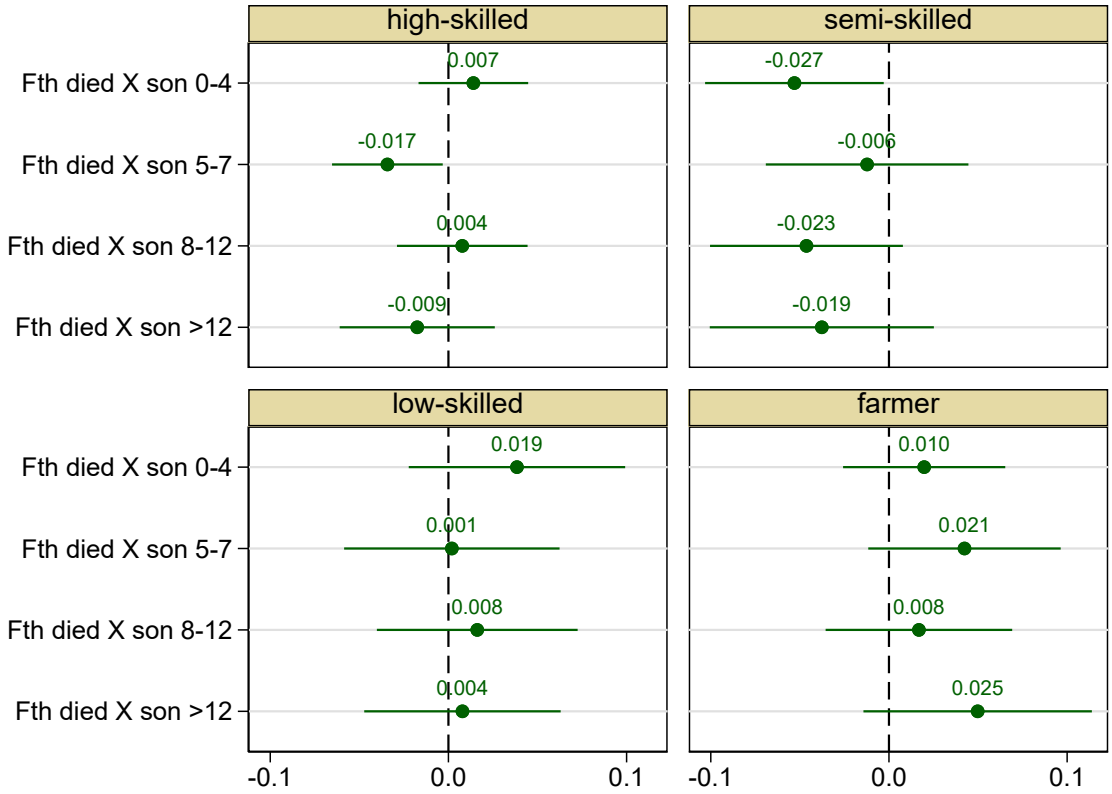
Note: occupational income score normalized to have mean zero and unit variance. OLS estimation (equation 1). Father death is interacted with the four quartiles of the distribution of age at father enlistment. Lines represent 95% confidence intervals.

We find mixed evidence regarding heterogeneity with respect to age at father enlistment. Figure 6 and 7 present the results of estimating the baseline model of equation (1), interacting the dummy for father death with each quartile of age at father enlistment. The effect does not primarily come from younger children, and even appears to be stronger for children older than 12, though results for each skill dummy reveal that this is mostly coming from an increase in the probability to be a farmer — maybe as older

[46]We were interested in whether mothers had to start to work and therefore might have lacked the time to care for their children. Due to surname changes upon marriage we were not able to construct a large enough sample to investigate this channel. Related to this is the replacement of deceased fathers with a new husband if the mother remarries.

sons who lose a father are more likely to take on the family farm. When we consider heterogeneous effect in the IV (appendix table F.3), we do find that effects are concentrated at younger ages, especially for the probability to be semi-skilled and to be low-skilled. The discrepancy between the two results is explained by heterogeneous effects on the compliers and non-compliers: appendix table F.4 estimates the interacted OLS model on the subset of "compliers" only (see section 5.3 above): results as qualitatively similar to IV results.

Figure 7: Heterogeneity by age at father enlistment, skill dummies



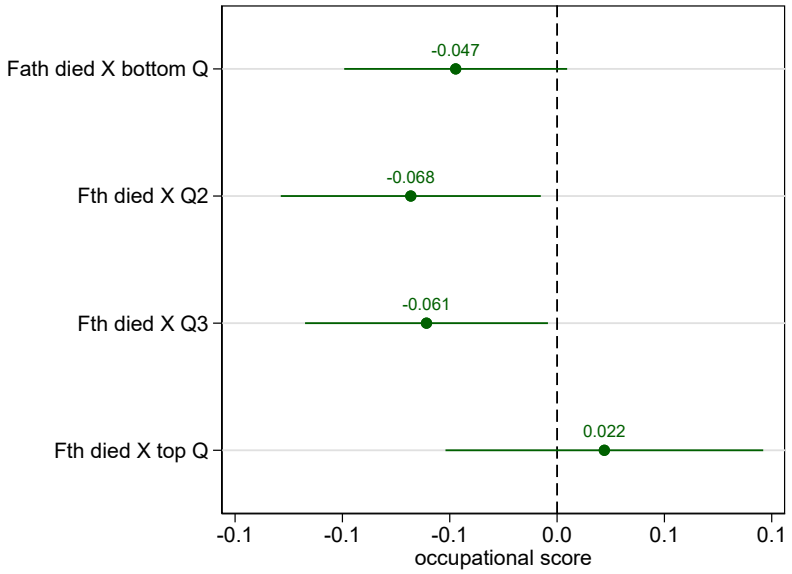
Note: OLS estimation (equation 1). Father death is interacted with the four quartiles of the distribution of age at father enlistment. Lines represent 95% confidence intervals.

We find more evidence in support of the income channel. Figure 8 interacts the dummy for father death with each quartile of the 1860 father wealth distribution. The effect of father death is similar in the three bottom quartiles of father wealth distribution, but there is no effect on the occupational income score in the top of the distribution.[47] Figure

[47]The effect in the top of the distribution is statistically different from the effect in the third quartile (p-values=0.08) and in the second quartile (p-value=0.07), but not from the effect in the bottom quartile (p-value=0.14).

9 looks at the effect on skill dummies and reveals that, at the top of the distribution, children who lost their father during the war are 3 percentage points less likely to be high-skilled and 4 percentage points more likely to become farmers.

Figure 8: Heterogeneity by father wealth, occupational income score

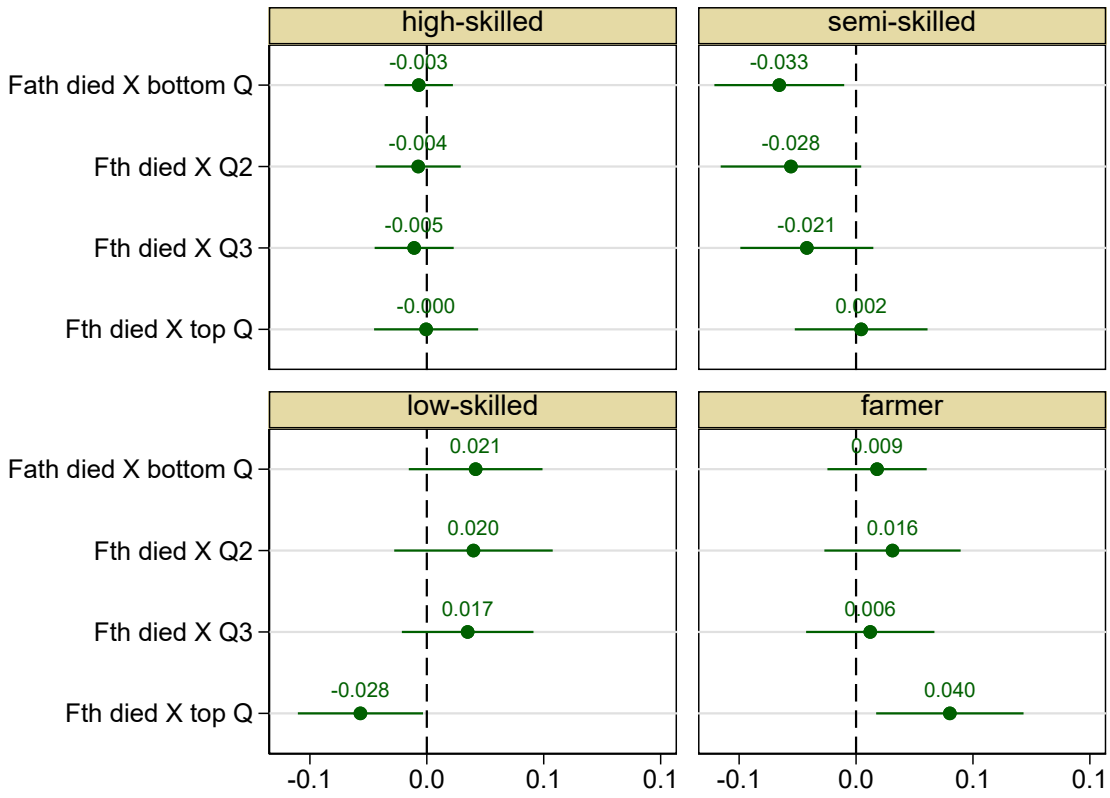


Note: occupational income score normalized to have mean zero and unit variance. OLS estimation (equation 1). Father death is interacted with the four quartiles of the distribution of father wealth in 1860. The effect in the top quartile is statistically different from the effect in the third quartile (p-value=0.08) and in the second quartile (p-value=0.07), but not from the effect in the first quartile (p-value=0.14). Lines represent 95% confidence intervals.

