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

Yannick Dupraz, Andreas Ferrara

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

Yannick Dupraz
Aix-Marseille University, CNRS, AMSE

Andreas Ferrara
University of Pittsburgh

This version: January 7, 2023

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 the 1880 and 1900 census. Sons of soldiers who died had lower occupational income scores and were less likely to work in a high- or semi-skilled job as opposed to being low-skilled or farmers. These effects persisted at least until the 1900 census. Our results are robust to instrumenting paternal death with the mortality rate of the father's regiment, which we argue was driven by military strategy that did not take into account the social origins of soldiers. 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.

Keywords: U.S. CIVIL WAR; ORPHANS; INTERGENERATIONAL MOBILITY
JEL Classification: N11, J13, J62

*Yannick Dupraz is a research assistant professor at Aix-Marseille University, CNRS, AMSE. Contact: yannick.dupraz@univ-amu.fr. Andreas Ferrara is an assistant professor at the University of Pittsburgh. Contact: a.ferrara@pitt.edu The data and code to replicate the results of the paper are available online: Dupraz, Yannick, and Ferrara, Andreas. Fatherless: The Long-Term Effects of Losing a Father in the U.S. Civil War. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2023-01-07. https://doi.org/10.3886/E183863V1. We gratefully acknowledge financial support for this project from CAGE, the French National Research Agency Grant ANR-17-EURE-0020, the Excellence Initiative of Aix-Marseille University - A*MIDEX (Dupraz), the Royal Economic Society, and the Leverhulme Trust (Ferrara). We thank Martha Bailey, Sascha O. Becker, James Feigenbaum, James Fenske, Victor Gay, Christopher Jepsen, Matt Nelson and Allison Shertzer, 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 are grateful to the American Battlefield Trust (https://www.battlefields.org/) for granting written permission to use their maps and reproduce the original battle map of Iuka in our appendix.

1 Introduction

The U.S. Civil War was the deadliest conflict in the history of the country (Costa and Kahn, 2008). More than 720,000 soldiers died out of a population of 31 million in the 1860 census (Hacker, 2011). Consequently, around 363,000 children had to grow up without their father because of the war.[1] Parental inputs at early ages of children have been shown to be an important determinant of later-life outcomes (Heckman, Pinto and Savelyev, 2013). While the consequences of divorce and incarceration of a parent are better understood,[2] the long-term impact of losing a parent permanently are less well known.[3] Given the persistence of income, wealth, or educational inequalities across generations (Black and Devereux, 2011), and the fact that social welfare systems in the second half of the 19th century were largely inadequate,[4] a key question is how these 363,000 children fared in the long-run and whether they ever recovered from losing their father in terms of their economic outcomes.

In this paper we show that the Civil War provides a unique opportunity to study the causal effect of losing a parent on their children's economic outcomes throughout their entire working life. In particular, we study the effects of losing a father in the U.S. Civil War on their sons' later-life incomes, proxied by occupational income scores, their occupations, marital status, and geographic mobility.[5] We collect data on the 2.2 million Union Army soldiers which we link to the full-count 1860 U.S. census to identify soldiers with children aged zero to 20. We then track the sons into the 1880 census when they are aged 20 to 40, comparing the outcomes of children whose fathers did or did not return from the war.[6] This yields a sample of almost 30,000 children observed in 1860 and 1880. Focusing on children of the Union, where relatively little fighting occurred, also rules out other direct effects of the war on their later-life outcomes such

[1] There were at least 360,000 deaths on the Union's side, the average number of children per family was 2.8 in the 1860 census, and in our data approximately 36% of soldiers were fathers. The true number of orphans is likely higher since the conventionally used number of Union Army deaths is underestimated (Hacker, 2011).

[2] For the long-term effects of divorces on children's outcomes see Painter and Levine (2000); Corak (2001); Gruber (2004), and Bhuller, Dahl, Loken and Mogstad (2018); Dobbie, Gronqvist, Niknami, Palme and Priks (2018) for the effects of parental incarceration.

[3] The published orphan study with the longest time dimension we found was Beegle, De Weerdt and Dercon (2010), who observed children for 13 years after losing a parent. A working paper by Adda, Björklund and Holmlund (2011) observes children up to thirty years after the death of their parent. We discuss the orphan literature in more detail below.

[4] A basic pension for orphans and widows existed in the states of the Union, however, it was equal to a third of a low-skilled worker's monthly wage and thus was not sufficient to support a family (McClintock, 1996).

[5] We focus on sons and fathers only because women tended to change their surnames upon marriage, complicating the linking of historic records which we heavily rely on.

[6] As 94% of all soldiers were volunteers, comparing children of soldiers as opposed to all children seeks to avoid issues of selection bias into the military.

as diseases brought by passing soldiers or destruction of physical capital (Feigenbaum, Lee and Mezzanotti, 2022).

Our OLS results show that children who lost their father in the war had 2.2% lower occupational income scores in 1880 compared to children whose father returned from the fighting, holding fixed a wide set of pre-war characteristics of the parents, the fathers' military units, and the sons.[7] We measure occupational income with the IPUMS occscore which assigns each occupation its median wage in the 1950 census, but our results are robust to using alternative measures of occupational standing using data closer in time to 1880. Affected children were also significantly less likely to be in a semi-skilled occupation as opposed to being low-skilled, and more likely to be farmers and have married earlier. While we do not find that children who lost their father were less likely to migrate out of their county of residence between 1860 and 1880, we find they migrated shorter distances, were less likely to move between states, and less likely to move West. The effect of losing a father on their sons' incomes are particularly pronounced in the lower quartiles of the 1860 wealth distribution. If a son's father had been in the top wealth quartile in 1860 and subsequently died in the war, then the son did not experience any reduction in his later-life income on average. We take this as evidence for wealth being an important mitigating factor. When we estimate the effect of paternal death on a sample of sons linked to the 1900 census, when they are between the age of 40 and 60, we find that the income difference relative to non-orphaned soldier sons is as big as it was in 1880. This suggests that affected children never recovered from the adverse effects of losing a father.

A potential concern with our estimates is that soldiers' deaths in the war are not completely random. Though our OLS estimates control for a large number of observed variables that could be correlated with both the probability of dying and a son's future income, such as father military rank and occupation, unobserved variables might bias the OLS estimate upward or downward. One example is height. Costa and Kahn (2007) show that taller soldiers were more likely to die in captivity as rations in Confederate prisons were fixed, hence shorter soldiers with a lower caloric intake had higher survival rates. However, taller fathers had on average taller sons, who would enjoy a wage premium in the labor market later on (see Lundborg, Nystedt and Rooth, 2014). Another example is the propensity to take risks, a personality trait that could be transmitted from father to son, and increase both the father's probability of dying

[7] These include the sons' and parents' age, county of residence fixed effects, and pre-war parental characteristics such as both mothers' and fathers' occupational income score, wealth, literacy, race, nativity, military controls, such as rank fixed effects, date of enlistment and ex ante service duration, and regiment type fixed effects.

and the son's later-life income. Failing to control for such unobserved variables could bias our estimate upwards, thus attenuating the negative effect of father loss in absolute value. The bias might also go in the opposite direction. For example, fathers with poorer health might have been more likely to die during the war while having transmitted their poor health to their sons, lowering their earning capacity as adults.[8]

To mitigate such concerns, we provide estimates from an instrumental variables framework where soldiers' deaths are instrumented with the mortality rate in their regiment.[9] We argue that the mortality rate in a unit was mostly driven by i) the duration and period when the regiment was active, and ii) chance and military strategy. Though duration and period of enlistment might be correlated with father characteristics, military strategy did not take the social composition of regiments into account. We digitize 128 battle maps to show that the socioeconomic composition of units did not determine their proximity to the nearest enemy unit in battle. We show that conditional on enlistment date, ex ante service duration and county fixed effects, soldier's 1860 characteristics, including literacy, age, occupation or wealth, do not predict the casualty rate of their regiment. The instrumental variable results confirm our baseline results, but they are larger: father death instrumented by the mortality rate of their regiment decreases occupational income by 11.6%. This difference can be explained by unobserved variables biasing the OLS estimate towards zero and by the fact that the effect of father loss is stronger for the children of fathers who died because of the increased risk in their regiment. We also provide a rationale for this discrepancy based on measurement error resulting from incorrect matches of individuals across data sets which would lead to an attenuation bias of OLS and an inflation bias of the IV estimates.[10]

We contribute to the literature that studies orphans and their socioeconomic outcomes. The majority of these papers focuses on orphaned children's health and educational outcomes in the context of developing countries, where parental deaths are often the result of a high HIV prevalence (see Case, Paxson and Ableidinger, 2004; Gertler, Levine and Ames, 2004; Ainsworth and Filmer, 2019; Evans and Miguel, 2007; Beegle, De Weerdt and Dercon, 2009; Beegle et

[8]Other sources of bias may exist in both directions, making it impossible to definitely sign the bias of the OLS estimator.

[9]When computing the unit-specific mortality for each soldier, we omit this soldier from the computation to avoid creating a mechanical correlation. While the typical regiment was 1,000 men strong and the contribution of a single soldier to their unit's mortality rate is small, this is the cleanest way in which we can measure our instrument.

[10]For a conventional linkage error rate, this can explain 54% of the difference between the OLS and IV results.

al., 2010).[11] The focus on health and education is usually motivated by the available data. For instance, Beegle et al. (2010) could observe children for 13 years. The lack of available long-term data tends to prohibit the study of long-run outcomes relating to incomes, occupational choices, geographic mobility, and similar variables that are of interest to economists (one exception is Adda et al. (2011) who find a wage drop of 6-7% for orphaned boys in Sweden).[12] In addition, it has been difficult to find plausibly exogenous variation in parental deaths. Despite being historical in nature, our paper speaks to both problems. First, our census linking approach allows us to estimate the effects of paternal deaths on their sons' later-life outcomes when they are aged 20 to 40 in 1880 and when they are aged 40 to 60 in 1900. This allows us to study the adverse effect over almost all of the working life of these children. Second, we provide an identification strategy to estimate the causal effect of parental deaths on their sons' later-life outcomes based on the wartime mortality rate in the fathers' military units.

There are two issues regarding the external validity of our results. The specific context of the Civil War and historical nature of our data make it difficult to draw comparisons for today. However, this is also true in modern studies where the effect of parental deaths on children's long-run outcomes may be different in a developing versus an industrialized country. More research is required to provide a broader picture across countries and time, though our income results are strikingly similar to Adda et al. (2011). We hope to contribute to this debate with our paper. The other concern is whether sons of Union Army soldiers are comparable to the general population. With 37% of northern men aged 15-44 fighting in the Union Army (Skocpol, 1993), they provide a broad cross section of the American northern population and we report additional results to show that selectivity into our estimation sample or differences to the underlying population do not drive the results.

We also contribute to the literature on the economics of the U.S. Civil War. Earlier work on the topic has focused on the soldiers themselves, their wartime experience, group cohesion or health outcomes after the war (see Costa and Kahn, 2003, 2007, 2008), or the Union Army pension as the first large-scale public assistance program in the U.S. and its impact on marriage markets and family structure (McClintock, 1996; Costa, 1997; Salisbury, 2017). A recent strand of the literature has studied the long-term and intergenerational effects of the war. This includes the trauma experienced by ex-POW or soldiers who were wounded in the war and

[11] Outside of the developing world, Kovac (2017) shows that children who lost their father in the Croatian-Serbian war had lower high school GPAs, absentee rates, and more behavioral problems.

[12] In a recent working paper, Bockerman, Haapanen and Jepsen (2021) study the short and long-run mental health consequences of parental death in Finland.

the transmission of these negative shocks to the next generation (Costa, Yetter and DeSomer, 2018, 2020). In the context of the Confederacy, Feigenbaum et al. (2022) show the negative long-term effects of the destruction during Sherman's march to the sea on agricultural investment and manufacturing activity in affected counties. Ager, Boustan and Eriksson (2021) find that children of former slave owners weathered the negative wealth shock of emancipation via their elite networks, better paying white collar occupations, and marrying up in the social ladder. Our paper provides a natural complement to these works by studying how children who lost their father in the war fared in the long-run.

2 Historical Background and Related Literature

When Amos Humiston of Portville, New York, left to fight for the Union Army, his wife and three children depended on the support of their neighbors. He served for the 154th New York Volunteer Infantry regiment and died at Gettysburg, which is when *his family's economic situation became destitute* (McClintock, 1996). He was found holding a picture of his children, which was recovered and later publicized by mere coincidence. On October 19, 1863, the Philadelphia Inquirer published the article 'Whose father was he?' together with the picture. It eventually reached the widow, Philinda Humiston, who had taken up a job as seamstress to support her family. The story of the unknown patriot father was not only widely publicized in the news, it also inspired the song *The Children of the Battle Field*. The proceeds from the sales of the picture and the song ensured that Amos' children did not grow up in poverty and they received an education with one of them even attending college (Dunkelman, 1999).

This was not the typical experience for most of the estimated 363,000 children who lost their father in the war. While there was a basic pension system for widows and orphans in place, take-up was low. Skocpol (1992) estimates that only 25% of the survivors of Union soldiers killed during the war received dependent pensions in 1875. Salisbury (2018) shows that only between one half and one third of *eligible* widows received a pension. Even when granted, pensions could be denied or rescinded for many reasons including remarriage or even living with another man, lack of birth, military, and marriage certificates, or 'immoral conduct' (McClintock, 1996). With a relatively small federal government at the time, applications could take years to be processed (Salisbury, 2017).[13] The slack in the processing of applications was also owed to the sheer size of the Union Army.

[13] In the sample of Salisbury (2017), the average processing time for a widow's pension is more than two years.

The widow of a private in the Union Army received eight dollars per month, below the $11 a Private would earn at the beginning of the war, or less than half the monthly wage of the average farm laborer in 1870 (Margo, 2000). She also received an additional two dollars per child under the age of 16. According to Salisbury (2017), "it was hardly enough to comfortably support a family" (p.3). This meant that many families who lost their main breadwinner would live impoverished for a substantial period. The system improved slowly over time as, in a context of patronage politics, the Union Army pension developed into a general old age and disability program for Union Army veterans (Eli, 2015). Initially, disability pensions were given only to the Union Army veterans who had been disabled in the war, but in 1890, the Invalid Pensions Act extended pensions to disabilities form other causes. Concerns that the increasing generosity of the Union Army pension towards veterans provided and income boost to the families of surviving soldiers, therefore treating the control group, are tempered by the fact that these developments happened decades after the end of the war, mostly after 1880, the date at which we observe the sons of Union Army soldiers for the first time.

Between the start of the war on April 12, 1861, when the Confederates attacked Fort Sumter, and the end of hostilities on April 9, 1865, the Union Army alone raised 2.2 million soldiers, 94% of whom volunteered. Among men born between 1838 and 1845, 81% served in the war (Costa and Kahn, 2008). Initially, the war was expected to last for six months at most but four years later more than 720,000 Americans had lost their lives (Hacker, 2011).[14] With a population of 31 million in 1860, the Civil War became the deadliest conflict in U.S. history both in absolute and relative terms. The number of Civil War deaths exceeds the sum of U.S. military deaths from both World Wars, the Korean, Vietnam, Gulf, Afghanistan, and Iraq wars taken together.

The introduction of a unique data set of Union Army soldiers by Fogel (2000) has enabled economists to study the Civil War and its soldiers in many different settings.[15] This includes the wartime experience and direct effects of the war on soldiers as well as post-war impacts on veterans, their families, and local communities. Work that focuses on soldiers, their military units, and their wartime experience directly has studied peer effects on survival rates in prisoner of war camps (Costa and Kahn, 2007), the impact of group homogeneity on desertions (Costa and Kahn, 2003, 2008), formation of social and human capital among Black soldiers

[14]For comprehensive reviews of the history of the American Civil War see McPherson (1988) or Selcer (2006).

[15]The online data base *Union Army Data - Early Indicators of Later Work Levels, Disease, and Death* by Fogel, Costa, Haines, Lee, Nguyen, Pope, Rosenberg, Scrimshaw, Trussell, Wilson, Wimmer, Kim, Bassett, Burton and Yetter (2000) is maintained at http://uadata.org/.

(Costa and Kahn, 2006), and health outcomes (Costa and Kahn, 2010), how leadership styles affected desertion rates (Costa and Kahn, 2014), as well as the economic factors that influenced individuals' decisions to join the conflict (Hall, Huff and Kuriwaki, 2019).

