Childhood Health and Human Capital: New Evidence from Genetic Brothers in Arms

John Parman*

June 6, 2013

Abstract

Negative shocks to childhood health can have a lasting impact on the economic success of an individual by altering families' schooling investment decisions. This paper introduces a new dataset of brothers serving in World War II and uses it to demonstrate that improvements in childhood health led to substantial increases in educational attainment in the first half of the twentieth century. By exploiting variation in health within families, the data show that this relationship between childhood health and educational attainment holds even after controlling for both observed and unobserved household and environmental characteristics.

*College of William & Mary and NBER, Department of Economics, P.O. Box 8795, Williamsburg, VA 23187-8795, jmparman@wm.edu; I have benefited greatly from discussions with Trevon Logan, Joseph Ferrie, Doug Miller, Hoyt Bleakley, Marianne Wannamaker and participants at the Economic History Association annual meeting, the UC-Davis economic history seminar and the Center for Population Economics workshop. I thank Kirstin Miller for her excellent research assistance.

1 Introduction

Health has an important role in the acquisition of human capital. Not only is health a primary determinant of a person's physical human capital, it also has a crucial role in human capital acquired through education. Health and education go hand in hand, with healthier individuals more able to attend school and more likely to benefit from their time in the classroom. It comes as little surprise then that the rapid improvements in American educational attainments during the first decades of the twentieth century were concurrent with dramatic improvements in health. This paper presents new evidence on the relationship between health and education between the late 1890s and early 1920s, a period when the United States experienced dramatic increases in secondary school attendance, in an effort to understand the extent to which the improvements in American health paved the way for the growth of the American human capital stock.

Using data on the educational attainment and height of World War II enlistees, I explore the link between childhood health and education. The evidence shows that at an aggregate level, the secular trend of rising educational attainment in the first decades of the twentieth century mirrored that of improvements in health. This is a relationship that holds across time as well as space, with higher average educational attainments across cities and states correlated with lower child morbidity rates and higher average heights.

To move beyond aggregate trends to an analysis of the impact of childhood health at the individual level, I link the military enlistment records to federal census data to construct a sample of sets of brothers. These sibling data afford a unique opportunity to look at variation in childhood health within households and its effects on educational attainment in a historical context. The sibling data are used to demonstrate that differences in childhood health across brothers within a household, proxied by differences in adult height, translated into significant differences in educational attainment. These results suggest that even when holding both observable and unobservable parental and environmental characteristics constant, negative shocks to childhood health had important consequences for long term outcomes in the early twentieth century. The link between health and education operated not just at an aggregate level across time and place but also within households. This relationship between height and educational attainment is substantially stronger in cities with worse childhood disease environments, suggesting that the relationship is being driven by differences in height resulting from differences in childhood health rather than simply being the product of height discrimination.

Given the dramatic public health improvements over this period, the results of the paper raise the possibility that a substantial portion of the increase in American educational attainments was made possible by improvements in health. The strength of the relationship between childhood health and educational attainment for the United States offers important lessons for education and health policies in developing countries today facing levels of childhood disease and school enrollment rates comparable to those of the United States in the 1920s and 1930s.The experience of the World War II enlistees suggests that improvements in health and education go hand in hand; attempts to improve educational institutions and increase an economy's human capital stock depend on a population healthy enough to take advantage of those improvements.

2 The Existing Literature on Childhood Health and Human Capital

A positive correlation between measures of childhood health and human capital is quite intuitive. Poor health as a child hinders both the ability to attend school and the ability to learn while in school. These consequences of poor health would translate into poor adult outcomes in the form of lower educational attainment, a potentially lower quality stock of human capital and lower earnings as a consequence. Reinforcing this negative relationship between childhood health and human capital formation and subsequent earnings is the correlation between childhood health and parental and environmental characteristics. Children from poorer households or less developed regions will tend to receive fewer investments in their human capital and be subject to harsher health conditions. Numerous studies have attempted to estimate the relationship between childhood health and human capital formation and to disentangle the role of health from the effects of parental characteristics and other environmental factors. Studies have used a broad range of measures of childhood health and a variety of identification strategies for isolating the causal impact of childhood health.

Birthweight has been the most commonly used measure of childhood health. It serves as a strong indicator of health in utero and is correlated with health outcomes throughout childhood.

Behrman and Rosenzweig (2004) use differences in the birthweights of twins to estimate a positive impact of fetal growth on educational attainment. In a similar use of twin data, Black, Devereux and Salvanes (2007) demonstrate a positive relationship between birthweight and a variety of adult outcomes including height, cognitive ability, earnings and educational attainment. Royer (2009) uses California twins to link differences in birthweight to differences in educational attainment and the birthweights of the next generation.

Several studies have explored alternative measures of childhood health. Oreopoulos et al. (2008) supplement data on birthweight with differences in Apgar scores and gestational length between twins and other siblings and find that infant health predicts both high school completion and social assistance takeup and length. Case, Fertig and Paxson (2005) measure childhood health with teenage height as well as the incidence of chronic illness as a child and find that poor childhood health leads to lower educational attainment and socioeconomic status for adults. This relationship between height and labor market outcomes is also demonstrated by Case and Paxson (2008) who point to the positive correlation between height and cognitive ability as an important part of the explanation for why measures of height are related to labor market outcomes.

The studies cited above rely on individual health measures to examine the relationship between childhood health and adult outcomes. To address issues of endogeneity, they typically focus on differences in these measures across siblings or twins to control for common household and environmental conditions, an approach made possible by modern panel studies and health surveys. A second strand of the literature on health and human capital focuses on aggregate measures of the health environment rather than individual measures of health outcomes to identify the effects of health on human capital formation. Such an approach makes identification possible without the detailed panel data needed for sibling or twin studies, broadening the scope of the time periods and regions that can be studied. Examples of modern studies taking this approach include Alderman et al. (2001) in which price shocks affecting childhood nutrition are used to estimate a positive relationship between childhood health and school enrollment and Alderman, Hoddinott and Kinsey (2006) in which shocks to childhood nutrition caused by drought and civil war are used to demonstrate a positive relationship between childhood nutrition and adult height and completed schooling.

Changes in disease environment have also been used to identify the effects of childhood health on human capital. The phasing in of deworming drugs to schools in Kenya was shown by Miguel and Kremer (2004) to substantially improve health and school participation. Bleakley (2007) finds that the eradication of hookworm led to increases in school enrollment, attendance and literacy in the American South. Almond (2006) shows that being in utero during the 1918 influenza pandemic had a negative effect on educational attainment and adult socioeconomic status. These studies by Bleakley and Almond represent some of the only work on the historical relationship between childhood health and human capital in the United States. However, because they rely on specific events in the history of hookworm and influenza for identification, the results speak to a narrow set of health issues and time periods. It is a tradeoff between clean identification and the scope of health issues considered that is an inescapable feature of studies exploring the intersection of health and human capital.

This study builds on the work of Bleakley and Almond, developing a broad picture of the secular trends in health and education over the first decades of the twentieth century and the extent to which improvements in health were driving aggregate improvements in educational attainment and individual educational investment decisions. Beyond the eradication of hookworm and the 1918 influenza pandemic, the late-nineteenth and early-twentieth century were a period of sweeping changes in both the health and human capital stock of the American population. In terms of health these decades witnessed extensive improvements in public health efforts in the form of better sewage and sanitation, better understanding of diseases and the ways to combat them, and improvements in the general awareness of good health practices. The results of these improvements were profound, with mortality rates falling, average adult heights rising and the mortality penalty associated with living in urban areas disappearing.[1]

The dramatic improvements in health were matched by equally impressive improvements in the educational attainments of Americans. The first decades of the twentieth century were a period in which public school systems improved and expanded throughout the country and increasing importance was placed on children attending and completing school. School attendance

[1]For descriptions of public health innovations in the late nineteenth and early twentieth centuries, see Cutler and Miller (2005), Meeker (1972), Condran and Crimmins-Gardner (1978) and Preston and Haines (1991). For an overview of trends in American health, see Costa and Steckel (1997). Komlos and Lauderdale (2007) demonstrate that health gains as proxied by height were rapid for pre-Depression birth cohorts with the rate dropping off after that. The height data used in this study afford a significantly more detailed view of this period than the data available to Komlos and Lauderdale.

rates rose and translated into higher literacy rates and greater numbers of high school graduates.[2] As the United States rapidly transformed into a healthier nation, it was simultaneously becoming a more educated nation. The findings of Bleakley and Almond suggests that in the cases of hookworm and influenza, changes in health had a direct impact on educational attainments at this time. The goal of this study is to test whether such a relationship existed more broadly over cohorts born from the late nineteenth century up to the Great Depression.

3 Data and Methodology

The dynamic nature of the early twentieth century both in terms of health and schooling makes it a fascinating focus of study but also introduces a number of complications in terms the data and methods that can be employed. With extensive heterogeneity in the quality and availability of schools, large differences in health environments across cities and regions and a wide range of attitudes toward investing in education, identifying the impact of health on educational attainment independent of all of these other factors is a difficult task. The general approach I will take is to create a sample of brothers and use the differences in health and educational attainment between brothers to estimate the effect of health on education holding both observed and unobserved family and environmental characteristics constant. The remainder of this section outlines the specific data sources and estimation issues associated with this empirical approach.

3.1 Historical Health and Education Data

First and foremost among the difficulties in executing a study of health and education for the early twentieth century is finding historical data on health and education. Childhood health and educational attainment throughout the history of the United States are topics that have received significant attention and a variety of interesting and useful data sources have been uncovered. However, these data are typically either aggregated at a level that makes them unusable for a study of individual outcomes or extremely narrow in their coverage.

Historical health measures used in past studies have typically been mortality statistics at

[2]See Goldin (1998) and Goldin and Katz (2008) for descriptions of the changes to the educational system and the trends in literacy, attendance, and graduation rates in the early decades of the twentieth century.

the city or state level. Such statistics provide an excellent way to capture the overall health of the population and geographic variation in conditions but make it difficult to estimate a meaningful relationship between health and educational investments. Consider a city with high mortality rates because of underinvestment in public works. If such a city shows a similar distaste for investment in educational institutions, the inhabitants may have both poor health and low educational attainment but the link between the two may signify nothing more than an aversion to public investment, not a causal link from poor health to low educational attainment. Far more useful would be data on individual health outcomes that could be related to individual educational attainment outcomes.

Several potential sources of individual health outcomes exist. Vital statistics from hospital records and information from death certificates can be used to obtain mortality data for individuals. Records from the federal census can offer only very limited information on morbidity (through a question on work related illness that was included in the 1880 federal census) and infant mortality (by comparing the number of surviving children to the number of children ever born at the household level). This lack of individual-level health data is a minor concern next to the difficulty in obtaining information on education. While certain details of health may be inferred from birth and death certificates as well as federal census data, comparable sources of individual-level educational attainment data are extremely rare. Information on educational attainment in federal censuses in the early decades of the twentieth century is limited to questions on literacy and whether individuals are attending school at the time of the census, both extraordinarily crude measures of education. State censuses in Iowa and South Dakota did collect information on years of educational attainment but reliance on these censuses would severely limit the scope of the analysis, losing much of the heterogeneity in health environments and school characteristics across the country and still not resolving the issue of a lack of good health data. It is this paucity of both health and education data at the individual level that has limited research on the historical relationship between health and education.

3.2 World War II Enlistment Records

There is one data source that is uniquely suited to resolving the problem of finding both health data and educational attainment data reported for individuals, the enlistment records of individuals serving in World War II. An enlistment card with personal details was filled out for each individual enlisting in the army for World War II and kept on file. The National Archives and Records Administration obtained these cards and digitized the information on them, creating a database of over nine million enlistees. The personal details on these cards include both years of secondary and post-secondary education, height and weight, and demographic information including race, year of birth, state of birth, state and city of residence at the time of enlistment and year of enlistment. Given the large number of individuals drafted into the army for World War II, these enlistment records offer a nationally representative sample of individuals with both education and health statistics reported at the individual level.[3] The records include individuals who enlisted between 1938 and 1943. Given this five year range of enlistment dates and the wide range of ages for individuals at the time of enlistment, the enlistment records include individuals born between the late 1890s and the early 1920s.

