THE UNIVERSITY OF CHICAGO

ESSAYS ON THE MEASUREMENT OF INCOME IN ECONOMIC ANALYSIS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY

EZRA KARGER

CHICAGO, ILLINOIS

JUNE 2021

Copyright © 2021 by Ezra Karger

All Rights Reserved

I would like to thank Martha Bailey for providing me with extensive data describing county-level demographic characteristics and public libraries from prior work with Brian Jacob, Michael Kevane, and William Sundstrom. I would also like to thank Dan Aaronson, Enrico Berkes, Dan Black, Doni Bloomfield, Celeste Carruthers, Kerwin Charles, Sergio Correia, Daniel Feenberg, Camilo Garcia-Jimeno, Rick Hornbeck, Max Kellogg, Bhash Mazumder, Valerie Michelman, Derek Neal, Peter Nencka, Matt Notowidigdo, Eliana Pfeffer, Devin Pope, Lauren Sartain, and Marianne Wanamaker for their feedback, assistance with data collection, and advice about framing these projects.

Table of Contents

LIST OF FIGURES v

LIST OF TABLES vi

ABSTRACT vii

1 THE LONG-RUN EFFECT OF PUBLIC LIBRARIES ON CHILDREN: EVIDENCE FROM THE EARLY 1900S 1
- 1.1 Introduction 1
- 1.2 Data 4
  - 1.2.1 Library Data 4
  - 1.2.2 Census Microdata 7
- 1.3 Methodology 11
- 1.4 Results 14
  - 1.4.1 Education 14
  - 1.4.2 Placebo Tests 16
  - 1.4.3 Income 18
  - 1.4.4 Imputed Income Measures 19
  - 1.4.5 Occupational Prestige 21
- 1.5 Compulsory Schooling Laws 22
  - 1.5.1 Occupational Choice 24
  - 1.5.2 Life Expectancy 26
- 1.6 Comparison of Libraries and Compulsory Schooling Laws 29
- 1.7 Conclusion 30
- 1.8 Data Appendix 32
  - 1.8.1 Decennial Census Data 32
  - 1.8.2 High School Data 34

2 THE BLACK-WHITE LIFETIME EARNINGS GAP 62
- 2.1 Introduction 62
- 2.2 Data 64
  - 2.2.1 Census Microdata 64
  - 2.2.2 Life Expectancy Data 67
- 2.3 Cross-Sectional Earnings Gaps 69
- 2.4 Lifetime Earnings Gaps 70
- 2.5 Model 73
- 2.6 Comparative Statics 75
- 2.7 Calibration 78
- 2.8 Conclusion 78

3 MATCHING CENSUS DATA 94
- 3.1 Introduction 94

REFERENCES 104

List of Figures

1.1 Carnegie Library Grants in United States 36
1.2 Public Libraries in United States as of 1929 37
1.3 Public Libraries in United States, 1880–1929 38
1.4 High School Entry by 1944 39
1.5 Comparison of Treatment Effects from Carnegie and CSL Regressions 40

2.1 Number of White Male Births by Year 80
2.2 Number of Black Male Births by Year 81
2.3 Male Life Expectancy by Race and Birth Year 82
2.4 Male P(Survival through Age 30) by Race and Birth Year 83
2.5 White/Black Average Earnings by Birth Year for 3 Age Groups 84
2.6 White/Black Average Earnings by Age for 4 Birth Years 85
2.7 White/Black Lifetime Earnings by Birth Year 86
2.8 White and Black Utility by Birth Year 87
2.9 White-Black Utility Gap by Birth Year 88

List of Tables

1.1 Comparison of All 1940 Men and Matched Sample 41
1.2 Fraction of Matched Sample in Largest Cities and Towns 42
1.3 Effect of Carnegie Grant on Years of Schooling 43
1.4 Effect of Carnegie Grant on High School Graduation Rate 44
1.5 Effect of Carnegie Grant on College Attendance 44
1.6 IV Effect of Carnegie Grant on Years of Schooling 45
1.7 Effect of Carnegie Grant on Log Annual Wage Income 45
1.8 Effect of Carnegie Grant on Probability $\geq$ 50 Dollars of Non-Wage Income 46
1.9 Effect of Carnegie Grant on Log Imputed Total Income from 1950 Data 46
1.10 Effect of Carnegie Grant on Log Imputed Wage Income from 1950 Data 47
1.11 Effect of Carnegie Grant on Log Imputed Non-Wage Income from 1950 Data 47
1.12 Effect of Carnegie Grant on Occupational Prestige 48
1.13 Effect of Education on Various Outcomes (Instrumented with Compulsory Schooling) . 49
1.14 Effect of Carnegie Grant on Occupational Choice 50
1.15 Effect of Education on Occupational Choice (Instrumented with Compulsory Schooling) 51
1.16 Projected Effect of Carnegie Grant on Life Expectancy, Conditional on Reaching Age 65 53
1.17 Effect of Carnegie Grant on Educational Attainment, State Cluster 54
1.18 Effect of Carnegie Grant on Educational Attainment, only Siblings 54
1.19 Effect of Carnegie Grant on Educational Attainment, Continuous Measure 55
1.20 Effect of Carnegie Grant on Educational Attainment, Midwest 55
1.21 Effect of Carnegie Grant on Educational Attainment, Whites 56
1.22 Effect of Carnegie Grant on Educational Attainment, High School Control 57
1.23 Effect of College Libraries on Educational Attainment 58
1.24 Effect of Carnegie Grant on Educational Attainment, Plus 20 Years 58
1.25 Effect of Carnegie Grant on Educational Attainment, Lagged 20 Years 59
1.26 Effect of Carnegie Grant on Educational Attainment, Perfect Matches 60
1.27 Effect of Carnegie Grant on Educational Attainment, All Matches 61

2.1 Average Earnings at Age 30 89
2.2 Lifetime Earnings (undiscounted) 90
2.3 Lifetime Earnings ($\beta = 0.96$) 91
2.4 White/Black Lifetime Earnings (undiscounted) 92
2.5 White/Black Lifetime Earnings ($\beta = 0.96$) 93

3.1 Parental Name Accuracy Measure by Loop Iteration 97
3.2 Birthplace Accuracy Measure by Loop Iteration 98
3.3 Parental Name Accuracy Measure by Loop Iteration for Matches from the Same Loop 100
3.4 Potential Matches for the Final Analysis Sample 102

ABSTRACT

Economists often evaluate public policies by measuring the effects of those policies on participants' wage income. I use two historical settings to argue that this focus on wage income biases estimates of the returns to education, the returns to local investment, and the measurement of inequality. I investigate a rich set of non-wage outcomes that are available in historical decennial Census data, and I show that by augmenting standard economic and econometric analyses with these new outcomes, economists and policymakers can develop a more complete understanding of long-run trends in inequality and the causal effects of public policies on children.

In the first chapter of my dissertation, I measure the causal effects of public libraries on children. Between 1890 and 1921, Andrew Carnegie funded the construction of 1,618 public libraries in cities and towns across the United States. I link these library construction grants to census data and measure the effect of childhood public library access on adult outcomes. Library construction grants increased children's educational attainment by 0.10 years, did not affect wage income, and increased non-wage income by 4%. These income effects are driven by occupational choice. Access to a public library caused children to shift away from occupations like manual labor, factory-work, and mining into safer and more prestigious occupations like farm-ownership, clerical, and technical jobs. I show that compulsory schooling laws had parallel effects on children, increasing educational attainment, non-wage income and occupational prestige without affecting wage income. Economists often rely solely on wage income to measure the returns to education. But public libraries and compulsory schooling laws in the early 1900s increased educational attainment and had positive effects on children's adult labor market outcomes without affecting wage income.

In the second chapter of my dissertation, I embed income and mortality in a unified measure of black-white inequality. In the 1940 census, white males earned twice as much as black males. But white males born in 1900 had lifetime earnings 3.4 times as large as black males born in 1900. The black-white lifetime earnings gap is so much larger than the cross-sectional earnings gap because of the large black-white gap in life expectancy. 48% of black males born in 1900 died

before the age of 30 as compared to just 26% of white males. Economists often use cross-sectional earnings gaps to measure inequality between groups, but a more complete measure of inequality combines income profiles and mortality risk. I develop a model of optimal consumption in a world with mortality and I calibrate the model separately for each cohort of black and white males born between 1900 and 1970. Using this model, I find that the black-white welfare gap shrank by 48% from the 1900 to the 1920 birth cohorts as black and white mortality rates converged, but the black-white welfare gap declined only modestly from the 1920 to 1970 birth cohorts.

CHAPTER 1

THE LONG-RUN EFFECT OF PUBLIC LIBRARIES ON CHILDREN: EVIDENCE FROM THE EARLY 1900S

1.1 Introduction

Andrew Carnegie immigrated with his family from Scotland to Pennsylvania at the age of 13. He immediately began work at a cotton mill as a low-skilled bobbin boy, earning $1.20 a week.[1] A nearby resident opened up his personal library to the "local working boys," giving Carnegie access to a large set of books. Carnegie attributed much of his success as a businessman to this private library, and he wanted to provide a similar experience to children everywhere by funding the construction of public libraries in underserved communities.[2]

After a successful career in business, Carnegie followed through on this resolution, funding the construction of thousands of libraries worldwide, including 1,618 in the United States. He first funded the construction of a library in 1880, in his birthplace of Dunfermline, Scotland. In the following years he paid for the construction of a handful of public libraries in his home state of Pennsylvania. He then formalized his grant program, expanding it to other communities across the United States. Carnegie's requirements were simple: a town had to agree to supply a public plot of land for the new library and fund the library annually at an amount greater than or equal to 10% of the initial grant value.

I use these Carnegie-funded public libraries to provide the first evidence that public libraries have long-run positive effects on children. I measure these effects by linking children from the 1900–1930 census to adult records in the 1940 census. I use an iterative matching procedure, increasing accuracy by relying on 1940 records that match to multiple early census years. I then use

1. The New York Times, "Obituary: Carnegie Started as a Bobbin Boy," August 12, 1919.

2. Carnegie wrote in his autobiography: "I resolved, if ever wealth came to me, that it should be used to establish free libraries, that other poor boys might receive opportunities similar to those for which we were indebted to that noble man" (Carnegie, 1901).

within-family variation in the timing of when children got access to a Carnegie Library to measure the causal effect of public library access on adult outcomes. My regression specifications are similar to those used by Aaronson and Mazumder (2011), who measure the causal effect of Rosenwald school construction grants on children's school attendance during a similar time period. I show that access to a Carnegie grant increased educational attainment, did not affect wage income, and increased non-wage income by causing children to shift into safer, more entrepreneurial, and more prestigious occupations.

To provide additional evidence that my results are not driven by confounders associated with each town's time-varying decision to apply for a library construction grant, I conduct a series of placebo tests and falsification exercises. I first show that the entry of college and university libraries in towns and cities across the United States had no effect on children's educational attainment. This is a falsification test because children likely did not have access to colleges and universities. I then perturb the date of each Carnegie grant by assuming that Carnegie offered the grant 20 years before (or after) it was actually offered. I show that these placebo grants had no effect on children.