Papers that study the medium to long-term impact of the war have considered the effects of the veteran pension program on political and migration outcomes, longer term health and mortality, and, to a lesser extent, its intergenerational effects. The impact of the Union Army pension, the first large scale social program in U.S. history, has been studied with respect to veterans' retirement decisions (Costa, 1995), family structure (McClintock, 1996; Costa, 1997), or marital outcomes and incomes of women after the war (Salisbury, 2017). Work on health outcomes has mainly focused on mortality rates later on and how those were affected by the war (Costa, 2003; Costa and Kahn, 2010; Costa et al., 2018; Eli, Logan and Miloucheva, forthcoming; Costa et al., 2020). Higher levels of group cohesion in soldiers' military units has been shown to reduce the mortality rates of veterans at older ages (Costa and Kahn, 2010). Black veterans experienced higher mortality rates later on due to bias among physicians that reduced their pension (Eli et al., forthcoming). The war also increased medical patenting and innovation, especially for prosthetic devices, both in terms of the quantity and quality of research (Clemens and Rogers, 2020). In terms of migration and political outcomes, Eli, Salisbury and Shertzer (2018) show that returning veterans from border states tended to migrate and sort along ideological lines after the conflict depending on which side they fought on. Dippel and Heblich (2021) study the influence of the so-called 48ers, German democrats who fled to the U.S. after the failed German revolution of 1848, on the raising of Union Army regiments and volunteering by Germans in the Civil War. They provide evidence that prominent individuals can have a strong effect on changing behaviors and beliefs within the social networks they act in.

Relatively fewer papers have studied the intergenerational effects of the war. Costa et al. (2020) examine how soldiers and their children fared after the war if they were wounded in combat. They show that younger veterans who were wounded tended to leave agriculture and become laborers whereas older veterans were less likely to switch and thus tended to experience a decline in wealth. They also find a negative relationship between war wounds of fathers and their daughters' socioeconomic status. Interestingly, such a relation did not exist for sons.[16] This paper is the closest to ours in the sense that it examines intergenerational effects of the war.

[16]In an earlier paper, Costa et al. (2018) find negative intergenerational health effects from wartime stress of Civil War POWs and their sons but not for their daughters.

Other work on intergenerational effects of losing slave wealth among Southern elites by Ager et al. (2021) has shown how the children of slave owners fared no worse after the war. They argue that this is because their networks helped these children to shift into white collar work and to marry up in the social hierarchy that preserved their rank in society. While not directly about intergenerational effects, Feigenbaum et al. (2022) show that capital destruction in the South had long-term negative effects on agricultural investment and manufacturing activity. They argue that the underdeveloped financial sector and lack of credit played a role in the slow recovery.

3 Data Sources

The main data sources used in this paper are the full-count U.S. census files for 1860, 1880, and 1900, as well as military records from the Union Army. The census was provided by IPUMS-USA via special licensing agreement. The military data were digitized from various printed volumes published after the war which were compiled by each state's Adjutant General.[17] The Adjutant Generals' reports were compiled after the war to keep a record of veterans and deceased soldiers to determine pension eligibility. An excerpt for the records of the 22nd Massachusetts volunteer infantry regiment is shown in Figure 1.

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 number of records is larger than the number of Union Army soldiers because of re-enlistments).[18]

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 initially collected by Fogel (2000), for which

[17] A full list of the different sources is provided in the appendix in Table A.1.

[18] 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.

they added further information on pensions, families, and later-life outcomes of soldiers.[19] The advantage of our longer 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.[20] The large scale of the data allows us to complete two separate record linking processes, the first being the linking of the military records to the 1860 census to identify soldiers who were fathers, and the second being the linking of the soldiers' children from the 1860 census into the 1880 or 1900 census, while maintaining sufficiently large sample sizes.

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.[21] Table A.2 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 (16.5%) because 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 because of re-enlistments).

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 — Figure 2 presents a schematic of the linking procedure used to build our data. 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 further increase comparability, we focus on children of Union Army soldiers in the core states of the Union where relatively little fighting occurred.[22]

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

[20] As stated by the Adjutant General of Massachusetts, a small fraction of soldiers could not be recovered and hence we do not have the entire universe of Union Army soldiers. This is due to the unprecedented scale of the war. Not everyone could receive a proper burial or be identified at all under those circumstances.

[21] The information on the exact reason of death should be treated with caution. Indeed, according to Lee (1999, p. 72) "*It should be kept in mind [...] that the distinction between deaths from disease and injury was not very meaningful in the Civil War because many of those who were injured eventually died from illness caused by infection in wounds or from disease contracted while being treated in hospitals.*"

[22] These are Connecticut, Illinois, Indiana, Iowa, Kansas, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Ohio, Pennsylvania, Rhode Island, Vermont, and Wisconsin. We exclude

This allows us to separate the effect of losing a father from other adverse effects of the war such as diseases brought by passing soldiers, the substantial changes of social and legal institutions that were experienced in the South, or the associated destruction and expropriation of physical capital (see Feigenbaum et al., 2022; Ager et al., 2021).

To build an intergenerational 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 and 1900 census. The second step is more easily described as it follows now commonly used approaches to historical record linking. We employ the algorithm of Ferrie (1996) which was further developed by Abramitzky, Boustan and Eriksson (2012) and Abramitzky, Boustan and Eriksson (2014). We link individuals exactly on first name, last name, and state of birth. We then keep links that have an absolute birth year difference of 2 years or less. In case of multiple links, we keep the link with the smallest age difference. If this does not resolve ties or 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, Cole, Henderson and Massey (2020b) show that this tends to increase false link rates.

It is more challenging to link the 1860 census to the Union Army records, where state of birth is never given and year of birth is missing in 60% of cases. Instead of state of birth, we use state of residence — this means that we cannot link men who changed states between 1860 and their enlistment in the Union Army (1861-1864).[23] To infer the correct link in case of multiple exact matches on first and last name, we use both age difference (when available) and distance between counties of residence. More precisely, for each man aged 10 to 60 in the 1860 census, we start by finding all the Union Army records matching exactly on first name, last name and state of residence. Then we distinguish two cases: i) if all potential links in the UA records give the birth year of the soldier, we use the same algorithm as for the census-to-census linking: we exclude links with an age difference larger than 2 years, keep the link with the smallest age difference, and decide we cannot determine the link in cases of ties;[24] ii) if the birth year is missing for at least one potential link, we discard the links with non-missing soldier birth year

frontier and border states due to data quality and comparability.

[23]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.

[24]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.

and an age difference larger than 5 years, and among the remaining links we choose the one with the shortest distance between the county of residence in the 1860 census and in the UA records.[25] We decide we cannot determine the link in cases of ties or if the only remaining link has an age difference larger than 2 years. Finally, if a soldier is linked to more than one record in the 1860 census, we exclude all links using this soldier.

In the first step, linking military records to the 1860 census produces 482,983 links which is 22% of all soldiers. 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. These figures are only slightly lower that those provided by Skocpol (1993) for the entire Union Army (death rate of 16.47% and disability rate of 12.73%).[26] Our figures imply a disability rate of 4.1% in the Northern male fighting-age population after the war (Skocpol's figures imply a 5% rate), while in the 1880 census, the disability rate in the Northern male population was 1.59% for affected cohorts.[27] The discrepancy between the two figures might be explained by mis-measurement of disability status in the 1880 census, or by the fact that a large fraction of disabled men died between their discharge and 1880.

Among the 482,983 linked soldiers, 422,299 (72%) lived in core states of the Union in 1860. Among these, 77,496 (18%) were living with at least one son younger than 20 in the 1860 census, for a total of 137,653 sons. We were able to link 37,560 (27%) of them to the 1880 census. We unfortunately lack information on death and disability for about 11% of fathers who fought. Our final sample is therefore slightly smaller, with 29,381 sons of Union Army soldiers observed in 1860 and 1880. 14.16% lost a father during the war, and 14.25% had a father who came back from the war with a severe disability. We also build a sample linked from the 1860 census to the 1900 census, which allows us to observe sons of Civil War soldiers when they are between 40 and 60. This is a sample of 24,846 men, 14.5% of whom

[25] 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 Union Army was very local, so that soldiers usually enlisted in the vicinity of their place of residence (see Costa and Kahn, 2008). We compute the geodesic distance between the county of residence in the 1860 census and the county of residence/enlistment in the Union Army 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).

[26] Skocpol (1993), using the *Historical Statistics of the United States, Colonial Times to 1970*, writes that 2,213,000 men served in the Union Army, of which 364,511 died and 281,881 were wounded and survived.

[27] Skocpol (1993) finds that 37% of northern men aged 15-44 in 1860 fought in the Union Army. Having information on disability rate and death rate in the Union Army, we estimate the disability rate in the northern male fighting-age population after the war to be $\frac{\text{disability rate} \times 0.37}{1 - \text{death rate} \times 0.37}$. In the 1880 census, 1.59% of men aged 35-64 and residing in core Union States were "maimed, crippled, bedridden, or otherwise disabled."

lost a father during the war.[28]

Figure 6 plots the geographic distribution of the soldiers who we linked to the 1860 census and the number of those who died in the war in panels a and b, respectively, both of which also represent the approximate pattern of the population distribution in the Union. Figure 7 shows the number of soldier sons in our sample (panel a), as well as the number of sons who lost their father (panel b).

We also compare whether and how fathers in our sample differ from the average father in the 1860 U.S. population. There are three main reasons why they would be different: 1) selection into the Union Army, 2) selection due to linking fathers by name between the 1860 census and Union Army records, 3) selection due to linking children by name between the 1860 and 1880 censuses. 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., 2020b). We cannot disentangle the first two reasons, because record linking by name between the census and Union Army records is the only way for us to infer whether a man observed in 1860 later enrolled into the Union Army. However, we can separate the third source of selection from the first two.

In Table 1, we first consider the universe of fathers residing in core Union states in the 1860 census and we assess how the fathers we linked to Union Army records differ from those we could not link (columns 1 to 3). Fathers linked to Union Army records are younger, less likely to be born abroad, have lower wealth and are less likely to be farmers. These differences combine the effect of selection into the army and selection due to rare name linking. In columns (4) to (6), we consider selection due to linking sons of soldiers to the 1880 census. We consider all fathers linked to Union Army records and we compare those not in the final sample to those in the final sample (who had a son we could link to the 1880 census). They are more similar, probably because linking to Union Army records already restricted the sample to rare names. Interestingly, when we consider socio-economic characteristics, selection due to linking sons between 1860 and 1880 seems to compensate for selection due to the combined effect of enrollment in the Union Army and linking the 1860 census to Union Army records, so that our final sample of fathers is remarkably similar to the population of fathers in the 1860 census on wealth, income score and occupational dummies (columns 7 to 9). The only important differences are that the fathers in our sample are almost five years younger and 12 percentage

[28] A son observed in the 1860-1880 linked sample is not necessarily observed in the 1860-1900 linked sample, and vice versa. Only 15,402 sons are observed in both samples.

points less likely to be born abroad. In Appendix B, we show the robustness of our results to re-weighting observations to make our sample more representative of the population of fathers in 1860 or of the population of fathers linked to Union Army records, following the method of Bailey, Cole and Massey (2020a).

Our sample over-represents Whites, who are 99.61% of father-soldiers. By the end of the war, there were 180,000 Black soldiers in the Union Army. However, 90,000 of them were freedmen, i.e. formerly enslaved persons, who were not enumerated in the 1860 census. Therefore we cannot link their military records to the census in the way we do for White soldiers to identify their sons. Another 45,000 Black soldiers came from the border states which are not in our sample. The remaining 45,000 Northern Black soldiers are too few to have a sufficiently large sample of soldier-sons after our linking steps to draw valid inference. Additionally, in this period, linking African Americans across censuses is particularly difficult because of more common names, and greater levels of illiteracy and innumeracy. As non-White soldiers are more likely to correspond to false links, we prefer to focus on the White sample in this paper. This leaves us with a sample of 29,269 sons whose father fought in the War.

4 Empirical Strategy and Estimation Results

4.1 Baseline OLS Regressions

We are interested in estimating the effect of fathers' deaths in the U.S. Civil War on their sons' later-life socioeconomic outcomes. The main outcome we consider is the log occupational income score. This score assigns the median income in 1950 of a given occupation. Since actual incomes were enumerated for the first time in the 1940 census, occupational income scores are frequently used as income proxies for earlier census years (see Olivetti and Paserman, 2015; Inwood, Minns and Summerfield, 2019; Connor and Storper, 2020; Ward, 2020; Ager et al., 2021; Collins and Wanamaker, 2022). We explore below the robustness of our result to using other measures of occupational standing, using data closer in time to 1880.

We also consider labor market outcomes which include indicators for having a high-skilled, semi-skilled, or low-skilled occupation, as well as for being a farmer.[29] 40.1% of the soldiers-fathers in our database were farmers and 15% had a low skilled occupation in 1860. Compared to this, 25.5% of their sons were farmers and 28.3% were low-skilled workers in 1880. Es-

[29]The more detailed breakdown of our occupational groupings is as follows: high skilled (professional, technical, manager, craftsmen, officials, and proprietors), semi-skilled (sales, operatives), low skilled (service workers, laborers, including farm laborers), farmers (farmers - owners and tenants, farm managers).

pecially the shift from farm into non-farm occupations at different occupational levels is of interest for two reasons. First, the significant intergenerational change in the occupational structure raises the question of how losing a father affected sons' probability of moving up the occupational ladder. Second, the majority of prior orphan studies in a modern setting cannot follow children long enough to observe labor market outcomes, hence the effect of losing a father on these types of outcomes are of particular interest to us.[30] Finally, we also consider geographic mobility and whether a son has ever been married by 1880 or 1900.[31]

To estimate the impact of losing a father in the Civil War on their sons' outcomes, we use the cross sectional data we constructed for the sons in 1880 and 1900 and regress,

$$y_{ict} = \beta \text{ father died}_i + X'_{ict}\lambda + \text{parents}'_{i,1860}\theta + \text{military}'_i\phi + \alpha_{c,1860} + \varepsilon_{ict} \tag{1}$$

where y_{ict} is the outcome of son i in county c in year t (either 1880 or 1900). The main coefficient of interest is β which estimates the average difference in later-life outcomes of soldier sons whose *father died* in the war and those whose father returned. While having limited the issue of selection into the military by comparing children of soldiers only, a remaining concern is the selection into death. According to Lee (1999), the probability of dying while in service during the Civil War was affected by age, place of birth and socioeconomic status, not because the military missions of Union Army recruits were determined on the basis of class, but because higher-rank positions were filled with people with higher human capital, and because disease death affected disproportionately inhabitants of rural areas, who were less likely to have acquired immunity against common infectious diseases. Equation (1) controls for characteristics of the sons, their parents, and the fathers' military unit. In particular, we control for the age and age squared of sons measured in year t. The vector parents$'$ includes pre-war characteristics of child i's parents measured in 1860. We control for both father's and mother's age, age squared, an illiteracy indicator, a foreign born indicator, their occupational income scores, the inverse hyperbolic sine of real estate and personal wealth,[32] and literacy.[33] The vector military$'$ con-

[30] Aizer, Eli, Ferrie and Lleras-Muney (2016) are able to follow into adulthood the recipients of the Mothers' Pension program of 1911-1935, but their estimate the effect of receiving a cash transfer in childhood, not the effect of losing a parent.

[31] The slightly unconventional wording *ever married* indicates that we do not exclude a small number of widowers and divorcees as we are interested in the marriage decision per se and not necessarily the survival of the relationship.

[32] 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 transformed variable but without having to replace zeroes with an arbitrarily chosen constant (see Friedline, Masa and Chowa, 2015).

[33] When a mother is missing (15% of cases), we set these controls to zero and include an indicator for unobserved

tains variables relating to i's father experience during the war: enlistment date and its square, ex ante service duration and its square (the number of days between the enlistment date and the date of disbandment of the regiment),[34] and enlistment rank fixed effects. It also contains the characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting) and shares of different ranks (privates, low-level officers and higher-level officers). We also control for 1860 county of residence fixed effects $\alpha_{c,1860}$, which allows us to remove the influence of unobserved local and environmental factors that might have affected both the probability of fathers' deaths due to local enlistment as well as sons' later-life outcomes by comparing sons within the same counties. Finally, all remaining outcome variation remains in the error term ε_{ict}.

We cluster our standard errors by father's last regiment of service.[35] Given the local nature of the enlistment process, the regiment was the primary unit of combat and group cohesion (Costa and Kahn, 2008), hence the assignment of the 'treatment', i.e. a father's death indicator, was largely determined by where regiments went, who they fought, and the morale and health conditions in their group of peers. This becomes particularly apparent later when we use the mortality rate in a father's regiment to instrument for his death indicator. Hence we account for arbitrary shocks that potentially correlated across soldiers of the same unit who in most cases hailed from the same communities. We show in robustness checks that the results retain statistical significance when we vary the clustering unit or use standard errors accounting for spatial autocorrelation.