The education information in these enlistment records is as detailed as one could hope for in a large historical sample. Short of having measures of cognitive ability, years of educational attainment is the best measure of human capital accumulation typically available even in modern studies. The value of the height and weight variables as a measure of health and in particular childhood health is less clear. Weight clearly varies for reasons unrelated to childhood health. Heterogeneity in adult weight reflects differences in adult behaviors as much if not more than childhood health conditions. Adult height, however, is invariant to adult behavior but influenced by childhood health conditions. The height reported in the enlistment records, as the cumulative product of childhood health experiences, offers a potentially useful proxy of childhood health.

In a review of previous scientific studies, Silventoinen (2003) finds that roughly 80 percent of the variation in height in modern Western societies is due to genetics while the remaining 20

[3] There are two important respects in which the sample of enlistees is not representative of the population as a whole. First, the enlistee records offer a representative sample of *males*. There are female enlistees in the database but the number of females is quite small. Both the limited number of female observations and the empirical strategy of looking at differences in height across siblings (which requires siblings be of the same gender) limit this study to males. Second, there were minimum physical requirements for enlisting although these requirements were relaxed over the course of the war. Consequently, the enlistee sample is underrepresentative of the shortest and least healthy members of the population. This can be seen in Figure 7 in the appendix giving the height distribution for veterans of World War II and civilians. The upper tails of the height distributions appear identical but there are clearly more individuals in the lower tail of the distribution for civilians. The most direct consequence of the sample being overrepresentative of healthy individuals is that the random variance in heights relative to the variance due to childhood health differences will be larger in the sample than in the population as a whole. As Section 5 discusses, this will lead to an attenuation bias and underestimates of the effects of health on educational attainment.

percent of variation is due to environmental factors. It is this 20 percent that has the potential to capture an individual's childhood health history and is a product of such things as nutrition and childhood disease. Several studies confirm that a variety of measures of childhood health are indeed correlated with adult stature. Behrman and Rosenzweig (2004) and Black, Devereux and Salvanes (2007) both find a positive relationship between birthweight and adult height. Bozzoli, Deaton and Quintana-Domeque (2008) find a positive relationship between infant mortality rates and adult heights for the United States and Europe in the second half of the twentieth century. In terms of the impact of specific childhood diseases on height, Voth and Leunig (1996) argue that smallpox had a strong negative effect on the heights achieved by survivors, estimating that smallpox reduced adult height by as much as one inch.

While height is a promising proxy for childhood health, it is not without its problems. First and foremost is the 80 percent of variation in height that is due to genetics. Given this large component of height that is unrelated to childhood health, adult height is a noisy measure of childhood health. This raises standard issues of a mismeasured variable, most importantly an attenuation bias. Beyond this econometric issue is a set of more fundamental concerns that height fails to capture important differences in childhood health experiences and can in fact be inversely related to childhood health in certain circumstances. These concerns stem from two important features of childhood development: children with negative shocks to their health early in life often experience catchup growth later in life and the children who survive negative health shocks early in life may have done so because they had a high health endowment to begin with.

The issue of catchup growth is raised in Oxley's (2003) critique of Voth and Leunig's study of smallpox. Oxley notes that while short term diseases do have a direct effect on nutritional status at the time of the disease, these effects are often limited to the duration of the disease and do not necessarily lead to long-term stunting. Steckel's (1986) study of American slaves indeed found that slaves could overcome even protracted periods of poor nutrition and still attain healthy adult heights. However, Steckel notes that this catch-up growth was unusual; other developing countries with short child populations tended to have short adult populations. To the extent that catch-up growth does occur, height differences across individuals as adults will tend to understate differences in their childhood health histories.[4]

[4]While this is true of final adult heights, there is one feature of catch-up growth that helps make height a more useful measure of childhood health in this paper than in other studies. The sample consists of fairly young adults, with

The selection issue is more problematic. Suppose that children have an unobserved health endowment that is drawn at random from some distribution. A child with a better health endowment would both achieve a greater adult height and be more resistant to disease. If diseases strike children randomly, the set of individuals not experiencing a disease would have health endowments matching the distribution from which they were initially drawn. Individuals that did experience a childhood disease would have a different distribution of health endowments as the children with the lowest health endowments may die from the disease. Those children that survive would have a health endowment that is higher on average than the mean of the distribution at birth.

If mortality rates from childhood diseases are relatively low, we can still expect the average height of adults that experienced a childhood disease to be lower than adults that had healthy childhoods. However, in cases of very high infant and child mortality rates, it is possible for the survivors of childhood disease to be taller on average than children with disease-free childhoods, a phenomenon documented by Bozzoli, Deaton and Quintana-Domeque (2008) in extremely poor countries. This does not render height a meaningless indicator of childhood health, it simply requires a more nuanced interpretation of differences in height. Specifically, height is a measure of both the health shocks a person was subjected to as a child as well as the person's unobserved health endowment. In the next section, I will provide evidence that in the case of the United States, height is negatively correlated with childhood diseases suggesting that in the enlistee data the negative impact of disease on survivors' heights dominates the selection effect of the least healthy children not surviving to adulthood. These data on childhood diseases come from a panel of state-level mortality data collected by Grant Miller and city-level mortality and morbidity data collected for this project from the Public Health Reports of the United States Public Health Service. Details on these data sources are provided in the appendix.

Imperfect as it may be, adult height will serve as the measure of childhood health. The sample of enlistees then offers several million observations of adult males born between the late 1890s to

many of the enlistees in their late teens or early twenties. As Case and Paxson (2008) note, one feature of catch-up growth is that it causes individuals who experience negative health shocks as children to achieve their full height at a later age than individuals who had healthy childhoods. This suggests that the differences in heights due to differences in childhood health observed for the young adults in the sample used in this study may be larger than if the same individuals were observed later in life. The enlistee sample may afford more heterogeneity in height driven by differences in childhood health than a study using a sample of older individuals.

the early 1920s with information on educational attainment, a proxy for childhood health and a set of demographic variables including race, state of birth, state and county of residence at the time of enlistment, year of birth and year of enlistment. Given the young ages of enlistees, there is a concern that their educational careers were interrupted by the war. To limit the sample to those individuals who have completed their educational careers, I calculate the number of years out of school as age minus years of secondary and post-secondary education minus fourteen. I then keep only those individuals who have been out of school for two or more years.[5]

3.3 Matching Enlistee Records to the Federal Census

With these data alone, one could regress educational attainment on height as well as age, race and geographical controls to estimate the relationship between education and childhood health but the resulting coefficient on height would suffer from a wide range of biases. A variety of variables correlated with both education and health would be omitted from the analysis. Relevant household characteristics including parental income, parental education and family size would be missing as would important information about the local disease environment and local school resources. To overcome these problems of omitted variable bias, I match the enlistment records to federal census data to incorporate information on the enlistees' childhood households and identify sets of brothers in the sample to control for unobserved household and environmental characteristics. By matching individuals from the enlistment records to their childhood households in the federal census, I can control for observed household characteristics that are correlated with both health and educational attainment including household location, family size and family income. By using the family information from the census to establish sets of brothers in the enlistment records, I can further control for unobserved household and environmental characteristics by looking at how differences in health between brothers translate into differences in educational attainment.

[5] I am assuming that all years of schooling are completed consecutively and that high school is started at age 14. To the extent that high school may have been started at a later age or time off was taken in between years of schooling, some of the men in the sample may not have actually completed their educational careers. For these men, I will be underestimating their overall educational attainment. An alternative approach would be to redefine educational attainment as educational attainment achieved by a particular age. While this would allow me to drop the assumption that educational careers have been completed, it would still require the assumption that schooling years are being completed in consecutive years and consequently does not seem to offer any significant advantages over the chosen approach.

This matching is a time consuming process and requires sampling the enlistees rather than attempting to match the full sample to census records. Given that I ultimately want to use brothers to control for family characteristics, I first limit the set of enlistees to those that have a potential brother in the records. Individuals are defined to be potential brothers if they share the same last name, were born in the same state, lived in the same state and county at the time of enlistment and were born within three years of each other.[6] The restriction on the difference in ages is used to ensure that the brothers grew up in similar environments. The greater the difference in ages between brothers, the more likely it is that household income, educational resources or some other important but potentially unobserved variable differed across the two brothers. I then take a ten percent sample of these sets of potential brothers and attempt to link them to the federal census.

Linking to the federal census is done by searching for individuals by first name and last name, state of birth and year of birth in an electronic database of the 1920 or 1930 federal census records.[7] If two potential brothers have unique matches in the federal census I then check images of the original census records to determine whether the individuals had the same parents.[8] If they did, the brothers are confirmed as a match and included in the final dataset. Information on their parents' occupations, ages and places of birth are transcribed from the census records as well as the number of children in the family, the number of male children and the birth orders of the children.[9]

The end result is a sample of roughly 4,000 sets of brothers that contains information on the height, weight and educational attainment of each brother from the World War II enlistment records as well as information on the location of their childhood household, birth order, family size and parental occupations and incomes from the federal census.[10] Table 1 provides summary

[6]Clearly this definition does not cover all pairs of brothers. These strict criteria to identify brothers are chosen in order to have the highest rate of successfully matching brothers in the federal census. The drawback of this approach is that the sample is biased toward brothers that have remained geographically close to one another. If brothers who are more geographically mobile have a different relationship between childhood health and educational attainment than less mobile individuals, the regressions will produce biased estimates. The direction of this bias is ambiguous.

[7]For each set of potential brothers, I search the earliest census in which all of the potential brothers in the set were alive. This gives the highest probability that the brothers are still living in their parents' household.

[8]A unique match is defined as finding a single individual with the exact same name and childhood state of birth as the enlistee and a birthyear that is within one year of the birthyear given on the enlistment records. If there are no matches meeting these criteria or multiple matches meeting this criteria, the enlistee in question is dropped from the sample as a failed match.

[9]A more detailed description of the entire linking procedure is provided in the appendix.

[10]Income was not reported in the federal censuses used in this study. Parental incomes are imputed from the

statistics for the sample of matched brothers. For comparison, the height, weight and education statistics are also provided for the entire set of male enlistees in the World War II enlistment records and household statistics are given for all households with at least one child from a one percent sample of the 1930 federal census.[11] In terms of height, weight, educational attainment and father's income the sample of successfully linked brothers looks representative of the general population. What does differ substantially for the brother sample relative to the population is family size and composition. The matched brothers come from substantially larger families that tend to have more sons than daughters relative to the general population. This feature of the data is unsurprising given that families have to have at least two sons to be in the sample. It does raise a concern about the generalizability of any results from the matched brothers. It is quite possible that the relationship between health and educational investments varies by family size. Indeed, the estimates of the relationship between height and educational attainment presented later in the paper are sensitive to whether controls for family size and birth order are included. Consequently, while the sample appears representative of the population in terms of education, height, weight and family income, the results of this study may not be applicable to smaller families consisting of only one or two children.

When looking at the differences between brothers, brothers look similar to one another on average but there is still substantial variation within families. Height, weight and particularly educational attainment are highly correlated between brothers, with the standard deviation of these characteristics within a family typically half as large as the overall standard deviation of the variables in the population. However, the within family standard deviations are still large enough to suggest sufficient variation between brothers to estimate the effects of health on education within families. The following section will use the data for all enlistees to establish a strong relationship between childhood health and height as well as between height and educational outcomes across locations and over time. Section 5 then turns to the matched brothers data to examine the effects of household characteristics and within-family variation in health on educational attainment.

occupations using the median income for the occupation in 1950. These median incomes are taken from the Integrated Public Use Microseries (IPUMS) website.

[11]The sample of the 1930 federal census is the public use sample available through IPUMS.