Finally, figure 10 and 11 look at heterogeneity by father skill category. The sons most economically affected by the loss of their father are the ones whose father was a semi-skilled worker, with an occupational income score lower by 9.1% of a standard deviation. In fact, the effect is largely driven by the downward mobility of the sons of semi-skilled fathers, who are 6.1 percentage points less likely to be semi-skilled workers and 3 percentage points more likely to be low-skilled workers. This is consistent with a mechanism where the son of semi-skilled workers, who, if their father had been alive, would have accumulated skills in school or through an apprenticeship with their father, but who had to start working earlier to compensate for the loss of father income.

If the income effect channel is indeed the most important one, then was the negative effect of father loss on socioeconomic outcomes partly compensated by receiving a pension? The Union Army pension program was America's first social insurance program

Figure 9: Heterogeneity by father wealth, skill dummies



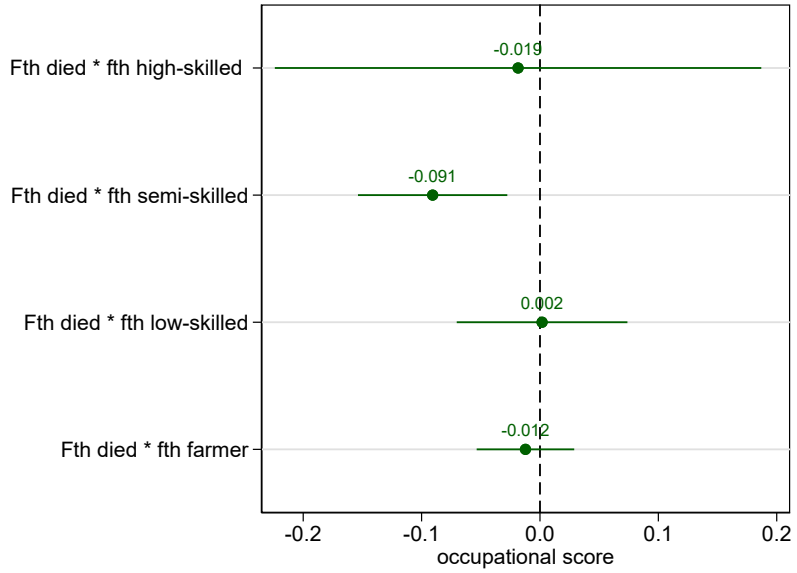
Note: OLS estimation (equation 1). Father death is interacted with the four quartiles of the distribution of father wealth in 1860. Lines represent 95% confidence intervals.

(Skocpol, 1992; Costa, 1995, 1997; Salisbury, 2017). Though it slowly became a universal disability and old-age pension program for veterans of the Union Army, it began as a restricted program compensating soldiers disabled by the war as well as the relatives of deceased soldiers.

Compensation amounts were determined by a series of laws passed between 1862 and 1873. Pensions for disability depended on rank and type of injury. Originally a private totally disabled for manual labor received $8 per month, and widows of deceased soldiers received the same amount. This represented less than 1/2 of the monthly income of a farm laborer (Salisbury, 2017). If a widow remarried, the pension was given to the minor children (younger than 16) of the deceased soldier (Salisbury, 2017).

Pensions were gradually increased in the decades following the war. In 1870, the average monthly pension received by invalid veterans was $ 8.7 (about 20% of the unskilled wage) and the average monthly pension received by widows and dependents was $ 14

Figure 10: Heterogeneity by father occupation, occupational income score



Note: occupational income score normalized to have mean zero and unit variance. OLS estimation (equation 1). Lines represent 95% confidence intervals.

(about a third of the unskilled wage) (Glasson, 1918; Long, 1960).[48]

Unfortunately, our data does not allow us to know who received a pension and who did not.[49] However, previous work has shown that take-up was extremely low: Skocpol (1992) estimates that only 25% of the survivors of Union soldiers killed during the war received dependent pensions in 1875, and that only 43% of wounded men claimed a pension. This low take-up surely explain why we find a big role for the income effect of parental loss in the context of the United States in 1880, while other more recent studies like Kovac (2017) focus on setting were surviving spouses were well compensated.

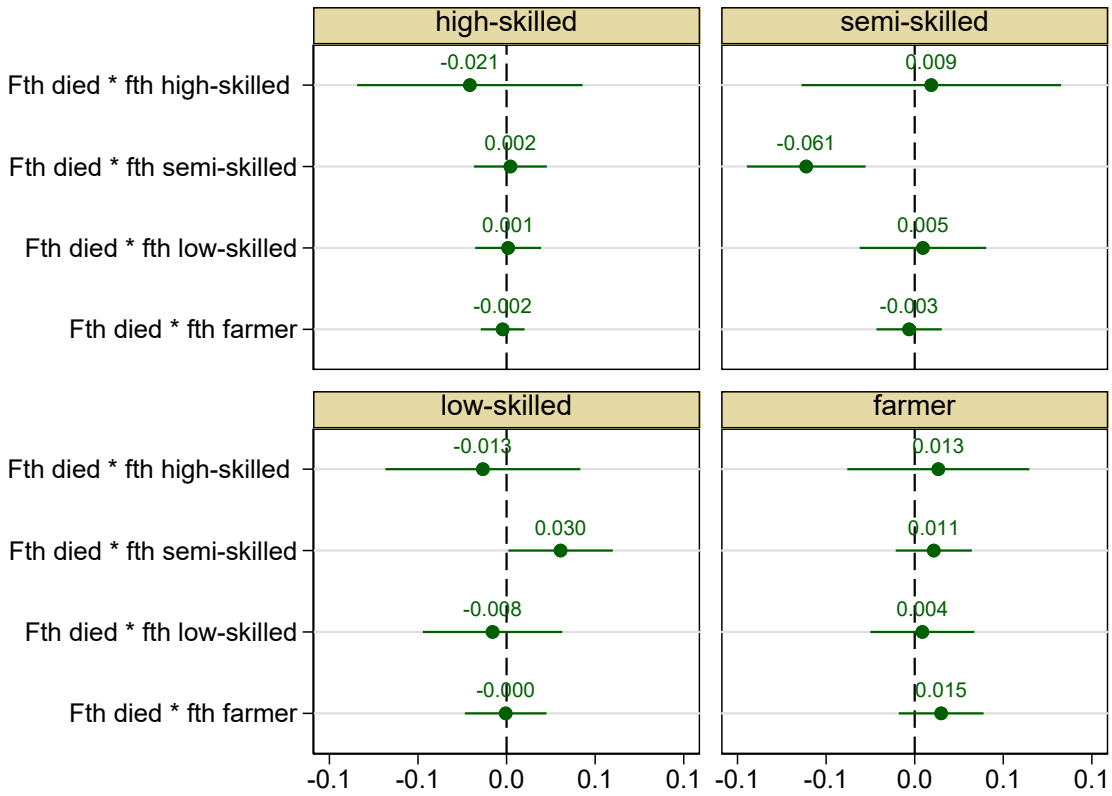
7 Conclusion

This paper studies the long-term effects of losing a father on children's later-life socioeconomic outcomes using the example of the U.S. Civil War. Having linked military records from the 2.2 million Union Army soldiers to the 1860 Census, and following their

[48]Total amounts and number of pensioner from Glasson (1918), average daily unskilled wage from Long (1960). We assume workers work 26 days in a month.

[49]Ancestry.com is currently digitizing the universe of Civil War "Widow's Pensions" records, but it is only 21% complete as of October 2019. Even if these data were available, adding another step of automated linking to our dataset would probably be very costly in terms of the measurement error introduced.