4.1.1 Results

Table 2 reports the results from OLS estimations of equation (1). The results in panel a show that losing a father during the Civil War was associated with a 2.2% decrease in affected sons' occupational income scores in 1880. In Appendix C, we provide an approximation for the aggregate human capital cost from the war suffered by the North in the spirit of the seminal accounting exercise by Goldin and Lewis (1975).

We also find that affected sons had worse labor market outcomes more generally. In particular, they were 2 percentage points less likely to be employed in a semi-skilled job and were instead 1.6 percentage points more likely to be farmers. The coefficients for high- and low-

mothers.

[34] We do not use actual service duration because it is mechanically correlated with the death variable, as fathers who died served for a shorter period than fathers who survived.

[35] Transfers were uncommon: only 5.4% of fathers in our sample finished the war in a different regiment than the one they enlisted in.

skilled occupations have the expected sign, which is consistent with the other results, however, they are not statistically significant.

Did lower geographic mobility explain these worse labor market outcomes? We find no evidence that paternal orphans were more or less likely to migrate out of their county of residence between 1860 and 1880, but we show in Table 3 below that they migrated shorter distances. In terms of marriage market outcomes, we find that paternal orphans were 1.6 percentage points more likely to have married by 1880. 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. Indeed, the effect is entirely driven by sons who were younger than 30 in 1880 and disappears when we focus on men older than 30.[36] Marrying early might have been a way to mitigate the loss of father income. In the Southern context, Ager et al. (2021) show that marrying up the social ladder was a way for sons of former slave owners to mitigate the loss of slave wealth. While their study focuses on Southern elites, this is also a potential explanation for our finding on the elevated marriage rates among paternal orphans.

Panel b shows results obtained on the 1860-1900 linked sample, when the sons of Union Army soldiers were between 40 and 60 years of age. The most striking result is that paternal orphans still experienced a similar income penalty compared to twenty years earlier. On average, their occupational income score was approximately 1.8% lower than that of sons whose father returned from the war. We find that bereaved sons are 1 percentage points more likely to be employed in a low-skilled occupation and 0.9 percentage points less likely to be employed in a high-skilled occupation, but these effects are not statistically different from zero. The effects on the probability to be semi-skilled and farmer are reduced compared to 1880, and not significant. The marriage effect is divided by two and becomes insignificant. This confirms that the 1880 coefficient mainly acted through an earlier age at marriage for paternal orphans as opposed to a generally higher tendency of getting married.

Table 3 presents additional results on migration between 1860 and 1880. While sons who lost their father during the war were not less likely to have migrated to a different county between 1860 and 1880, they migrated shorter distances: they were 2.7 percentage points less likely to have migrated to a different state (column 2). Conditional on migrating, they migrated about 40 km closer to home (column 3). In the last 3 columns of Table 3, we investigate the places where children of Union Army soldiers migrated. We find that those who lost their father

[36]On the sample of men aged 30 and below (22,878 observations): coefficient of 0.018 and standard error of 0.008. 0n the sample of men older than 30 (5,712 observations): coefficient of 0.002 and standard error of 0.014.

were 1 percentage point less likely to migrate West.[37] However, we do not find that they were less likely to migrate to an urbanized county or a county with a high manufacturing output per head. However, economic opportunities at the local level are hard to measure, and it seems likely that losing a father did limit sons' opportunities to better their socioeconomic condition by migrating.

In terms of effect heterogeneity, we find that wealth is an important mitigating factor and that the children of semi-skilled fathers were affected the most. We repeat the previous analysis by interacting the indicator for father death with quartiles of the 1860 wealth distribution. The main outcome we consider is the sons' occupational income score in 1880. We then also interact the indicator for father death with dummies for the father's skill group, for the quartiles of children's age at father enlistment distribution, and for quartiles of the family size distribution in 1860. Figure 3 plots the coefficients from this exercise. Panel a shows that the effect of losing a father are negative, significant, and similar across the first three quartiles of the wealth distribution. Sons of fathers in the top quartile of the wealth distribution, however, saw no negative effect from losing their father in the war.[38] This implies that wealth is an important mitigating factor that could absorb this large shock at least in economic terms. It could be because family wealth compensated for the loss of the deceased father's income and allowed the mother to invest in their son's education.[39] It could also be because family wealth improved the mother's remarriage prospects. Panel b shows that the negative effects of losing a father are concentrated among sons of fathers with semi-skilled pre-war occupations. Panel c does not reveal a clear pattern with respect to age at father enlistment. Though the estimated effects are larger in the second and top quartile of age at father enlistment, the four interactions are not statistically different from each other. On the one hand, one could expect that younger children suffered more from the absence of a father, on the other hand, sons who were of working age when they lost their father had to drop out of school to support their families. It is possible that the two effects are compensating each other. We do not see substantial differences between different household sizes in 1860 in panel d. The effect is larger for household with 6 to 7 members in 1860, but the four interactions are not statistically different from each other.

Another interesting mitigating channel to test for is whether a mother managed to re-marry

[37]The West was also a more favored destination for former Confederates, and Northerners would generally not sort into the same locations (Bazzi, Ferrara, Fiszbein, Pearson and Testa, 2021). It is intuitive that children would avoid such areas, especially if they lost a father fighting the Confederacy.

[38]The top quartile interaction is statistically different from the bottom and second quartiles at the 10% level.

[39]Unfortunately, the 1880 census does not contain information about educational attainment.

an observationally equivalent man compared to the child's father or if she changed labor force status between 1860 and 1870. Unfortunately, we did not have much success in linking women and their children over time in this way given that women tend to change their surname upon marriage. Also, for a substantial part of our sample the affected children would have moved out by 1870 and thus would be in a separate household in which we cannot observe the mother. However, this might be a potential avenue for future research.

In Figure 4, we explore heterogeneity by 1860 place of residence. In panel a, we divide individuals into three groups: those living in rural counties, with an 1860 urbanization rate of 0% (50.1% of individuals), those living in mixed counties, with a strictly positive urbanization rate below 50% (35.8% of individuals) and those living in urban counties, with an urbanization rate above 50% (14.1% of individuals). We find that the negative effect of father loss comes predominantly from mixed counties, maybe because paternal orphans from urban counties came from richer families whose wealth protected them from the negative socioeconomic effects of father loss, while in rural counties, economic opportunities were limited for everybody, whether they lost a father or not. In panel b of Figure 4, we divide the sample into the Northeast (48% of our sample) and the Midwest (52% of our sample), but we find that the negative effect of father loss has the same magnitude in both regions.

4.1.2 Robustness

Alternative measures of occupational income. Our main measure of occupational income is the IPUMS occscore which assigns a given occupation the median income of this occupation in the 1950 census (one of the first to give information on income). One concern with this often-used measure is that it assumes that the income distribution of occupations did not change between 1880 (or 1900) and 1950. In particular, farmer was a higher status occupation in the late nineteenth century than in 1950. Another problem with occupational income scores is that they give the same income to all farmers in the census, though their economic standing is in reality very heterogeneous. We want to check that the negative effect of father loss on occupational income is not solely explained by the shift towards farming (Table 2, column 5) and the fact that farmer had low incomes on average in 1950. Appendix Table A.3 explores the robustness of our result to alternative measures of occupational income. We start by considering alternative transformations of the 1950 occupational score. Column (2) shows the effect of father death on the occupational score in 1950 dollars rather than in logs. This allows considering the 7.5% of sons with no occupation in 1880 (we assign them an occupational income score of zero).

Losing a father during the war decreases occupational income by 1950$ 44 per year (about $ 500 today). Column (3) displays the effect on occupational income ranking: sons who lost their father during the War are 1.5 percentiles lower in the occupational income distribution. We then turn to measures of occupational income built using data closer to 1880. In column (4), we use the log of the occupational income score built by Feigenbaum (2018) using the 1915 Iowa census. In column (5), we use the log of the occupational income score built by Olivetti and Paserman (2015) using the 1900 occupational earnings distribution in Preston and Haines (1991).[40] In columns (6) and (7) we follow Olivetti and Paserman (2015) and Collins and Zimran (forthcoming) and we build an occupational wealth score from the 1870 census, which gives personal and real estate wealth along with occupation.[41] In column (6), we assign each occupation its median wealth in the 1870 census. Since an important share of occupations have a median wealth of zero, and since we are taking logs, this results in a much lower sample of 16,092. In column (7), we assign each occupation its average wealth in 1870, which allows us to consider a larger sample. Using these 4 alternative occupational income score produces estimates ranging between -1.3% (Iowa 1915 occupational score) and -2.4% (1870 average wealth score). In panel b of Appendix Table A.3, we drop the sons of farmers from the sample (38% of sons). This brings down the percentage of farmers in the son's generations from 24 to 15% (directly dropping the sons who are farmers would amount to selecting on the outcome). In this sample, the result using the 1950 occupational score does not change, and the results using alternative occupational income score are more tightly distributed, between -1.6% (1900 occupational score) and -2.3% (1870 average wealth score). This confirms that results are sensitive to the exact value chosen for farmer's income, but shows that the negative effect of father loss on occupational income is not driven primarily by the shift towards farming.

Inference: As stated before, we cluster standard errors at the level of fathers' military unit due to the local nature of enlistment during the Civil War. Other error correlation structures are possible and we probe for the sensitivity of our results to alternative clustering variables and methods. Appendix Table A.4 reports our OLS results together with standard errors that are clustered by fathers' unique identifier to allow for a within-family error correlation. We further

[40] Preston and Haines (1991) do not give average income for generic farm owners and tenants. Farmers are assigned the average income of occupations in the 1910 census that were coded as farmers in the 1950 occupational classification.

[41] Following Olivetti and Paserman (2015), we adjust farmers' personal property downwards by the average value of farm equipment in the 1870 census of Agriculture and we adjust real estate property by subtracting the average cash value of farms in the 1870 census of agriculture.

cluster by county of residence in 1860 to account for unobserved correlated local shocks, and we also report Conley (1999) standard errors with a distance cutoff of 50 and 100km. In all cases, the standard errors remain comparable to those in the main specification.

Selection on observables and functional form issues: The estimates are robust to the inclusion of additional dimensions of fixed effects to capture unobserved characteristics of fathers. Table A.5 shows that the baseline results barely change when we replace the county fixed effects with more demanding geographical fixed effects like township/ward fixed effects or post office fixed effects — columns (2) and (3). Likewise, the baseline estimate is barely affected when we include, along with county fixed effects, military unit fixed effects — columns (4) and (5). Finally, in columns (6) and (7) we also control for last name fixed effects and first name fixed effects.[42] Last names and first names capture unobserved socioeconomic status (like country of origin and assimilation).[43] In the bottom two line of Table A.5, we follow Pei, Pischke and Schwandt (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) to (7) differ substantially from the baseline results.

To test for potential issues of selection on observables and functional form, we apply the post-double machine learning selection algorithm by Belloni, Chernozhukov and Hansen (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 observable controls. Results are reported in the appendix in Table A.6 and show that the baseline results remain unchanged.

Alternative linkage methods: Table A.7 displays the effect of a father's death on their sons' occupational score in 1880 using samples obtained from different linking techniques. Results are robust to excluding records that produce multiple links in a 5-year window, linking on rare names only as in Ferrie (1996), excluding Union Army records without birth year, and to expanding the possible age range from plus or minus 2 years to plus or minus 5 years.

[42]The number of the respective fixed effects is as follows. There are 9,534 different townships/wards, 8,051 different post office areas, 2,413 regiments and 12,992 companies, 8,925 different last names and 1,019 different first names.

[43]Regarding the relation between last name and socioeconomic status, see for example Clark and Cummins (2015), on first names and social status, see Lieberson (2000).

Generalizability: Linked samples from the census are arguably unlikely to generate a random subsample of the population (see Bailey et al., 2020a). In Appendix B we re-weight our regressions to be more representative of 1) the entire northern population of fathers in 1860 and 2) the population of fathers we could link to Union Army records. The exercise shows that our weighted and unweighted results are similar in both the OLS and IV regressions, which we describe in more detail in the next section, and hence results do not appear to be driven by sample composition issues caused by the record linkage procedures we apply.

4.2 Instrumental Variables Regressions

Despite having controlled for a wide set of family and military characteristics in the estimation of equation (1), a major concern are unobservable variables that affect both parental deaths in the U.S. Civil War and sons' later-life outcomes. Height might be such a variable: taller men were more likely to die in POW camps due to the fixed size of rations (Costa and Kahn, 2007). However, taller fathers were more likely to have taller sons who would enjoy a wage premium in the labor market later on (Lundborg et al., 2014). If height is positively correlated with the probability of dying and sons' later-life outcomes, this would bias our estimates upwards (making the estimated negative effect of father loss smaller in absolute value).[44] Risk-taking or recklessness is another unobserved variable that could correlate within families and increase both the probability of the father dying in the War and the son's income, biasing the OLS estimates upwards. The bias might go in the other direction, if for example unobserved health characteristics of fathers were correlated positively with the probability of dying and negatively with their children's health and income later in life.

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 which comes from the fact that regiments fought and camped together. Lee (1999) states that men from the same town, "*were often recruited to the same company, were sent to the same battleground, and fought in battles side by side. Therefore, the chances of survival of a recruit would have been greatly influenced by the missions given to the company he belonged to.*" (p. 72). We argue that the assignment of such missions and battles was done by officers and generals who tended to follow military strategy and who were unlikely to be concerned with the later-life outcomes of their soldiers'

[44] Soldiers' height is not available in our data. It is available only in the Early Indicators data by Fogel et al. (2000).

sons. Hence the death rate of a father's military unit should only affect their sons' future outcomes through the death or survival of the father.[45]

Even though contemporaries argued that the Civil War was 'a rich man's war, but a poor man's fight', Lee (1999) finds evidence that largely speaks against this claim while analyzing the socioeconomic composition of units, their allocation of military tasks, and promotions. We also test this idea in Appendix D. We digitized 128 battle maps and linked units, their movement on the battle field, and their distance to the enemy to their socioeconomic composition. The tightly estimated zero effects of a whole set of socioeconomic regiment characteristics indicate that poorer units were indeed no more or less likely to be placed in the front lines. This complements the conclusion by Lee (1999) that, "*the socioeconomic variables explain very little of the variation in the degree of risk faced by a recruit, measured by the wartime mortality rate of his company.*" (p. 85).[46]

Our instrument for the death of child i's father is the "leave-one-out" death rate of the father's regiment, computed as

$$\text{death rate}_{ir} = \frac{\sum_{s \neq i} \text{died}_s}{N_r - 1} \tag{2}$$

where died is an indicator for whether a given soldier s died, and N_r is the number of men who served in regiment r. Note that this computation excludes i's father to avoid producing a mechanical correlation between the instrument and a father's probability of dying. In practice, because the typical infantry regiment had around 1,000 men, the contribution of a single soldier to the regimental death rate should be negligible. 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. Figure 5 displays the distribution of regimental death rates in the sample of father-soldiers. The average regimental death rate is 13% with a standard deviation of 7.5 percentage points. Note that we focus on the regimental death rate instead of the company death rate because the regiment was the primary battle unit (McPherson, 1988; Costa and Kahn, 2008).

In the first column of Table 4, we regress the instrument on county fixed effects and a

[45] We discuss below the possibility of fathers returning disabled, which likely would also affect their sons' socioeconomic outcomes in the long-run, as well as the effect of the regimental death rate on other men in the sons' network.

[46] As stated before, white collar job status had an impact through potential promotions, hence we control for each father's pre-war occupational income score which captures the variation in occupations and socioeconomic standing beyond the inclusion of skill dummies. Also note that a white collar job increased the chances of promotion but was by no means a guarantee for being promoted.

vector of 16 father pre-war characteristics, including occupational income, wealth and literacy in 1860. Though only the coefficient on wealth appears statistically significant, we cannot reject the joint significance of all father pre-war characteristics. However, this first specification fails to control for the most important determinants of regimental death rate: the duration and period when the regiment was active (Lee, 1999, p. 85). Some regiments were active for only three or six months, while others were active for three years. The exact time when the regiment was active also matters, since during the U.S. Civil War, periods of relative calm (like the fall of 1861) alternated with periods of intense fighting (like the spring and summer of 1863). Because men of different socioeconomic background enlisted at different moments of the war (as recruiting men was becoming harder, states and towns began offering bounties for enlistment), it is particularly important to take the date and duration of enlistment into account. In column (2) of Table 4, we control for a quadratic polynomial in date of enlistment (in days) and a quadratic polynomial in ex ante duration of service, that is the number of days the father would have served if he had not died or deserted. We obtain it by subtracting the enlistment date from the date of disbandment of the last regiment of service. We do not use the actual number of days of service because it is mechanically correlated with the death variable. The addition of these controls causes the adjusted R^2 to jump from 17.7% to 51.5%. Conditional on these controls, we cannot reject the hypothesis that none of the pre-war characteristics have a statistically significant effect on the regimental death rate.[47] In column (3), we add additional military controls to the regression. The pre-war observable characteristics of fathers remain jointly not significantly different from zero.