4 Health, Height and Education in the Aggregate

Figure 1 offers a simple but telling picture of the relationship between health and education among World War II enlistees. The figure plots the mean height and educational attainment by year of birth for enlistees born between 1892 and 1923. The large means for height and educational attainment in the earliest years are a result of officers from World War I reenlisting for World War II. These officers were more highly educated and taller than the other enlistees of the same age, highlighting the correlation between height and education within cohorts. Removing these officers produces steadily increasing heights and educational attainments across the entire time span, demonstrating a similarly strong correlation in height and education over time.[12] The increases in average heights and average educational attainment over these decades are remarkable both for their rapid pace and for their similarity to one another. It is clear that the first decades of the twentieth century were a period of steady gains in both health and human capital for the population of the United States.

An alternative view of the aggregate relationship between health and education is given by Figure 2 which shows mean height and educational attainment by county. These maps reveal substantial heterogeneity across locations in terms of both height and educational attainment, with a range in mean heights of over three inches and the mean educational attainment varying from under one year of high school to four full years of high school. Regional differences are stark and highlight the value of having data for the nation as a whole rather than extrapolating from a small number states. The Midwest and in particular the more recently settled areas of the Midwest clearly led the country in both educational attainment and height. The South is striking for its relatively tall but poorly educated population. Looking within the South it becomes clear that despite the counterintuitive relationship between Southern heights and educational attainment relative to the rest of the country, a positive relationship between height and educational attainment exists within the region, with northern Texas and Oklahoma leading the region in both average height and average educational attainment.[13]

[12] Figure 4 in the appendix plots mean height by birth cohort excluding officers and individuals who had not yet completed their educational careers by the time of enlistment.

[13] This is a relationship that can be confirmed quantitatively. Regressing average educational attainment on average height where the unit of observation is either state or census region yields a positive but statistically insignificant coefficient on average height. Running the same regression with counties as the unit of observation produces a positive coefficient significant at the one percent level.

These trends are suggestive of a strong link between health and education. While the spatial variation in height may be due to genetic variation, the change in average height over time shown in Figure 1 could not realistically be driven by changes in genetics. The secular trend in height is far more likely the result of improvements in health due to environmental reasons (better sanitation, higher average incomes, etc.) that were concurrent with improvements in educational attainment. However, the empirical approach of this study relies on height also capturing variations in health across siblings. If the aggregate trends in height are the product of changes in nutrition, parental health, or any other factor that varies across but not within households, height will not allow for identifying the effects of childhood health separately from these other household and community level variables correlated with height. It is necessary to demonstrate that heights in the sample are also a function of health factors that could vary across brothers. The main source of such variation would be childhood disease, something that could differentially impact brothers who otherwise share the same environmental and household characteristics. If height varies with the incidence of childhood diseases in the sample, it remains a good candidate to be a proxy for childhood health differences between brothers. If height does not vary with levels of childhood disease, it would be unlikely to serve as a useful measure of the health differences between brothers.

Ideally one would test whether the heights of individuals vary significantly with their own histories of childhood disease to confirm that height is a useful proxy for childhood health. Unfortunately, there is no information on the health histories of the individual enlistees. What is possible is to test whether variation in heights across cities and states is correlated with the disease environments of those cities and states.[14] To measure disease environment at the state level I use an average of state level mortality rates by disease for 1910, 1920 and 1930.[15] To measure the disease environment at the city level, I collected data on disease specific morbidity and mortality rates from the Public Health Reports of the United States Public Health Service. At the city level, it is possible to look at average morbidity and mortality rates between 1918 and

[14]City and state refer to the city and state of residence at the time of enlistment. While I have information on the state in which enlistees were born, I do not know the city in which they were born.

[15]These data are taken from Grant Miller's database of state mortality rates for 1900 to 1936. While Miller's data is annual, only the 1910, 1920 and 1930 data are used because the state populations from the federal censuses in those years can be used to convert the number of deaths by disease into mortality rates.

1924.[16] I have collected these data for both the reports of large cities (defined as those with a population greater than 100,000 in 1925) and the reports of small cities (cities with a population between 10,000 and 100,000 in 1925). The analysis will focus on the large city data as they are more reliable than the small city data.[17]

The morbidity and mortality rates can be separated by disease into three categories: those primarily affecting infants, those primarily affecting older children and those primarily affecting adults. This allows for distinguishing between high disease environments that would lead to stunting (places with high rates of diseases targeting infants and young children) as opposed to high disease environments that would lead to poor adult health and high mortality but not necessarily stunting (places with high rates of diseases targeting adults). I infer the age distribution of diseases from the 1880 federal census which asked about sickness on the day of the census. The responses to this question reveal the incidence of specific diseases by age. Table 2 reports the mean age, median age and skewness for the age distribution of cases for major diseases from a one percent sample of the census.[18] Supplementing these data are mortality rates by disease from 1921 to 1925, a time period when most of the enlistees were children, collected from the annual mortality statistics published by the census bureau. These mortality data confirm that the age distributions of cases from 1880 match up quite well with the age distributions of deaths at the time that the enlistees are children.[19] Based on these age distributions, two different ways to group diseases are used, one based on the median age and one based on the skewness. The grouping by median age defines diseases affecting infants as those with a median age of less than 10, diseases affecting older children as those with a median age between 10 and 20, and diseases affecting adults as those with a median age above 20. The grouping by skewness defines diseases

[16]It is unclear whether morbidity rates or mortality rates are better measures to use for assessing the childhood disease environment. Table 3 gives the correlations between morbidity and mortality rates at the city level. For all diseases, these correlations are positive but not nearly as strong as one might expect. Given that I am interested in children that survive to adulthood, morbidity rates seem the most reasonable measure of the childhood disease environment. However, mortality rates indicate something about the severity of diseases not captured by the morbidity rates alone. They also have the advantage of being more consistently reported than morbidity rates over time and across locations.

[17]The main problems with the small city data are that there are large measurement error issues with the number of cases by disease and estimated populations, missing data for several diseases and far fewer enlistees per city. The combination of these factors makes it difficult to obtain precise results regarding the relationship between average height and morbidity or mortality rates.

[18]Histograms for the age distributions by disease are given in the appendix in Figure 5.

[19]I use the 1880 census data despite the fact that they are gathered well before the enlistees are born because they offer the incidence of cases rather than deaths. Cases may be the more relevant figure given that all of the enlistees have survived to adulthood.

affecting infants as those with a skewness of greater than 2, diseases affecting older children as those with a skewness between 1 and 2, and diseases affecting adults as those with a skewness of less than 1. Both ways of grouping diseases yield similar results. For the sake of brevity, only the results based on the skewness definition are presented here. Summary statistics for the state-level mortality rates and city-level morbidity rates by disease type are given in Table 3.

Figure 3 plots average height as a function of mortality rates for diseases targeting infants and diseases targeting older children.[20] The figure shows that the substantial variation in disease environment seen in the summary statistics is highly correlated with the variation in height across states. As one would expect, high levels of childhood disease are associated with lower average height.

Table 4 reports the results of regressing height on local levels of disease incidence controlling for enlistee race and a quadratic in age. The regression results are consistent with height being a function of childhood disease environment. For both the state-level and city-level disease data, an increase in the incidence of childhood diseases is associated with a statistically significant decrease in adult height. These effects are quite large. A one standard deviation increase in the city-level infant disease mortality rate decreases the expected height of an enlistee by roughly 0.2 inches while a one standard deviation increase in the state-level infant disease mortality rate decreases expected height by 0.3 inches. Furthermore, the levels of disease incidence actually explain a large portion of the variation in average height across cities and states. When city- and state-level average height is regressed on the disease incidence variables, the R^2 of the regressions are 0.29 and 0.82 respectively, suggesting that variation in disease incidence can account for nearly a third of the variation in average heights across cities and the majority of the variation in average heights across states.

At the aggregate level, the enlistee records reveal that both the health and human capital stock of the US population experienced steady and substantial increases during the first decades of the twentieth century. The disease incidence regressions suggest that the observed height differences among enlistees are due at least in part to differences in childhood disease environment, making difference in height between brothers a useful proxy for differences in their childhood

[20]An alternative depiction of the correlation between disease environment and height is presented in Figure 6 of the appendix. This figure presents maps of the United States showing disease incidence, average height and average educational attainment by state.

health histories. The next sections test whether this correlation between health and education at the aggregate level held as strongly at the individual level when controlling for observed and unobserved household characteristics.

5 Height and Education Within Families

5.1 Estimation Strategy

Two different approaches to estimating the relationship between height and educational attainment at the individual level will be used. The first will be a simple linear regression of educational attainment on height and other characteristics of the following form

$$E_{i,j} = \alpha + X'_{i,j}\beta + Z'_j\gamma + \theta_j + \varepsilon_{i,j} \tag{1}$$

where $E_{i,j}$ is the educational attainment of individual i from family j, $X_{i,j}$ is a vector of observable characteristics for individual i, Z_j is a set of observable characteristics for his family, θ_j is a term capturing unobservable family characteristics and $\varepsilon_{i,j}$ is an individual-specific error term assumed to be independent of the other terms in the equation. The individual characteristics include the primary variable of interest height as well as birth order among siblings and a quadratic in age. Z_j includes household location, family size, race and family income. The problem with estimating equation (1) is that θ_j, capturing such variables as local school quality and parental tastes for investment in children, is unobserved and potentially correlated with height leading to an omitted variable bias for the estimated coefficient on height.

To address this problem, the variation within households can be exploited by including family fixed effects. Estimating Equation (1) with the inclusion of family fixed effects will produce an unbiased estimate of the coefficient on height. The drawback of using a fixed effects estimator is that the fixed effects will absorb any variables that do not vary across brothers such as family size, father's income, and local disease environment.

The second approach to estimating the relationship between height and educational attainment is motivated by the distribution of educational attainment in the sample of brothers. Over 25 percent of the brothers report zero years of secondary and post-secondary education. Over 30 percent report being high school graduates with exactly four years of secondary education and no postsecondary education.[21] With such clear bunching at zero and four years of secondary education, a nonlinear functional form for modeling the relationship between height and education seems reasonable. Of particular interest are the effects of height on the probability of attending high school, the probability of graduating high school and the probability of attending college. The difficulty with using these binary dependent variables and a nonlinear functional form is that incorporating family fixed effects presents an incidental parameters problem potentially leading to inconsistent estimates of all of the coefficients, not just the fixed effects.[22]

One way to avoid the incidental parameters problem and still estimate a nonlinear relationship between education and height is to use a conditional logit specification. First consider the probability of graduating high school modeled as a function of personal characteristics and a family fixed effect using the logistic cumulative density function $\Lambda(\cdot)$:

$$Pr(HS_{i,j} = 1|X_{i,j}) = \Lambda\left(\alpha + X'_{i,j}\beta + \alpha_j\right) \tag{2}$$

where $HS_{i,j}$ is an indicator variable equal to one if individual i from family j graduated from high school, $X_{i,j}$ is a vector of observable individual characteristics, and α_j is the fixed effect for family j. If this equation was estimated by including family dummy variables for the fixed effects, the estimated coefficients would be inconsistent due to the incidental parameters problem. However, following the approach proposed by Chamberlain (1980), it is possible to condition on the sum of $HS_{i,j}$ within a family to rewrite the probability of high school graduation in a way that no longer depends on the family fixed effect. In the case of a pair of brothers for which only one brother graduated high school, the reduces equation (3) to:

$$Pr(HS_{1,j} = 1|X_{1,j}, X_{2,j}, \sum_{i=1}^{2} HS_{i,j} = 1) = \Lambda((X_{1,j} - X_{2,j})'\beta) \tag{3}$$

The above is a conditional logit model in which the estimation of β through maximum likelihood

[21]See Figure 8 in the appendix for a histogram of the educational attainment distribution.

[22]See Greene (2004) for discussion of the incidental parameters problem in the context of a wide range of nonlinear fixed-effects models.

does not depend on the family fixed effects.[23] This conditional logit model offers a way to obtain an estimate of the effect of height on the binary educational outcomes that does not suffer from the omitted variable bias problems of the basic logit model or the incidental parameters problem of the logit model with family fixed effects. The drawback of the conditional logit model is that it requires the sum of the indicator variable across brothers to be equal to one, substantially reducing the sample size.