Figure 11: Heterogeneity by father occupation, skill dummies



Note: OLS estimation (equation 1). Lines represent 95% confidence intervals.

sons into adulthood in the 1880 Census, we provide one of the first long-term studies on the topic with a substantially sized sample of paternal orphans. Comparing the children of soldiers who fought in the Civil War and died to the children of the soldiers who fought and returned, we find that losing a father decreased the occupational income score of sons observed in 1880, reduced their probability of having a high- or semi-skilled occupation, and increased their probability of being in a low-skilled or farm job instead.

To deal with the potential endogeneity of paternal death in the war, we instrument father's death with the mortality rate in their military unit. This arguable generates plausibly exogenous variation on the probability of dying for fathers as death rates were mainly influenced by military strategy which is unlikely to have taken into account the future outcomes of the soldiers' children. To substantiate this argument, we digitized battle maps to show that the distance to the nearest enemy unit did not correlate with any measure of regiments' socioeconomic composition. This implies that poorer regiments were not placed in the front line. The IV results confirm the OLS findings, despite being

significantly larger. We provide potential reasons for the difference in magnitude based on heterogeneous treatment effects and linkage errors changing the treatment status of sons.

We find some support for an income channel, notably the fact that wealth acts as a protection. Does it mean that the negative long-term of father death can be partly compensated by monetary compensation? In this regard, analyzing the effect of Widow's Pension is a fruitful avenue for future research. Likewise, whether mothers entered the labor market to generate household income at the expense of caring for their children and the mitigating effects of mother remarriage remain questions to be explored in future work.

References

Abramitzky, R., Boustan, L., and Eriksson, K. (2012). Europe's Tired, Poor, and Huddled Masses: Self Selection and Economic Outcomes in the Age of Mass Migrations. *American Economic Review*, 102(5):1832–1856.

Abramitzky, R., Boustan, L., and Eriksson, K. (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migrations. *Journal of Political Economy*, 122(3):467–506.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Perez, S. (2019). Best practices for automated linking using historical data: A progress report. mimeo.

Abramitzky, R., Mill, R., and Perez, S. (2018). Linking individuals across historical sources: A fully automated approach. Working Paper 24324, National Bureau of Economic Research, Cambridge, MA.

Adda, J., Björklund, A., and Holmlund, H. (2011). The role of mothers and fathers in providing skills: Evidence from parental deaths. DP 5425, IZA.

Ager, P., Boustan, L., and Eriksson, K. (2019). The intergenerational effects of a large wealth shock: White southerners after the civil war. *NBER Working Paper No. 25700*.

Aigner, D. J. (1973). Regression with a Binary Independent Variable Subject to Errors of Observation. *Journal of Econometrics*, 1:49–60.

Ainsworth, M. and Filmer, D. (2006). Inequalities in children's schooling: Aids, orphanhood, poverty, and gender. *World Development*, 34(6):1099–1128.

Almond, D. and Currie, J. (2011). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspective*, 25(3):153–172.

Angrist, J. and Krueger, A. B. (1994). Why do world war ii veterans earn more than nonveterans?. *Journal of Labor Economics*, 12(1):74 – 97.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313 – 336.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249 – 288.

Bailey, M., Cole, C., Henderson, M., and Massey, C. (2017). How well do automated methods perform in historical samples? Evidence from new ground truth. Working Paper 24019, National Bureau of Economic Research, Cambridge, MA.

Bailey, M., Cole, C., and Massey, C. (2019). Simple strategies for improving inference with linked data: A case study of the 1850-1930 ipums linked representative historical samples. Working Paper.

Beegle, K., De Weerdt, J., and Dercon, S. (2009). The intergenerational impact of the african orphans crisis: a cohort study from an hiv/aids affected area. *International Journal of Epidemiology*, 38:561–568.

Beegle, K., De Weerdt, J., and Dercon, S. (2010). Orphanhood and human capital destruction: Is there persistence into adulthood? *Demography*, 4(1):163–180.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Bhuller, M., Dahl, G. B., Loken, K. V., and Mogstad, M. (2018). Intergenerational effects of incarceration. *AEA Papers and Proceedings*, 108:234–240.

Bingley, P. and Martinello, A. (2017). Measurement error in income and schooling, and the bias for linear estimation. *Journal of Labor Economics*, 35(4):1117–1148.

Black, S. and Devereux, P. (2011). Handbook of labor economics. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*. North Holland, Amsterdam.

Case, A., Paxson, C., and Ableidinger, J. (2002). Orphans in africa. Working Paper 9213, National Bureau of Economic Research, Cambridge, MA.

Chamarbagwala, R. and Moran, H. E. (2011). The human capital consequences of civil war: Evidence from Guatemala. *Journal of Development Economics*, 94(1):41 – 61.

Chambers, J. (1987). *To Raise an Army: The Draft Comes to Modern America*. Free Press.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics*, 129(4):1553–1623.

Clark, G. and Cummins, N. (2015). Intergenerational wealth mobility in england, 1858-2012: Surnames and social mobility. *Economic Journal*, 125(582):61–85.

Corak, M. (2001). Death and divorce: The long-term consequences of parenta loss on adolescents. *Journal of Labor Economics*, 19(3):682–715.

Costa, D. L. (1995). Pensions and retirement: Evidence from Union Army veterans. *The Quarterly Journal of Economics*, 110(2):297–319.

Costa, D. L. (1997). Displacing the family: Union Army pensions and elderly living arrangements. *Journal of Political Economy*, 105(6):1269–1292.

Costa, D. L. and Kahn, M. E. (2003). Cowards and Heroes: Group Loyalty in the American Civil War. *Quarterly Journal of Economics*, 118(2):519–548.

Costa, D. L. and Kahn, M. E. (2006). Forging a new identity: The cost and benefits of diversity in civil war combat units. *The Journal of Economic History*, 66(4):936–962.

Costa, D. L. and Kahn, M. E. (2007). Surviving Andersonville: The Benefits of Social Networks in POW Camps. *American Economic Review*, 97(4):1467–1487.

Costa, D. L. and Kahn, M. E. (2008). *Heroes and Cowards: The Social Face of War*. Princeton University Press.

Costa, D. L. and Kahn, M. E. (2010). Health, wartime stress, and unit cohesion: Evidence from union army veterans. *Demography*, 47(1):45 – 66.

Cunha, F. and Heckman, J. (2007). The Technology of Skill Formation. *AEA Papers and Proceedings*, 97(2):31–47.

Cunha, F., Heckman, J. J., and Lochner, L. (2006). *Interpreting the Evidence on Life Cycle Skill Formation*, volume 1 of *Handbook of the Economics of Education*, chapter 12, pages 697–812. Elsevier.

Dobbie, W., Gronqvist, H., Niknami, S., Palme, M., and Priks, M. (2018). The intergenerational effects of parental incarceration. Working Paper 24186, National Bureau of Economic Research, Cambridge, MA.

Domingues, P. and Barre, T. (2013). The health consequences of the Mozambican civil war: An anthropometric approach. *Economic Development and Cultural Change*, 61(4):755 – 788.

Evans, D. K. and Miguel, E. (2007). Orphans and schooling in africa: A longitudinal analysis. *Demography*, 44(1):35–57.

Feigenbaum, J. J. (2015). Intergenerational Mobility during the Great Depression. Working Paper.

Feigenbaum, J. J. (2016). A Machine Learning Approach to Census Record Linking. Working Paper.

Feigenbaum, J. J., Lee, J., and Mezzanotti, F. (2018). Capital destruction and economic growth: The effects of sherman's march, 1850-1920. *NBER Working Paper No. 25392*.

Ferrie, J. P. (1996). A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedule. Working Paper.

Fogel, R. (2000). Public Use Tape on the Aging of Veterans of the Union Army: U.S. Federal Census Records, 1850, 1860, 1900, 1910. Technical Report C3, Center for Population Economics, University of Chicago Graduate School of Business, and Department of Economics, Brigham Young University.

Friedline, T., Masa, R. D., and Chowa, G. A. (2015). Transforming wealth: Using the inverse hyperbolic sine (IHS) and splines to predict youth's math achievement. *Social Science Research*, 49:264–287.

Galdo, J. (2013). The long-run labor-market consequences of civil war: Evidence from the Shining Path in Peru. *Economic Development and Cultural Change*, 61(4):789 – 823.

Gertler, P., Levine, D. I., and Ames, M. (2004). Schooling and Parental Death. *The Review of Economics and Statistics*, 86(1):211–225.

Glasson, W. H. (1918). *Federal Military Pensions in the United States*. Oxford University Press, New York.

Goldin, C. D. and Lewis, F. D. (1975). The economic cost of the american civil war: Estimates and implications. *Journal of Economic History*, 35(2):299–326.