Appendix Table A.8 shows balance with the instrument on the right-hand side: we regress each pre-war characteristic on the regimental death rate (and the controls) and display the coefficient on the regimental death rate. Only one coefficient is statistically different from zero: the instrument is correlated negatively with age, maybe because regiments composed of younger recruits were taking more risk. In Table A.14 below, we show that our results are insensitive to controlling for father age.

[47]The coefficients on enlistment date and enlistment date squared are not individually statistically significant from zero, but the polynomial is jointly significant (F-test of 27.81, p-value of 0.00).

4.2.1 First Stage Regression

In order to instrument fathers' probability of dying in the Civil War, we estimate the following first stage equation,

$$\text{father died}_i = \delta \text{death rate}_{ir} + X'_{ict}\tau + \text{parents}'_{i,1860}\psi + \text{military}'_i\kappa + \alpha_{c,1860} + \text{regiment}_{ir} + \nu_{ir} \quad (3)$$

where father died$_i$ is an indicator for whether the father of son i died during the war as before. The death rate$_{ir}$ in i's father's last regiment of service r is defined in equation (2). The vectors X', parents$'$, military$'$, and $\alpha_{c,1860}$ contain the same variables and fixed effects as previously defined for the estimation of equation (1), while regiment$_{ir}$ is a vector of socioeconomic characteristics of the regiment computed from information on counties of enlistment of soldiers.[48]

Table 5 presents the result from estimating (3). A one-percentage point increase in the regimental death rate of a father's regiment increases his probability to die by about 0.85 percentage point. This implies that if everyone else in the father's regiment died (remember that he himself is excluded from the computation of the regimental death rate), then he would have died with his fellow soldiers almost surely. The strength of this relationship is not only reflected by the very high first stage F-statistic but also in the historic narrative. The combination of relatively modern weapons and old-style Napoleonic tactics of close line formations meant that casualties in battle were high (Costa and Kahn, 2008).[49]

In columns (2) to (7), controls are added progressively. In line with the results of Table 4, the only controls that reduce the relationship between death rate and the probability of dying are the quadratic polynomial in enlistment date, and the quadratic polynomial in ex ante duration of service. 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 increases the probability to die by 0.864 percentage points.[50] All coefficients are highly significant.

[48] We link soldiers' residence counties to economic and population data from the 1860 county-level census. For each county variable x_c, we compute the regimental average weighted by the number of soldiers belonging to each county. The variables are access to water, access to railways, share of urban population, improved acreage in agriculture, farm values, farm machinery values, livestock values, employment share in manufacturing, average real estate wealth, churches per capita, ratio of foreigners to natives, share of young men in the county. These are the variables used in Appendix D, where they are described in more detail.

[49] Close line formation means that soldiers march towards one another in a fixed formation and begin to fire at each other when reaching a certain distance. However, new repeater rifles not only shot faster but also further, which led to high casualty rates both among rank and file soldiers but also sergeants and officers as nobody could escape the range of the new weapons (McPherson, 1988).

[50] A coefficient below one does not necessarily mean that fathers had a lower probability of dying than the rest of their regiment. Conditional on military controls, the relationship between fathers' death and their regimental mortality rates becomes concave which we show in Appendix Table A.9 where we also include the square of the

Appendix Table A.10 shows robustness to alternative standard error clustering.

4.2.2 Results

Table 6 reports the results from the instrumental variables regressions. Panel a shows the results of a parsimonious specification controlling only for county fixed effects and quadratic polynomials in enlistment date and ex-ante duration of service (like in Table 4, column 2). Panel b shows the results of our preferred specification with the full set of controls (like in equation 3). The two specifications yield very similar estimates, reassuring us on the validity of our instrument. In our preferred specification (panel b), the death of a father, which is instrumented by the mortality rate in his last regiment of service (excluding the father himself from this rate), reduces their sons' incomes in 1880 by 11.6% (significant at the 10%-level).[51] It decreases the probability for the son to have a high-skilled occupation (by 6.5 percentage points, not significant), and to have a semi-skilled occupation (by 15.7%, significant at the 5%-level), and increases the probability for the son to have a low-skilled occupation (by 11 percentage points, significant at the 10%-level) and to be a farmer (by 2.6 percentage points, not significant). Results are qualitatively similar to the OLS estimates of Table 2, but effects sizes are larger (5.6 times for the occupational income score).

The larger size of IV estimates can be explained by unobserved omitted variables positively correlated with the father probability of dying and with the son's income, which would bias the OLS estimates upwards. Variables such as height or the propensity to take risks can correlate within families and increase both the father's probability to die on the battlefield or in POW. camps and the son's future income (Costa and Kahn, 2007, on how taller men were more likely to die in POW camps). Another explanation is that the instrumental variable specification estimates a local average treatment effect (LATE), which may be different in the sub-population that changed treatment status because of the instrument. In our case, the compliers are the sons of fathers who died because they were in a regiment with a higher mortality rate and who would not have died had they been assigned to a regiment with a lower mortality rate. In the next section we will provide another explanation based on potential record linkage errors when matching individuals across census years. We show that if such errors flip individuals' treatment status, OLS is downward biased while IV is upward biased. We also provide an

regimental mortality rate. This is likely because units with very high mortality rates would surrender as opposed to fight until the last soldier. For our first stage, this implies a slight loss in precision due to the omitted squared term but given the already high F-statistic, we prefer a parsimonious specification that is easier to interpret.

[51] For a coefficient of -0.123, we computed the marginal effect as $100\times(e^{\beta}-1) = -11.57\%$.

argument for how both estimates can be used to set-identify the true treatment effect via a bounding exercise.

Appendix Table A.11 displays results estimated on the 1900 linked sample. They are qualitatively similar to results estimated on the 1880 linked sample, but only the negative effect on log income score is statistically significant. It falls within the confidence interval of the 1880 coefficient, though it is larger (father death decreases income by 15.6% in 1900 versus 11.6% in 1880).

Losing a father in the war is not the only outcome. Many Union Army soldiers returned severely wounded or with disabilities, oftentimes developing habits of drug or alcohol abuse due to lacking pain medicine (Jones, 2020). This means that disabled fathers had a potentially negative effect on their sons' later-life outcomes too and thus some individuals in the control group are partially treated. This would lead to a downward bias in our estimates. However, disabilities and deaths within a regiment may also positively correlate and hence we might attribute too much of the negative impact on sons to their fathers' deaths. To test this, Table 7 shows the OLS and IV estimates using the 1880 linked sample of sons, including an indicator for whether a father returned from the war with a disability. The OLS results in panel a do not change from the baseline results, nor do the IV results in panel b. In both cases, the disability coefficients move together with the death coefficients but are much smaller relative to the death effect. So while a disabled father also has a negative impact on sons' later-life outcomes, this effect is not as severe relative to a deceased father. This might be explained by the increasing generosity of the Union Army pension system towards disabled veterans during the 1870s and 1880s (Eli, 2015).

A related issue concerns the loss of other family members besides the father, and the labor market spillover effects of a large number of men dying in the community. Since regiments were largely from the same town, a father's high regimental death rate might have affected his son through the death of other men in his close family circle and in the broader community. The average death rate in the Union Army was 16%, and 37% of Northern fighting-age men enrolled (Skocpol, 1993), which means that, on average, about 6% of the male labor force died in the conflict (and a similar number returned wounded). However, given the variability in regimental death rates (Figure 5), some towns were affected much more than others, and might have suffered a lasting post-war economic downturn due to the loss of working age men. We should start by underlining that all our effects are estimated within-county; we are therefore

never using the cross-county variation in death rates in our estimation. However, within the same county, some towns paid a larger demographic price than others. To understand the extent to which our instrument, regimental death rate, captures the effect of deaths besides the father's, Table 8 repeats our previous IV estimation together with additional controls for the number of brothers and other male family members who died in the war, and the share of men in the same neighborhood who died in the war.[52] We also control for the number of brothers and other male family members who fought in the war, and for the share of men who fought in the local neighborhood, and we include the same observable controls for these three groups that we used for the fathers. As with the disability results in the previous table, the coefficients on the death of other men cannot receive a causal interpretation, but we do estimate negative effects of losing a brother or another family member, and of high mortality rates in the neighborhood. However, these do not explain away the estimated effect of father loss, which remains similar to the baseline result.

To show that our IV results are unlikely to be driven by the spillover effects of a large number of men dying in the community, we also replicate our specification with increasingly stringent geographic fixed effects. In Appendix Table A.12, we test the robustness of our results to controlling for sub-county fixed effects, such as town, post office district, and neighborhood fixed effects, respectively.[53] This means we compare individuals from the same town/post office district/neighborhood whose fathers served in different regiments (for example because they enrolled at different dates — note that we control for date of enlistment). If regimental death rates affected the outcomes of children through local general equilibrium effects or shocks to the social network, we would expect the estimated effect of father death to attenuate as we add increasingly stringent fixed effects, as we are increasingly comparing individuals who belong to the same local communities and social networks, meaning that control children would also experience a negative effect thus making them indistinguishable from treated children. However, as can be seen in Appendix Table A.12, this is not the case: the effect of father death does not decrease with increasingly stringent geographic fixed effects.

[52]Other male family members are men living in the same household as the son in 1860 but who are neither the father nor a brother. We call neighborhood the smallest geographical unit that can be identified in the 1860 census using the post office district and the town/ward (some post office districts contain several towns/wards, some towns/wards contain several post office districts). There are 11,705 neighborhoods in our data, with an average population in 1860 of 2,719, and average enlistment rate of 10% and an average death rate of 1.6%.

[53]We call neighborhood the smallest geographical unit that can be identified in the 1860 census using the post office district and the town/ward (some post office districts contain several towns/wards, some towns/wards contain several post office districts).

Lastly, while our results are likely to have high internal validity given the quasi-experimental setting of the Civil War, there are two issues regarding external validity. First, our results may not generalize to contexts that are too different from the setting of the U.S. in the second half of the 19th century. This is a general issue faced by other orphan studies as the estimated effects of losing a parent in Sweden (Adda et al., 2011) may be different from the effect in Tanzania (Beegle et al., 2010). While the U.S. in the second half of the 19th century were more akin to a developing country today, there are striking similarities in the effect sizes found here compared to other studies such as the 6-7% wage drop observed for orphaned sons in Adda et al. (2011). More research is needed to provide a fuller picture of the more general pattern of the effects of parental loss on their children's later-life economic outcomes but we hope to contribute a useful piece of causally identified empirical evidence to this debate. Second, a related question is how comparable the children of Union Army soldiers are relative to the rest of the population at the time and whether the automated record linking methods we use produce a sample that is not representative of this population. We take this issue up in more detail in Appendix B, where we employ inverse propensity score reweighting methods to make our sample more similar to the broader underlying population.

4.2.3 Linkage Errors and the Difference between OLS and IV

Even though the IV results recovered a similar pattern across outcomes and time periods as the OLS results, we noted before that they are larger in magnitude. The discrepancy can be explained by omitted variables biasing the OLS estimate towards zero and by the fact that IV estimates a local average treatment effect. In this section, we provide another explanation based on potential measurement errors coming from mistakes in the record linkage algorithm. To provide a closed form solution to the corresponding biases in the OLS and IV estimates, we abstract from other potential endogeneity concerns in this section and consider the issue of measurement error from linking errors in isolation.

It is well known that measurement error in a binary variable, such as our *father died* indicator, cannot be classical by construction and therefore cannot be removed by using an instrument (Bingley and Martinello, 2017).[54] Since we link records in two steps, first from the soldier data to the 1860 census to identify fathers and then again from 1860 to the 1880 (or 1900) census to track their sons over time, this increases the potential for incorrect links. If we link a deceased

[54]The assumption behind classical measurement error is that the error is uncorrelated with the true variable value. However, for a binary variable, the error is always going to be the opposite of the true value and thus correlates negatively with it by construction.

soldier to the 1860 census, when in fact that soldier survived but was not matched, or if an orphaned son was linked from 1860 to an incorrect individual in 1880 whose father survived, then such linkage errors introduce non-classical measurement error into our main treatment of interest. Related to work by Aigner (1973) and Bingley and Martinello (2017), the bias in our estimates would be

$$\text{plim}\ \widehat{\beta}_{OLS} = \beta(1-\nu-\eta) \qquad , \qquad \text{plim}\ \widehat{\beta}_{IV} = \beta\frac{1}{1-\nu-\eta} \tag{4}$$

where ν is the share of false positives (children who did not lose their father but who were coded as paternal orphans), and η the share of false negatives (children who lost their father but were not coded as paternal orphans). We provide derivations, further discussion, and a simulation exercise in Appendix E. Equation (4) shows that for $\nu + \eta < 1$, such errors lead to an attenuation bias for OLS and an inflation bias for IV. Bailey et al. (2020b) find that the rate of linkage 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 OLS estimate is 70% and IV estimate is 143% of the true coefficient. This error rate can explain 54% of the difference in the OLS and IV coefficients for the occupational income score under the stated assumptions.[55]

Equation (4) also provides the possibility of applying a parametric bias correction to the OLS and IV estimates. Assuming that the true error rate is 0.3 and that all mislinked individuals also switch treatment status, multiplying the IV coefficient by $(1-\nu-\eta)$ could in principle recover the correct estimate. In our case, this would reduce the IV coefficient from an income reduction of 11.6% to a reduction of 8.2%. Vice versa, multiplying the OLS estimate by $(1-\nu-\eta)^{-1}$ would revise the income loss estimate of 2.2 up to 3.1%. However, this approach requires strong assumptions.

Can our robustness exercise using different linking algorithms be used to assess the effect of measurement error on the OLS and IV coefficient? On the one hand, a more stringent algorithm, like using only rare names, should reduce measurement error, and therefore decrease the OLS estimate and increase the IV estimate. On the other hand, samples produced using different linking algorithms are not directly comparable. In particular, (Bailey et al., 2020b) argue that rare name linking tends to select individuals with longer names, which positively

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

correlated with income and education. In Appendix Table A.7, we see that, as expected, the OLS coefficient on father death decreases when we use the Ferrie rare name algorithm (-2.8 percentage points instead of -2.2) and increases when we use the large sample size linking, likely to produce more measurement error (-1.5 instead of -2.2).[56] In Table A.15, we explore the sensitivity of our IV estimates to different linking techniques. As expected, the coefficient on father death increases when we use the Ferrie rare name algorithm (it also flips sign, but the large confidence interval does not allow to reject a substantial negative effect). However, when we use the large sample size linking, the coefficient also increases. At the end of the day, the lower precision of IV coefficients, together with the fact that different linking techniques produce differently selected samples, makes it hard to use Appendix Tables A.7 and A.15 to assess the effect of measurement error on OLS and IV estimates.

Under certain assumptions, the OLS and IV coefficients can be used in a bounding exercise. While under the described setting neither estimator would recover a point estimate of the true average treatment effect, it is still possible to set identify it. For a valid and strong instrument, and in the absence of strong treatment heterogeneity, the IV estimator provides an upper bound (in absolute value). The OLS estimate provides a lower bound (in absolute value) if any additional endogeneity issues on top of the measurement issue maintain that $|\widehat{\beta}_{OLS}| < |\beta|$. The true treatment effect then is bounded by $|\beta| \in \left(|\widehat{\beta}_{OLS}|, |\widehat{\beta}_{IV}|\right)$,[57] which in our case would imply that the income loss experienced by sons whose father died in the Civil War was between 2.2 and 11.6%.[58] To put this into perspective, Adda et al. (2011) find with modern administrative data from Sweden that the death of a father reduces sons' earnings by 6 to 7%. Note though while they rely on OLS and as the context of their study is much different from the late 19th century U.S., this shows that our estimated bounds reasonably include similar estimates from the related literature.

[56] We do not consider the "only non-missing birthyear" linking, which clearly select a different sample because whether or not birthyear is missing is very correlated across states.