Lastly, readers may worry that towns and cities invested in other local institutions that positively affected children at the same time that they constructed a public library. In particular, Goldin (1994, 1998, 1999) discusses the expansion of public high schools across the United States in the early 1900s, showing that this expansion had large effects on children. I construct a complete panel of all public high schools in the United States from 1890–1951. I use this data to show that controlling for the expansion of public high schools across the United States has no effect on my library results.

I connect my results to the findings of Stephens and Yang (2014) who provide evidence that compulsory schooling laws in the early and mid-1900s increased educational attainment but had null or negative effects on wage income. I replicate their results and extend them to additional outcomes, showing that compulsory schooling laws increased non-wage income and caused children to move into safer, more entrepreneurial, and more prestigious occupations. Both public library access and compulsory schooling laws increased non-wage income and made children more likely to enter occupations such as farm-ownership and technical work instead of occupations like

machine-operation and manual labor. My results highlight the importance of non-wage income and non-pecuniary outcomes in measuring the returns to education and the causal effects of local institutions on children. This may be especially important in a historical context where a larger fraction of the population was self-employed and engaged in highly unsafe occupations.

A handful of papers measure the causal effect of public libraries on individuals or communities. Kevane and Sundstrom (2014, 2016) measure the expansion of public libraries across the United States in the early 1900s and use aggregated data to argue that public libraries had no effect on county-level voting behavior. And contemporaneous work by Berkes and Nencka (2019) shows that Carnegie's library construction grants increased patenting activity when compared to towns receiving a grant offer that never materialized. Examining a more recent time period, Bhatt (2010) uses distance to a public library as an instrument for library access and argues that library access increases the amount of time children spend reading, the amount of time parents spend reading to their children, and homework completion rates. Bhatt also finds that access to a public library decreases the amount of time people spend watching television. And Gilpin, Karger, and Nencka (2020) show that sharp increases in public library investment causes increases in library usage and children's test scores. Lastly, Neto (2019) uses data from a modern public library census to argue that public library resources in Appalachia are not correlated with local employment and labor force participation rates.

My work draws on a large literature from the library sciences field, which ethnographically profiled and gathered information about Andrew Carnegie's library construction grants. Most notably, Bobinksi (1969) digitized and standardized information from the Carnegie Corporation's microfilm archives, producing a table of all of the grants Carnegie made to towns and cities. In other important work, Daniel (1961) discusses the expansion of public libraries across the United States; Martin (1993) collects detailed descriptions of why dozens of cities turned down Carnegie's promised grants; and Klinenberg (2018) traces the importance of social infrastructure like public libraries in the functioning of society. These sources provided important historical context for my empirical findings.

1.2 Data

1.2.1 Library Data

Andrew Carnegie did not publicize his interest in funding the construction of new public libraries. Still, after news spread of his first grants to construct public libraries in Southwestern Pennsylvania, politicians and officials in hundreds of cities and towns sent Carnegie unsolicited requests for funds to build new public libraries. Carnegie did not have time to respond to the requests he received for library funding, so after directing and overseeing the first few grants himself, he quickly put his personal secretary, James Bertram, in charge of the application process. Bertram would only seriously consider requests from city and town officials, but many members of the general public sent him letters requesting funds for a library. When he received a letter from a local resident of a town, he asked that they find an elected official who could submit an official request for a library (Bobinski 1969). In most cases, Bertram conducted the entire application and grant process by mail. Bertram also ensured that library construction grants only went to towns that committed to spending 10% of the grant amount on annual upkeep of the library. Carnegie scaled his grant amounts by the population of each town, targeting a grant amount of $2 per person in most cases. So in letters to cities and towns requesting funds, Bertram requested current population counts, which he often verified using publicly available tabulations from decennial census data (Bobinski 1969).

Carnegie allowed Bertram to give away the vast majority of $37 million in the early 1900s with little oversight.[3] After Bertram corresponded by mail with city or town officials, he would forward applications satisfying the basic requirements to Andrew Carnegie, who rubberstamped dozens of grant applications in short meetings. Carnegie trusted Bertram completely, and for good reason. Bertram was well-known for being fair in his assessment of applicants. As George Bobinski ex-

3. $37 million is the nominal dollar amount Carnegie used to fund library construction in the United States between 1890 and 1921. Using a CPI inflator, that is equivalent to $1 billion in 2019. But U.S. GDP was only $34 billion in 1910, so as a constant fraction of GDP, $37 million in 1910 is equivalent to close to $21 billion today.

plained in his authoritative book about the Carnegie library grant program, "[Bertram's] aloofness was attributed by one observer to a desire to maintain a strict impersonal and disinterested attitude toward each and every applicant... He judged proposals strictly on their merit. Personal relations or considerations never influenced his judgment. No worthy applicant was to be rejected, and, yet, no unworthy one was to be accepted" (Bobinski 1969, p. 30).[4] After Carnegie approved each grant, it took an average of 2.4 years for the town to build a Carnegie-funded public library. And 90% of the Carnegie-funded public libraries opened within four years of the initial grant.

The Carnegie Corporation preserved the correspondence between Bertram and community leaders. Drawing on these documents, Bobinski painstakingly collected exact dates, locations, and grant amounts for the grants Carnegie gave to cities and towns. In total, Bobinski calculated that Carnegie funded the construction of 1,618 libraries in 1,417 cities in the United States.[5] Almost all places that received a grant got funding to build one library, but a small number of cities received funding for an entire branch system. For example, New York City received funding to construct 66 libraries and Philadelphia received funding to construct 25 libraries. I hand-collect the opening date of each Carnegie library from archival sources.

In Figure 1.1, I link the digitized data from Bobinski to cities and I mark the modern-day zip codes of cities that received money from Carnegie.[6] Figure 1.1 shows several important patterns. First, New England was densely populated but received few grants. This is because it already had hundreds of public libraries before Carnegie began his philanthropic endeavors (see Figure 1.2, described below). Second, grants were concentrated in the Midwest, close to Carnegie's hometown in southwest Pennsylvania. Third, communities in the South applied for grants at lower

4. While Bertram funded most applications, Bobinski (1969) compiled an additional list of 225 failed grants. Carnegie offered these 225 grants to cities and towns, but the grant offer fell through for various reasons. In many cases the grant fell through because a local philanthropist offered to fund a new public library on his or her own. And even when a town rejected a Carnegie grant, they often built a new public library soon after. So I do not use these failed grants in my analysis.

5. Carnegie also funded a significant number of libraries outside of the United States. In this paper, I do not consider those grants.

6. Kevane and Sundstrom (2014) digitized this information from a table in Bobinski's book.

rates because they worried that Carnegie would require the libraries he funded to racially integrate (Bobinski, 1969).[7] Fourth, as Figure 1.1 shows, some cities received grants in the late 1800s while other cities received grants in the early 1900s. My econometric specifications will rely on spatial and intertemporal variation in the availability of Carnegie libraries, and Figure 1.1 confirms that there is a significant amount of variation within states and over time in the availability of Carnegie libraries.

I augment this data with information from library censuses. Between 1875 and 1929, the U.S. Bureau of Education surveyed libraries in the United States to gather information about the name, location, type, and founding year of every library in the United States with a non-negligible number of volumes. In some years, that number was as low as 300 books.[8] The type variable is not consistent across survey years, so it is difficult to differentiate public libraries from subscription libraries. I follow the methodology of Kevane and Sundstrom (2014) to identify public libraries by looking for keywords in the library name like "public library" or "free library."[9] In Figure 1.2, I use this data to map the founding date of the first public library in each town in the United States, through 1929. And in Figure 1.3, I plot the evolution of public libraries and Carnegie grants over time. Figure 1.3 shows that by 1920, more than 30% of the towns and cities in the U.S. with a public library had constructed a Carnegie-funded public library.

7. This worry was unfounded. Carnegie did not require that towns integrate the public libraries that he funded.

8. Kevane and Sundstrom (2014) digitized this data and Bailey, Jacob, Kevane, and Sundstrom (2011) used the data in unpublished work linking the expansion of public libraries across the United States with county-level economic outcomes.

9. More specifically, I call a library a **public** library if the library name contains one of these phrases: 'public library,' 'county library,' 'city library,' 'village library,' or 'free library.' The library name also must not contain any of these phrases: 'school,' 'department of public instruction,' 'college,' or 'university.' The founding dates in the public library data are unreliable, and often misreported. Still, I show that Carnegie grants are highly correlated with the founding dates of public libraries from this survey data.

1.2.2 Census Microdata

I use complete count census data to measure the causal effect of public library access on children. Researchers at the University of Minnesota's Institute for Social Research and Data Innovation cleaned and standardized this data, and the 1900–1940 datasets are publicly available at `usa.ipums.org` (Ruggles et al., 2020). I access these datasets as part of a license from the National Bureau of Economic Research (NBER), which hosts a restricted copy of the datasets that includes the full name of each respondent. I produce all of the results in this paper on NBER's servers. I measure childhood demographic characteristics in the 1900–1930 census data and I measure adult outcomes in the 1940 census, which was the first national census in the U.S. to contain individual-level educational attainment and income information. I restrict my attention to men, because many women's last names changed from childhood to adulthood, making it difficult to match women across decennial census years.

I link the 1900–1930 census records to 1940 census records using an iterative matching technique, drawing from research, programs, and thoughtful descriptions by Ferrie (1996), Abramitzky, Boustan, and Erikkson (2012, 2014), and Bailey et al. (2019). See the Matching Appendix for more details about the procedure I use. To summarize the method, I use 36 combinations of exact, cleaned, and phonetic first and last name, birthplace, race, and age to search for possible matches between 0–25 year old men in each of the four early census years (1900–1930) and 15–65 year-old men in the 1940 census. In each of the 36 iterations, I relax an exact matching requirement and drop any uniquely identified child who matches to multiple possible adults. As an example, in the first pass I look for unique perfect matches between records from the 1900 and 1940 census using phonetic name codes, race, birthplace, and age. In the second iteration, I allow age to be one year higher in the 1940 data than in the 1900–1930 data, and in the third iteration, I allow age to be one year lower in the 1940 data. After checking for matches within one- and two-year age bands, I relax additional restrictions. This method of matching closely follows Abramitzky, Boustan, and Erikkson (2012, 2014), except that after finding high-quality matches, I search the remaining un-

matched records to attempt to find lower-quality matches so that I can better analyze adult census records in 1940 that match to multiple early childhood records.

In the set of 39.2 million black, white, and Native-American men in the 1940 census born between 1875 and 1920, 25.7 million (66%) match to a unique record in at least one earlier census year. 12.3 million adult men in 1940 match to one childhood census record and 13.3 million adult men in 1940 match to unique record in multiple early census years. I use two tests of match accuracy to further refine my sample. First, in the 1940 decennial census, sample-line respondents reported their parents' birthplaces. I test link accuracy by comparing these responses to reported parental birthplaces from the matched childhood census record (following Bailey et al. 2019). Second, I focus on the 13.3 million adult records in 1940 that match to unique records in two early censuses. If a 45 year-old John Smith in the 1940 census matches to the record of a 5 year-old John Smith in 1900 and a 15 year-old John Smith in 1910, then I can compare the reported names of John Smith's parents in 1900 and 1910. If the first two letters of John Smith's mother's and father's first names in 1900 are different from the first two letters of the mother's and father's first names in 1910, then it is likely that one of those early census records is not a correct match.[10] In those cases, I do not include that 1940 record in my matched sample. For 1940 records that match to multiple early census records, I keep the earliest matched census record in my matched sample.