5.2 Results

Table 5 presents regression results in which individual educational attainment is regressed on height and a set of household characteristics. While these regressions clearly do not control for unobserved household characteristics and will produce biased coefficient estimates, they do offer insight into how household characteristics related to educational attainment in the first half of the twentieth century, something previous studies have not been able to examine due to data limitations. The most striking results are the large and highly significant coefficients on height across all specifications. A one standard deviation increase in height is associated with an additional 0.1 to 0.15 years of secondary and post-secondary education, a large increase given that the average level of secondary and post-secondary schooling is only 2.4 years. This is comparable in magnitude to the marginal effect of a ten percent increase in father's annual income on schooling. Interestingly, the inclusion of father's log income in the regression has only a small effect on the height coefficient. Family income would have been one of the primary concerns in terms of omitted variable bias with a regression of education on height and yet the data suggest that the exclusion of father's log income leads to only a very small bias on the height coefficient. Beyond height and family income, family structure also proves important in explaining differences in educational attainment, with family size having a large and significant negative coefficient and the coefficients on sibling order suggesting that younger siblings receive more education than their older brothers and sisters.

[23] The derivation of equation (5) can be shown easily by first writing $Pr(HS_{1,j} = 1 | X_{1,j}, Z_j, \sum_{i=1}^{2} HS_{i,j} = 1)$ as

$$\frac{Pr(HS_{1,j} = 1, HS_{2,j} = 0 | X_{1,j}, X_{2,j}, Z_j)}{Pr(HS_{1,j} = 1, HS_{2,j} = 0 | X_{1,j}, X_{2,j}, Z_j) + Pr(HS_{1,j} = 0, HS_{2,j} = 1 | X_{1,j}, X_{2,j}, Z_j)}, \tag{4}$$

then rewriting the joint probabilies as the products of the individual probabilities of high school graduation and substituting the logit functional form for the probabilities.

The logit estimates in Table 6 produce very similar results when looking at the set of binary educational outcomes (high school attendance, high school completion and college attendance). As with the years of educational attainment regressions, increases in height and father's log income are both associated with improvements in educational outcomes. Once again a larger family size makes it less likely an individual will graduate from high school or attend college but conditional on a given family size, the later a child is in birth order, the more likely he is to graduate high school and attend college. This is an important distinction and one that underscores the value of having detailed data on families. The correlation of educational attainment and birth order in the data is negative. However, these results suggest that this negative correlation is driven by the effects of family size while within families younger children actually receive more educational investment than their older siblings.

The large height coefficients from the regressions in Table 5 and Table 6, while possibly an indicator of childhood health's direct impact on educational attainment, are also likely to be a product of better parents investing more in both education and health or areas with better health environments also having better schooling resources. To control for these unobserved characteristics, Table 7 presents the results from including family fixed effects. The results show that once unobserved family characteristics are controlled for, the magnitude of the height coefficient is cut roughly by roughly two thirds but the coefficient remains positive and significant. This suggests that even once controlling for household and environmental characteristics, childhood health as proxied by height has a significant impact on overall educational attainment.[24]

Variation in height within families also proves to be important in explaining differences in high school graduation and college attendance. Table 8 gives the results of the conditional logit regressions. While the coefficients on height are small and statistically insignificant for the high school attendance regressions, the coefficients on height in the high school graduation and college attendance regressions are positive and significant suggesting that taller brothers are more likely to graduate high school and attend college.

[24]It is worth contrasting this with the findings of Magnusson, Rasmussen and Gyllensten (2006), who find that controlling for family characteristics leads to a much smaller drop in the marginal effect of height on educational attainment. They focus on Swedish males born between 1950 and 1975, suggesting that for these more recent cohorts, variation within the family accounts for a much larger proportion of the relationship between height and educational attainment.

5.3 Interpreting the Effects of Height on Educational Attainment

These fixed effect and conditional logit results provide strong evidence that childhood health has a significant impact on educational attainment even after controlling for unobserved household and environmental characteristics. However, it is difficult to use them to assess just how large the effect of a childhood health shock is on educational outcomes. Differences in height are clearly correlated with differences in childhood health as demonstrated by the disease incidence regressions presented earlier. However, there is also a great deal of random variation in height that is not due to differences in childhood health. It is important to consider the downward bias introduced by this random variation in height when assessing the magnitude of the estimated coefficients and consequently the strength of the relationship between health and education. Suppose that height for individual i is given by

$$h_i = h_0 + \alpha health_i + \gamma family_i + \varepsilon_i \tag{5}$$

where h_0 is a baseline height, $health_i$ captures an individual-specific components of health that influence height, $family_i$ captures traits specific to the individual's family that influence height (which may be health-related), and ε_i is a random component of height with mean zero and uncorrelated with the other terms. While it is the effect of the health component of height on educational attainment that is of interest, the estimated coefficient on height is also picking up the effects of ε_i ($\gamma family_i$ is absorbed by the family fixed effects). The observed height can be thought of as the health component of height (the variable of interest) plus mean zero measurement error in the form of ε_i.

This measurement error will lead to an attenuation bias for the estimated height coefficient, the magnitude of which will depend on the ratio of the variance in height due to variation in health relative to the overall variance in height after removing the family-specific component. This overall variance in height can be approximated with the sample variance of height within families (approximately 2 in^2). The variance in height due to health is harder to assess. As a very rough approximation, one can take the stunting of one inch estimated by Voth and Leunig (1996) for smallpox as the effect of a severe childhood disease on height and write the variance due to health in terms of the probability p of being afflicted with such a disease. The variance of the health related component would then be $p(1-p)$.Even in the extreme case of p equal to 0.5, minimizing the attenuation bias, the ratio of the variance of height due to health relative to the overall variance in height would be 0.125. With a ratio this small, the estimated coefficient on height will be roughly an order of magnitude smaller than the true coefficient on the health component of height, suggesting that the seemingly small coefficients on height in Table 7 and Table 8 actually imply sizeable effects.

One major criticism of this approach to interpreting the coefficients, and to interpreting height as a proxy of childhood health in general, is that the the coefficient on height may simply be picking up the effects of height discrimination, either by parents or by educators. If this is the case, then the attentuation bias described above is not relevant and the coefficient would represent a small effect and one not necessarily related to childhood health. The city and state disease environment data offer a way to test whether this relationship between height and educational attainment is being driven by height discrimination rather than by the effects of childhood health. If the relationship is driven by variation in childhood health, the relationship should be stronger in those cities and states with worse childhood disease environments (variation in height due to health would be more substantial in these locations leading to a smaller attenuation bias for the height coefficient). If the relationship is instead driven by height discrimination, the effects of height on educational attainment should not be correlated with the disease environment. By including an interaction between height and the local disease environment, it is possible to test whether the effect of height is stronger in cities where childhood disease is more prevalent.

Table 9 and Table 10 present the results from incorporating these interaction terms into the fixed effect and conditional logit regressions. In 19 out of the 20 specifications and in all of the cases where the interaction term is statistically significant, the coefficient on the interaction term is positive suggesting that the marginal effect of height on educational attainment is larger is cities with higher levels of disease. This is evidence that the coefficient on height is indeed picking up the effects of childhood health on educational attainment rather than simply height discrimination. In an attempt to bolster this interpretation, pairwise correlations were calculated between the standard deviation of heights within brother pairs and the various measures of city- and state-level disease environment. The one correlation that proved to be significant was between the standard deviation of brother heights and the number of infant deaths from disease at the city-level, which had a positive correlation of 0.04, significant at the five percent level. This positive correlation suggests that there is greater variation in height between brothers in cities with worse disease environments, offering further support that exposure to disease is driving at least a portion of the variation in heights in the brothers sample.[25]

6 Conclusions

The unique features of the World War II enlistment records offer a detailed picture of the links between childhood health and human capital both over time and across individuals. The steady growth in the American human capital stock over the first decades of the twentieth century was matched by equally impressive improvements in health. Younger cohorts of enlistees enjoyed greater health and educational attainment than their older counterparts as did enlistees from healthier cities and states.

These aggregate trends in health and educational attainment, while impressive in terms of their rapid pace and high correlation, do not come as a surprise. As national income rises, we generally expect improvements in both health and education. Where the enlistment records offer a unique insight is the relationship between health and education within families. The matched brothers data reveal that even after controlling for observed and unobserved family characteristics, differences in height between brothers predict differences in educational attainment implying that differences in childhood health across brothers had long term consequences in terms of human capital formation. This suggests that improvements in health in the early twentieth century may very well have been a necessary prerequisite for the United States' remarkable gains in educational attainment over that period.

The exact mechanisms through which childhood health influenced educational attainment remain unclear. The higher educational attainments of healthier brothers could simply be a

[25] This finding is also consistent with the findings of Lauderdale and Rathouz (1999) for the Civil War showing that the variability of brothers' heights was greater in more densely populated areas in which disease environments would be worse. In related work, Alter and Oris (2008) demonstrate that nineteenth-century Belgian brothers' heights were more strongly correlated for brothers from high socioeconomic status families, which Alter and Oris point to as evidence that elite families were better able to control the environmental challenges to childhood development. The World War II brothers also demonstrate this positive correlation between family income and correlation in heights but it is not a statistically significant correlation.

product of having more days on which they were healthy enough to attend school but their brothers were not. If this were the case, improvements in public health would directly lead to increased educational attainment. However, if the differences in educational attainment were the product of parents choosing to invest more in their healthier sons it is less clear how improvements in public health would translate into improvements in educational attainment. How parents redistribute resources across children when the health of those children improve would determine the extent to which health advances impact educational attainments. If health improvements save resources that could then be transferred to educational investment, educational attainments would rise. However, if they simply lead to a redistribution of resources already devoted to educational attainment, declines in child morbidity may lead to a more equitable distribution of resources across children without a clear impact on overall average educational investment.

These are issues that warrant further exploration. Understanding how childhood health impacts the ability to attend school, the returns to that schooling and family decisions about educational investments will advance our understanding of how the United States achieved its gains in education over the first half of the twentieth century. Such an understanding of the evolution of the health and human capital of the American population would provide insight into not only how the United States economy evolved but also how educational and health policy reforms would influence the economic growth of countries still struggling with high levels of childhood morbidity today.

References

Alderman, H., J. Hoddinott, and B. Kinsey. 2006. "Long Term Consequences of Early Childhood Malnutrition." *Oxford Economic Papers*, 58(3): 450–474.

Alderman, H., J.R. Behrman, V. Lavy, and R. Menon. 2001. "Child Health and School Enrollment: A Longitudinal Analysis." *Journal of Human Resources*, 36(1): 185–205.

Almond, D. 2006. "Is the 1918 Influenza Pandemic Over? Long-term Effects of In Utero Influenza Exposure in the Post-1940 US Population." *Journal of Political Economy*, 114(4): 672–712.

Alter, George, and Michel Oris. 2008. "Effects of inheritance and environment on the heights of brothers in nineteenth-century Belgium." *Human Nature*, 19(1): 44–55.

Behrman, J.R., and M.R. Rosenzweig. 2004. "Returns to Birthweight." *Review of Economics and Statistics*, 86(2): 586–601.

Black, S.E., P.J. Devereux, and K.G. Salvanes. 2007. "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes." *The Quarterly Journal of Economics*, 122(1): 409–439.

Bleakley, H. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *The Quarterly Journal of Economics*, 122(1): 73–117.

Bozzoli, C., A. Deaton, and C. Quintana-Domeque. 2008. "Adult Height and Childhood Disease." *Demography*, 46(4): 647–669.

Case, A., A. Fertig, and C. Paxson. 2005. "The Lasting Impact of Childhood Health and Circumstance." *Journal of Health Economics*, 24(2): 365–389.

Case, A., and C. Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy*, 116(3): 499–532.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies*, 47(1): 225–238.

Condran, G.A., and E. Crimmins-Gardner. 1978. "Public Health Measures and Mortality in US Cities in the Late Nineteenth Century." *Human Ecology*, 6(1): 27–54.

Costa, D.L., and R.H. Steckel. 1997. "Long-term Trends in Health, Welfare, and Economic Growth in the United States." In *Health and welfare during industrialization.* 47–89. University Of Chicago Press.