Gruber, J. (2004). Is making divorce easier bad for children? the long-run implications of unilateral divorce. *Journal of Labor Economics*, 22(4):799–833.

Heckman, J., Pinto, R., and Savelyev, P. (2016). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–2086.

Imbens, G. and van der Klaauw, W. (1995). Evaluating the cost of conscription in the netherlands. *Journal of Business and Economic Statistics*, 13(2):207 – 215.

Kalil, A., Mogstad, M., Rege, M., and Votruba, M. (2016). Father presence and the intergenerational transmission of educational attainment. *Journal of Human Resources*, 51(4):869–899.

Kennedy, S., Kidd, M. P., McDonald, J. T., and Biddle, N. (2015). The healthy immigrant effect: Pattern and evidence from four countries. *Journal of International Migration and Integration*, 16(2):317–332.

Kovac, D. (2017). Do Fathers Matter Paternal Mortality and Childrens Long-Run Outcomes. Working Paper 609, Princeton University, Industrial Relations Section.

Lieberson, S. (2000). *A Matter of Taste: How Names, Fashions and Culture Change*. Yale University Press, New Haven and London.

Long, C. D. (1960). Wages by Occupational and Individual Characteristics. In *Wages and Earnings in the United States, 1860-1890*, pages 94–108.

Maloney, T. N. and Smith, K. R. (2018). Parental loss, economic inequality, and intergenerational mobility in the long run: Evidence from the utah population database. Technical report, mimeo.

McDonald, J. T. and Kennedy, S. (2004). Insights into the 'healthy immigrant effect': health status and health service use of immigrants to Canada. *Social Science & Medicine*, 59(8):1613–1627.

McPherson, J. M. (1988). *Battle Cry of Freedom: The American Civil War*. Oxford University Press, Oxford.

Meyer, B. D. and Mittag, N. (2017). Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295–311.

Miguel, E. and Roland, G. (2011). The long-run impact of bombing Vietnam. *Journal of Development Economics*, 96(1):1 – 15.

Newbold, K. B. (2005). Self-rated health within the Canadian immigrant population: risk and the healthy immigrant effect. *Social Science & Medicine*, 60(6):1359–1370.

Painter, G. and Levine, D. I. (2000). Family structure and youths' outcomes: Which correlations are causal? *The Journal of Human Resources*, 35(3):524–549.

Pei, Z., Pischke, J.-S., and Schwandt, H. (2018). Poorly measured confounders are more useful on the left than on the right. *Journal of Business and Economic Statistics*, 37(2):205–216.

Salisbury, L. (2017). Women's income and marriage markets in the United States: Evidence from the civil war pension. *The Journal of Economic History*, 77(1):1–37.

Selcer, R. F. (2006). *Civil War America, 1850 to 1875*. Almanacs of American life. Facts On File.

Senne, J.-N. (2014). Death and schooling decisions over the short and long run in rural Madagascar. *Journal of Population Economics*, 27(2):497–528.

Serneels, P. and Verpoorten, M. (2015). The impact of armed conflict on economic performance evidence from rwanda. *Journal of Conflict Resolution*, 59(4).

Skocpol, T. (1992). *Protecting soldiers and mothers: The political origins of social policy in the United States*. Belknap Press, Cambridge (MA) and London.

UNICEF Press Center (2017). Orphans. `https://www.unicef.org/media/media_45279.html`. Accessed: 2017-09-15.

A Data Sources

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

- **California**: Orton, R.H. (1890) "Records of California Men in the War of the Rebellion 1861 to 1867", State Office, J. D. Young, Supt. State Printing, Sacramento, CA
- **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
- **Nebraska**: Dudley, E.S. (1888) "Rosters of Nebraska Volunteers from 1861 to 1869", Wigton & Evans State Printers, Hastings, NB
- **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
- **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

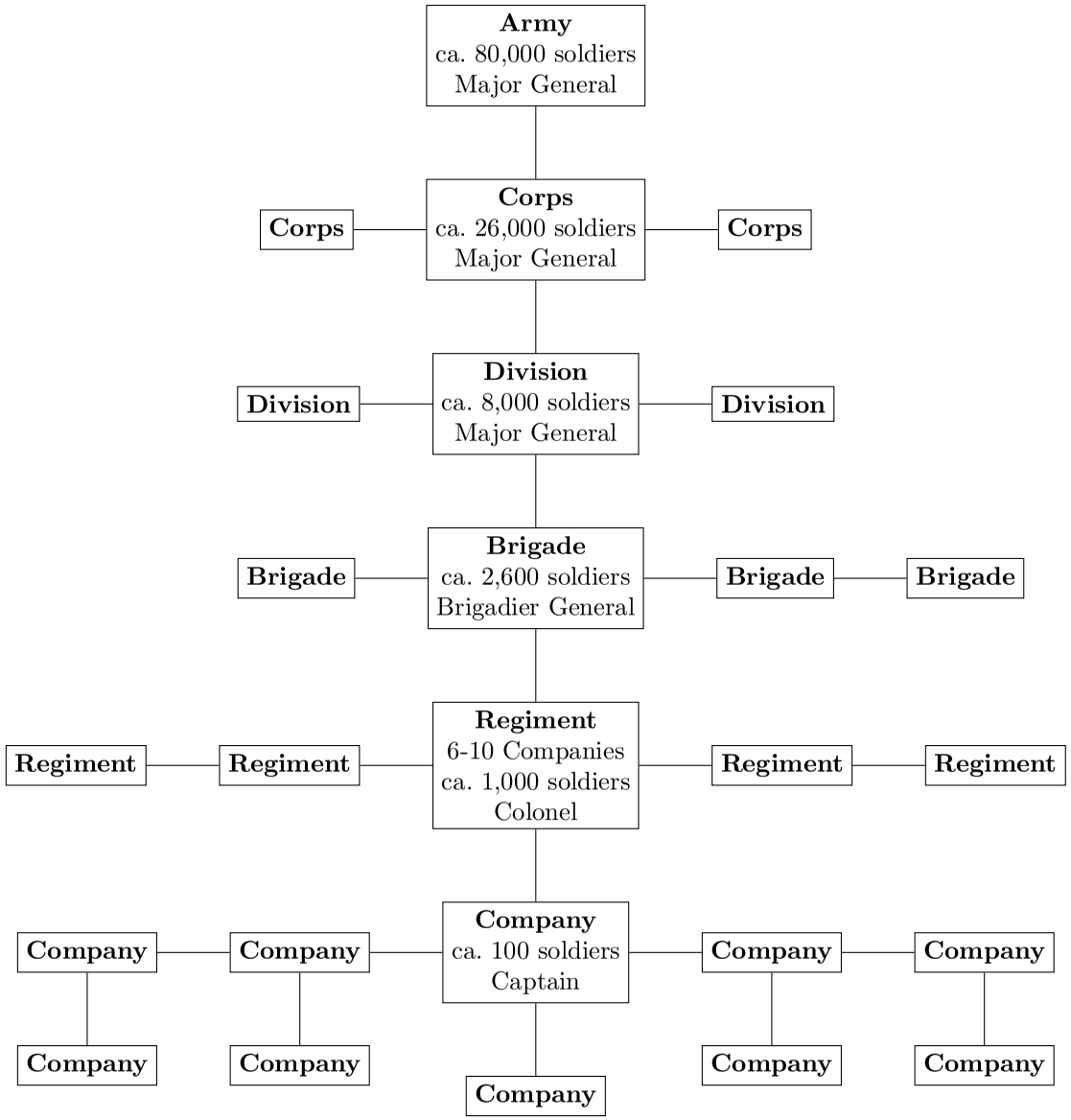
B Additional figures and descriptive statistics

Table B.1: Final sample of fathers: summary statistics and representativeness (without Southern states)

	(1) All free men 10-60 in 18600	(2) UA soldiers linked to 1860 csus	(3) Diff.	(4) All free fathers in 1860	(5) Final sample of father	(6) Diff.
Age	28.50	23.95	-4.81***	39.36	35.52	-3.88***
			(0.02)			(0.06)
Born abroad	0.23	0.15	-0.09***	0.28	0.19	-0.10***
			(0.00)			(0.00)
Non white	0.02	0.01	-0.02***	0.01	0.00	-0.01***
			(0.00)			(0.00)
Illiterate	0.05	0.03	-0.02***	0.07	0.05	-0.02***
			(0.00)			(0.00)
Wealth	1,458	672	-831***	3,111	2,701	-414
			(125)			(800)
Occ. score	12.25	11.77	-0.51***	17.95	17.82	-0.13*
			(0.02)			(0.07)
High-skilled	0.05	0.04	-0.01***	0.08	0.07	-0.01***
			(0.00)			(0.00)
Low-skilled	0.18	0.21	0.03***	0.16	0.16	-0.00**
			(0.00)			(0.00)
Semi-skilled	0.20	0.22	0.02***	0.27	0.29	0.03***
			(0.00)			(0.00)
Farmer	0.20	0.15	-0.05***	0.39	0.37	-0.02***
			(0.00)			(0.00)
Observations	7,653,111	427,646	7,653,111	2,618,474	27,331	2,618,474

Notes: the final samples are soldiers/fathers with non-missing survival information. Excluded states are Alabama, Arkansas, North and South Carolina, Florida, Georgia, Louisiana, Mississippi, Oklahoma, Tennessee, and Texas. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure B.1: Union Army Organizational Chart



Note: Typical structure of an Infantry Army of the Union. In total there were 16 Armies. One Army typically contained 2-3 Corps, a Corps had 2-3 Divisions, a Division had 2-4 Brigades, a Brigade had 2-5 Regiments, and a Regiment had 6-10 companies of a 100 men each. Armies, Corps, and Divisions were commanded by Major Generals, Brigades by Brigade Generals, Regiments by Colonels, and Companies by Captains.