My final matched sample consists of 17.8 million people with both an adult census record in 1940 and a childhood census record in 1900, 1910, 1920, or 1930. In Table 1.1, I show average characteristics of the full set of 1940 adult men, my matched sample, and a subset of the matched sample consisting of 5.8 million people for whom I additionally match at least one brother to adult

10. This is unlikely to happen by chance. The most common first two letters of fathers' first name in my matched sample are 'JO' and the most common first two letters of mothers' first name in my matched sample are 'MA'. Each occurs around 15% of the time because of the large number of people named John and Mary in the early 1900s. So assuming independence, the probability that two randomly selected boys will have a father or mother with the same first two letters of their first names is at most $2 * 0.15^2 = 4.5\%$. I rely on first letters rather than exact matches of parent names across childhood census records to allow for transcription errors, which are quite common in digitized early census records. And in cases where a child is reported in a household with only one parent, I use only that one parent for this comparison.

records in the 1940 census. I use this sample for my within-family analysis. My overall match rate of 1940 records to childhood records, at 45.5%, is at the high end when compared with other linking strategies (Abramitzky et al. 2020; Bailey et al. 2019).[11] And I estimate that roughly 80% of these matches are correct,[12] which is a high level of accuracy relative to other matching techniques with a similar overall match rate (see Abramitzky et al. 2020 and Bailey et al. 2019 for accuracy rates using different linking methods.). See the Matching Appendix for details about these calculations.

Each early census record identifies the respondent's enumeration district, and I use these enumeration districts to link each adult to the city or town in which he lived as a child. Enumeration districts in the early 1900s defined a contiguous geographic area where one census enumerator could knock on every door and collect census information for every household within 2–4 weeks (Census Bureau, 1940). So cities like Chicago contained many enumeration districts, while sparsely populated enumeration districts could contain an entire county. There were around 65,000 enumeration districts in each census year. In the census data, there are also place names associated with most respondents. For example, respondents living in one enumeration district have a place name of "Precinct 16, Cahaba," referring to the 16th precinct of Cahaba, Alabama: a town with a population of only a few hundred people in 1900. The average enumeration district had a population of roughly 1,200, making enumeration districts a granular measure of geographic place. See the Data Appendix for more details about the raw geographic variables I use in this paper.

I clean and standardize these raw places names and I use the cleaned place names to match census records to Carnegie grants and public libraries. In other words, I link each of Carnegie's grant to children in the 1900–1930 census living in towns and cities with the exact names corresponding to each grant. Of the 1,417 successful grants, I map all but seven to at least one enumeration

11. This is unsurprising, because for each adult in 1940 I often have two attempts to find their childhood census record in an early census year. Most current approaches to census linking attempt to link records between only two decennial censuses.

12. In other words, for every 100 records in 1940 that are in my final matched sample, 80 are matched to a childhood census record corresponding to the same person, and 20 are not.

district. And many grants match to a large number of enumeration districts (like the grant Andrew Carnegie gave to New York). In Table 1.8.2, I list the most populous cities and towns in my sample of matched census records, along with the date Carnegie offered a grant to that city, if a grant was offered. The matched sample is not concentrated in a small number of places. Chicago contained 2.3% of the matched sample, and the boroughs of New York City together make up 3.3% of my matched sample. My results are robust to the exclusion of cities and towns containing more than 0.2% of my matched sample, which excludes all major cities.

Cities and towns provide the best unit of analysis for this project because Carnegie grants targeted individual cities and towns. In a textbook for students interested in working in library circulation occupations, Flexner (1927) differentiates between patrons who could use the public library for free, and patrons who paid a subscription fee. Flexner explains that local students, residents of the local service area, and non-residents with business addresses in the "town or county" would normally receive "free access" to the local public library. But "all other applicants living outside the legal boundaries of the library service area" would be given library access only "on payment of an annual, semi-annual, or monthly fee." In other words, the public library was freely available only to local residents. And since Carnegie gave library construction grants to cities and towns, which raised revenue to fund the library, in most cases only residents of the city or town who paid for library maintenance would have had free access to the Carnegie library. Another reason to rely on cities and towns as the unit of analysis is distance. I am measuring the effect of library access on children; if children lived in the same county as a Carnegie library, but a different town in that county, getting to the library would have been more difficult. So when I compare children in neighboring towns within a county, I will attribute differences in adult outcomes to this differential library access.

1.3 Methodology

For each person in my matched sample, I construct two measures of Carnegie grant exposure. First, in my main specifications I use the binary measure $1(GrantByAge5)$–an indicator for whether each child received a Carnegie grant by the age of 5. Almost all Carnegie-funded libraries were constructed within four years of when the grant was given. So, this treatment indicator is an indicator of whether each child had access to a Carnegie library during their childhood. Second, in robustness checks I analyze *ExposureCarnegie*, a measure of the fraction of years between age 5 and 20 when a child had access to a Carnegie grant. I use the data collected by Bobinski (1969) to calculate these two measures. As an example, if the decennial census records John Smith as an eight year-old living in New York City in 1900, I look in the Carnegie grant data and see that New York City received a Carnegie grant in 1899. So I calculate John Smith's exposure to a Carnegie grant as $1(GrantByAge5) = 0$ and $ExposureCarnegie = \frac{14}{16} = 0.88$. Or more generally, for a child born in year y and living in city c as a child,

$$ExposureCarnegie_{y(i),c(i)} = \frac{1}{16}\sum_{t=5}^{20} 1(CarnegieGrant_{y+t,c})$$

where $1(CarnegieGrant_{y+t,c})$ is an indicator for whether city c received a Carnegie grant during or before year $y+t$.

I can now analyze the effect of Carnegie grant exposure on adult outcomes of children. I model that relationship using specifications of the form:

$$Outcome_i = \beta_0 + \beta_1 * 1(GrantByAge5)_{y(i),c(i)} + \Pi X_i + \varepsilon \tag{1.1}$$

where i indexes individuals, $Outcome_i$ is an outcome measured in the 1940 census, and X_i is a matrix of controls. My use of cohort*city-level variation in exposure to Carnegie grants draws from several papers measuring the effect of school construction on children. Duflo (2001) uses the exposure of different cohorts of students in regions of Indonesia to a school construction program

and measures the effect of the program on educational attainment. In related work, Neilson and Zimmerman (2014) link a staggered school construction program in New Haven, Connecticut to neighborhood and child outcomes to measure the effect of the program on children. And Aaronson and Mazumder (2011) link Julius Rosenwald's funding for black schools in the South in the early 1900s to children's school attendance using a methodology that I follow closely.

My choice of specifications draws from Aaronson and Mazumder (2011), who measure the effect of exposure to a constructed Rosenwald school between ages 7 and 13 on school attendance. I follow their strategy of beginning with a model that controls for baseline characteristics, and progressively adding controls until my final specification includes combinations of family fixed effects, county-by-birth year fixed effects, and birth order fixed effects. In my most basic specification, X_i includes birth year, state of birth, and birth order fixed effects. I then progressively add more control variables, including parental characteristics, enumeration district fixed effects, county-by-birth year fixed effects, and census microfilm page fixed effects. Because enumerators traveled door-to-door collecting data, census microfilm pages often list households in the order of neighboring houses; this makes census microfilm pages a granular measure of place (Logan and Parman, 2017; Magnuson and King, 2010). In my within-family specification, X_i also includes household fixed effects, allowing me to compare siblings within a household to each other and to similarly situated children in the same county who received no (or less) access to a Carnegie grant. I interact all of these fixed effects, and the fixed effects included in specifications throughout this paper, with race-by-childhood census year fixed effects. The coefficient of interest is β_1, which measures the effect of exposure to a Carnegie grant on children's adult outcomes.

In my within-family specification, I perform a more complicated version of this thought experiment: Consider two households (A and B), living in different towns in the same county in 1910. Each household has two sons, ages 0 and 16, who I will index A_{oldest}, $A_{youngest}$, B_{oldest}, and $B_{youngest}$. Household A lives in a town that receives a Carnegie library grant in 1915. Household B's town never received any grant. This means that $A_{youngest}$ has $1(GrantByAge5) = 1$ and the other three children have $1(GrantByAge5) = 0$. Let Y_i be years of educational attainment in 1940.

β_1 measures the difference-in-difference of educational attainment between these four children. Or algebraically, β_1 would be $(Y_{A_{youngest}} - Y_{A_{oldest}}) - (Y_{B_{youngest}} - Y_{B_{oldest}})$. And if the youngest child in household A obtains an abnormally high level of educational attainment, I will attribute this to his exposure to a Carnegie grant. In the regression specifications that I present, there are more than two million families and β_1 represents a more complicated weighted average over the within-family differences in grant exposure and outcomes.

To theoretically motivate the within-family model, consider a framework where unobserved heterogeneity at the family-level is correlated with exposure to a Carnegie grant. Following Todd (2008), let Y_{ij} be the outcome for person i from family j and assume that the data-generating process is:

$$Y_{ij} = \phi(X_{ij}) + D_{ij}\beta + \theta_j + \nu_{ij}.$$

Where D_{ij} is an indicator for whether person i in family j was treated, X_{ij} is a series of controls, like birth cohort or birthplace, and β is the coefficient of interest. θ_j is a family effect, which could represent genetics, family tradition, or the decision of the family to live in a particular place, all of which may be correlated with access to a Carnegie grant. For example, parents may move to a town because they expect that it will receive a Carnegie grant in the near future. Or, families with a high θ_j might be more likely to live in towns with a Carnegie-funded library. ν_{ij} is an unobserved error term that is uncorrelated with θ_j, D_{ij}, and X_{ij}. Controlling for family fixed effects allows me to control for any family-level characteristics that are correlated with Carnegie grant exposure. If I did not include θ_j in my regression specifications, and it was correlated with X_{ij} or D_{ij}, we would have a standard omitted variable problem. Chetty and Hendren (2018) rely on a similar framework to measure the causal effect of neighborhoods on adult outcomes of children. If there is heterogeneity in the treatment effect so that β_i varies with i, then the within-family specification recovers a weighted average of the individual-level treatment effects, where the weights correspond to within-family variation in treatment.

1.4 Results

1.4.1 Education

I begin by looking at the effect of Carnegie grant exposure on years of educational attainment, high school graduation rates, and college attendance rates. These results provide strong evidence that access to a Carnegie grant increased educational attainment. In Table 1.3, I present the main set of regression results (see Equation 1) with years of schooling as an outcome. Access to a Carnegie grant is consistently associated with increased educational attainment across specifications. With a baseline set of controls, children with access to a Carnegie grant by age five had 0.57 more years of educational attainment than children without access to a Carnegie grant. But most of this gap is driven by selection on observables. When I control for neighborhood fixed effects, the difference drops to 0.18 years, and the effect of a Carnegie grant on years of educational attainment settles between 0.08 and 0.13 years when I add additional controls. In my within-family specification, a Carnegie grant received by age five increases educational attainment by 0.10 years.[13]

The difference between the magnitude of the treatment effect in the first five models and the sixth within-family model is not because of the sample's changing composition. I can use only 30% of my sample to estimate the within-family model because the other 10 million children do not have a brother in my matched sample. But when I re-run all models with only the sample of matched brothers from the within-family model, identical patterns emerge (see Appendix Table 1.18). Exposure to a Carnegie grant had a positive and statistically significant effect on educational attainment in all models using this sample. The magnitude of the treatment effect remains between 0.10 and 0.12 for this sample of brothers in all specifications containing enumeration district fixed effects and parental controls.