Cutler, D., and G. Miller. 2005. "The Role of Public Health Improvements in Health Advances: The Twentieth-Century United States." *Demography*, 42(1): 1–22.

Department of Public Health. 1926*a*. "The Notifiable Diseases: Prevalence During 1925 in Cities of 10,000 to 100,000 Population." *Public Health Reports*, 41(42): 2239–2348.

Department of Public Health. 1926*b*. "The Notifiable Diseases: Prevalence During 1925 in Cities over 100,000." *Public Health Reports*, 41(38): 1997–2029.

Goldin, C. 1998. "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century." *Journal of Economic History*, 58(2): 345–374.

Goldin, C.D., and L.F. Katz. 2008. *The Race Between Education and Technology.* Harvard University Press.

Greene, W. 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *The Econometrics Journal*, 7(1): 98–119.

Komlos, J., and B.E. Lauderdale. 2007. "The Mysterious Trend in American Heights in the 20th Century." *Annals of Human Biology*, 34(2): 206–215.

Lauderdale, Diane S, and Paul J Rathouz. 1999. "Evidence of environmental suppression of familial resemblance: Height among US Civil War brothers." *Annals of human biology*, 26(5): 413–426.

Magnusson, P.K.E., F. Rasmussen, and U.B. Gyllensten. 2006. "Height at Age 18 Years is a Strong Predictor of Attained Education Later in Life: Cohort Study of Over 950,000 Swedish Men."

Meeker, E. 1972. "The Improving Health of the United States, 1850-1915." *Explorations in Economic History*, 9(4): 353–373.

Miguel, E., and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1): 159–217.

Oreopoulos, P., M. Stabile, R. Walld, and L.L. Roos. 2008. "Short-, Medium-, and Long-Term Consequences of Poor Infant Health: An Analysis Using Siblings and Twins." *Journal of Human Resources*, 43(1): 88.

Oxley, D. 2003. "'The Seat of Death and Terror': Urbanization, Stunting, and Smallpox." *Economic History Review*, 56(4): 623–656.

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

Royer, H. 2009. "Separated at Girth: US Twin Estimates of the Effects of Birth Weight." *American Economic Journal: Applied Economics*, 1(1): 49–85.

Ruggles, S., M. Sobek, T. Alexander, C. Fitch, R. Goeken, P. Hall, M. King, and C. Ronnander. 2009. "Integrated Public Use Microdata Series Sample of the 1880 Federal Census." *Accessed throgh usa.ipums.org/usa/.*

Silventoinen, K. 2003. "Determinants of Variation in Adult Body Height." *Journal of Biosocial Science*, 35(02): 263–285.

Steckel, R.H. 1986. "A Peculiar Population: The Nutrition, Health, and Mortality of American Slaves from Childhood to Maturity." *Journal of Economic History*, 46(3): 721–741.

United States Army. 1946. "Army Serial Number Electronic File, ca. 1938-1946." *Electronic file from the National Archives and Records Administration, Washington, D.C.*

U.S. Bureau of the Census. 1880. "Tenth Census of the United States, 1880, Population Schedule." *Digital scans of original records in the National Archives, Washington, D.C., accessed through www.ancestry.com.*

U.S. Bureau of the Census. 1920. "Fourteenth Census of the United States, 1920, Population Schedule." *Digital scans of original records in the National Archives, Washington, D.C., accessed through www.ancestry.com.*

U.S. Bureau of the Census. 1930. "Fifteenth Census of the United States, 1930, Population Schedule." *Digital scans of original records in the National Archives, Washington, D.C., accessed through www.ancestry.com.*

U.S. Department of Commerce. 1921. “Mortality Statistics, 1921-1925.” *Government Printing Office, Washington, D.C.*

Voth, H.J., and T. Leunig. 1996. “Did Smallpox Reduce Height? Stature and the Standard of Living in London, 1770-1873.” *Economic History Review*, 49(3): 541–560.

Table 1: Summary statistics for the sample of World War II enlistees matched to their childhood households.

	Sample of matched brothers			Population	
	Mean	Standard deviation	Standard deviation within families	Mean	Standard deviation
Height (inches)	68.11	2.74	1.39	68.21	2.76
Weight (pounds)	147.89	21.02	11.57	148.08	21.72
Years of secondary and postsecondary education	2.42	1.94	0.84	2.38	2.01
Age	22.17	2.43	1.32	22.83	3.47
Enlistment year	1941.77	0.78	0.48	1941.81	0.80
Birth order among siblings	3.19	1.79	0.70	--	--
Birth order among brothers	2.38	1.26	0.60	--	--
Father's log income	3.10	0.42	--	3.12	0.43
Number of siblings	5.23	2.17	--	2.58	1.74
Number of brothers	3.68	3.68	--	1.38	1.25

Notes: Father's income is measured in hundreds of 1950 dollars. Population is defined as all potential brothers in the enlistment records for the individual characteristics and all households in a 1% sample of the 1930 census with at least one child for the household characteristics.

Table 2: Age distribution of cases and deaths for major diseases.

	Cases reported in the 1880 federal census				Deaths reported in federal mortality statistics, 1921-1925	
Disease	Mean age	Median age	Skewness	Cases	% of deaths under 2 years old	% of deaths under 10 years old
Diabetes	49.7	54	-0.45	35	0.33%	2.33%
Nephritis	48.1	50	-0.27	619	0.74	1.82
Circulatory disease	42.0	43	0.02	591	0.68	1.72
Diarrhea	31.9	30.5	0.2	210	--	--
Smallpox	26.8	30	0.27	19	11.60	18.10
Influenza	33.8	32	0.27	301	17.92	24.95
Pneumonia	35.7	34	0.28	253	40.10	48.31
Typhus	29.0	30	0.43	9	0.00	6.25
Tuberculosis	35.2	32	0.45	2389	3.25	6.46
Malaria	30.4	28	0.53	917	15.84	34.09
Meningitis	29.2	21	0.73	13	36.28	59.87
Typhoid	26.5	22	0.92	313	1.72	12.05
Mumps	18.9	14.5	1.58	60	22.57	52.60
Diphtheria	16.8	13	1.59	123	20.03	85.60
Scarlet fever	9.3	6	1.67	143	13.86	72.48
Measles	10.7	8	1.82	1184	55.35	87.16
Chicken pox	12.1	7	2.12	16	--	--
Whooping cough	5.8	4	4.71	338	82.26	98.97

Notes: Data on cases are compiled from a 1% sample of the 1880 federal census. The number of cases are the number of individuals reported as having that particular illness on the day of the census (see the appendix for more details). Data on mortality are taken from the annual mortality statistics reports of the Census Bureau.

Table 3: Summary statistics for morbidity and mortality rates by state and city.

	Deaths per 100,000 people (state level)	Deaths per 100,000 people (city level)	Cases per 1,000 people (city level)	Correlation between deaths and cases (city level)
From diseases targeting	51.21	11.54	9.22	0.08
infants	(25.96)	(7.99)	(5.23)	
From diseases targeting	20.87	9.37	2.94	0.22
older children	(11.08)	(5.77)	(1.35)	
From diseases targeting	419.01	7.18	0.65	0.45
adults	(82.22)	(6.91)	(0.70)	

Notes: Standard deviations are given in parentheses. The state level figures include data for 47 states (data are not available for Alaska, Hawaii and Nevada). The city level figures include data for the 74 largest cities. Note that the sets of diseases differ between the state and city level data so the means cannot be directly compared.

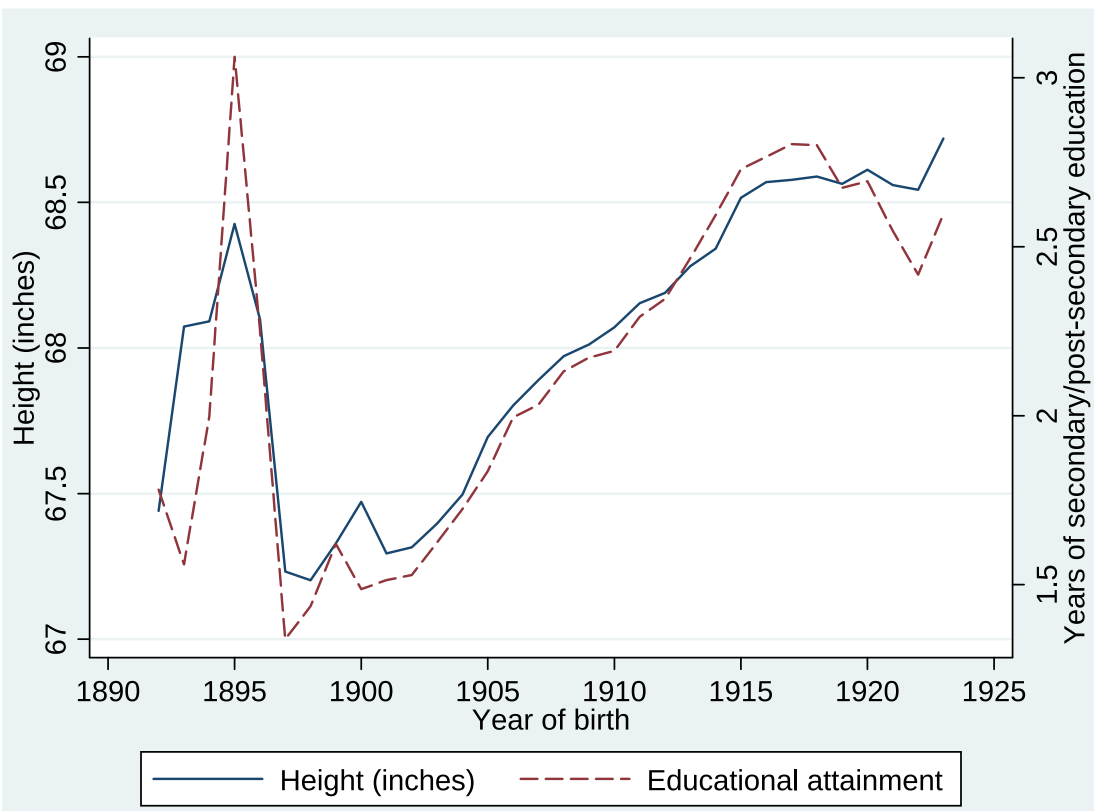


Figure 1: Mean height and educational attainment by cohort, 1892-1923.

Table 4: The effect of mortality rates on stature, height in inches as dependent variable.

	City level disease data, deaths per 100,000 people				City level disease data, cases per 1,000 people				State level disease data, deaths per 100,000 people			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Diseases targeting	-0.023***	-0.004***			-0.009***	-0.004***			-0.012***	-0.012***		
infants	(0.000)	(0.000)			(0.001)	(0.001)			(0.000)	(0.000)		
Diseases targeting	-0.018***	-0.007***			-0.017***	-0.014***			-0.006***	-0.002***		
older children	(0.001)	(0.001)			(0.003)	(0.003)			(0.000)	(0.000)		
Diseases targeting			-0.021***	-0.005***			-0.011***	-0.006***			-0.010***	-0.010***
all childran			(0.000)	(0.000)			(0.001)	(0.001)			(0.000)	(0.000)
Diseases targeting	0.030***	0.008***	0.031***	0.008***	0.440***	0.156***	0.442***	0.157***	-0.001***	-0.001***	-0.001***	-0.001***
adults	(0.000)	(0.000)	(0.000)	(0.000)	(0.005)	(0.006)	(0.005)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)
Region dummies:												
Northeast		-0.526***		-0.525***		-0.514***		-0.517***		-0.012*		0.012*
		(0.006)		(0.006)		(0.006)		(0.006)		(0.007)		(0.007)
South		0.224***		0.230***		0.161***		0.160***		0.111***		0.038***
		(0.010)		(0.009)		(0.011)		(0.011)		(0.005)		(0.005)
West		0.349***		0.347***		0.357***		0.353***		0.234***		0.196***
		(0.008)		(0.008)		(0.008)		(0.008)		(0.007)		(0.007)
Observations	1344525	1344525	1344525	1344525	1289257	1289257	1289257	1289257	3042439	3042439	3042439	3042439
R-squared	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.03	0.03	0.03	0.03	0.03

Notes: Robust standard errors in parentheses. Omitted region is the Midwest. All regressions control for race and a quadratic in age. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: OLS regression results for the effects of height and household characteristics on education, years of secondary and post-secondary education as dependent variable.