Figure B.2: Map of Civil War battles

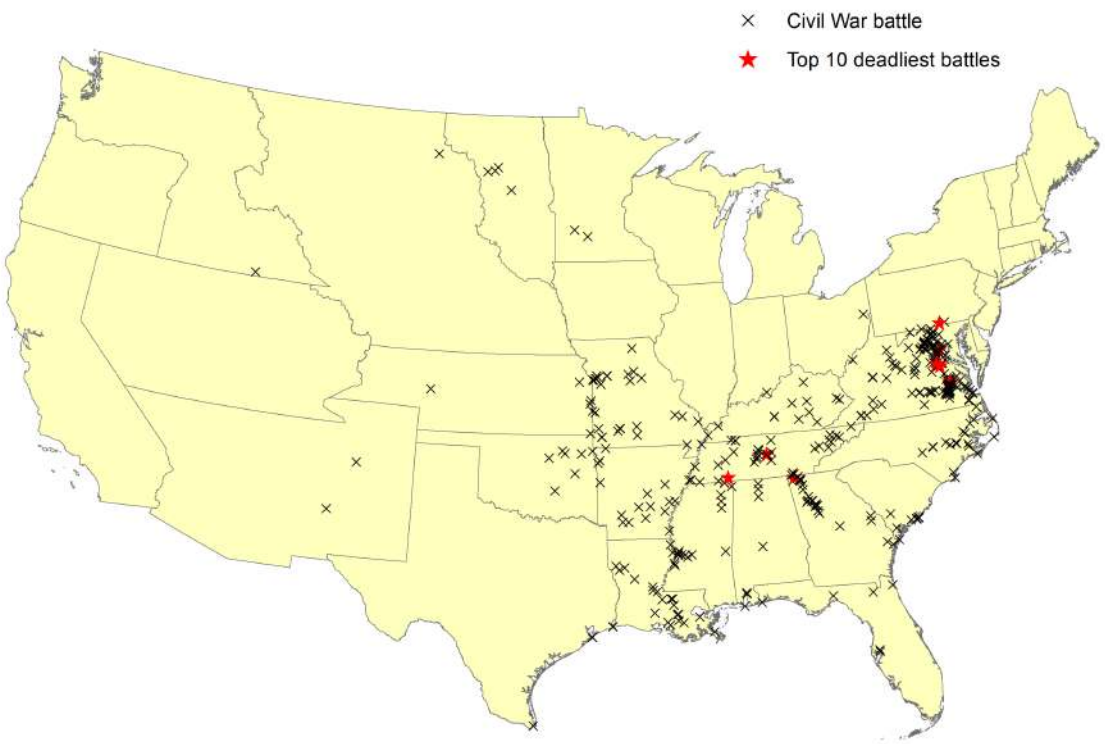
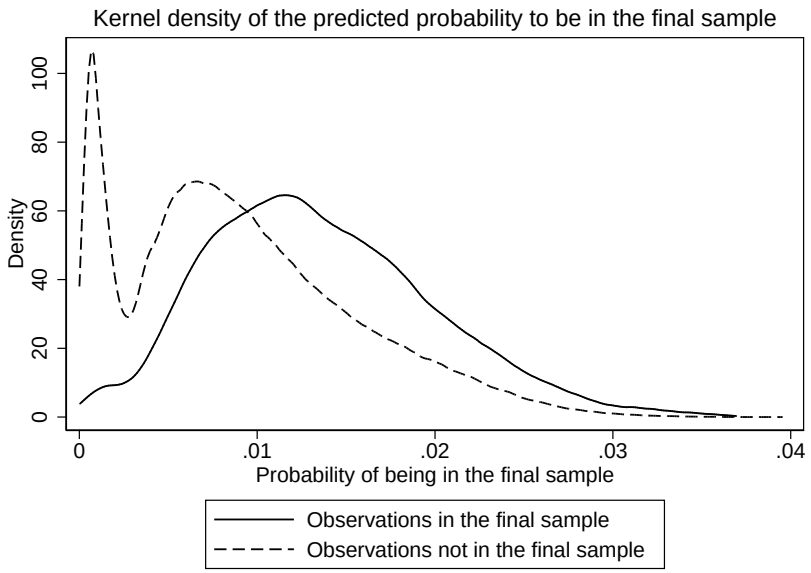


Figure B.3: The predicted probabilities to be in the final sample have a broad common support



C 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.[50] 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.[51]

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 C.1 provides an example.

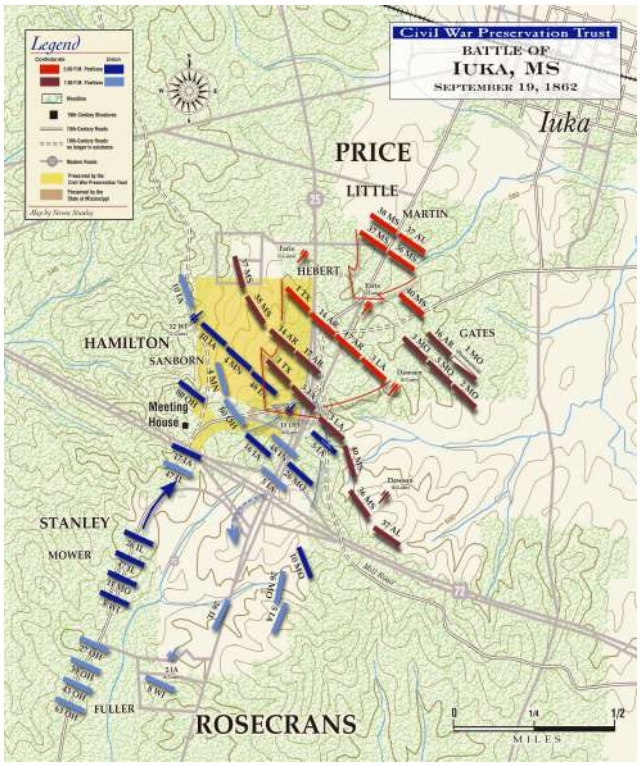
This resulted in 4,147 unit-battle-stage locations for a total of 128 battles and 799 unique Union units. Battles tend to be large with an average number of 20.5 Union units where a typical infantry regiment consists of 1,000 men. To compute the economic

[50] The maps were retrieved from: `https://www.battlefields.org/learn/maps` on May 27th, 2018.

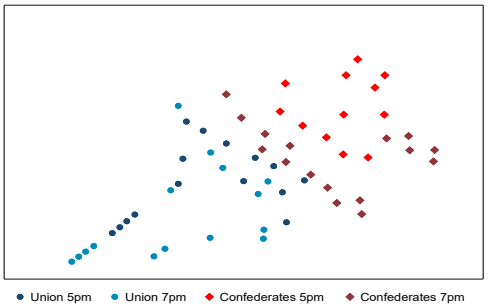
[51] 88 of the 128 maps show unit positions for different stages of a battle. This means that there is within-battle variation in the location of regiments. The average battle has 1.45 stages with a maximum of 5.

Figure C.1: Digitizing Civil War Battle Maps

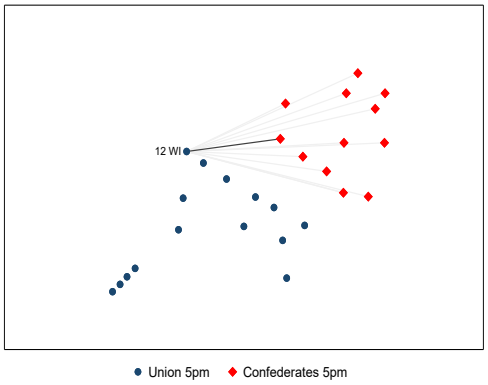
(a) Raw Battle Map



(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.