[57] For a graphical example using simulated data, see Appendix Figure E.1.

[58] How can we interpret our results in light of *both* linkage errors and potential endogeneity? Absent endogeneity issues, linkage errors bound the effect between the IV and the OLS estimates — the effect of father death is definitely negative. Absent linkage error issues, the difference between IV and OLS suggests that OLS is biased towards zero (or that there is treatment effect heterogeneity), so that we can also conclude that the effect of father death is negative. To generate our finding if the true effect of father death was positive or null, there would have to be some implausible interaction between linkage error and endogeneity bias.

4.2.4 Robustness

Placebo IV regression: Table A.13 displays the results of placebo regressions where, instead of sons' 1880 socioeconomic outcomes, we regress pre-war 1860 father outcomes on an indicator variable for whether the father dies, instrumented by the death rate in his regiment. None of the estimated coefficients are statistically different from zero at conventional levels.

Sensitivity to father characteristics: we show the insensitivity of the IV estimates to fathers' observable characteristics in Appendix Table A.14. None of the observables, including age, nativity, literacy, income, or wealth, individual or jointly change our main estimate of the effect of losing a father on sons' log occupational income score in 1880.

Alternative linkage methods: Appendix Table A.15 repeats the IV analysis using different record linking methods. We replicate our results in all cases, except when using the rare name linking method by Ferrie (1996),[59] however, this method also leads to a substantial reduction in sample size, and the large confidence interval does not allow to reject a substantial negative effect.

Changing the instrument: Since many men during the war died from disease, which might have affected fathers with initially worse health status more, we use an alternative instrument excluding variation from disease death in a father's military unit: we recompute the death rate in equation (2) excluding disease death from the numerator (that is, replacing the variable died_s by zero if the soldier died of disease).[60] Getting shot in battle may arguably be more exogenous than dying of disease, especially given the evidence in Appendix B which showed that socioeconomic status did not determine units' location on the battle field. Table A.16 shows results for the 1880 linked sample of sons with this new instrument. The instrument loses some of its initial strength, with a F-statistic divided by three, but remaining well above 10. Estimated coefficients are less precise than using the total regimental death rate, but effect sizes do not drop and are, if anything, slightly larger.

Functional form: Appendix Table A.17 reports the IV results using the post-double selection algorithm by Belloni et al. (2014). Similar to before in the OLS, the algorithm uses LASSO regressions to select the most significant predictors of the instrument and the outcomes from the set of covariates, their squares, and cross-term interactions. This rules out that our IV results

[59] A name is rare if it occurs less than 10 times in the country in a given census year.

[60] As stated above, the information on the exact reason of death should be treated with caution. Indeed, according to Lee (1999, p. 72) "*It should be kept in mind [...] that the distinction between deaths from disease and injury was not very meaningful in the Civil War because many of those who were injured eventually died from illness caused by infection in wounds or from disease contracted while being treated in hospitals.*"

are driven by issues of functional form.[61] Results are very similar to the baseline IV results in panel b of Table 6.

5 Conclusion

The Civil War, as the bloodiest conflict in U.S. history, has attracted much scholarly attention. We provide new evidence on the long-run effects of losing a father in the conflict on children of fallen soldiers in the Union Army. Our OLS results show that sons who lost their father earned around 2.2% lower incomes than comparable sons in 1880. They were also less likely to be employed in semi-skilled occupations, and more likely to be employed as low-skilled workers or farmers. While they were not more likely to relocate, they migrated longer distances when they did. They also had a higher probability of having married by 1880. All these comparisons hold fixed a wide range of son and parental pre-war characteristics, observables of the fathers' military units, and county fixed effects. This is in line with a previous literature that documents the importance of parental inputs in the early years of children (Heckman et al., 2013) and that the lack of such inputs typically leads to negative outcomes for the affected children (Gruber, 2004; Kalil, Mogstad, Rege and Votruba, 2016; Bhuller et al., 2018; Dobbie et al., 2018). The more interesting finding is that these effects are not transitory. When we estimate the effect of father loss on a sample of sons linked between 1860 and 1900, i.e. when they are between 40 and 60 years of age, we still estimate economic outcomes significantly worse than those of non-orphans. This suggests that these sons did not recover from the adverse effects of losing a father throughout their working life.

This long-term view of the effects of paternal losses on children's later-life outcomes is one of two main contributions of this paper. Previous orphan studies have mainly focused on health and education outcomes due to lack of longer-run data to track children over time (see Case et al., 2004; Gertler et al., 2004; Ainsworth and Filmer, 2019; Beegle et al., 2009, 2010; Kovac, 2017). We therefore provide new estimates on the long-term effects of losing a father on economic variables such as income, occupations, mobility, and marriage decisions. The second contribution we make is to provide a causal estimate of this long-term effect. One might be concerned that the death of a father is not random and finding plausibly exogenous variation in paternal deaths (or parental deaths in general) has been challenging. The historic setting

[61] The quadratic polynomial in enlistment date and the quadratic polynomial in ex ante service duration (the difference between enlistment date and the date of disbandment of the regiment) are always included as controls because they are important predictors of regimental death rate that could be correlated with soldier characteristics (see Table 4).

allows us to exploit the mortality rates in father's military units as one of the rare opportunities that provides plausibly exogenous variation in paternal deaths. We argue that the regimental casualty rate was mainly driven by military rationale and bad luck (see Lee, 1999), which did not take into account the future outcomes of the soldiers' children. We show that the socioeconomic composition of units did not affect their location on the battlefield and that, conditional on county fixed effects, enlistment date and ex ante service duration, pre-war characteristics of the fathers cannot predict their regiment's wartime mortality rate. Our IV results confirm the OLS results but are about 6 times larger, partly because errors in the linking of historical records over different census years attenuate the OLS estimates and inflate the IV estimates.

We also contribute to a substantial literature on the economics of the Civil War. This includes studies on the wartime experience of soldiers, their group cohesion, or the health outcomes of veterans (Costa and Kahn, 2003, 2007, 2008; Dippel and Heblich, 2021), or the effects of the military pension system (Costa, 1995; McClintock, 1996; Salisbury, 2017), among others. We mostly relate to a new strand of the Civil War literature which studies the long-run and intergenerational economic consequences of the war. This includes the transmission of wartime trauma of fathers to the next generation Costa et al. (2018, 2020), the consequences of the large loss of slave wealth on the sons of Southern farmers Ager et al. (2021), or the destruction of productive capital during military campaigns such as Sherman's march to the sea Feigenbaum et al. (2022). We also add to this literature by providing a long-term view of the effects of the sons of soldiers who lost their father in the war. A potential future avenue for research is the mitigating effect of remarriage or the labor market response of the paternal orphans' mothers, and whether an observationally equivalent new husband or a working mother can negate (or worsen) the adverse effects coming from the death of the children's biological father.

References

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

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migrations," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , _ , James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

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

Adda, Jerome, Anders Björklund, and Helena Holmlund, "The Role of Mothers and Fathers in Providing Skills: Evidence from Parental Deaths," *IZA Working Paper 5425*, 2011.

Ager, Philipp, Leah Boustan, and Katherine Eriksson, "The Intergenerational Effects of a Large Wealth Shock: White Southerners after the Civil War," *American Economic Review*, 2021, *111* (11), 3767–3794.

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

Ainsworth, Martha and Deon Filmer, "Inequalities in Children's Schooling: AIDS, Orphanhood, Poverty, and Gender," *World Development*, 2019, *34* (6), 1099–1128.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 2016, *106* (4), 935–971.

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

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

Bazzi, Samuel, Andreas Ferrara, Martin Fiszbein, Thomas P. Pearson, and Patrick A. Testa, "The Other Great Migration: Southern Whites and the New Right," *NBER Working Paper 29506*, 2021.

Beegle, Kathleen, Joachim De Weerdt, and Stefan Dercon, "The intergenerational impact of the African orphans crisis: a cohort study from an HIV/AIDS affected area," *International Journal of Epidemiology*, 2009, *38* (2), 561–568.

_ , _ , and _ , "Orphanhood and human capital destruction: Is there persistence into adulthood?," *Demography*, 2010, *47*, 163–180.

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

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Loken, and Magne Mogstad, "Intergenerational Effects of Incarceration," *AEA Papers and Proceedings*, 2018, *108*, 234–240.

Bingley, Paul and Alessandro Martinello, "Measurement Error in Income and Schooling, and the Bias for Linear Estimation," *Journal of Labor Economics*, 2017, *35* (4), 1117–1148.

Black, Sandra and Paul Devereux, *Recent Developments in Intergenerational Mobility*, in Ashenfelter O. and Card, D. (eds.) 'Handbook of the Economics of Education', Vol. 4, Ch. 16, pp. 1487-1541. Elsevier, North Holland, NL, 2011.

Bockerman, Petri, Mika Haapanen, and Christopher Jepsen, "Dark Passage: Mental Health Conse-

quences of Parental Death," Discussion Paper 14385, IZA 2021.

Case, Anne, Christina Paxson, and Joseph Ableidinger, "Orphans in Africa: parental death, poverty, and school enrollment," *Demography*, 2004, *41*, 483–508.

Clark, Gregory and Neil Cummins, "Intergenerational Wealth Mobility in England, 1858-2012: Surnames and Social Mobility," *Economic Journal*, 2015, *125* (582), 61–85.

Clemens, Jeffrey and Parker Rogers, "Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents," *NBER Working Paper No. 26679*, 2020.

Collins, William J. and Ariell Zimran, "Working their Way Up? US Immigrants' Changing Labor Market Assimilation in the Age of Great Migration," *American Economics Journal: Applied Economics*, forthcoming.

_ **and Marianne H. Wanamaker**, "African American Intergenerational Economic Mobility since 1880," *American Economic Journal: Applied Economics*, 2022, *14* (3).

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

Connor, Dylan Shane and Michael Storper, "The changing geography of social mobility in the United States," *PNAS*, 2020, *117* (48), 30309–30317.

Corak, Miles, "Death and Divorce: The Long-Term Consequences of Parenta Loss on Adolescents," *Journal of Labor Economics*, 2001, *19* (3), 682–715.

Costa, Dora L., "Pensions and Retirement: Evidence from Union Army Veterans," *The Quarterly Journal of Economics*, 1995, *110* (2), 297–319.

_ , "Displacing the Family: Union Army Pensions and Elderly Living Arrangements," *Journal of Political Economy*, 1997, *105* (6), 1269–1292.

_ , "Understanding Mid-Life and Older Age Mortality Declines: Evidence from Union Army Veterans," *Journal of Econometrics*, 2003, *112* (2), 175–192.

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

_ **and** _ , "Forging a New Identity: The Cost and Benefits of Diversity in Civil War Combat Units," *The Journal of Economic History*, 2006, *66* (4), 936–962.

_ **and** _ , "Surviving Andersonville: The Benefits of Social Networks in POW Camps," *American Economic Review*, 2007, *97* (4), 1467–1487.

_ **and** _ , *Heroes and Cowards: The Social Face of War*, Princeton University Press, Princeton, NJ, 2008.

_ **and** _ , "Health, Wartime Stress, and Unit Cohesion: Evidence from Union Army Veterans," *Demography*, 2010, *47* (1), 45–66.

_ **and** _ , "Leaders: Privilege, Sacrifice, Opportunity, and Personnel Economics in the American Civil War," *Journal of Law and Economics*, 2014, *30* (3), 437–462.

_ **, Noelle Yetter, and Heather DeSomer**, "Intergenerational transmission of paternal trauma among US Civil War ex-POWs," *Proceedings of the National Academy of Sciences of the United States of America*, 2018, *115* (44), 11215–11220.

_ **,** _ **, and** _ , "Wartime Health Shocks and the Postwar Socioeconomic Status and Mortality of Union Army Veterans and their Children," *Journal of Health Economics*, 2020, *70*, 1–16.

Dippel, Christian and Stephan Heblich, "Leadership in Social Networks: Evidence from the Forty-Eighters in the Civil War," *American Economic Review*, 2021, *111* (2), 472–505.

Dobbie, Will, Hans Gronqvist, Susan Niknami, Marten Palme, and Mikael Priks, "The Intergenerational Effects of Parental Incarceration," *NBER Working Paper 24186*, 2018.

Dunkelman, Mark H., *Gettysburg's Unknown Soldier: The Life, Death, and Celebrity of Amos Humiston*, Praeger Publishers, Westport, CT, 1999.

Eli, Shari, "Income Effects on Health: Evidence from Union Army Pensions," *Journal of Economic History*, 2015, *75* (2), 448–478.

_ , Laura Salisbury, and Allison Shertzer, "Ideology and Migration after the American Civil War," *Journal of Economic History*, 2018, *78* (3), 822–861.

_ , Trevon D. Logan, and Boriana Miloucheva, "The Enduring Effects of Racial Discrimination on Income and Health," *Journal of Economic Literature*, forthcoming.

Evans, David K. and Edward Miguel, "Orphans and Schooling in Africa: A Longitudinal Analysis," *Demography*, 2007, *44*, 35–57.

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

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

_ , James Lee, and Filippo Mezzanotti, "Capital Destruction and Economic Growth: The Effects of Sherman's March, 1850-1920," *American Economic Journal: Applied Economics*, 2022, *14* (4), 301–342.

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

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

_ , Dora L. Costa, Michael Haines, Chulhee Lee, Louis Nguyen, Clayne Pope, Irvin Rosenberg, Nevin Scrimshaw, James Trussell, Sven Wilson, Larry T. Wimmer, John Kim, Julene Bassett, Joseph Burton, and Noelle Yetter, "Aging of Veterans of the Union Army: Version M-5," *Chicago: Center for Population Economics, University of Chicago Graduate School of Business, Department of Economics, Brigham Young University, and The National Bureau of Economic Research*, 2000.

Friedline, Terri, Rainier D. Masa, and Gina A.N. Chowa, "Transforming wealth: Using the inverse hyperbolic sine (IHS) and splines to predict youth's math achievement," *Social Science Research*, 2015, *49*, 264–287.

Gertler, Paul, David I. Levine, and Minnies Ames, "Schooling and Parental Death," *The Review of Economics and Statistics*, 2004, *86* (1), 211–225.

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

Gruber, Jonathan, "Is Making Divorce Easier Bad for Children? The Long-Run Implications of Unilateral Divorce," *Journal of Labor Economics*, 2004, *22* (4), 799–833.

Hacker, David J., "A Census-Based Count of the Civil War Dead," *Civil War History*, 2011, *57* (4), 307–348.

Hall, Andrew B., Connor Huff, and Shiro Kuriwaki, "Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War," *American Political Science Review*, 2019, *113* (3), 658–673.

Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Re-*

view, 2013, *103* (6), 2052–2086.

Inwood, Kris, Chris Minns, and Fraser Summerfield, "Occupational income scores and immigrant assimilation. Evidence from the Canadian census," *Explorations in Economic History*, 2019, *72*, 114–122.

Jones, Jonathan, "Opium Slavery: Civil War Veterans and Opiate Addiction," *The Journal of the Civil War Era*, 2020, *10* (2), 185–212.

Kalil, Ariel, Magne Mogstad, Mari Rege, and Mark Votruba, "Father Presence and the Intergenerational Transmission of Educational Attainment," *Journal of Human Resources*, 2016, *51* (4), 869–899.

Kovac, Dejan, "Do Fathers Matter?: Paternal Mortality and Children's Long-Run Outcomes," *Princeton University Working Papers 609*, 2017.

Lee, Chulhee, "Selective Assignment of Military Positions in the Union Army: Implications for the Impact of the Civil War," *Social Science History*, 1999, *23* (1), 67–97.

Lieberson, Stanley, *A Matter of Taste: How Names, Fashions and Culture Change*, Yale University Press, New Haven, CT, 2000.

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

Lundborg, Petter, Paul Nystedt, and Dan-Olof Rooth, "Height and Earnings: The Role of Cognitive and Noncognitive Skills," *Journal of Human Resources*, 2014, *49* (1), 141–166.

Margo, Robert A., *Wages and Labor Markets in the United States, 1820–1860*, University of Chicago Press, Chicago, IL, 2000.

McClintock, Megan J., "Civil War Pensions and the Reconstruction of Union Families," *Journal of American History*, 1996, *83* (2), 456–480.

McPherson, James M., *Battle Cry of Freedom: The American Civil War*, Oxford University Press, Oxford, UK, 1988.

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

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

Painter, Gary and David I. Levine, "Family Structure and Youths' Outcomes: Which Correlations are Causal?," *The Journal of Human Resources*, 2000, *35* (3), 524–549.