13. Throughout this paper, I cluster standard errors at the level of the childhood county of residence, following Aaronson and Mazumder (2011). The United States contains over 3,000 counties. Clustering at the state of residence level (51 clusters) only marginally increases the size of standard errors in some specifications, and does not change the statistical significance of my coefficient of interest (see Appendix Table 1.17). I calculate large numbers of multi-way fixed effects using the reghdfe Stata package, introduced in Correia (2016).

In Appendix Table 1.19, I present the same baseline models measuring the effect of Carnegie grants on educational attainment, but I use the continuous measure of exposure (*ExposureCarnegie*) instead of the binary measure of exposure that I use throughout the rest of this paper. In the within-family model, exposure to a Carnegie grant throughout childhood increased educational attainment by 0.20 years. This treatment effect estimate is larger than the estimate using a binary indicator because it reflects the causal effect of moving a child from no exposure to a Carnegie grant to complete exposure (from ages 5–20). But the within-family variation in exposure to a Carnegie grant is well below one because very few families have children born 15 or more years apart.

There are several clear concerns about exogeneity in this model. First, the results may be driven by sharply improving conditions for children in the South or by higher-quality public libraries in the Northeast. The county-by-birth year fixed effects should capture most of this variation, but in Appendix Table 1.20 I show that the education results are unchanged when I subset the sample to only children in the Midwest, the location that received the most Carnegie grants. My results also do not change when I subset the sample to white children (see Appendix Table 1.21).

Another concern is that parents may have differentially migrated with their children to towns that received a Carnegie library. And the younger children in families that moved may have had particularly high innate ability or expected educational attainment. In Appendix Table 1.8.2, I remove from my sample any child whom I see in the census only after a Carnegie grant was given to the town. For example, if a town received a grant in 1905, I remove all children in that town who I observe in the 1910, 1920, and 1930 decennial censuses from the analysis sample. This reduces my sample size significantly, and the within-family treatment effect is slightly less precisely estimated than in my main specification, but still shows that access to a Carnegie grant by age five causes children to obtain a statistically significant 0.10 additional years of educational attainment.

In Tables 1.4 and 1.5, I present linear probability models in the same regression framework from Equation 1, with high school graduation and college attendance as outcomes, instead of years of educational attainment. Access to a Carnegie grant increased high school graduation rates across

specifications. In the within-family specification, access to a Carnegie grant increased high school graduation rates by 1.4 percentage points, from a baseline of 26 percent in the matched sample of siblings. Carnegie grants also increased the probability that a child would attend at least some college by 1.5 percentage points from a baseline of 11 percent in the matched sample of siblings used in the regression model. These are large effects, and they imply a positive effect of Carnegie-funded public libraries on the right tail of educational attainment.

I now present two-stage-least-squares estimates of the same model, but instead of measuring the effect of Carnegie grants on educational attainment, I instrument for the availability of an opened Carnegie library by the age of 5 (as an indicator) with the year of each Carnegie grant. As Table 1.6 shows, the first stage is highly significant, with a coefficient of 0.43 in the within-family specification and an F-statistic of 352. Access to a Carnegie-funded public library induced by a Carnegie grant leads to 0.24 additional years of educational attainment in the within-family specification. I present only the 2SLS estimates for this main outcome to show the strong first stage. In the rest of the paper, I focus on reduced form results that rely on the precise Carnegie grant dates and the fact that the construction of Carnegie-funded public libraries took an average of 2.4 years.

1.4.2 Placebo Tests

There are three clear concerns with my interpretation of the coefficients in Table 1.3 as causal effects of library exposure on children. First, cities and towns may have invested in public libraries at the same time as they invested in other local institutions. In dozens of state reports on education systems from the early 1900s, officials discuss the expansion of public libraries and public high schools across their state. Goldin (1994, 1998, 1999) presents state-level data on the dramatic expansion of high schools across the United States in the early 1900s, and it is possible that towns constructed public high schools at the same time that they accepted Carnegie's library construction grants. To address this concern, I construct the first complete panel of public high schools in the

United States. I collect information about each high school's founding date and exact location from 8,000 pages of books and reports that I scan, digitize, and standardize. For more details about the sources, see the Data Appendix. In Figure 1.4 I plot the first year in which each town and city in the United States opened its first public high school, and as with public libraries, we can see that high schools spread across the Northeast and Midwest much earlier than in the South. In Appendix Table 1.22, I take my main educational attainment result (Table 1.3) and I include as a control variable county fixed effects interacted with fixed effects indicating the age of each child when a public high school first entered their town. I calculate these ages at high school entry using the complete panel of public high schools. Controlling flexibly for the entry of each town's first public high school has no effect on my coefficient of interest. Access to a Carnegie grant increased educational attainment by 0.10 years.

A second related concern is that when a town received a Carnegie grant, it could function as a proxy for contemporaneous town-level financial success. When a town can support an institution like a library, it may be a sign of time-varying growth in population, financial independence, or tax revenue that would positively affect children in a way that correlates with, but is not a function of, public library access. To address this concern, I use the library censuses described in the Data section to create a dataset of all college and university libraries in the United States, along with their founding date. I link 1,126 of these libraries to census microdata and I construct an indicator for whether each child had access to a college or university library by the age of 5. In Appendix Table 1.23, I re-run my main specification from Table 1.3, looking at the effect of exposure to a college or university library on children's educational attainment as adults. Exposure to a college library by the age of 5 is associated with educational attainment until I control for neighborhood (enumeration district) fixed effects and parental characteristics. In addition, when I add county-by-birth year fixed effects, census reel fixed effects, and household fixed effects to my main specification, we can see that exposure to a college library has a precise null effect on the educational attainment of affected children.

Finally, Andrew Carnegie may have given library construction grants to towns that were reg-

ularly investing in local institutions more than neighboring towns. If this were the case, then younger children living in the Carnegie-funded towns might always achieve higher levels of educational attainment than their older brothers, even if Carnegie-funded libraries had no effect on children. To test this, I perturb the Carnegie grant dates. In Appendix Table 1.24, I assume that Carnegie gave grants to towns 20 years after he actually did. In the baseline specification, these placebo Carnegie grants are still associated with children's educational attainment. But once I add parental characteristics (column 3), exposure to these placebo grants has no effect on the final educational attainment of children. In Appendix Table 1.25, I perform the same analysis, but I assume that Carnegie gave grants to towns 20 years before he actually did. In the within-family specification (column 6), access to these lagged Carnegie grants has a precise null effect on children's educational attainment.

1.4.3 Income

Now that I have established that Andrew Carnegie's library construction grants increased children's educational attainment, I analyze the effect of Carnegie grants on income. In Table 1.8.2, I measure the effect of Carnegie grants on log annual wage income. I treat anyone who reports zero wage income as having zero log wage income and I use the same set of specifications discussed above.[14] The first column shows that Carnegie gave grants to communities where children would grow up to have more wage income than children from communities which did not receive a grant. But once I control for neighborhood fixed effects (in column 2), exposure to a Carnegie grant is associated with a 2 log point decline in adult wage income. Adding additional controls implies that Carnegie grants had a negative effect on wage income, and in the within-family specification, Carnegie grants had a precise null effect on wage income.

The wage effects may be puzzling, because economists often find that human capital interventions increase wage income. But wage income is only one component of total income. Wage

14. While these results use the full sample of 20–65 year old men in the 1940 census, all results are unchanged if I subset to prime-age 25-55 year old men.

income includes all money a worker earned as an employee, but wage income excludes "the earnings of businessmen, farmers, or professional persons derived from business profits, sale of crops, or fees" (Census Bureau, 1940). In 1940, the Census Bureau did not ask respondents to report their non-wage income. However, the census did ask all non-institutionalized respondents to state whether or not they had received at least $50 of non-wage income in the reference year. 33% of men in my matched sample of siblings reported having at least $50 of non-wage income. In Table 1.8, I show that Carnegie grants had large positive effects on non-wage income. The baseline specification (column 1) shows that Carnegie gave grants to communities where children would grow up to have less non-wage income. But once I add granular geographic controls, we see that Carnegie grants increased the probability that children had at least $50 of non-wage income by 0.4–0.7 percentage points. And in my within-family specification (column 6), we see that Carnegie grants increased the probability of having at least $50 of non-wage income by 0.7 percentage point relative to a baseline of 33 percent in this sample (see Table 1.1).

To summarize the income effects, Andrew Carnegie's library construction grants had no effect on wage income and increased non-wage income. But because of how the Census Bureau collected income information in the 1940 census, it is impossible to directly measure the effect of Carnegie-funded public libraries on total income.

1.4.4 Imputed Income Measures

In 1950, the Census Bureau asked respondents to state the dollar value of their non-wage income in the decennial census for the first time. I use the non-wage income distribution from 1950 to impute a measure of total income in 1940. Because the Census Bureau did not collect income information before 1940 and collected only some income information in 1940, many economists impute wage and non-wage income in early census years using the wage distribution in 1950.[15]

15. For examples of this type of imputation, see Bailey and Collins (2006), Feigenbaum (2015), and Bayer and Charles (2018). A recent paper by Saavedra and Twinam (2020) discusses many more examples, and proposes a LASSO-based technique to construct an income measure for early

I begin my imputation procedure with a sample of 20–65 year old white, black, and Native-American men in the 1950 census who were asked to report wage, non-wage, and total annual income. I then regress each of those income measures on an indicator for whether the respondent earned more than $120 of non-wage income,[16] indicators for 20 quantiles of wage income, and fixed effects for age, adult state of residence, industry, occupation, years of educational attainment, and annual hours worked. I interact these fixed effects with race. I use the coefficients from this regression in 1950 to predict wage income, non-wage income, and total income for each respondent in the 1940 census. When imputed values are missing because fixed effect cells contained only one individual, I use predicted values from a simpler secondary regression that only includes these fixed effects as regressors: state, race, industry, occupation, age, weekly hours worked, and annual weeks worked.[17]

In Tables 1.9–1.11, I measure the effect of Carnegie grants on imputed total income, imputed wage income, and imputed non-wage income. First, we see in Table 1.9 that access to a Carnegie grant by age five had a positive effect on imputed total income in the within-family specification. This coefficient is a precisely estimated 2.5 log points. In Table 1.10, we can see that exposure to a Carnegie grant had no effect on wage income once I include county-by-birth year fixed effects as controls. This is consistent with the actual effect of Carnegie exposure on wage income in 1940 (see Table 1.8.2), which is unsurprising because my imputation procedure conditions on reported wage income percentiles. Lastly, in Table 1.11 we see that Carnegie grants had large

censuses using income measures from the 1950 census. Although Saavedra and Twinam's method is ideal for imputing income measures when no income information exists, in 1940 the Census Bureau collected exact wage income information and an indicator for whether each respondent earned more than $50 of non-wage income. I therefore use these two variables as inputs into my imputation procedure. This improves accuracy, and is feasible because I can condition on those variables in the 1940 and 1950 census microdata.