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 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), indicators 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 C.1.

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

$$\ln(\text{distance})_{rbs} = \delta_b + \phi_s + \overline{x}_r'\gamma + m_r'\beta + \eta_{rbs} \tag{8}$$

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 C.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 regression specification.

For the majority of variables these coefficients are tightly estimated zeroes and are not just insignificant due to measurement error in the outcome. The only coefficients with

Table C.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.

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 C.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)
County diversity			0.000		0.000
			(0.001)		(0.001)
% 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)
Mean value of livestock				-0.000	-0.000
				(0.000)	(0.000)
% employed in manufact.				0.001	0.003
				(0.007)	(0.007)
Manufact. capital value				0.000	0.000
				(0.000)	(0.000)
Manufact. output value				-0.000	-0.000
				(0.000)	(0.000)
Mean real estate value				-0.000	-0.000
				(0.000)	(0.000)
Churches per 1k pop.				0.023	0.027
				(0.028)	(0.030)
Value of church property				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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D Additional results

Sample selection is not a problem for our identification strategy, because we never compare the sons of fathers in our linked sample to the son of fathers in the unlinked population. However, it might be a concern for external validity, especially in the presence of very heterogeneous effects. To alleviate this concern, we create customized weights following the method of (Bailey et al., 2019): on the whole sample of fathers in the 1860 census, we create a variable l_j equal to j 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.[52] This gives us, for each father in the 1860 census, a probability $\hat{p}$ to be in the final sample predicted from observable. Appendix figure B.3 displays the kernel density of this predicted probability for fathers in the final sample and not in the final sample. As expected, fathers not in the final sample have, on average, a lower predicted probability to be linked, but the two distribution have a fairly large common support, which means that we can re-weight fathers in the final sample to be more representative of the broader population of fathers (Bailey et al.). When 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. We show below than weighted and unweighted results are very similar.

Table D.3: Effect of father death on socioeconomic characteristics of sons in 1880 with customized weights

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.044**	-0.004	-0.018**	0.010	0.014*	-0.005	0.014*
	(0.017)	(0.005)	(0.008)	(0.009)	(0.008)	(0.009)	(0.008)
Observations	32,121	32,121	32,121	32,121	32,121	32,121	31,378
Own controls	✓	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓	✓
Father other controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The method to obtain the customized weights is taken from Bailey et al. (2019) and is explained in the data section.

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

Table D.4: OLS estimation: 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.: occupational income score (normalized)				
Father died	-0.040***	-0.035**	-0.062***	-0.054**	-0.029**
	(0.015)	(0.017)	(0.023)	(0.021)	(0.012)
Observations	32,121	25,015	16,021	14,811	54,312

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. 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.

Table D.5: OLS Robustness to Double ML Covariate Selection

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.032**	-0.003	-0.018**	0.009	0.015**	0.002	0.018**
	(0.015)	(0.004)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Observations	32,121	32,121	32,121	32,121	32,121	32,121	31,378

Notes: OLS estimation (equation 1) 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. This provides a robustness check with respect to selection on observables and potential functional form issues. Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D.6: IV results: sensitivity to father controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dep. var.: occ. score (normalized)								
Father died	-0.295***	-0.293***	-0.295***	-0.295***	-0.300***	-0.296***	-0.281***	-0.296***	-0.283***
	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)	(0.089)
Fth age		-0.020***							-0.040***
		(0.007)							(0.010)
Fth age squared			0.000						0.000***
			(0.000)						(0.000)
Fth non white				-0.212					-0.176
				(0.140)					(0.134)
Fth foreign born					0.113***				0.115***
					(0.017)				(0.017)
Fth cannot read						-0.066***			-0.060**
						(0.023)			(0.023)
Fth occ. score							0.128***		0.131***
							(0.008)		(0.009)
Fth wealth (IHS)								-0.005**	-0.008***
								(0.002)	(0.002)
Observations	31,460	31,460	31,460	31,460	31,460	31,460	31,460	31,460	31,460
K-P F-stat	935.32	934.97	934.44	935.78	938.37	935.39	932.81	934.64	933.53

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" death rate of the father's last regiment of service.

Table D.7: First stage using regiment "leave-one-out" combat death rate

	Dependent variable: probability of dying						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Regimental combat death rate	1.502***	1.495***	1.384***	1.323***	1.328***	1.338***	1.337***
	(0.067)	(0.068)	(0.080)	(0.081)	(0.081)	(0.077)	(0.077)
State F.E.		✓	✓	✓	✓		
Regiment controls			✓	✓	✓	✓	✓
Enl. date poly				✓	✓	✓	✓
Enl. rank F.E.					✓	✓	✓
County F.E.						✓	✓
Father controls							✓
Mother controls							✓
Own controls							✓
Observations	31,460	31,460	31,460	31,460	31,460	31,460	31,460
F-stat	501.75	484.58	298.11	268.69	271.55	302.27	298.69

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D.8: First stage using regiment "leave-one-out" disease death rate

	Dependent variable: probability of dying						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Regimental disease death rate	1.088***	1.425***	1.271***	1.194***	1.195***	1.206***	1.205***
	(0.058)	(0.076)	(0.075)	(0.074)	(0.074)	(0.074)	(0.074)
State F.E.		✓	✓	✓	✓		
Regiment controls			✓	✓	✓	✓	✓
Enl. date poly				✓	✓	✓	✓
Enl. rank F.E.					✓	✓	✓
County F.E.						✓	✓
Father controls							✓
Mother controls							✓
Own controls							✓
Observations	31,460	31,460	31,460	31,460	31,460	31,460	31,460
F-stat	348.76	353.47	286.27	262.40	262.33	264.43	263.82

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D.9: IV results using combat death rate as an instrument

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.288**	-0.003	-0.211***	0.116*	0.031	0.107	-0.009
	(0.133)	(0.041)	(0.065)	(0.060)	(0.053)	(0.072)	(0.055)
Mean dep. var.		0.091	0.310	0.283	0.241	0.561	0.468
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,731
K-P F-stat	298.76	298.76	298.76	298.76	298.76	298.76	292.90

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" combat death rate of the father's last regiment of service.

Table D.10: IV results using disease death rate as an instrument

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.182	-0.075**	-0.102*	0.120**	0.039	0.112	0.036
	(0.125)	(0.037)	(0.060)	(0.060)	(0.050)	(0.070)	(0.049)
Mean dep. var.		0.091	0.310	0.283	0.241	0.561	0.468
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,731
K-P F-stat	263.81	263.81	263.81	263.81	263.81	263.81	251.69

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" disease death rate of the father's last regiment of service.

Table D.11: IV results excluding foreign-born soldiers

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.309***	-0.028	-0.138***	0.079*	0.030	0.054	0.063*
	(0.091)	(0.027)	(0.041)	(0.043)	(0.038)	(0.047)	(0.035)
Mean dep. var.		0.091	0.310	0.283	0.241	0.561	0.468
Observations	25,760	25,760	25,760	25,760	25,760	25,760	25,183
K-P F-stat	861.25	861.25	861.25	861.25	861.25	861.25	835.63

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" death rate of the father's last regiment of service.

Table D.12: IV Robustness to Double ML Covariate Selection

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.141*	-0.048**	-0.074**	0.074**	0.032	0.094**	0.036
	(0.077)	(0.024)	(0.035)	(0.036)	(0.031)	(0.041)	(0.031)
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,734
K-P F-stat	1,160	1,159	1,160	1,159	1,160	1,160	1,125

Notes: IV estimation (equation 1) 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 instrument and the outcome and then runs the original regression with the set of selected controls in either step. This provides a robustness check with respect to selection on observables and potential functional form issues. Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D.13: IV estimation with customized weights

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.349***	-0.072**	-0.075	0.085*	0.010	0.104*	0.028
	(0.110)	(0.033)	(0.047)	(0.047)	(0.044)	(0.055)	(0.040)
Mean dep. var.		.091	.310	.283	.241	.561	.468
Observations	31,460	31,460	31,460	31,460	31,460	31,460	30,731
K-P F-stat	680.15	680.15	680.15	680.15	680.15	680.15	667.94

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Father death is instrumented using the "leave-one-out" death rate of the father's last regiment of service. The method to obtain the customized weights is taken from Bailey et al. (2019) and is explained in the data section.

Table D.14: IV estimation: 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.: occupational income score (normalized)				
Father died	-0.283***	-0.332***	-0.171	-0.197	-0.202***
	(0.089)	(0.100)	(0.131)	(0.139)	(0.067)
Observations	31,460	24,481	15,607	14,590	53,244

Notes: Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. 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.