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

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

Salisbury, Laura, "Women's Income and Marriage Markets in the United States: Evidence from the Civil War Pension," *The Journal of Economic History*, 2017, *77* (1), 1–37.

_ , "Union Army Widows and the Historical Take-up of Social Benefits," *working paper*, 2018.

Selcer, Richard F., *Civil War America, 1850 to 1875*, Facts On File, New York, NY, 2006.

Skocpol, Theda, *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*, Belknap Press, Cambridge, MA, 1992.

_ , "America's First Social Security System: The Expansion of Benefits for Civil War Veterans," *Political Science Quarterly*, 1993, *108* (1), 85–116.

Ward, Zachary, "The Not-So-Hot Melting Pot: The Persistence of Outcomes for Descendants of the Age of Mass Migration," *American Economic Journal: Applied Economics*, 2020, *12* (4), 73–102.

Tables

Table 1: Summary Statistics - Linked Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Fathers in 1860 Census			Fathers in 1860 Census linked to UA records			Fathers in 1860 Census		
	Not linked to UA records	Linked to UA records	Diff.	Not in final sample	In final sample	Diff.	Not in final sample	In final sample	Diff.
Age	39.57	34.10	-5.47***	33.65	35.43	1.78***	39.37	35.43	-3.94***
			(0.60)			(0.09)			(0.60)
Born abroad	0.31	0.21	-0.10***	0.22	0.19	-0.03**	0.31	0.19	-0.12***
			(0.01)			(0.01)			(0.01)
Non white	0.01	0.01	-0.00***	0.01	0.00	-0.00***	0.01	0.00	-0.01***
			(0.00)			(0.00)			(0.00)
Illiterate	0.06	0.05	-0.01**	0.05	0.04	-0.01***	0.06	0.04	-0.01***
			(0.00)			(0.00)			(0.00)
Wealth	2,984	1,853	-1,130***	1,570	2,703	1,132*	2,936	2,703	-234
			(248)			(597)			(600)
Log income score	3.22	3.25	0.03*	3.26	3.21	-0.05***	3.22	3.21	-0.01
			(0.01)			(0.01)			(0.02)
High-skilled	0.08	0.07	-0.01	0.07	0.07	-0.00	0.08	0.07	-0.01*
			(0.00)			(0.00)			(0.00)
Low-skilled	0.17	0.17	0.01	0.18	0.16	-0.02***	0.17	0.16	-0.01
			(0.01)			(0.01)			(0.01)
Semi-skilled	0.28	0.33	0.05***	0.34	0.30	-0.04***	0.28	0.30	0.02
			(0.02)			(0.01)			(0.02)
Farmer	0.38	0.32	-0.06***	0.31	0.37	0.06***	0.38	0.37	-0.01
			(0.01)			(0.01)			(0.01)
Observations	2,090,019	101,765	2,191,784	76,291	25,474	101,765	2,166,310	25,474	2,191,784

Note: In columns (1)-(3), the population of interest is fathers residing in core Union states in the 1860 census — we can identify fathers when they co-reside with their children, therefore fathers who do not live with their (perhaps adult) children are absent. We compare the mean characteristics of fathers not linked vs. linked to the Union Army Records. In columns (4)-(6), the population of interest is fathers linked to UA records (the population of column 3). We compare the mean characteristics of fathers not in our sample to fathers in our sample (because they had a son we could link to the 1880 census). In columns (7)-(9), the population of interest is fathers residing in core Union states in the 1860 census (like in columns 1 to 3). We compare the mean characteristics of fathers not in our sample to fathers in our sample ((because we could link them to Union Army records and they had a son we could link to the 1880 census). Core Union states are Connecticut, Illinois, Indiana, Iowa, Kansas, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Ohio, Pennsylvania, Rhode Island, Vermont, and Wisconsin. Occupational skill groups follow the 1950 occupation definition of the U.S. Census Bureau. Standard errors are clustered by state of residence in 1860 and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Effect of Soldiers' Deaths on their Son's Socioeconomic Outcomes

	Panel a: Son's outcomes in 1880						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.022***	-0.003	-0.020***	0.007	0.016**	-0.007	0.016**
	(0.007)	(0.005)	(0.008)	(0.008)	(0.007)	(0.009)	(0.007)
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	27,081	29,269	29,269	29,269	29,269	29,269	28,590
	Panel b: Son's outcomes in 1900						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.018**	-0.009	-0.001	0.010	0.004	-0.007	0.007
	(0.009)	(0.007)	(0.009)	(0.007)	(0.008)	(0.009)	(0.006)
Mean dep. var.	2.995	0.165	0.324	0.159	0.291	0.678	0.895
Observations	23,501	25,012	25,012	25,012	25,012	25,012	24,993
Son controls	✓	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓	✓
Father controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓

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

Table 3: Additional Results on Migration in 1880

	Full sample		**Sample of migrants**			
	(1)	(2)	(3)	(4)	(5)	(6)
	migrant	out-of-state migrant	distance migrated (km)	West	1880 county urb. rate	1880 county manuf. output per head ($)
Father died	-0.007	-0.027***	-38.733***	-0.010**	0.007	3.453
	(0.009)	(0.008)	(14.951)	(0.004)	(0.007)	(2.570)
Mean dep. var.	0.550	0.287	477.516	0.040	0.289	121.926
Observations	29,269	29,269	15,663	16,087	15,407	15,405
Son controls	✓	✓	✓	✓	✓	✓
Father military controls	✓	✓	✓	✓	✓	✓
Father controls	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓

Note: Regressions of sons' migration outcomes in 1880 on an indicator for whether their father died in the U.S. Civil War. Migrant is an indicator for whether the individual moved county between 1860 and 1880. Out-of-state migrant is an indicator for whether the individual moved state between 1860 and 1880. Distance migrated is the distance between the centroid of the 1860 county and the centroid of the 1880 county. West is an indicator for having migrated to a state in the Western region of the U.S. (Arizona Territory, California, Colorado, Idaho, Montana Territory, Nevada, New Mexico Territory, Oregon, Utah Territory, Wyoming, Washington Territory). 1880 county urbanization rate is the percentage of residents of the county living in towns of more than 2,500 inhabitants. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. The IHS transformation was chosen to account for zeros in the wealth data. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. Standard errors clustered by the father's last regiment of service and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Instrument Balance Test

	(1) Death rate ×100	(2) Death rate×100	(3) Death rate×100
Pre-war variables			
Age	0.0024	-0.0143*	-0.0150**
	(0.0096)	(0.0075)	(0.0070)
Foreign born	-0.0253	0.0307	0.0788
	(0.1384)	(0.1078)	(0.0971)
Occupational income (ihs)	0.0613	-0.0759	0.0105
	(0.2222)	(0.1719)	(0.1536)
High-skilled	-0.3816	0.4218	0.0051
	(1.0182)	(0.7792)	(0.6944)
Semi-skilled	-0.1516	0.2596	-0.1020
	(0.8815)	(0.6793)	(0.6067)
Low-skilled	0.2342	0.4146	-0.0168
	(0.7516)	(0.5785)	(0.5159)
Farmer	-0.2529	0.3276	-0.0990
	(0.7669)	(0.5888)	(0.5264)
Illiterate	0.1209	0.0073	0.0381
	(0.2532)	(0.1933)	(0.1728)
Wealth (ihs)	-0.0421**	0.0149	0.0272**
	(0.0182)	(0.0142)	(0.0128)
Wife age	-0.0065	0.0016	0.0053
	(0.0099)	(0.0075)	(0.0067)
Wife wealth (ihs)	0.0477	0.0581	0.0568
	(0.0561)	(0.0468)	(0.0423)
Wife illiterate	0.1488	-0.1138	-0.1050
	(0.2480)	(0.1819)	(0.1631)
Wife occupational income (ihs)	-0.0270	-0.0431	-0.0457
	(0.1139)	(0.0895)	(0.0789)
Wife not in household	-0.2457	0.0438	0.1746
	(0.3295)	(0.2563)	(0.2348)
Regiment controls			
Enlistment date		-0.0137	0.0225
		(0.0240)	(0.0262)
Enlistment date squared		0.0000	-0.0000
		(0.0000)	(0.0000)
Theoretical days of service		0.0291***	0.0274***
		(0.0010)	(0.0011)
Theoretical days of service squared		-0.0000***	-0.0000***
		(0.0000)	(0.0000)
Observations	28,911	28,911	28,911
F-stat joint significance balance var.	1.94	.731	.833
p-value	.0191	.745	.634
Adjusted R^2	.177	.515	.603
County fixed effects	✓	✓	✓
Additional military controls			✓

Note: regression of the mortality rate of each father's regiment on the father's pre-war characteristics in 1860. All regressions include county fixed effects. Additional military controls: enlistment rank fixed effects and regiment characteristics: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers, as well as regiment socioeconomic controls (like in Table 5). Standard errors clustered by last regiment of service in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: First Stage Regression of Fathers' Probability to Die on the Regiment Death Rate

	Dependent variable: Pr(Father died)=1						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Death rate	1.047***	1.068***	0.990***	0.875***	0.844***	0.864***	0.865***
	(0.029)	(0.031)	(0.033)	(0.043)	(0.047)	(0.047)	(0.047)
County F.E.		✓	✓	✓	✓	✓	✓
Enlistment date poly			✓	✓	✓	✓	✓
Ex ante service duration poly				✓	✓	✓	✓
Other military controls					✓	✓	✓
Regmt socioeconomic ctrls						✓	✓
Father controls							✓
Mother controls							✓
Son controls							✓
Observations	28,911	28,911	28,911	28,911	28,911	28,911	28,911
F-stat	1,272.52	1,192.97	926.31	404.93	323.76	341.08	343.87

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

Table 6: IV Results for Father's Death and Son's Socioeconomic Outcomes in 1880

	Panel a: Parsimonious specification						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.134**	-0.088**	-0.142**	0.117**	0.057	0.024	-0.065
	(0.059)	(0.039)	(0.061)	(0.058)	(0.050)	(0.069)	(0.065)
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	399.83	404.93	404.93	404.93	404.93	404.93	392.56
	Panel b: Full set of controls						
	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.123*	-0.065	-0.157**	0.110*	0.026	0.054	-0.028
	(0.064)	(0.043)	(0.068)	(0.063)	(0.054)	(0.078)	(0.057)
Mean dep. var.	2.906	0.092	0.318	0.280	0.236	0.550	0.464
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	346.99	343.87	343.87	343.87	343.87	343.87	334.14

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

Table 7: Results for the Effect of Losing a Father vs Disabled Father in 1880

Panel a: OLS results

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.022***	-0.004	-0.019**	0.009	0.014*	-0.007	0.016**
	(0.008)	(0.005)	(0.008)	(0.008)	(0.008)	(0.009)	(0.007)
Father disabled	-0.004	-0.007	0.002	0.010	-0.007	-0.001	0.000
	(0.008)	(0.005)	(0.008)	(0.008)	(0.007)	(0.010)	(0.008)
Observations	27,081	29,269	29,269	29,269	29,269	29,269	28,590

Panel b: IV results

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.119*	-0.061	-0.152**	0.104*	0.027	0.051	-0.026
	(0.061)	(0.042)	(0.065)	(0.061)	(0.051)	(0.075)	(0.055)
Father disabled	-0.027	-0.021*	-0.029*	0.033**	-0.004	0.013	-0.009
	(0.017)	(0.011)	(0.017)	(0.017)	(0.014)	(0.021)	(0.016)
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P F-stat	378.40	381.28	381.28	381.28	381.28	381.28	369.27

Note: in panel a, son's socioeconomic outcomes in 1880 are regressed on a dummy equal to one if the father died in the war and a dummy equal to 1 if the father returned from the war with a disability. The controls are the same as in Table 2. In panel b, father death is instrumented using the "leave-one-out" death rate of his last regiment of service, and we control linearly for a dummy equal to one if the father returned from the war with a disability. The controls are the same as in Table 6, panel b (full control set). Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Migrant is an indicator for whether the individual moved county between 1860 and 1880, and ever married is an indicator for having been married in 1880 including those who became widowers or divorcees before the enumeration date. Standard errors clustered by the father's last regiment of service and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

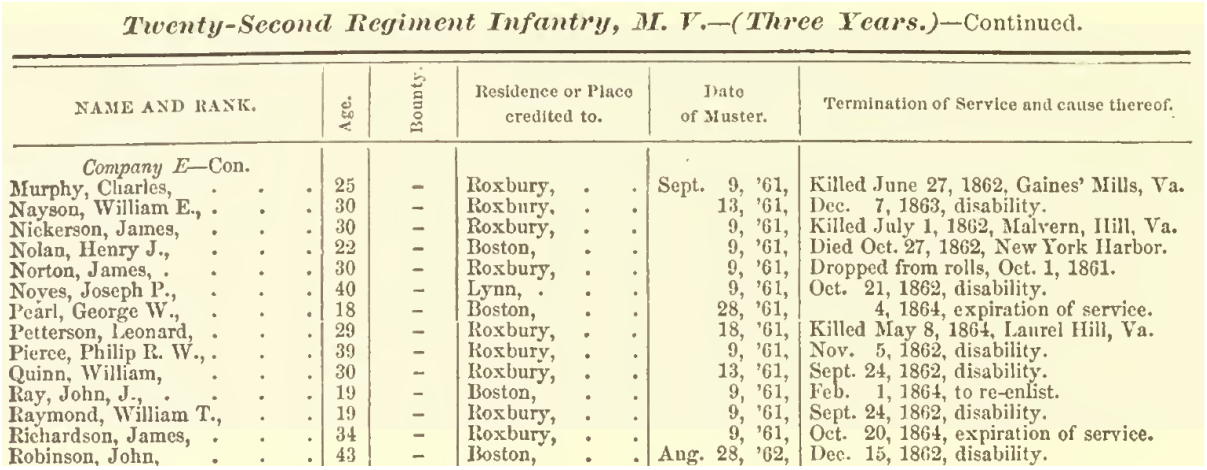
Table 8: IV Results for the Effects of Losing Family Members and Neighbors on Sons in 1880

	(1) log income score	(2) high-skilled	(3) semi-skilled	(4) low-skilled	(5) farmer	(6) migrant	(7) ever married
Father died	-0.120*	-0.062	-0.153**	0.105*	0.025	0.049	-0.030
	(0.063)	(0.043)	(0.068)	(0.064)	(0.054)	(0.078)	(0.058)
# brothers died	-0.014	-0.003	-0.048**	-0.038	0.074***	-0.013	-0.025
	(0.023)	(0.015)	(0.022)	(0.024)	(0.025)	(0.028)	(0.024)
# other family died	-0.012	0.030	-0.065	-0.029	0.063	0.028	-0.063*
	(0.048)	(0.035)	(0.044)	(0.044)	(0.039)	(0.049)	(0.038)
% neighbors died	-0.003	0.000	-0.005***	0.003	0.003*	0.001	-0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Son controls	✓	✓	✓	✓	✓	✓	✓
Father controls	✓	✓	✓	✓	✓	✓	✓
Mother controls	✓	✓	✓	✓	✓	✓	✓
Brother controls	✓	✓	✓	✓	✓	✓	✓
Other family controls	✓	✓	✓	✓	✓	✓	✓
Neighbor controls	✓	✓	✓	✓	✓	✓	✓
County F.E.	✓	✓	✓	✓	✓	✓	✓
Observations	26,753	28,911	28,911	28,911	28,911	28,911	28,244
K-P f-stat	347.09	342.91	342.91	342.91	342.91	342.91	333.69

Note: instrumental variable regression of sons' socio-economic outcomes in 1880 on father death during the war, instrument using the "leave-one-out" death rate of his last regiment of service, controlling for the number of brothers and other family members who died in the war, as well as the death rate among men who served in the local neighborhood. In our sample of men who had a father in the Union Army, 2,519 also had one or more brothers who fought (425 had a brother who died), and 824 an other family member who fought (146 had an other family member who died). All men in our sample had UA soldiers in the neighborhood. We call neighborhood the smallest geographical unit that can be identified in the 1860 census using the post office district and the town/ward (some post office districts contains several towns/wards, some towns/wards contain several post office districts). There are 11,705 neighborhoods in our data, with an average population in 1860 of 2,719, and average enlistment rate of 10% and an average death rate of 1.6%. Other male family members are men living in the same household as the son but who are not the father or a brother. The sample is the same as in Table 6, so all fathers fought in the war, but not all brothers and other family members. We therefore control for the number of brothers and other family members who fought in the war, as well as the share of men who fought in the local neighborhood. Son controls, mother controls and father controls are the same as in Table 6. Brother controls, other family controls and neighbor controls are the same as for the father, including military controls (averages for neighbors and when more than one brother or other family members fought in the war). Standard errors are clustered by the last regiment of service of the father and are reported in parentheses. Significance levels are denoted by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figures

Figure 1: 22nd MA Volunteer Infantry Regiment Records Example

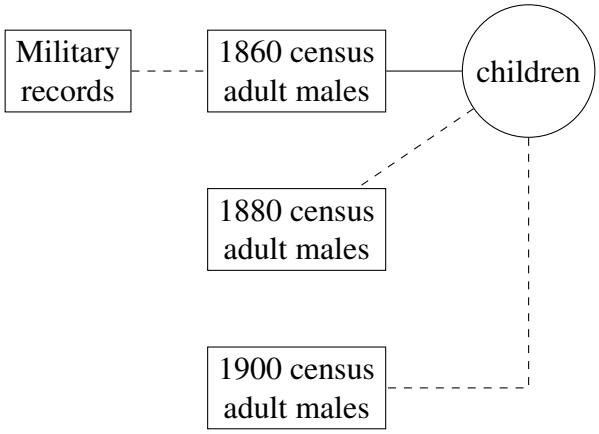


Twenty-Second Regiment Infantry, M. V.—(Three Years.)—Continued.