16. $120 in 1950 is equivalent to $50 in 1940, using the CPI to inflate 1940 dollars to 1950 dollars.

17. There are roughly 92,000 observations in these regressions of income on demographic characteristics in the 1950 census, 4,000 indicator variables in the first regression, and 700 indicator variables in the second regression.

positive effects on imputed non-wage income once I include a basic set of controls. The point estimate ranges from 4–5 log points across models 3-6. In column 6, the within-family estimate implies that access to a Carnegie grant by age five increased non-wage income by 4 log points. To summarize, Carnegie grants had no effect on wage income, increased non-wage income by 4 log points, and either had no effect on total income or increased total income by 2.5 log points (in the within-family model). These imputed income measures are based on the 1950 wage distribution.

1.4.5 Occupational Prestige

In Table 1.12, I measure the effect of Carnegie grants on occupational prestige. As my measure of occupational prestige, I use a coding by Siegel (1971) of 412 occupations, which is the only direct measure of occupational standing harmonized by IPUMS to match occupation codes in the census.[18] Siegel constructed these measures using a 1963 study that asked respondents to rate the prestige of different occupations. While this question is amorphous, Siegel (1971) discusses a study of 490 respondents that found a correlation of ≥ 0.9 in their ranking of 30 occupations along any of these dimensions: "opportunity for advancement," "security," "influence over others," "responsibility for supervising others," "work calls for originality and creativity," "education required," "training required," "interesting and challenging work," and "regarded as desirable to associate with" (Garbin and Bates, 1966). Although Siegel performed his study well after the 1940 census, a large literature shows that occupational prestige remained stable between the 1940s and the 1960s. Prestige rankings from a 1947 study and a 1963 replication had a correlation of 0.99 (Hodge, Siegel, and Rossi 1964).

The main problem with occupational prestige metrics is that the units are not easily interpretable. Siegel put his study participants in front of a numbered nine-step cardboard ladder. He

18. Many other measures of occupational standing rely on the average or median income or educational attainment from a set of respondents in an occupation or industry in one census year. These occupational scores can then be used to impute some measure of income or education, based on occupation, in analyses using pre-1940 census data. But my 1940 data contains individual income and educational attainment for respondents, so those measures are duplicative.

then gave them small paper cards, each with an occupation. Siegel told respondents to put occupation cards on the top position in the ladder if it "has the highest possible social standing," at the bottom of the ladder if the respondent thought that the occupation had "the lowest possible social standing," and otherwise, "somewhere in between." Siegel gave each rung an equally spaced numeric score, ranging from 0 for the lowest rung through 100 for the highest rung of the ladder. The prestige score for an occupation was then the weighted average of individual respondent ratings for each occupation. As Table 1.1 shows, the occupation-weighted average score in 1940 was around a 34 for my matched sample, and the standard deviation of occupational prestige was 18. But the meaning of the effect size here is difficult to interpret.

I standardize the occupational prestige measure to be mean zero and standard deviation one in my sample. As I show in Table 1.12, exposure to a Carnegie grant increased occupational prestige across specifications. In my baseline model, Carnegie grants increased occupational prestige by 0.126 standard deviations, but most of this effect is driven by selection on observables. When I add neighborhood fixed effects, the point estimate attenuates to 0.050. And in my within-family specification (column 6), exposure to a Carnegie grant increases occupational prestige by 0.020 standard deviations in the matched sample of siblings.

1.5 Compulsory Schooling Laws

I have shown that access to a public library in the early 1900s increased educational attainment, had no effect on wage income, and increased non-wage income. These results mirror the findings of Stephens and Yang (2014), who showed that compulsory schooling laws (CSLs) increased educational attainment but did not affect the wages of adults in the 1960–1980 censuses.

Stephens and Yang wrote their paper in response to a large literature measuring the effect of CSLs on children. For example, Angrist and Krueger (1991) and Oreopoulos and Salvanes (2011) show that CSLs increase educational attainment and increase wage income. But Stephens and Yang demonstrate that the positive effect of CSLs on wages is driven by differential changes in income

by census region, over time. They add region*birth year fixed effects to the standard regression models linking education to income, instrumenting for education with compulsory schooling laws. The addition of these fixed effects attenuates (to zero) the effect of education on income in their models.

In Table 1.13, I use the main specification from Stephens and Yang's paper to measure the effect of educational attainment on respondents' weekly wage, annual wage income, annual non-wage, and annual total income. I also measure the effect of educational attainment on occupational prestige, and housing values. The specification, taken from their paper, is:

$$Outcome_{st,i} = \alpha Educ_{st,i} + \chi_s + \delta_t + \beta x_{st,i} + \varepsilon_{st,i} \tag{1.2}$$

where i indexes individuals, s indexes state of birth, and t indexes year of birth. I follow their paper and instrument for years of educational attainment ($Educ_{st,i}$), using three indicators (RS7, RS8 and RS9) for whether each adult in the 1960–1980 census was required to attend school for seven, eight, or nine years. In other words, the first stage is:

$$Educ_{st,i} = \pi_1 RS7_{st} + \pi_2 RS8_{st} + \pi_3 RS9_{st} + \lambda_s + \theta_t + \mu x_{st,i} + v_{st,i} \tag{1.3}$$

I use Stephens' and Yang's dataset and replication programs to produce these results, but I merge on additional outcomes from IPUMS microdata (Ruggles et al., 2020). In Panel A of Table 1.13, I present results for all adults. In Panel B, I present results for only men. The first column of Table 1.13 replicates the 2SLS result from Table 1 of Stephens and Yang (2014). Stephens and Yang use weekly wage income as their main outcome, but the weeks worked variable is quite noisy, so I also present results using annual income measures. Similar to my library results, educational attainment had negative (albeit noisy) effect on annual wage income, decreasing the annual wage income of men by 24 log points. Education also dramatically increased non-wage income by 52 log points, but this outcome is zero around 70% of the time, making the point estimate quite sensitive to how I treat zeroes. In the fourth column, I show that an additional year of education increases

the probability a respondent had a significant amount of non-wage income by 5 percentage points, relative to a baseline of 34%. Educational attainment also increased occupational prestige by 0.09 standard deviations. These effects on wage and non-wage income combine to imply that educational attainment had a noisy but positive effect on total income of 7 log points for men and women (and 4 log points for men).

The 1960–1980 censuses also contain housing price values, and in the seventh column of Table 1.13, I show that an additional year of education, induced by compulsory schooling laws, increases respondents' home value by 9 log points, although this outcome is only measured for home-owners. To summarize these results, Carnegie-funded public libraries and compulsory school laws had null or negative effects on wage income, positive effects on non-wage income, and positive effects on occupational prestige.

1.5.1 Occupational Choice

I now show that occupational choices drove this large increase in non-wage income in both the Carnegie library and compulsory schooling law settings. In Table 1.14, I use a series of linear probability models to measure the effect of a Carnegie grant on occupational choice. The sample is the set of workers who report a valid occupation in 1940.

Census enumerators in 1940 asked respondents to state their occupation. IPUMS aggregated these tens of thousands of occupations into roughly 200 granular categories and nine broad categories. I examine the effect of access to a Carnegie grant on the probability that children chose occupations in one of these nine broad categories as an adult. The broad categories, along with some granular examples from each broad category, are:

1. Professional and Technical: accountants, dentists, professors, nurses, and engineers

2. Farm Owner: farm owners and tenant farmers[19]

19. It is difficult to differentiate farm owners and tenant farmers in the census data, since both responses were often enumerated as "farmers" in the 1940 decennial census. But these two occupa-

3. Managers, Officials, and Proprietors: buyers for stores and building managers

4. Clerical: banktellers, clerks, and telephone operators

5. Sales: advertising agents, insurance brokers, and salesmen

6. Craftsmen: bakers, mechanics, plumbers, and tailors

7. Operatives: electricians, machinists, furnacemen, miners, and taxi drivers

8. Service: hospital attendants, barbers, housekeepers, and waiters

9. Laborers: farm laborers, fisherman, and lumbermen

In Table 1.14, we see that access to a Carnegie grant increased the probability of becoming a professional worker, a farm owner, a clerical worker, and a salesman by a combined 2.5 percentage points. This is relative to 31% of the matched sample of siblings in those four occupational groups. Access to a Carnegie grant decreased the probability of becoming a manager, a craftsman, an operative, or a laborer by 2.5 percentage points from a baseline of 62 percentage points in the matched sample of siblings. The effect on managers may be surprising, but most managers in 1940 were small-scale proprietors of stores and businesses. While these were not dangerous or low-paid occupations, clerical and technical occupations were better-paid and more prestigious.

In Table 1.15, I measure the effect of educational attainment on occupational choice, instrumenting for educational attainment using compulsory schooling laws and the main regression

tions were more similar than readers might expect. For example, Hodge, Siegel, and Rossi (1964) discuss a survey from 1947 which asked people for their opinions about the occupational standing of 90 occupations. The most prestigious occupations were U.S. Supreme Court justice, physician, and nuclear scientist. The least prestigious were garbage collector, street sweeper, and shoe shiner. 'Farm owner' ranked as the 43rd most prestigious occupation, and 'tenant farmer' ranked as the 51st most prestigious occupation. Farm owners ranked directly below trained machinists and above undertakers. Tenant farmers ranked directly below bookkeeper and above insurance agent. So while farm owners did have higher occupational standing than tenant farmers in the 1940s, they had similar occupational standing and were both considered to be decent middle-class occupations. Also of note is that in 1940, 38% of farm operators were tenant farmers, and not land-owners (Census Bureau, 1950).

model described above (Equations 2–3). As in Table 1.14, the outcomes are indicators for whether an adult worked in the given broad occupational category. These effects are quite similar to the Carnegie grant effects from Table 1.14. Increases in educational attainment induced by compulsory schooling laws led to large shifts in occupational choice. An additional year of education increased the probability a person engaged in professional or technical work, or became a farm-owner. An additional year of education also caused fewer children to work as craftsmen or operatives, although the effect on entering professional/technical work and moving away from craftsmanship are not statistically significant when I focus on prime-age men in my sample.

1.5.2 Life Expectancy

It is difficult to place a direct monetary value on increases in educational attainment, changes in occupational choice, or increases in occupational prestige, but in this section of the paper I use the Social Security Death Master File (SSDMF) to measure the effect of occupational choice on mortality. I then use these occupational safety measures to show that Carnegie grants moved people into safer occupations.

The 2013 SSDMF is a dataset containing the birth date, death date, and social security number for 55 million people who died between 1899 and 2013.[20] The SSDMF is not comprehensive, but contains more than 70% of deaths among 65+ year-olds who died between 1967 and 1972, and more than 95% of deaths among 65+ year-olds who died after 1972 (Hill and Rosenwaike 2001). Coverage is significantly lower for deaths at ages below 65.