	(1)	(2)	(3)	(4)	(6)	(7)	(8)
Race and state							
controls included:	no	no	yes	yes	no	yes	yes
Height (inches)	0.0740***	0.0582***	0.0659***	0.0612***	0.0679***	0.0751***	0.0711***
	(0.0101)	(0.0095)	(0.0091)	(0.0089)	(0.0098)	(0.0096)	(0.0093)
Number of siblings		-0.2921***	-0.2653***	-0.2451***			
		(0.0166)	(0.0159)	(0.0187)			
Birth order among		0.1203***	0.1100***	0.1046***			
all siblings		(0.0174)	(0.0161)	(0.0158)			
Number of brothers					-0.3329***	-0.3007***	-0.2704***
					(0.0219)	(0.0222)	(0.0244)
Birth order among					0.0972***	0.0911***	0.0927***
brothers					(0.0246)	(0.0235)	(0.0236)
Ln(father's income)				1.0057***			1.0190***
				(0.0774)			(0.0761)
Observations	8457	8249	8205	7507	8198	8154	7459
R-squared	0.03	0.09	0.15	0.19	0.07	0.14	0.18

Robust standard errors clustered by childhood state of residence in parentheses. All regressions control for a quadratic in age. Only individuals with completed educational careers are included in the regression sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Logit estimates for the effects of height and household characteristics on educational outcomes.

Dependent variable:	Attended at least one year of high school (yes=1)		High school graduate (yes=1)		Attended at least one year of college (yes=1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Height (inches)	0.0490***	0.0594***	0.0735***	0.0836***	0.0951***	0.1038***
	(0.0126)	(0.0129)	(0.0106)	(0.0109)	(0.0120)	(0.0125)
Number of siblings	-0.2272***		-0.2597***		-0.4033***	
	(0.0238)		(0.0219)		(0.0475)	
Birth order among all siblings	0.0968***		0.1023***		0.1865***	
	(0.0205)		(0.0195)		(0.0565)	
Number of brothers		-0.2431***		-0.2922***		-0.4462***
		(0.0349)		(0.0252)		(0.0642)
Birth order among brothers		0.0693**		0.1135***		0.1214
		(0.0286)		(0.0249)		(0.0817)
Ln(father's income)	1.1834***	1.1845***	0.8696***	0.8771***	1.3030***	1.3289***
	(0.0980)	(0.0978)	(0.0955)	(0.0937)	(0.1365)	(0.1300)
Observations	7499	7451	7503	7455	5999	5951

Robust standard errors clustered by childhood state of residence in parentheses. All regressions control for race, birth state and a quadratic in age. Only individuals with completed educational careers are included in the regression sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Estimates of the effect of height on educational attainment with family fixed effects, years of secondary and post-secondary education as dependent variable.

	(1)	(2)	(3)	(4)	(5)
Height (inches)	0.0213**	0.0256**	0.0243	0.0294***	0.0297
	(0.0100)	(0.0103)	(0.0243)	(0.0104)	(0.0239)
Height x ln(father's income)			-0.0022		-0.0028
			(0.0070)		(0.0069)
Birth order among siblings		0.0796**	0.0842**		
		(0.0390)	(0.0419)		
Birth order among brothers				0.0577	0.0612
				(0.0428)	(0.0458)
Observations	8491	8276	7532	8225	7484
Number of families	4478	4378	4016	4374	4013
R-squared	0.01	0.01	0.01	0.01	0.01

Robust standard errors clustered by family in parentheses. All regressions control for a quadratic in age. Regression sample includes only those individuals with completed educational careers.

Note that number individuals is less than twice the number of families because there are families for which all but one son were dropped for not having completed educational careers.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 8: Conditional logit estimates of the effect of height on educational outcomes for brothers.

Dependent variable:	Attended at least one year of high school (yes=1)		High school graduate (yes=1)		Attended at least one year of college (yes=1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Height (inches)	-0.0132	-0.0064	0.0612**	0.0734***	0.0632**	0.0631**
	(0.0205)	(0.0209)	(0.0243)	(0.0264)	(0.0311)	(0.0310)
Birth order among all siblings	0.0838		0.1659*		0.2361	
	(0.1029)		(0.0939)		(0.1840)	
Birth order among brothers		0.0531		0.2198***		0.1726
		(0.0964)		(0.0828)		(0.2175)
Observations	1733	1724	1947	1913	574	569

Robust standard errors clustered by childhood state of residence in parentheses. All regressions control for a quadratic in age. Only individuals with completed educational careers are included in the regression sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 9: Estimates of the effect of height on educational attainment with family fixed effects and disease environment interactions, years of secondary and post-secondary education as dependent variable.

Disease environment measure:	City level disease data				State level disease data			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Height (inches)	0.0190	0.0301	0.0490	0.0466	-0.0047	-0.0013	-0.0726	-0.0679
	(0.0405)	(0.0409)	(0.0657)	(0.0672)	(0.0263)	(0.0267)	(0.0567)	(0.0570)
Height x deaths per 100,000 from	0.2644	0.2291			0.0488	0.0495		
diseases targeting infants	(0.2827)	(0.2882)			(0.0418)	(0.0426)		
Height x deaths per 100,000 from			0.0085	0.0437			0.0183*	0.0181*
all diseases			(0.2289)	(0.2324)			(0.0106)	(0.0107)
Birth order among siblings	0.1893***		0.2502***		0.0914**		0.0926**	
	(0.0719)		(0.0762)		(0.0397)		(0.0396)	
Birth order among brothers		0.1718**		0.2420***		0.0699		0.0714
		(0.0810)		(0.0864)		(0.0435)		(0.0435)
Observations	2875	2855	2562	2542	7988	7939	7988	7939
Number of families	1531	1530	1364	1363	4223	4219	4223	4219
R-squared	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01

Robust standard errors clustered by family in parentheses. All regressions control for a quadratic in age. Regression sample includes only those individuals with completed educational careers.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 10: Conditional logit estimates of the effect of height on educational outcomes for brothers with disease environment interactions.

Disease environment measure:	City level disease data						State level disease data					
Dependent variable:	Attended high school (yes=1)		Graduated high school (yes=1)		Attended college (yes=1)		Attended high school (yes=1)		Graduated high school (yes=1)		Attended college (yes=1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Height (inches)	0.0427	0.0582	0.0026	0.0311	-0.2371	-0.2062	-0.0704	-0.0821	0.0175	0.0355	0.0143	0.0178
	(0.0737)	(0.0808)	(0.0554)	(0.0528)	(0.1666)	(0.1751)	(0.0622)	(0.0628)	(0.0698)	(0.0722)	(0.0941)	(0.0994)
Height x deaths per 100,000 from	-0.0407	0.0052	0.7085*	0.6032	3.1951**	2.9803**	0.0919	0.1228	0.0689	0.0601	0.0797	0.0736
diseases targeting infants	(0.3099)	(0.3432)	(0.4033)	(0.3841)	(1.3266)	(1.3856)	(0.0911)	(0.0918)	(0.1050)	(0.1132)	(0.1312)	(0.1374)
Birth order among siblings	0.2531		0.2964		0.2306		0.1319		0.1819*		0.2283	
	(0.1869)		(0.1924)		(0.2694)		(0.1029)		(0.0974)		(0.1830)	
Birth order among brothers		0.2838*		0.4037**		0.1018		0.1081		0.2237***		0.1590
		(0.1641)		(0.1569)		(0.3360)		(0.0942)		(0.0843)		(0.2198)
Observations	523	517	818	805	267	264	1673	1664	1894	1860	568	563

Robust standard errors clustered by childhood state of residence in parentheses. All regressions control for a quadratic in age. Regression sample includes only those individuals with completed educational careers.

* significant at 10%; ** significant at 5%; *** significant at 1%

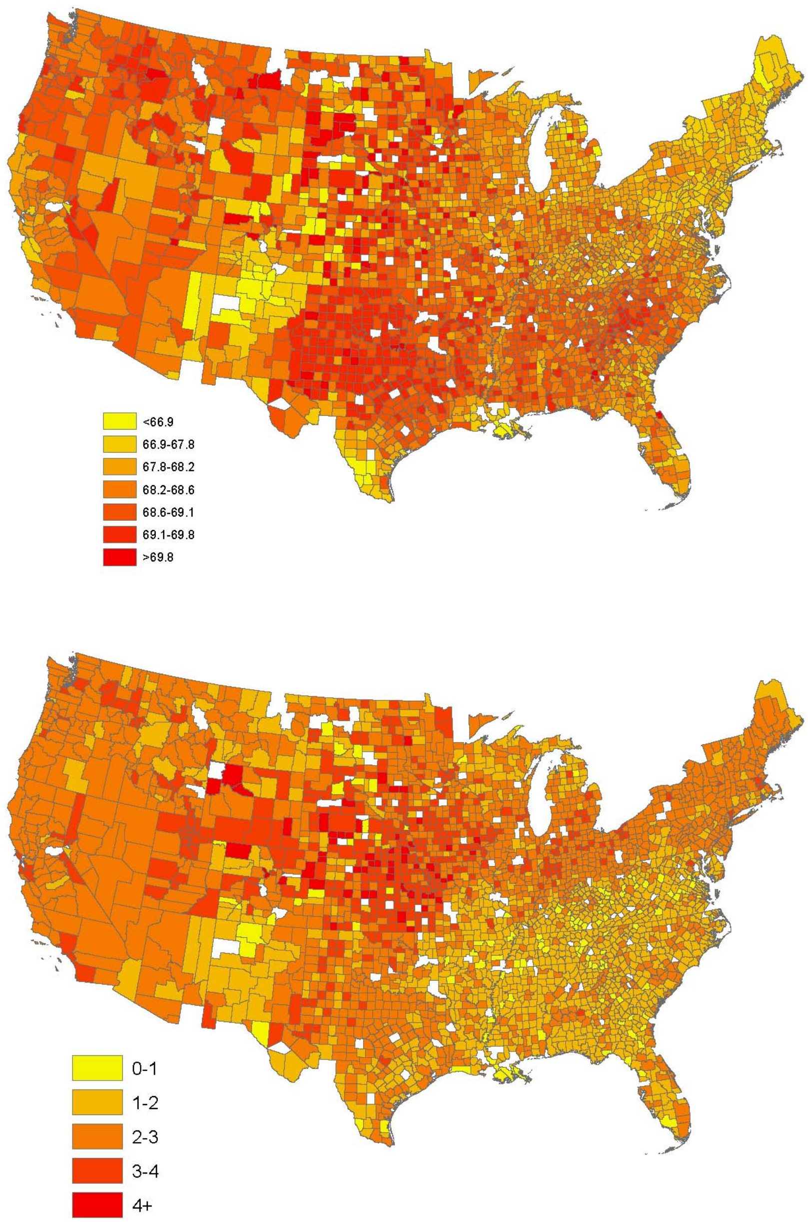


Figure 2: Mean height in inches (upper panel) and mean years of secondary and post-secondary education (lower panel) by county for World War II enlistees.