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.

Note: Sample record for the 22nd Massachusetts Infantry Regiment showing the typical information contained in the original sources published by the Adjutant Generals of each state after the war. Some of the information include the regiment, company, terms of service (three years here) and individual information such as each soldier's full name, their age, enlistment bounty if any, place of residence, muster date, and the date and reasons for why service ended.

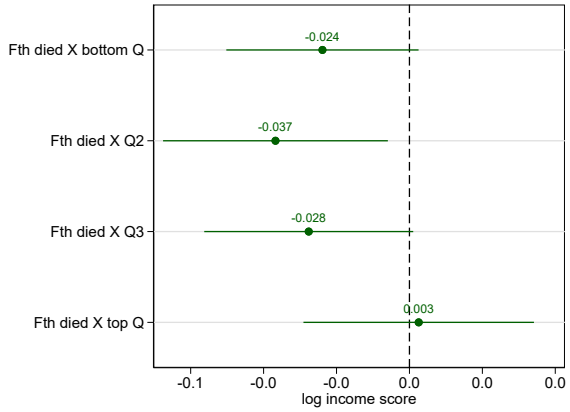
Figure 2: Schematic of the Linking Procedure



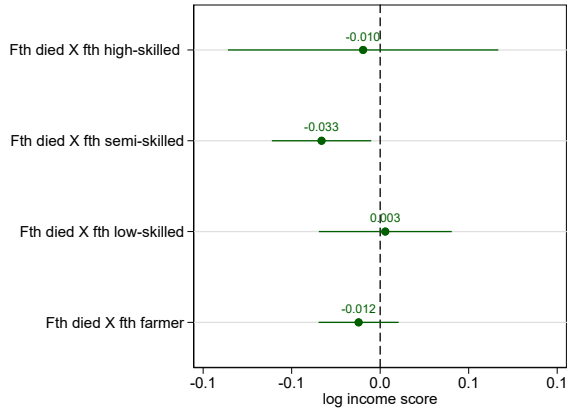
Note: Dashed lines indicate record linking between two different data sources (see section 3 for the exact algorithms used). The solid line indicate that the link comes from the data source itself (children live in the same household as their father).

Figure 3: Effect Heterogeneity by 1860 Characteristics

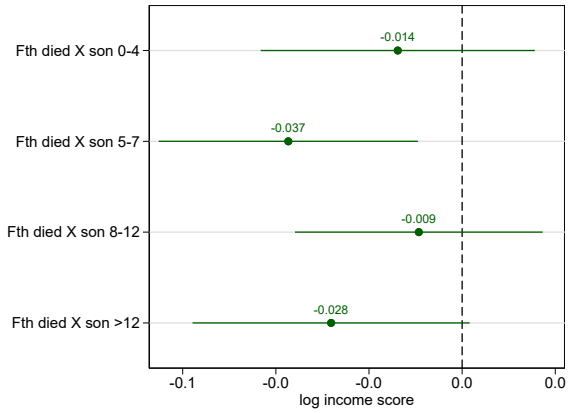
(a) by quartile of the 1860 wealth distribution



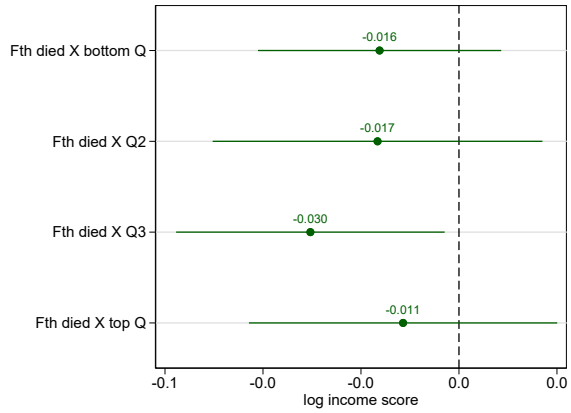
(b) by father's occupational group



(c) by child age at father enlistment quartile

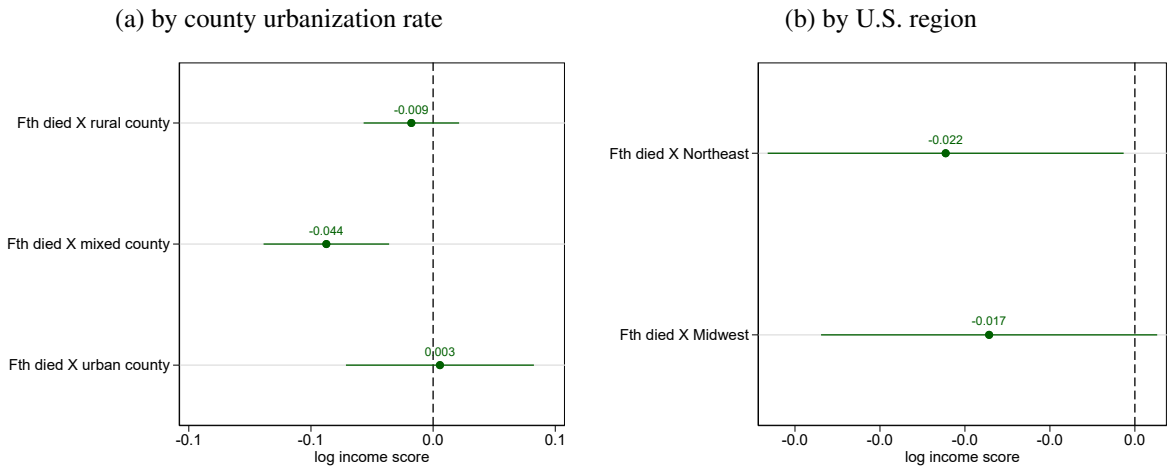


(d) by family size quartile in 1860



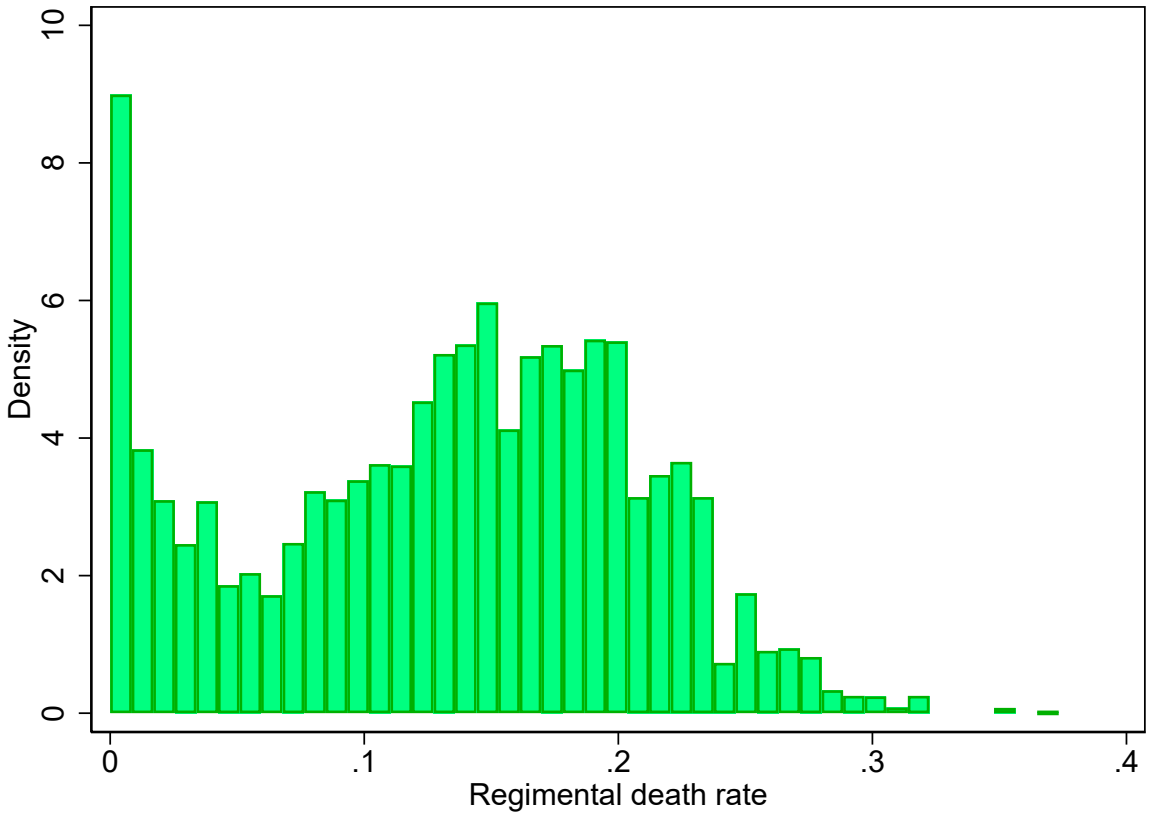
Note: Regressions of sons' log occupational income score in 1880 on an indicator for whether their father died in the U.S. Civil War interacted with the quantities indicated in panels a, b, c, and d. Skill-group classifications follow the 1950 definitions of the U.S. Census Bureau. Controls for sons' characteristics include their age and age squared in 1880. Father military controls include their enlistment date and enlistment date squared, ex ante service duration and ex ante service duration squared (ex ante service duration is the number of days between enlistment date and the date of disbandment of the regiment), fixed effects for their rank at enlistment, as well as characteristics of their last regiment of service: size, unit type fixed effects (infantry, cavalry, artillery, specialized fighting), share of privates, low level officers (captain and sergeant) and higher level officers. Father controls include 1860 baseline characteristics such as their age and age squared, occupational income score, indicators for being illiterate and foreign-born, and the inverse hyperbolic sine (IHS) of wealth. Mother controls include the same baseline variables measured in 1860 and an indicator for whether there was a mother present in the household. County fixed effects pertain to the county of residence of the father and son in 1860. The quartiles for son's age at father enlistment are 0-4, 5-7, 8-12, and 13+ years. The quartiles for family size in 1860 are 1-4, 5, 6-7, and 8+ household members. Standard errors clustered by the father's last regiment of service. Error bars display 95% confidence intervals. The p-values from the F tests for joint significance are 0.0192 (panel a), 0.1500 (panel b), 0.0274 (panel c), and 0.1051 (panel d). In panel a, the top quartile interaction is statistically different from the second quartile at the 10% level. In panels b to d, the interactions are not statistically different from one another at the 10% level.

Figure 4: Heterogeneity by 1860 place of residence



Note: Regressions of sons' log occupational income score in 1880 on an indicator for whether their father died in the U.S. Civil War interacted with binaries indicating the urbanization rate of the county of residence in 1860 (panel a), and the region of residence in 1860 (panel b). Rural counties have an urbanization rate of 0% in 1860 (50.1% of the sample). Mixed counties have a strictly positive urbanization rate below 50% (35.8% of observations). Urban counties have an urbanization rate above 50% (14.1% of observations). Urban population is population agglomerated in places of 2,500 inhabitants or more. In 1860, 47.8% of our sample live in Northeastern states (Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island & Vermont) and 52.2% live in Midwestern states (Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio & Wisconsin). Controls like in Table 2 and Figure 3. Standard errors clustered by the father's last regiment of service. Error bars display 95% confidence intervals. In panel a, the interaction with mixed counties is statistically different from the interactions with urban and with rural counties at the 5% level.

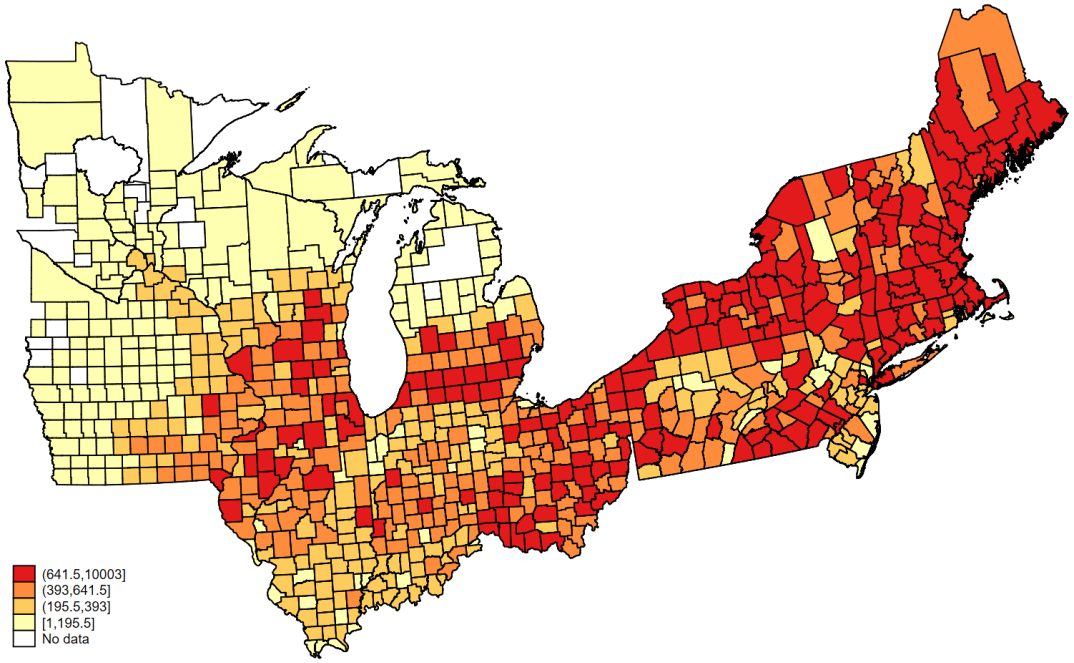
Figure 5: Distribution of Regimental Death Rate for our Sample of Linked Soldiers



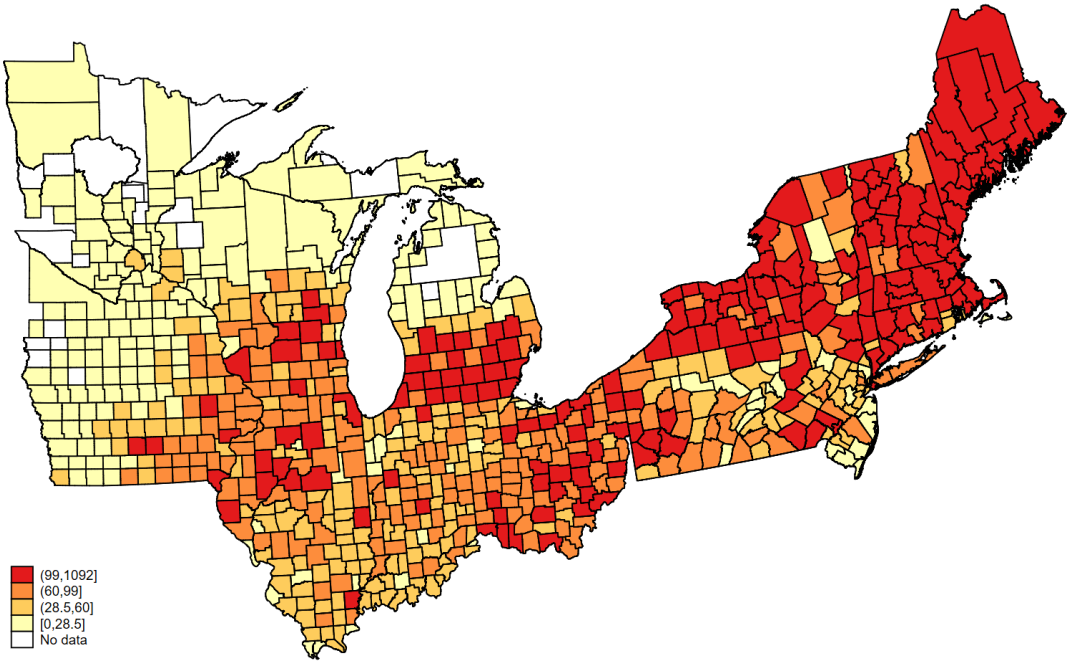
Note: Distribution of regimental mortality rates in the sample of soldiers who had children in 1860 and who we linked to the 1860 census. The spike at zero is explained by short-term regiments such as 1- and 3-months regiments that were mustered mainly for guard duty and never saw any action, regiments mustered towards the end of the war, as well as administrative and other supporting non-combat units. The average regimental death rate is 13% with a standard deviation of 7.5 percentage points.