Table D.15: Descriptive statistics for the compliers and non-compliers

	(1)	(2)	(3)	(4)	(5)
	non-compliers				Diff.
	never-takers	always-takers	total	compliers	(3)-(4)
Private soldier	0.78	0.85	0.79	0.80	0.02***
					(0.00)
Volunteer	0.88	0.94	0.88	0.92	0.04***
					(0.00)
Infantry	0.87	0.67	0.85	0.77	-0.09***
					(0.00)
Age	35.82	36.61	35.89	35.97	0.08
					(0.09)
Foreign-born	0.17	0.22	0.18	0.19	0.01**
					(0.00)
High-skilled	0.06	0.06	0.06	0.07	0.01***
					(0.00)
Semi-skilled	0.27	0.28	0.27	0.30	0.03***
					(0.01)
Low-skilled	0.15	0.17	0.16	0.15	-0.00
					(0.00)
Farmer	0.41	0.38	0.41	0.37	-0.04***
					(0.01)
Occ. score	-0.02	-0.06	-0.03	0.02	0.05***
					(0.01)
Wealth	1,777	1,997	1,796	3,315	1,519*
					(892)
Urban district	0.15	0.23	0.16	0.22	0.06***
					(0.00)
Observations	12,394	1,166	13,560	17,536	31,096

Notes: Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The "compliers" are the soldiers who were exposed to a high risk (larger than the median probability of dying predicted by the instrument) and died, as well as those who were exposed to a low risk (lower than the median) and survived. The "never-takers" were exposed to a high risk but survived. The "always-takers" were exposed to a low risk but died.

Table D.16: Effect of father death for the compliers and non-compliers

	(1) occupational score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
				Sample of non-compliers			
Father died	-0.016	0.001	-0.009	-0.001	0.017	-0.004	0.013
	(0.030)	(0.009)	(0.015)	(0.015)	(0.013)	(0.018)	(0.014)
Observations	13,653	13,653	13,653	13,653	13,653	13,653	13,321
				Sample of compliers			
Father died	-0.063***	-0.009	-0.028***	0.025**	0.010	0.002	0.023**
	(0.021)	(0.006)	(0.010)	(0.010)	(0.009)	(0.012)	(0.009)
Observations	17,807	17,807	17,807	17,807	17,807	17,807	17,413

Notes: OLS estimation — equation (1). Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The "compliers" are the soldiers who were exposed to a high risk (larger than the median probability of dying predicted by the instrument) and died, as well as those who were exposed to a low risk (lower than the median) and survived. The "never-takers" were exposed to a high risk but survived. The "always-takers" were exposed to a low risk but died.

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 et al., 2018) machine learning techniques for automated record linkage, as well as evaluations of the performance of such algorithms (Bailey et al., 2017). 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 et al., 2019), we know relatively little about what happens to our OLS and IV estimates when we get those links wrong. Abramitzky et al. (2018) 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,[53] 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.[54] 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

[53] 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., 2017).

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

value that the 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,

$$T^* = (1 - \nu)\tilde{T} + \eta\tilde{C} \tag{9}$$

and the mis-measured orphan status can be expressed as

$$x = x^* + u \tag{10}$$

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,

$$\begin{aligned}
\widehat{\beta}_{\text{OLS}} &= \frac{Cov(\alpha + \beta x^* + \epsilon, x^* + u)}{Var(x)} \\
&= \beta \left[\frac{Var(x^*) + Cov(x^*, u)}{Var(x)} \right] \\
&= \beta \left[\frac{T(1 - T) + Cov(x, u) - Var(u)}{\tilde{T}(1 - \tilde{T})} \right]
\end{aligned} \tag{11}$$

Now substitute the following quantities into (11),

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

to derive the OLS bias as,

$$\begin{aligned} \widehat{\beta}_{\text{OLS}} &= \beta\left[\frac{T(1-T) + Cov(x,u) - Var(u)}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[\frac{\left[(1-\nu)\tilde{T} + \eta\tilde{C}\right]\left[1-(1-\nu)\tilde{T} - \eta\tilde{C}\right] + (\nu\tilde{T}\tilde{C} + \eta\tilde{T}\tilde{C})}{\tilde{T}(1-\tilde{T})}\right] \\ &+ \beta\left[\frac{-\nu\tilde{T} - \eta\tilde{C} + \left[\nu\tilde{T}\right]^2 - 2\eta\nu\tilde{T}\tilde{C} + \left[\eta\tilde{C}\right]^2}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[\frac{\tilde{T} - \nu\tilde{T} - \tilde{T}^2 + 2\nu\tilde{T} - \left[\nu\tilde{T}\right]^2 + 2\eta\nu\tilde{T}\tilde{C} - 2\eta\tilde{T}\tilde{C} + \eta\tilde{C} - \left[\eta\tilde{C}\right]^2 + \nu\tilde{T}\tilde{C} + \eta\tilde{T}\tilde{C}}{\tilde{T}(1-\tilde{T})}\right] \\ &+ \beta\left[\frac{-\nu\tilde{T} - \eta\tilde{C} + \left[\nu\tilde{T}\right]^2 - 2\nu\eta\tilde{T}\tilde{C} + \left[\eta\tilde{C}\right]^2}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[\frac{\tilde{T} - \tilde{T}^2 - 2\nu\tilde{T} + 2\nu\tilde{T}^2 - \eta\tilde{T}\tilde{C} + \nu\tilde{T}\tilde{C}}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[\frac{\tilde{T} - \tilde{T}^2 - 2\nu\tilde{T} + 2\nu\tilde{T}^2 - \eta\tilde{T}(1-\tilde{T}) + \nu\tilde{T}(1-\tilde{T})}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[\frac{\tilde{T}(1-\tilde{T}) - \nu\tilde{T}(1-\tilde{T}) - \eta\tilde{T}(1-\tilde{T})}{\tilde{T}(1-\tilde{T})}\right] \\ &= \beta\left[1-\nu-\eta\right] \end{aligned} \tag{12}$$

It follows from (12) 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,

$$x^* = \pi_0 + \pi_{x^*z} z + \xi \tag{13}$$

and that satisfies the exclusion restriction. Let $\delta_{yz} = \frac{Cov(y,z)}{Var(z)}$ denote the reduced form coefficient from the regression of y on z. An IV estimate can then be constructed as,

$$\widehat{\beta}_{\text{IV}} = \frac{\delta_{yz}}{\pi_{x^*z}} \tag{14}$$

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,

$$\begin{aligned} \widehat{\beta}_{\text{IV}} &= \frac{\delta_{yz}}{\pi_{xz}} \\ &= \frac{\delta_{yz}}{(1 - \nu - \eta)\pi_{x^*z}} \\ &= \frac{1}{1 - \nu - \eta}\beta_{\text{IV}} \end{aligned} \tag{15}$$

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. (2017) 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.

Future work could potentially extend this framework to more realistic scenarios which allow for

1. different distributional assumptions on the outcome and treatment (binary, discrete, continuous)

2. changes in the outcome, the treatment, or both due to linkage errors where wrong links change these quantities only with a certain probability (e.g. we might have wrongly linked to individuals but both of them are true orphans, hence $x^* = x$ despite the linkage error)

3. multiple links across different data sets

4. linkage error when the instrument comes from another data set for binary, discrete, or continuous instruments, where linkage errors in the instrument may correlate with linkage errors in the treatment (which would violate the exclusion restriction)

5. differential linkage errors by observed and unobserved individual characteristics such as name length, the complexity of a name (e.g. the number of rare letters per name), rarity of the name, among others

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 rougly the type I error rate found for the Ferrie (1996) rare-name linkage algorithm in Bailey et al. (2017). The observed treatment status x is then generated as described above with $x = x^* + u$.