I link the SSDMF to my matched census dataset, searching for exact matches using first name, last name, gender, birth year, and birth month when possible. Of the 12.9 million records in my final matched sample, 4.0 million match to a record in the SSDMF with an age at death of 65 or older. I do not have enough statistical power to directly measure the causal effect of Carnegie

20. The dataset is no longer publicly available, but was previously purchasable through the Social Security Administration for around $2,000. A blogger made his copy of the 2013 dataset public for use by researchers. I downloaded the data from `http://ssdmf.info/download.html`

grants on individual life expectancy, but I use these linked records to measure the effect of entering different occupations and industries on life expectancy. I then show that Carnegie grants and compulsory schooling laws moved children into physically safer occupations and industries.

I begin with a hedonic regression, decomposing respondents' age at death according to this linear model:

$$AgeAtDeath_i = \beta_0 + \alpha_{O(i)} Occupation_i + \delta_{I(i)} Industry_i + \beta X_i + u_i \tag{1.4}$$

where α_O is a fixed effect for occupation O and δ_I is a fixed effect for industry I. X_i is a matrix of controls including household fixed effects, county*birth year fixed effects, birth order, and parental characteristics.[21]

I run this regression on a sample of roughly 600,000 people who lived in households as children where two or more siblings match to the SSDMF data, and who matched to records in the 1940 census. I extract these occupation and industry fixed effects (there are around 200 of each), and I use these fixed effects as measures of occupation and industry safety. In this within-family specification, each fixed effect measures the effect of occupation or industry on life expectancy, conditional on living to age 65. This strategy is similar to the literature calculating the value of a statistical life. In that literature, researchers begin with a worker-level dataset and regress wages on demographic characteristics and occupational fatality risk. The coefficient on fatality risk is used to determine the difference in wages that similar workers expect to receive when choosing jobs with higher or lower fatality risk (Gentry and Viscusi, 2016). The household and county*birth year fixed effects in my regression are particularly granular demographic characteristics which are not available in the standard studies using hedonic analysis to estimate the value of a statistical life. In recent work, in the same spirit, Aldy (2019) uses within-couple variation in occupational risk

21. The parental characteristics are mother and father's age, industry, occupation, and birthplace. These are not perfectly collinear with the household fixed effects because some households contain children with different mothers or fathers. All fixed effects, except for the occupation and industry fixed effects, are interacted with childhood census year and race.

choices to measure the value of a statistical life.

In Table 1.8.2, I list the 24 occupations in my 1940 matched sample with more than 100,000 workers. 13.5 million of the 17.8 million men (76%) in my matched sample worked in one of these occupations. I rank these jobs by my measure of occupational safety (α_O). The safest occupation is teaching, which increased life expectancy by 1.45 years relative to the group of people with no listed occupation. The riskiest is mining, which decreased life expectancy by 0.99 years relative to teaching. In Table 1.8.2, I also present two external measures of occupational fatality risk. First, I present a measure of age-adjusted mortality risk for these occupations using occupations reported on death certificates for men in 1950 (Guralnick, 1963). Second, I present the 2017 occupational fatality rate for the most similar occupation listed in the summary tables of the 2017 Census of Fatal Occupational Injuries (CFOI). The correlation between my measure of each occupation's causal effect on life expectancy and the 1950 measure of excess deaths is -0.61, and this variation is mainly driven by the extremes: both my life expectancy measure and the 1950 excess mortality measure identify teaching as a safe occupation and laborers, mine operatives, and painters as quite unsafe occupations.

For each respondent i in the 1940 matched sample, I now have an estimate of the value-added life expectancy of the industry and occupation they chose to work in. That measure is the sum of the industry fixed effect and occupation fixed effect ($\alpha_O(i) + \delta_I(i)$) from Equation 4. In Table 1.16, I regress this measure of occupation and industry life expectancy for each respondent on my measure of exposure to a Carnegie grant, removing any measure of occupational safety in the top or bottom 1% of the distribution so that I can increase precision by avoiding outliers due to noisy estimates of $\alpha_O(i)$ and $\delta_I(i)$. After adding neighborhood controls, access to a Carnegie grant moved people into occupations and industries that were safer by between 0.006 years (two days) and 0.027 years (ten days). Though this effect may seem small, if we interpret the estimates from Equation 4 as causal effects of occupational choice on life expectancy, an additional two days of life, summed over the millions of children who received access to a Carnegie grant at some point, implies a significant increase in total expected life years.

In the third-to-last column of Table 1.13, I use the main specification from Stephens and Yang (2014) to measure the effect of a year of educational attainment on occupational safety, instrumenting for educational attainment with CSLs. In a sample of both men and women, education had a positive effect on occupational safety, shifting people into occupations which increased life expectancy by 0.025–0.033 years (9–12 days).

1.6 Comparison of Libraries and Compulsory Schooling Laws

I can now quantitatively compare the effect of Carnegie library access and the effect of compulsory schooling laws on children. I do this by comparing two sets of coefficients: the effect of education on outcomes measured in the 1960–1980 census, instrumenting for education with CSLs; and the effect of Carnegie library exposure on those same outcomes for adults in the 1940 census. The two sets of coefficients are quite similar. In Figure 1.5, I plot the treatment effects from the Carnegie library regressions discussed above against the treatment effects from my analysis of the effect of compulsory schooling laws on men.

Twelve outcomes from these regression specifications are directly comparable: the effect of each intervention on wage income, the probability of having $\geq$ \$50 of non-wage income, occupational prestige, and the probability of working in each of nine broad occupational categories. These outcomes are not all independent, but the strong correlation is still stark. For completely different cohorts, in disjoint datasets, subject to different interventions—public library access and compulsory schooling laws—we see remarkably similar effects on children. The raw correlation between these T-statistics is 0.52, with a rank-rank correlation of 0.81. Both Carnegie library grants and compulsory schooling laws increased educational attainment and increased non-wage income by giving children the opportunity to enter safer, more entrepreneurial, and more prestigious occupations.

1.7 Conclusion

Many eminent politicians, scientists, and writers attribute their professional success to public libraries. Supreme Court Justice Sonia Sotomayor stumbled across her local public library while out shopping with her mother. Though her English language skills were poor at the time, the librarian told her about library cards and Sotomayor proceeded to use the library to develop a love of reading and excel in school.[22] And Justice Clarence Thomas similarly attributed his love of books and learning to his time as a child in a Carnegie-funded Library.[23] But economists have largely ignored the institution.

In this paper, I link the rollout of Andrew Carnegie's public library construction grants to complete count census data. I show that access to a public library increased educational attainment, had no effect on wage income, and increased non-wage income. Children exposed to Carnegie grants shifted into safer, more entrepreneurial, and more prestigious occupations. The same patterns explain the puzzling results of Stephens and Yang (2014), who showed that compulsory schooling laws increased educational attainment and had no effect on children's wage income. Like Carnegie's library construction grants, compulsory schooling laws increased educational attainment and increased non-wage income by shifting people into safer, more entrepreneurial, and

22. Sotomayor, in a 2018 interview with David Axelrod, said: "I needed to escape from home and I fortuitously found the local library... The local library was in a shopping center [near to where I lived]... It was on the second floor of a building that housed the Macy's... I went to that Macy's [with my mom]... I saw the library... [My mom] walked in with me and I asked what it was. The [librarian] came to us and told us you could borrow books. My mom asked how and [the librarian] said: get a library card... I started to read... it became my escape... [I only excelled in school] after I found reading... I was a marginal student in school... I was having difficulty understanding English... It took a very long time for me to understand the different meanings and usages of words that sound the same..."

23. "U.S. Supreme Court Justice Clarence Thomas and Pulitzer Prize winning author James Alan McPherson both spent their time engrossed in the stacks of books [at the Carnegie Library]... When Thomas was growing up, he spent most of his free time in the Carnegie Library, on the black side of town. It wasn't until he was a teenager that integration gave him access to the Big Library, as he called it, but once access was granted Thomas took advantage of it... When he was a kid, Thomas told them, the library was how he expanded his world, using books to visit places that were beyond his reach" (The Washington Post, "Supreme Discomfort," by Kevin Merida and Michael A. Fletcher. August 4, 2002).

more prestigious occupations. Economists often use wage income to measure the returns to education and the causal effects of local institutions on children, but wage income is only one component of the returns to investment in human capital: non-wage income, occupational prestige, and occupational safety are other important outcomes to consider and in this paper I show that both Carnegie-funded public libraries and compulsory schooling laws had positive effects on adult outcomes that operated through these measures.

1.8 Data Appendix

1.8.1 Decennial Census Data

Below, I describe the construction of each variable I use from 1900–1940 Census microdata. While the data are accessed on an NBER server, they are in the process of being cleaned and standardized by IPUMS (Ruggles et al., 2020), so many of the variables have the same format as the variables in public IPUMS census microdata.

- Race: Constructed from the RACE variable. I rely on race as it was reported in the childhood census year. I treat anyone enumerated as "Mulatto" as black.
- Years of educational attainment: Constructed from the EDUC variable, represents level of educational attainment in 1940. I map adults who have not progressed past kindergarten and those without any schooling to 0. First grade is 1 year, second grade is 2 years,..., and a 4-year degree is 16. Lastly, I map 5+ years of college to 18.
- High school graduation: Constructed from the EDUC variable, is an indicator equal to 1 if respondent was a high school graduate or had attended some college in 1940.
- College attendance: Constructed from the EDUC variable, is an indicator equal to 1 if respondent attended at least one year of college by 1940.
- Occupational prestige: Constructed from the PRESGL variable. Assigns an occupational standing score (based on Siegel, 1971) to each occupation in the 1940 census microdata.
- Weeks worked: Constructed from the WKSWORK1 variable. Reports the number of weeks worked, ranging from 0 to 52, for respondents in the 1940 census.
- Hours worked: Constructed from the HRSWORK1 variable. Reports the number of hours worked during the week of March 24-30, 1940. Includes unpaid family work (eg. on a farm). Topcoded at 98 hours.

- Enumeration district: Constructed from the ENUMDIST variable. Not comparable across years.

- City name: Constructed by extensively cleaning and manipulating these variables:

  1. 1900: us1900m_0045 (ward), us1900m_0052 (city)
  2. 1910: us1910m_0052 (city) us1910m_0053 (city) us1910m_0063 (ward)
  3. 1920: us1920c_0057 (city) us1920c_0058 (city) us1920c_0068 (city) us1920c_0069 (ward) stdmcd (city)
  4. 1930: stdmcd (city)

  I remove common prefixes (eg. precinct) and suffixes (eg. district). I trim white space, remove special characters, standardize words like "mount" and "mt.", and enforce that respondents in each enumeration district must map to the same city. When this does not happen, I replace city name with the modal city name reported within that enumeration district. In some cases, respondents in enumeration districts will consistently be linked to multiple city names. For example, respondents from Brooklyn, New York would report living in Brooklyn and New York City. In these cases, I use both city names in an attempt to link respondents to Carnegie grants and public libraries.

- Occupation: Constructed from the OCC1950 variable. Represents Ruggles et al. (2020) work to standardize reported occupations across census microdata. I use the 200 granular occupations to estimate Equation 4. And I aggregate the granular occupations into nine broader groups based on the classification reported here, for use in the linear probability models in Tables 1.14 and 1.15: `usa.ipums.org/usa-action/variables/occ1950#codes_section`.

- Industry: Constructed from the IND1950 variable. Represents Ruggles et al. (2020) work to standardize reported industries across census microdata.