Figure 6: Geographic Distribution of Linked Soldiers and Mortality Rates

(a) Soldiers Linked to the 1860 Census



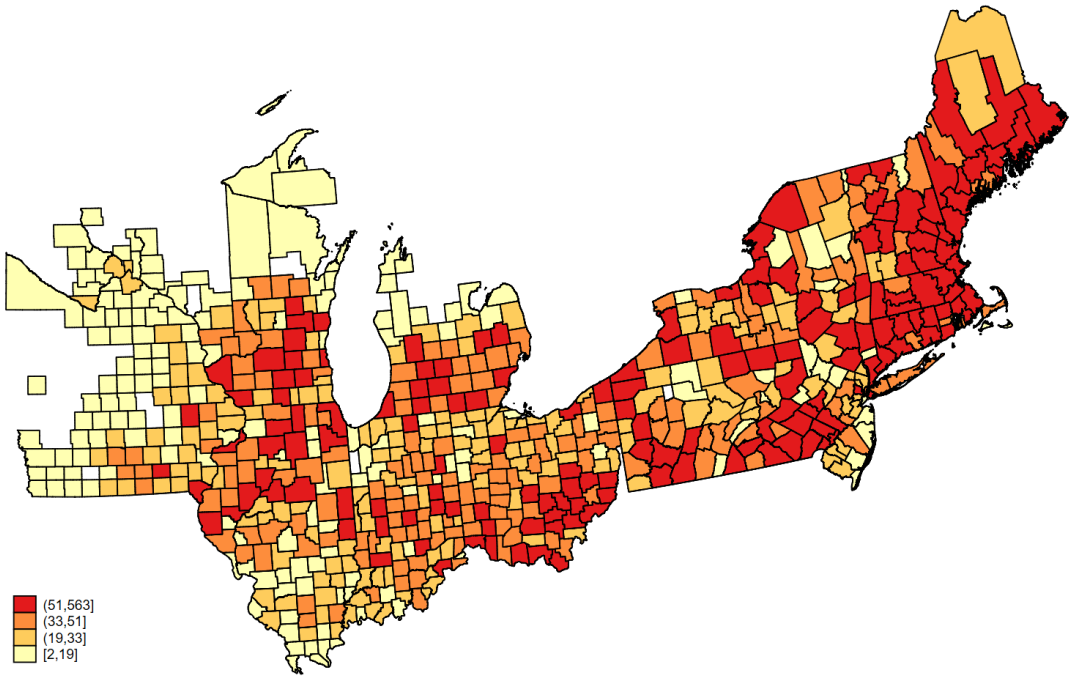
(b) Number of Fallen Soldiers



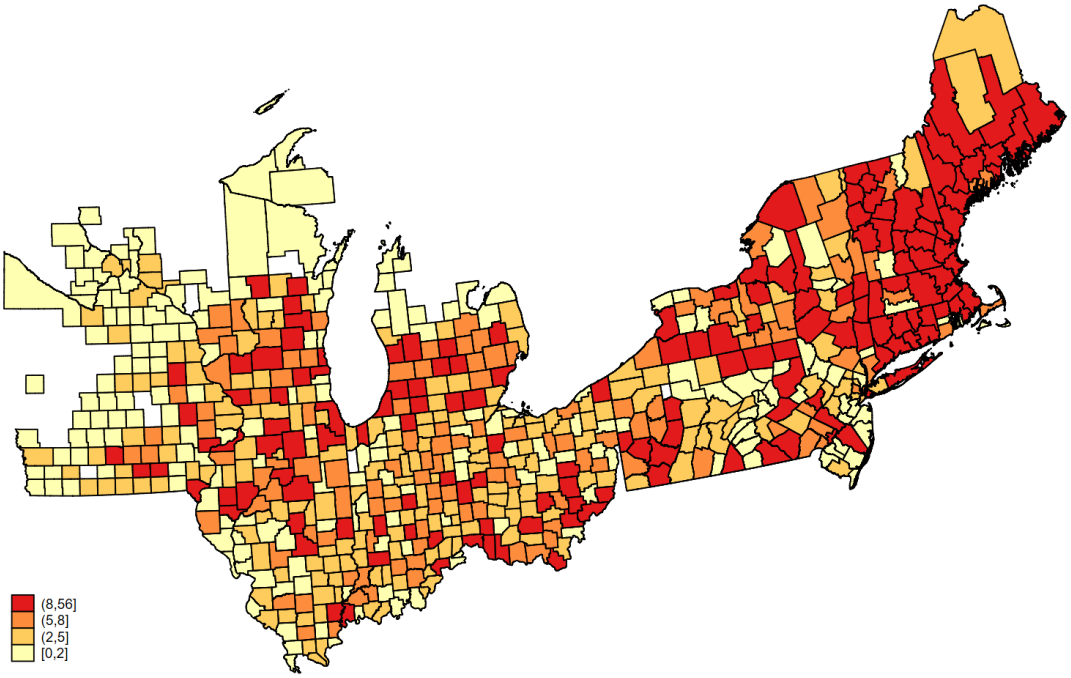
Note: Panel a shows the geographic distribution of the 482,983 soldiers who we managed to link to the 1860 U.S. census. Panel b plots the geographic distribution of the number of fallen soldiers among those we linked.

Figure 7: Geographic Distribution of Linked Sons

(a) Location of Sons in the Sample in 1860



(b) Location of Sample Sons who lost a Father



Note: Panel a shows the geographic distribution of the 29,558 sons in the sample for whom we could link their father from the military records to the 1860 census, and who we could track into the 1880 census. Panel b plots the geographic distribution of the sons in our sample who lost their father.

Appendix

List of Figures

List of Tables

A Appendix Tables

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

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

Table A.2: Military Records Summary Statistics

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table A.7: OLS Robustness to Different Linking Techniques

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table A.13: Placebo IV regressions

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

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

Table A.14: IV Sensitivity to Father Characteristics

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

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

Table A.15: IV Robustness to Different Linking Techniques

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

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

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

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

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

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

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

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

B External Validity and Weighting

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

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

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

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

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

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

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

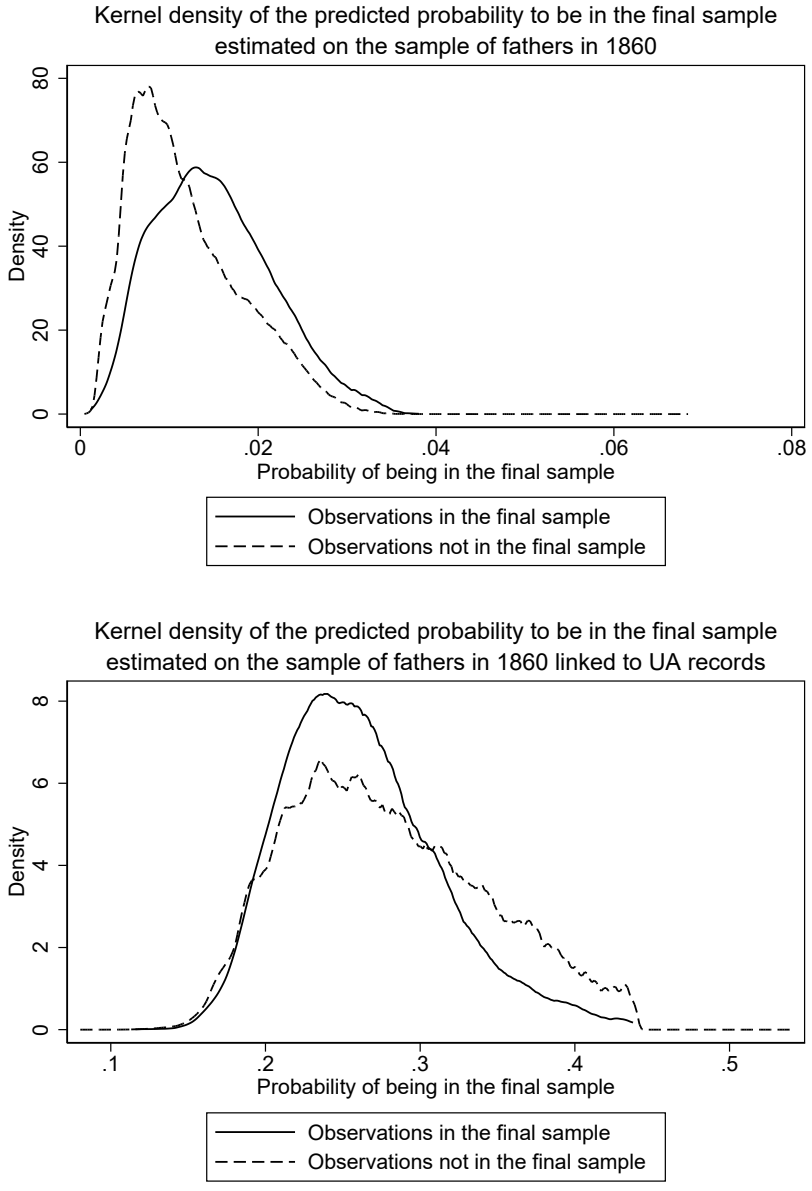


Table B.1: Effect of Father Death on Socioeconomic Characteristics of Sons in 1880 with Customized Weights

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

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

Table B.2: IV Estimation with Customized Weights

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

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

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

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

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

D Front Line Service and Socioeconomic Regiment Composition

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

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

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

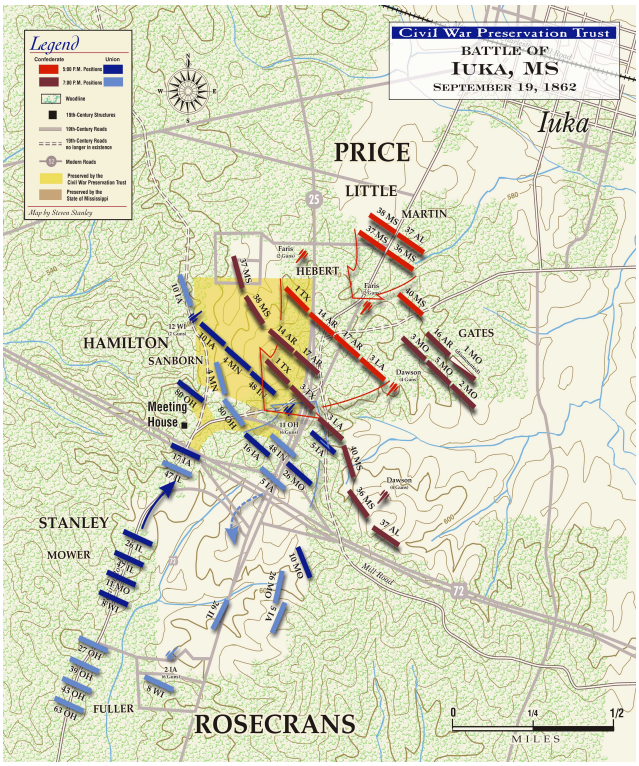
This resulted in 4,147 unit-battle-stage locations for a total of 128 battles and 799 unique Union units. Battles tend to be large with an average number of 20.5 Union units where a typical infantry regiment consists of 1,000 men. To compute the economic composition of each regiment, we used the individual-level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

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

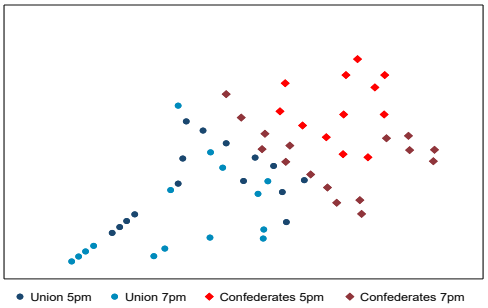
[3]88 of the 128 maps show unit positions for different stages of a battle. This means that there is within-battle variation in the location of regiments. The average battle has 1.45 stages with a maximum of 5.

Figure D.1: Digitizing Civil War Battle Maps

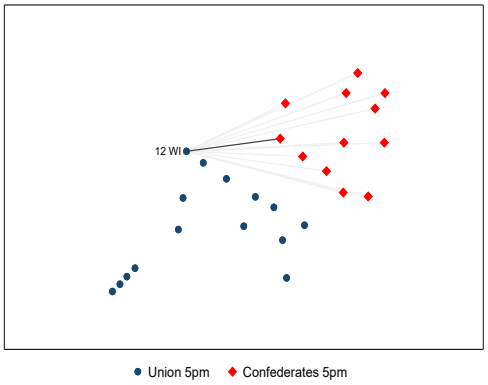
(a) Raw Battle Map



(b) Digitized Battle Map



(c) Minimum Distance to Enemy



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

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

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

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

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

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

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

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

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

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

Table D.1: Battle Distance Summary Statistics

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

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

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

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

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

E The Bias of OLS and IV Resulting from Linkage Errors

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

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

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

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

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

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

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

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

and the mis-measured orphan status can be expressed as

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

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

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

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

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

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

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

to derive the OLS bias as,

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

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

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

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

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

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

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

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

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

and therefore the bias of the IV estimator is,

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

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

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

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

E.1 Evidence from a Simulation Exercise

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

The true estimating equation is,

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

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

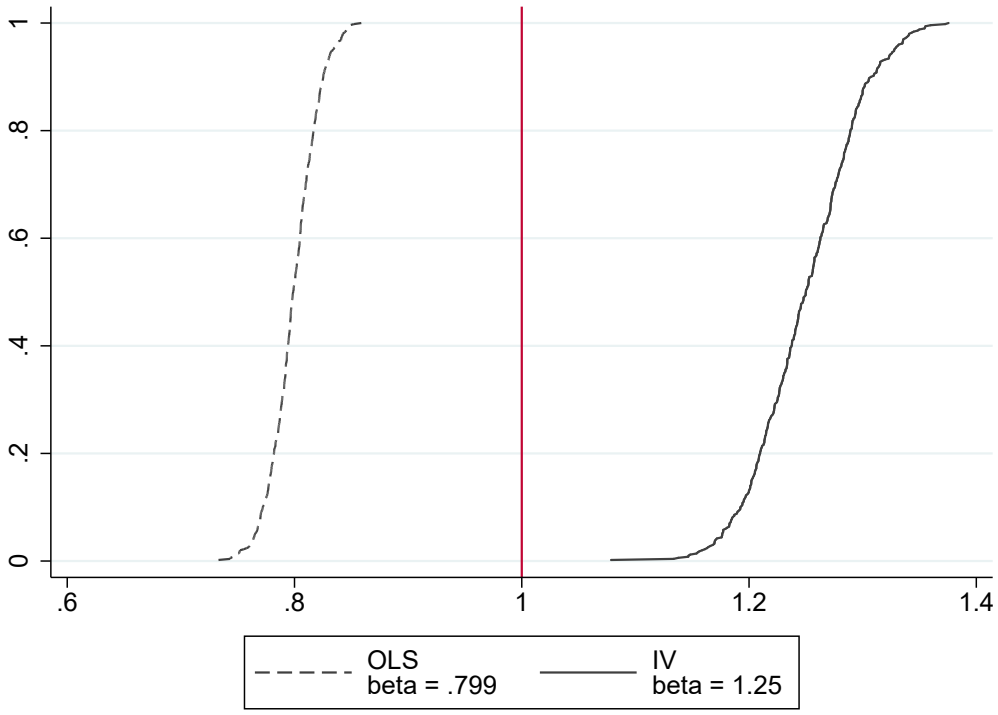
$$x_i^* = \frac{2}{3} z_i + \xi_i \tag{E.10}$$

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

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

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

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



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

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

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

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