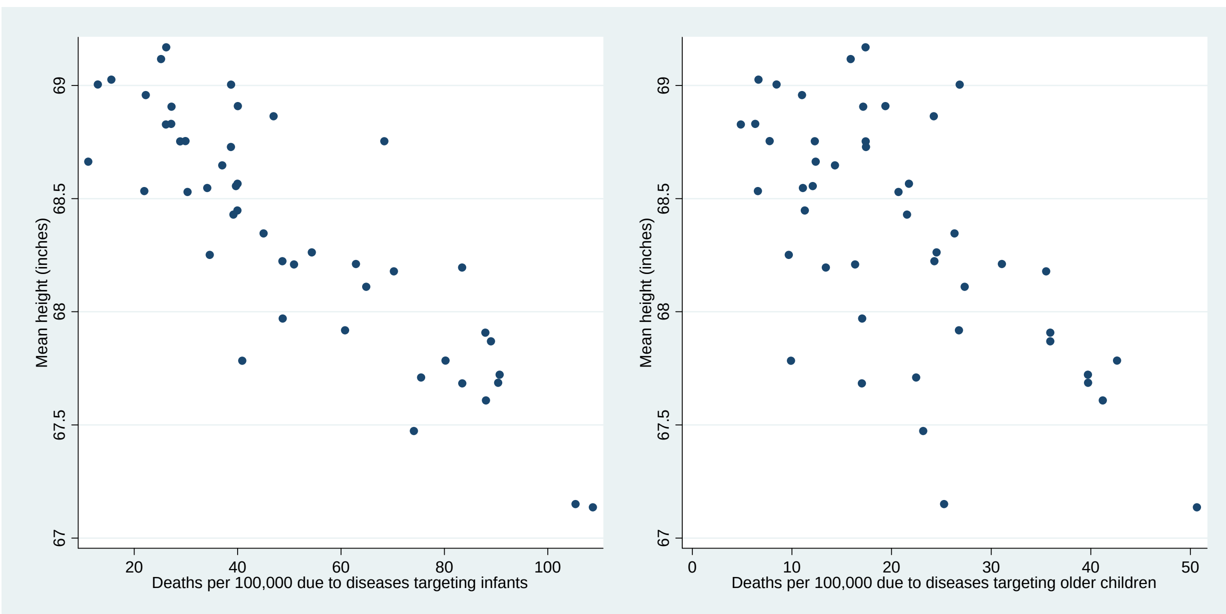


Figure 3: Mean height and mortality rate by disease type for states.

A Data Sources - For Online Publication

A.1 World War II Enlistment Records and the Matching of Brothers

The World War II enlistment records were obtained from the National Archives and Records Administration (NARA) as an electronic file. These electronic records were converted from the Army Serial Number microfilm of computer punchcards by the NARA and include records for roughly nine million men and women who enlisted in the United States Army between 1938 and 1946. The records are not complete both due to missing records for certain ranges of serial numbers and because several thousand records could not be interpreted by the NARA's scanning system.

The relevant variables reported in the enlistment records include: serial number, name, state and county of residence, place of enlistment, date of enlistment, military grade, military branch, nativity, year of birth, race, education, civilian occupation, marital status, height and weight. Not all variables were reported in all years. The most important change in the reporting over time for the purposes of this paper was the exclusion of height and weight information after 1943. Consequently, this study is restricted to individuals enlisting between 1938 and 1943. In some records, what replaced the height and weight information was actually the enlistee's score on the Army General Classification Test (AGCT), a test of cognitive ability. While this paper focuses on the relationship between height and educational attainment, a similar study comparing differences in educational attainment to differences in AGCT scores between brothers would certainly be worthwhile.

A further complication with the height and weight variables is that the reporting of those variables was inconsistent. The NARA notes that instructions for the use of the height and weight fields changed during the war and that some cards contain information on military occupation in the height and weight fields. However, there is no way to know for certain which cards report height and weight and which cards use the fields to report something else. In an attempt to restrict the sample to records reporting height and weight, I discard observations for which the stated height and weight imply an unrealistic body mass index for an individual. I compute the body mass index based on the stated height and weight and discard observations with a BMI of less than 15 (below which individuals are considered to be exhibiting starvation) or greater than 60 (above which individuals are considered hyper-obese). Despite these precautions, there may still be observations in the sample for which the height and weight fields do not actually contain information on height and weight.

For the purposes of documenting the secular trends in height and educational attainment, the sample is further restricted to include only those enlistees who were assigned the rank of private. Many of the individuals assigned higher ranks have ages that correspond to having served in World War I and are re-enlisting as officers for World War II (explaining their higher ranks). The army would discard an individual's old enlistment card and create a new card upon re-enlistment. Consequently, these officers from World War I re-enlisting to serve in World War II have an enlistment record that appears just like that of a draftee with the exception of the rank. These officers create a sample selection problem when it comes to documenting the secular trends in both height and education. Officers are on average significantly taller and more educated relative to other members of the army. The birth cohort that corresponds to World War I veterans has a disproportionate number of officers in the enlistment sample and therefore appears significantly taller and more educated than either the cohort before them or after them. To keep our samples of the birth cohorts comparable across birthyears, I restrict the sample to privates.

The following procedure was used to create the sample of brothers from the full enlistment records. First, as described above, all individuals with suspect height and weight data were discarded. Next, the individuals were sorted by last name, state of birth, state and county of residence and then age. Potential brothers were identified as individuals sharing the same last name, state of birth and state and county of residence and within three years of each other in age. Every tenth set of potential brothers was kept, creating a 10 percent sample of the enlistment records.

A Perl script was then used to search for every potential brother in the 10 percent sample in either the 1930 or the 1920 federal census. If all individuals in a group of potential brothers were born prior to 1920, the 1920 federal census was used for all brothers in the group. If any of the individuals in a group of potential brothers was born in 1920 or later, the 1930 federal census was used for all brothers in the group. For each individual, the Perl script searches ancestry.com's online database of census records using the individual's first and last names, state of birth and birthyear and returns the location of the person considered to be the best match in the federal census. All individuals in a group of potential brothers are then sorted by county of residence in the federal census and parents' names. Only potential brothers living in the same county in the federal census with identical parents' names are kept.

Next, these remaining individuals are then searched for by hand in the ancestry.com database to confirm that they have a unique match in the database. If there is not a unique match (multiple individuals have the same name and were born within one year of the enlistee's birthyear or no individuals exactly match both the name and birthyear of the enlistee record), the individual is discarded. For the remaining potential brothers with unique matches, images of the original census manuscripts containing the individuals are downloaded. From these images, it is possible to determine whether the potential brothers truly lived in the same household. If so, they are recorded as a confirmed match and information on the father and household structure are transcribed from the census image. If not, the individuals are dropped from the dataset. Roughly one third of the potential brothers have a unique match in the federal census. Of these uniquely matched potential brothers, roughly one quarter are actually in the same household as one of the other uniquely matched potential brothers.

A.2 Public Health Reports

The Public Health Reports are a weekly publication of the United States Public Health Service. They have been published since 1887. The typical weekly report contains articles on current public health issues and research findings and then a section on the prevalence of disease. The prevalence of disease section gives the number of cases and deaths reported for various diseases by states and cities in the previous week.

In the early 1920s, the Public Health Reports would include an annual summary of the prevalence of disease in the previous year. One issue presented the annual summary for cities with a population over 100,000 and a second issue presented the annual summary for cities with a population between 10,000 and 100,000. The morbidity and mortality data for cities used in the paper come from the 1926 annual summary, the last summary published for cities with populations greater than 100,000 (although weekly reports continued to be published). The small city data is of questionable quality, with many of the cities failing to report information for several of the diseases and warnings from the Public Health Service that reporting standards for the cities were changing a great deal over the period of interest.

The annual summary contains the total number of cases and deaths in the previous year for a variety of diseases including anthrax, cerebrospinal fever, chicken pox, dengue fever, diphtheria, influenza, lethargic encephalitis, malaria, measles, mumps, pellagra, pneumonia, poliomyelitis, rabies in animals, rabies in man, scarlet fever, septic sore throat, smallpox, tuberculosis, tyhpoid fever, typhus fever and whooping cough. Cases and deaths are reported in both absolute numbers and in per capita terms. In addition to the number of cases in the previous year, the summary includes what the Public Health Service calls the 'estimated expectancy'. This figure is the expected number of cases in a non-epidemic year and in most cases is calculated as the median number of annual cases reported between 1918 and 1924, inclusive. If epidemics occurred, those years are excluded and the estimated expectancy is calculated as the mean of the number of cases reported in non-epidemic years. The number of years of data used for each calculation is reported in the tables. No estimated expectancies were given for anthrax, influenza, lethargic encephalitis, malaria, pellagra, pneumonia, rabies, tuberulosis or typhus fever. For these diseases, I use the number of reported cases and deaths in 1925 in place of the missing estimated expectancies.

For several cities, the public health reports did not include a population estimate. In these cases, I have imputed the 1925 population by taking the average of the city populations reported in the 1920 and 1930 federal censuses. This was done for the following cities: Los Angeles (CA), Bridgeport (CT), Waterbury (CT), Atlanta (GA), Elizabeth (NJ), Akron (OH), Oklahoma City (OK), Portland (OR), Erie (PA), Houston (TX), Norfolk (VA) and Seattle (WA).

One major note of caution when using the Public Health Reports is that the ability to diagnose diseases and the efforts to report cases were changing over time. This makes it difficult to interpret changes over time in the number of cases, the number of deaths and in the ratio of deaths to cases. The Public Health Service included the following warning with each annual summary:

> "In comparing the figures for 1925 with the estimated expectancy, averages, or with reports for preceding years, it should be borne in mind that for several years there has been a gradual improvement in the reporting of communicable diseases. An increase

in the number of cases reported may be due to better reporting rather than to an increase in the number of cases occurring."

A.3 1920 and 1930 Federal Censuses

The 1920 and 1930 federal censuses are used to identify brothers and gather information on their families. The process of matching individuals to the 1920 and 1930 census is described in the section on the World War II enlistment records. The purpose of this section is to elaborate on the information available in the federal census, the differences between the censuses, and the limitations of the census data.

The forms for the 1920 and 1930 federal censuses are very similar. The 1930 census includes the following variables relevant to this study: full name, age, state or country of birth, occupation, industry, whether the household head owns or rents the residence, what the monthly rent or value of the home is, and relationship to head of household. The 1920 census includes all of these variables with the exception of the monthly rent or value of the home.

Once sets of brothers are found in the census, the number and ordering of siblings is recorded as is the information for the head of household. In nearly all cases, the head of household is the father of the brothers. In rare cases, the head of the household is a single mother, a grandparent, or another relative. In cases where one brother is listed as a son while the another is listed as a step-son, the brothers are dropped from the sample (the identification strategy depends in part on brothers having the same parents by birth).

To determine the income of the head of household, I match the listed occupation for the household head to the occupations from the 1950 federal census. This allows me to assign a 1950 occupational income score to the household head. The 1950 occupational income score is based on the median income in hundreds of 1950 dollars for each particular occupation. While the income distribution by occupation in 1950 is certainly different than that of 1920 or 1930, these 1950 occupational income scores offer the best income estimates available for the household heads in the sample. Reliance on the the 1950 occupational income scores does mean that the income variable may be a noisy proxy of actual household income.

More information on the construction of the occupational income scores and the occupational coding in the federal census can be found on the Integrated Public Use Microdata Series website (usa.ipums.org). The site also contains information on the full set of variables available in the 1920 and 1930 federal censuses.

A.4 1880 Federal Census

The 1880 federal census was unique for its collection of morbidity information. With the exception of the 1880 census, the federal census did not ask detailed questions about disability until 1970. While several decades removed from the period of interest in this paper (a major concern given the substantial improvements in health at the start of the twentieth century), it offers the only opportunity to get age distributions of morbidity rates for several diseases from a large, nationally representative sample of the population. The age distributions of deaths by disease, published annually by the census bureau, do demonstrate that the age distributions of cases in 1880 are quite similar to the age distributions of deaths in the 1920s, helping minimize concerns about the applicability of the 1880 data to the enlistees. These figures from the annual mortality statistics are included in Table 2 of the paper.

The census asked the following question: "Is the person (on the day of the enumerator's visit) sick, or temporarily disabled, so as to be unable to attend to ordinary business or duties? If so, what is the sickness or disability?" The phrasing of the question appears to focus on work-related disabilities raising concerns that it would not apply to children and would therefore not be useful to study the incidence of childhood disease. However, it appears from the age distribution of individuals reporting an illness that the question was treated more generally.

Of those reporting an illness or disability, 10 percent were below the age of 4 and 25 percent were below the age of 12. Given the large percentage of illnesses reported for individuals far too young to work, it seems likely that a sizable percentage of individuals were interpreting the question as asking about any illnesses on the day of the census, not simply illnesses that were interfering with work. However, these percentages are still lower than what we would expect and suggest that the morbidity data is providing an underestimate of childhood morbidity rates relative to adult morbidity rates.