- Birthplace: Constructed from the BPL variable, standardized by Ruggles et al. (2020).
- Wage: Constructed from the INCWAGE variable. Topcoded at $5,001 dollars in 1940.
- Probability $\geq$ $50 of Non-Wage Income: Constructed from INCNONWG in 1940.
- Non-wage income: Available in the 1950 census, but not the 1940 census. constructed as the difference between the INCTOT variable and the INCWAGE variable. I use this measure to impute non-wage income in 1940, as described in the text.

1.8.2 High School Data

I collect data about the founding dates of all high schools in the United States from 1890–1951 using four sources:

1. Annual and Biennial Bureau of Education reports from 1890–1940 listing the number of high school students and teachers in towns and cities with population greater than a threshold level of 2,500, 4,000, 5,000, or 10,000 depending on the year.
2. Censuses of all high schools in the United States, collected by the Bureau of Education every 1-2 years from 1890–1905 and in 1912. After 1912, the next census of high schools was published in 1951. These censuses contain information about the number of students, the number of teachers, the length of study, and the founding year of each high school; the founding year was only collected in a few of the censuses, including in 1903.
3. Lists of all accredited high schools in the United States, published by the Bureau of Education every 2-6 years from 1915 through 1944. Accreditation standards varied by state, and public colleges used these published bulletins to offer admission to local students with a diploma from a high school that met a set of criteria. These criteria often included (1) a requirement that the high school offer four years of study, and (2) that the high school offer at least a minimum number of math, English, and history credits. This source only contains

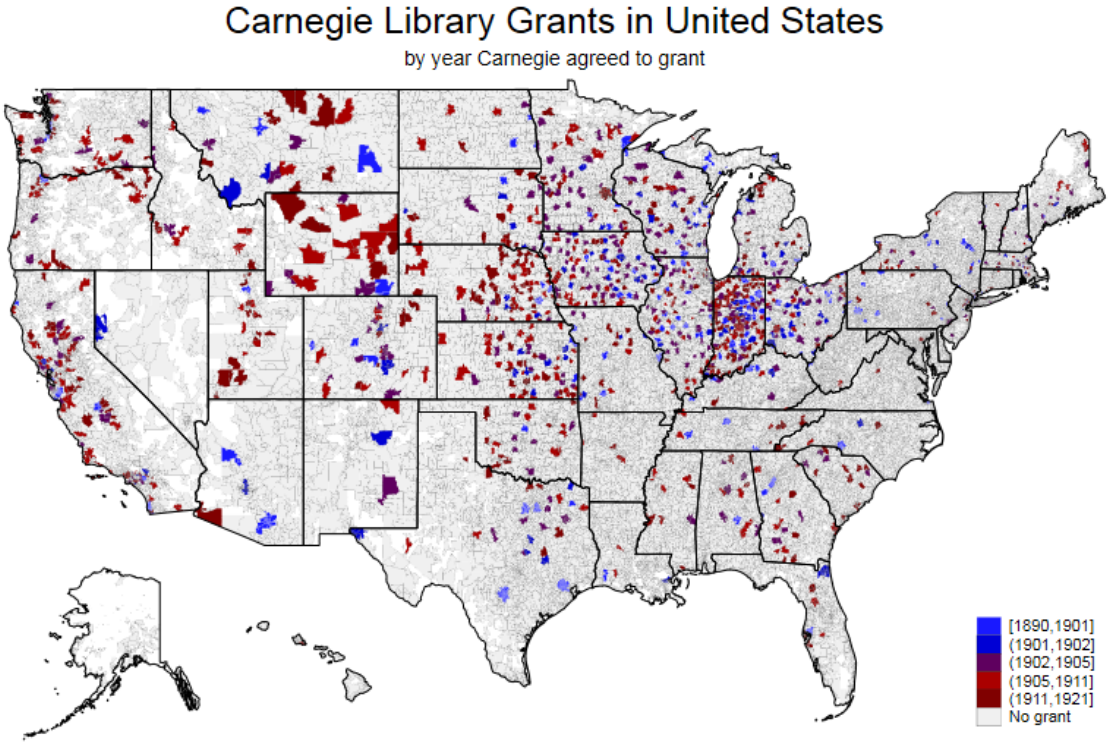
the name of each high school, the accrediting body, and the location of the school. There is no information about enrollment or school size in these lists of accredited high schools.

4. Patterson's College and School directory, available from 1904-1924 annually. These directories list the cities and towns that had high schools in each year, with almost no additional information beyond the location and the name of each principal.

I used a grant from the Institute of Education Sciences to digitize and standardize the documents described above. The data come from around 8,000 pages of tables that I scanned or extracted from hundreds of books that several university libraries have recently scanned and made available online. I worked with a group of data-enterers to transcribe these tables. I then standardized and validated the transcriptions, linking city and town names across years and sources.

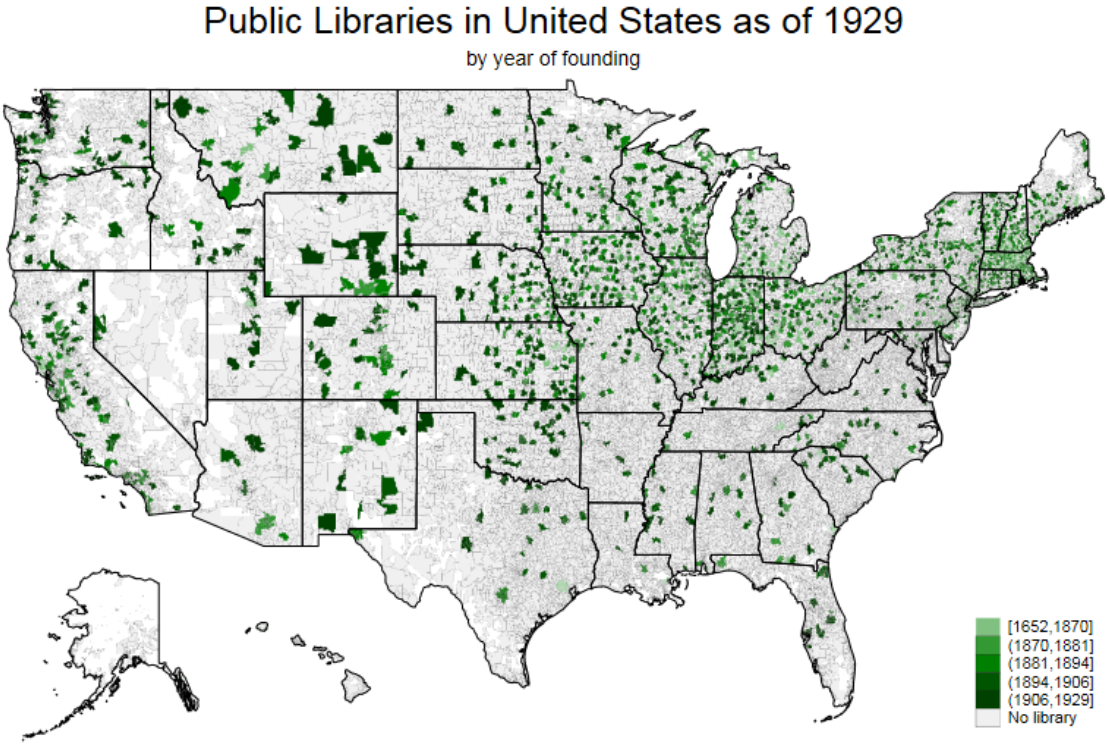
For each of the 23,000 cities and towns listed in these tables, I identify the first year when they had a public public high school. I match 76% of these cities and towns with at least one high school before 1940 to places in the 1900–1930 complete count censuses. In Figure 1.4, I plot the founding date of the earliest high school in each modern-day town in the United States, as of 1944.

Figure 1.1:



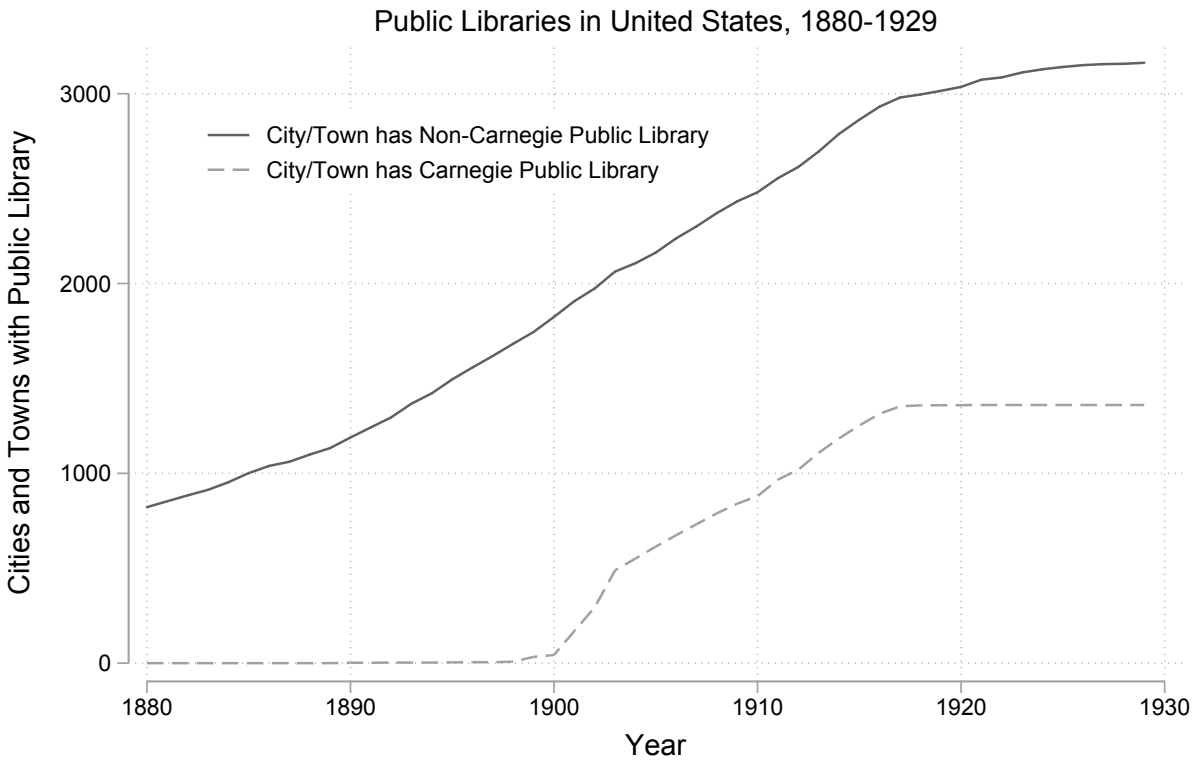
Note: I match Carnegie grants to city names in Census microdata. I then use a modern-day mapping of cities to zip codes to identify the zip codes associated with each Carnegie grant. Color indicates the year that Carnegie agreed to the grant. There is no complete list of Carnegie library opening dates, but anecdotal evidence suggests that the vast majority of communities receiving a grant constructed their library within four years of the grant announcement. Map constructed using the spmap package (Pisati 2018).