The true estimating equation is,

$$y_i = 1x_i^* + \epsilon_i \tag{16}$$

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,

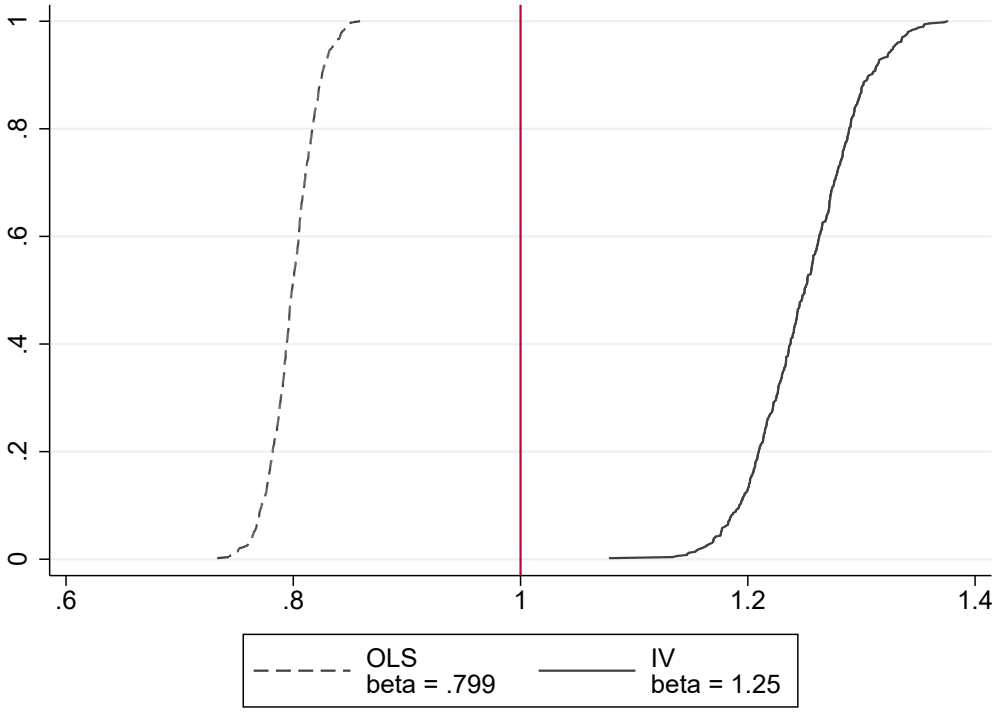
$$x_i^* = \frac{2}{3}z_i + \xi_i \tag{17}$$

with $\xi_i \sim N(0,1)$ *iid* errors, a first stage coefficient $\pi = \frac{2}{3}$, and $Corr(\epsilon, \xi) = 0$.[55] We simulate (16) 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.2 and numerically in table E.17.

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.

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

Figure E.2: 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.

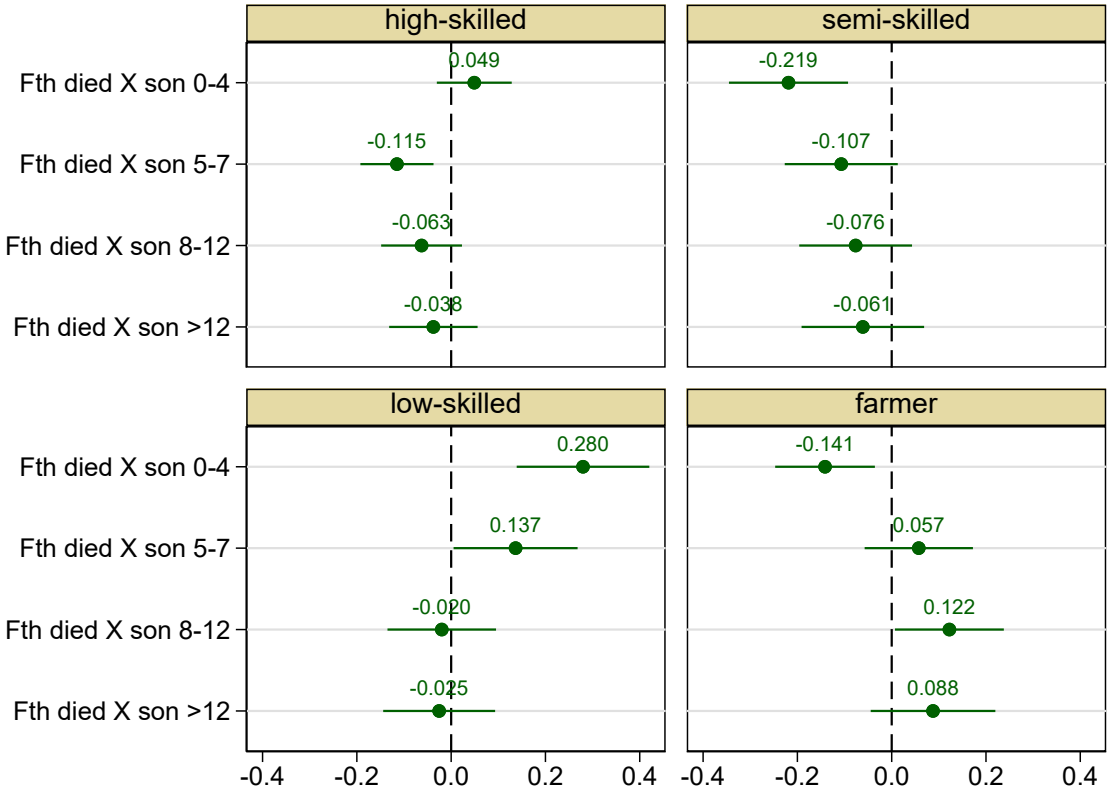
Table E.17: Summary Statistics for Simulated OLS and IV Estimations with a Mis-Measured Binary Treatment due to Linkage Errors

	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 $\widehat{\pi}_{xz}$.

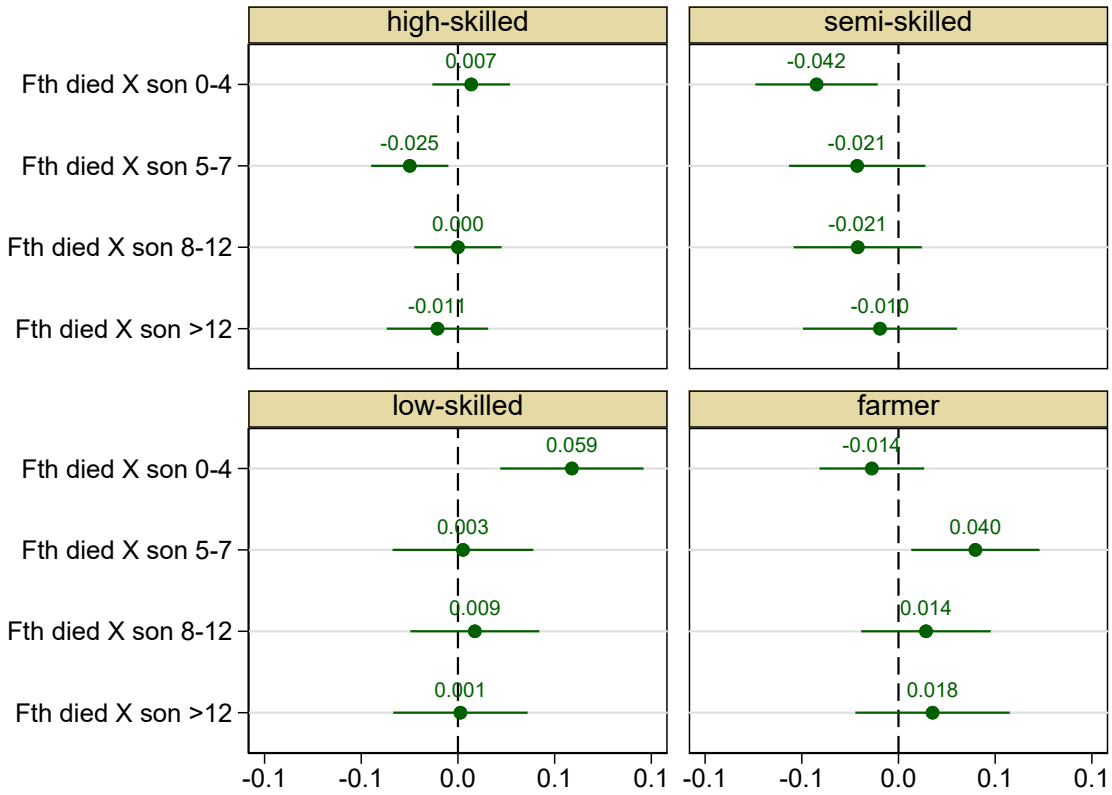
F Additional results on mechanisms

Figure F.3: Heterogeneity by age at father enlistment: IV estimation



Note: The indicator for father death and the instrument ("leave-one-out" regimental death rate) are interacted with each quartile of the age at father enlistment distribution. Lines represent 95% confidence intervals.

Figure F.4: Effect of father death on skill dummies: heterogeneity by age at father enlistment on the sample of "compliers"



Note: OLS estimation (equation 1) on the sample of "compliers". The "compliers" are the soldiers who were exposed to a high risk (larger than the median probability of dying predicted by the instrument) and died, as well as those who were exposed to a low risk (lower than the median) and survived. The "never-takers" were exposed to a high risk but survived. The "always-takers" were exposed to a low risk but died. Father death is interacted with the four quartiles of the distribution of age at father enlistment. Lines represent 95% confidence intervals.