A second caution about the interpretation of the 1880 census morbidity data is that all of the illnesses are self-reported (or reported by parents). It is certainly possible that individuals are misdiagnosing their illnesses, exaggerating their illnesses, or even hiding their illnesses. All of these possibilities contribute additional noise to the morbidity data.

B Additional Tables and Figures - For Online Publication

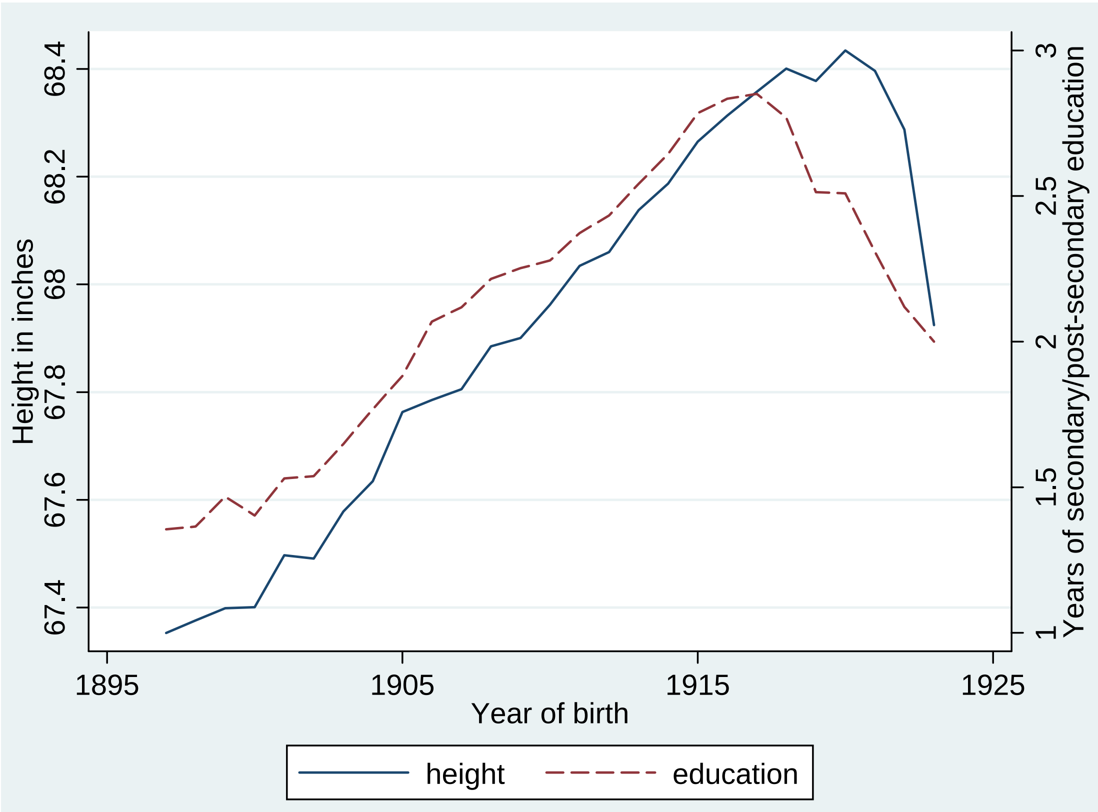
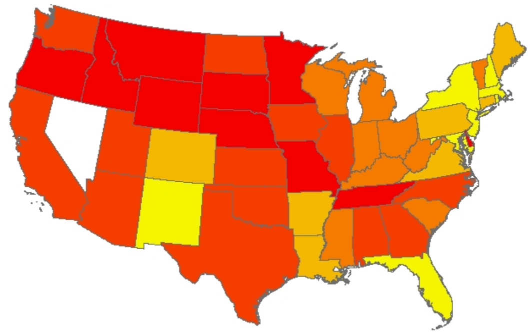
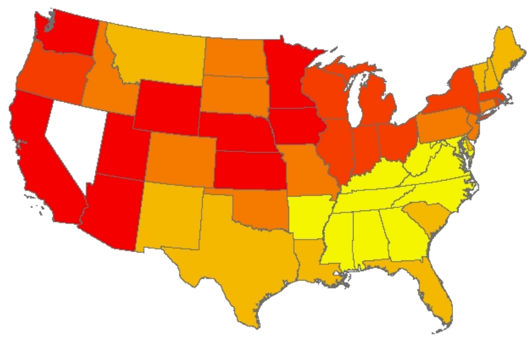


Figure 4: Mean height and educational attainment by cohort for privates with completed educational careers, 1897-1923. The dropoff in height and educational attainment for the youngest cohorts is a product of conditioning on completed educational careers: the youngest enlistees could only have completed their education if they received relatively few years of schooling.

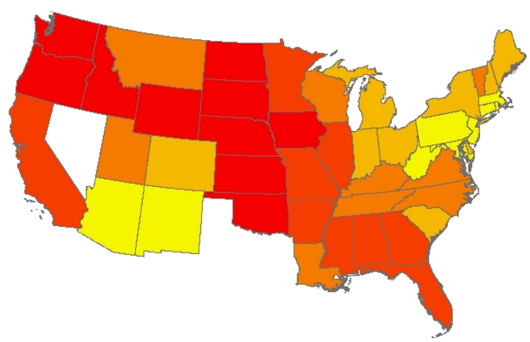
Figure 5: Age distributions for cases of major diseases as reported in the 1880 federal census.



(a)



(b)



(c)

Figure 6: Average height (a), education (b) and mortality rate due to diseases targeting infants (c) by state. Colors correspond to quintiles of each variable's distribution. Yellow (lightest shade) corresponds to the lowest height and education quintiles and the highest infant mortality quintile. Red (darkest shade) corresponds to the highest height and education quintiles and the lowest infant mortality quintile.

Table 11: The effect of mortality rates on height, city or state mean height as dependent variable.

	City level disease data		State level disease data	
	(1)	(2)	(3)	(4)
Mortality rate due to diseases	-0.025***	-0.008	-0.015***	-0.015***
targeting infants	(0.008)	(0.007)	(0.003)	(0.003)
Mortality rate due to diseases	0.004	0.003	0.005	0.008
targeting older children	(0.010)	(0.008)	(0.006)	(0.007)
Mortality rate due to diseases	0.029***	0.016***	-0.002***	-0.001**
targeting adults	(0.006)	(0.006)	(0.001)	(0.001)
Region dummies:				
Northeast		-0.636***		-0.141
		(0.099)		(0.107)
South		-0.154		0.040
		(0.128)		(0.071)
West		0.222*		0.103
		(0.116)		(0.068)
Constant	68.307***	68.428***	69.694***	69.465***
	(0.122)	(0.100)	(0.165)	(0.183)
Observations	64	64	47	47
R-squared	0.29	0.63	0.82	0.83

Notes: Robust standard errors in parentheses. Unit of observation is an individual city for columns (1) and (2) and an individual state for columns (3) and (4). Omitted region dummy is for the Midwest. All mortality rates are deaths per 100,000 people. * significant at 10%; ** significant at 5%; *** significant at 1%

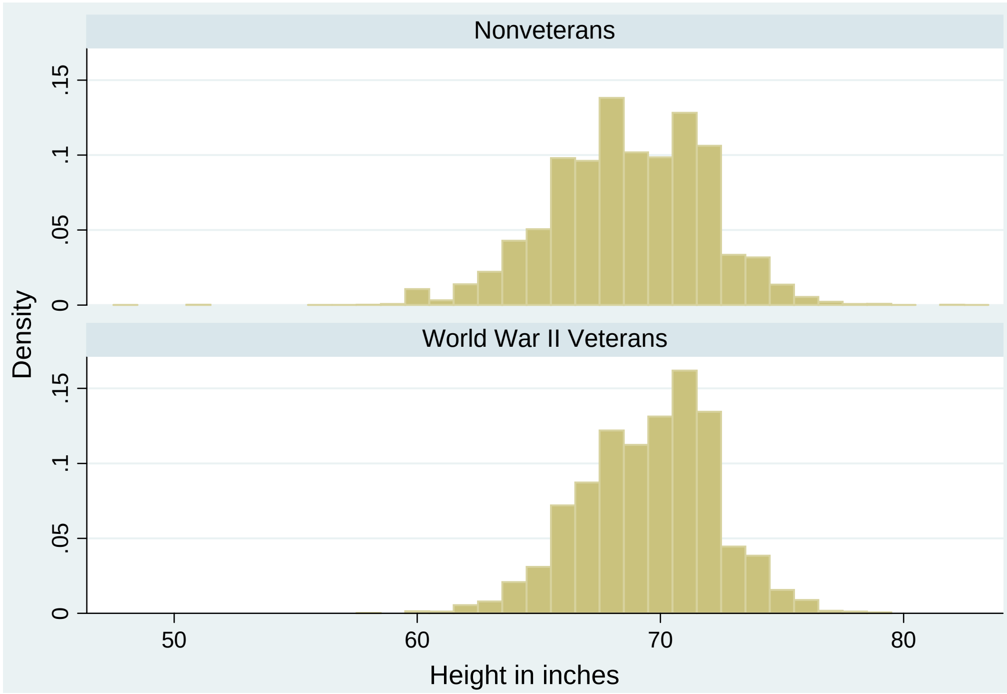


Figure 7: Distribution of male heights for WWII veterans and civilians in the 1976 Integrated Health Interview Series. Civilian observations are weighted to match the age distribution of veterans.

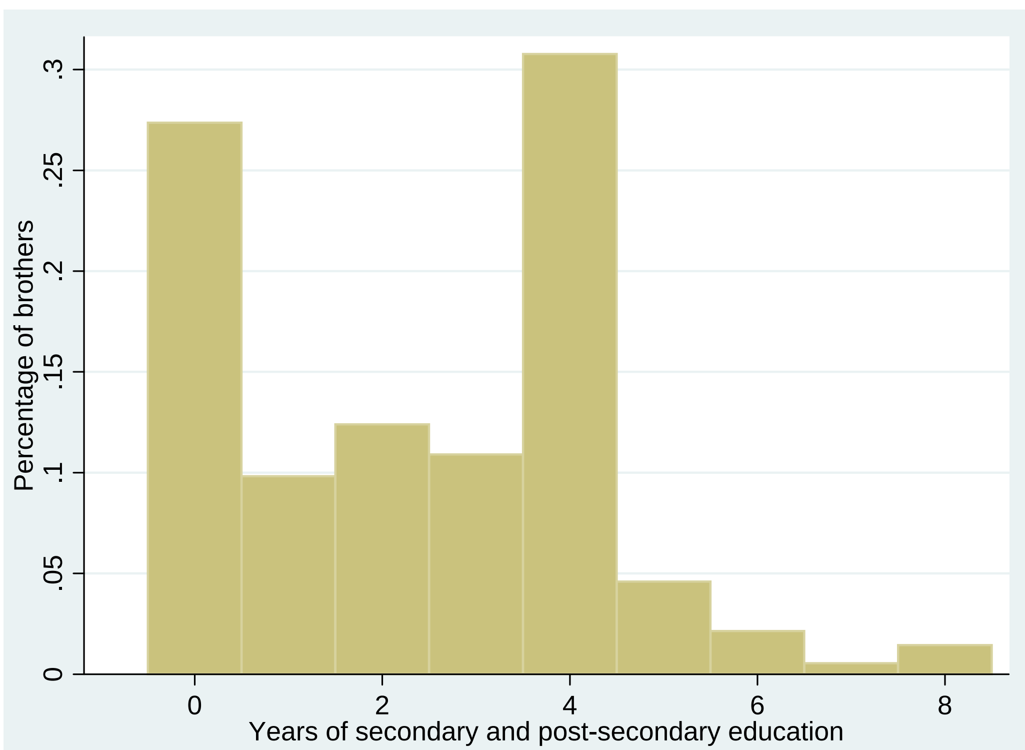


Figure 8: Distribution of educational attainment for all individuals with completed educational careers in the sample of brothers.