Figure 1.2:



Note: I match public libraries to city names in Census microdata. I then use a modern-day mapping of cities to zip codes to identify the zip codes associated with each public library. Color indicates the year that the public library was founded. Data come from the public library censuses produced by the Bureau of Education between 1875 and 1929. See Data section for more details. Map constructed using the spmap package (Pisati 2018).

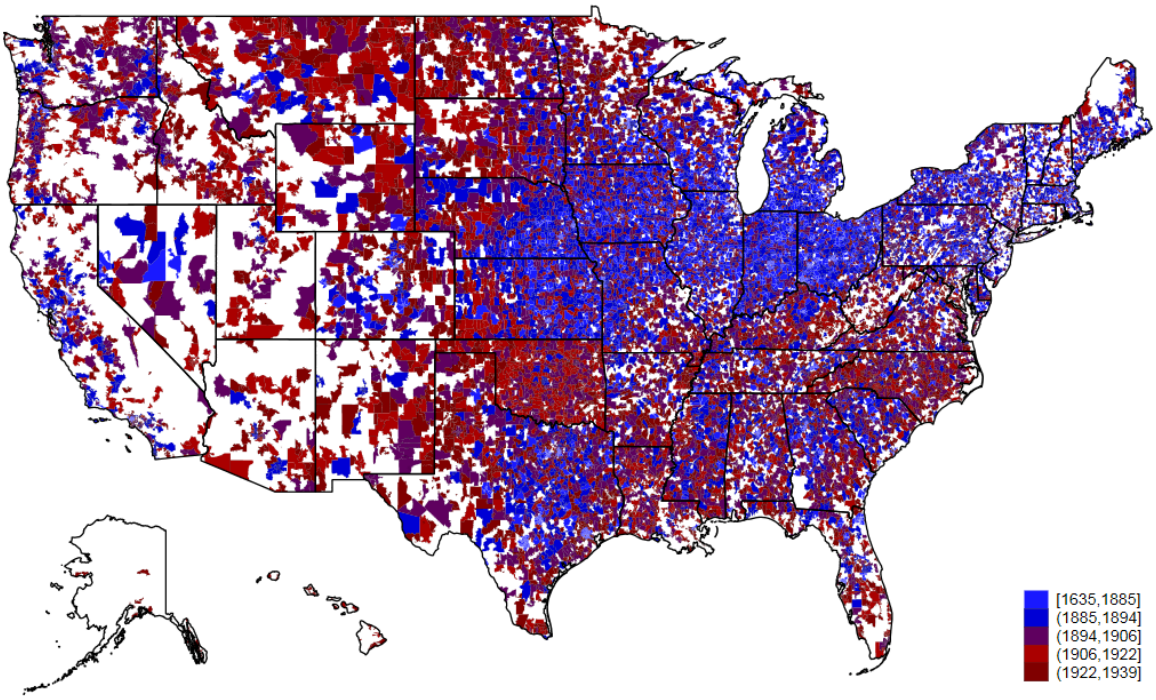
Figure 1.3:



Note: I match public libraries and Carnegie grants to city names in Census microdata. I then use a modern-day mapping of cities to zip codes to identify the zip codes associated with each public library and each Carnegie grant. This figure shows the number of cities and towns with a non-Carnegie grant as of each year and the number of cities and towns with a Carnegie grant promised as of each year. Carnegie grant data was originally collected by Bobinski (1969) and public library data come from the public library censuses produced by the Bureau of Education between 1875 and 1929. See Data section for more details.

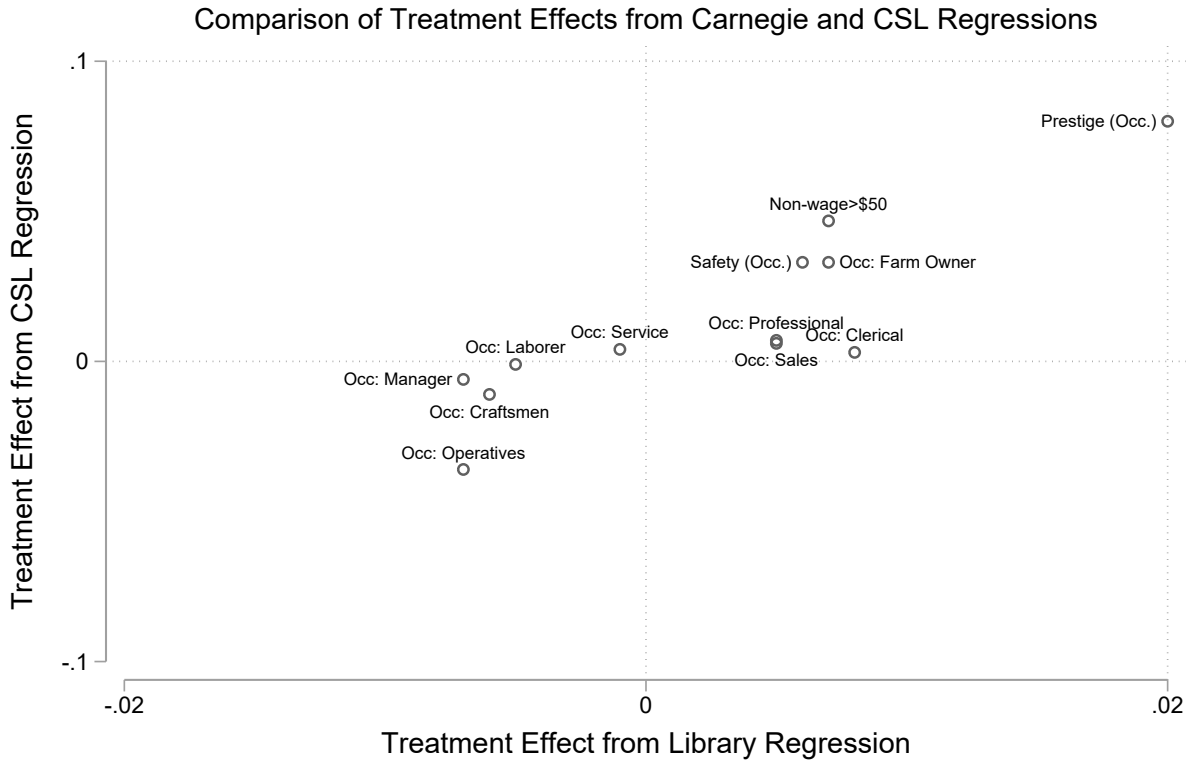
Figure 1.4:

High School Entry by 1944



Note: I match all public high schools from the public high school censuses described in the Data section to city names in Census microdata. I then use a modern-day mapping of cities to zip codes to identify the zip codes associated with each public high school. Color indicates the year that the first public high school was founded in each town. Map constructed using the spmap package (Pisati 2018).

Figure 1.5:



Notes:

1. Treatment effects from Compulsory Schooling Law (CSL) regressions are drawn from the models presented in Tables 1.13 and 1.15.

2. Treatment effects from the library regressions are from the within-family models in Tables 1.3, 1.8.2–1.9, and 1.14.

3. Data for the CSL regressions are the publicly available 1960–1980 IPUMS long-form census samples.

4. Data for the library regressions are the restricted 1900–1940 complete count census data, available through NBER.

5. Standard errors for the CSL regressions are clustered at the state of birth-by-year of birth level using the methods described in Stephens and Yang (2014).

6. Standard errors for the library regressions are clustered at the childhood county of residence level.

7. I suppress the wage income effect to ensure that the other coefficients can be easily viewed on the same figure.

Table 1.1: Comparison of All 1940 Men and Matched Sample

	All 1940	Matched Samples		All Matched, by Age of Carnegie Access			
Variable		All	Siblings	By Age 5	Age 6-15	Age>15	Never
N	39,175,147	17,818,121	5,844,571	3,171,477	912,487	730,274	13,003,883
Age in Childhood Census	-	10.37	9.65	9.01	12.78	17.93	10.12
Age in 1940	39.20	38.65	39.38	31.75	45.51	54.45	38.99
Fraction White	0.91	0.91	0.93	0.95	0.95	0.95	0.90
Fraction Black	0.09	0.09	0.07	0.05	0.05	0.05	0.10
Born in U.S.	0.87	0.94	0.94	0.93	0.84	0.84	0.95
Years of Education	8.66	8.95	8.79	10.16	9.01	8.57	8.68
High School Graduation Rate	0.26	0.28	0.25	0.40	0.27	0.23	0.25
College Attendance Rate	0.11	0.12	0.10	0.17	0.13	0.11	0.10
Prestige	34.43	34.91	34.85	35.37	37.81	37.57	34.45
Weeks Worked	38.20	39.16	39.56	38.92	40.59	37.26	39.22
Hours Worked	33.78	34.73	35.22	34.24	35.13	31.94	34.97
Exposure to Carnegie Grant	-	0.18	0.14	1	0	0	0
Exposure to Public Library	-	0.38	0.34	0.93	0.77	0.61	0.21
Occupation: Professional/Technical	0.06	0.06	0.05	0.08	0.08	0.07	0.05
Occupation: Farm owner	0.13	0.14	0.15	0.04	0.06	0.10	0.16
Occupation: Manager	0.09	0.09	0.09	0.09	0.15	0.15	0.08
Occupation: Clerical	0.06	0.07	0.06	0.11	0.08	0.07	0.06
Occupation: Sales	0.06	0.07	0.06	0.11	0.08	0.07	0.06
Occupation: Craftsman	0.16	0.16	0.16	0.17	0.21	0.20	0.15
Occupation: Operative	0.18	0.18	0.18	0.21	0.16	0.13	0.17
Occupation: Service	0.06	0.05	0.05	0.06	0.06	0.07	0.05
Occupation: Laborer	0.19	0.19	0.19	0.15	0.11	0.13	0.20
Wage	851	867	854	1,016	1,221	1,012	797
Wage (median)	600	600	600	900	1,000	644	520
Probability $\geq$ \$50 Non-Wage Income	0.31	0.31	0.33	0.21	0.33	0.40	0.33

Note: See Data Appendix for definition of each variable and description of processing.

Table 1.2: Fraction of Matched Sample in Largest Cities and Towns

Place Name	State	Count	% of Sample	Cumulative Percentage	Carnegie Grant	Public Library
Chicago	IL	395,768	2.2%	2.3%	1902	1872
Manhattan	NY	303,022	1.7%	3.9%	1899	1652
Philadelphia	PA	237,505	1.3%	5.3%	1903	1820
Brooklyn	NY	227,985	1.3%	6.5%	1899	1823
Detroit	MI	134,446	0.8%	7.3%	1901	1865
St. Louis	MO	121,565	0.7%	8.0%	1901	1860
Cleveland	OH	115,359	0.6%	8.6%	1903	1868
Baltimore	MD	108,224	0.6%	9.2%	1906	1874
Boston	MA	107,998	0.6%	9.8%		1848
Milwaukee	WI	82,650	0.5%	10.2%		1847
Buffalo	NY	79,618	0.4%	10.7%		1836
Pittsburgh	PA	73,498	0.4%	11.1%	1890	1872
Los Angeles	CA	70,593	0.4%	11.6%	1911	1872
Newark	NJ	66,317	0.4%	11.9%		1888
San Francisco	CA	63,506	0.5%	12.3%	1901	1866
Cincinnati	OH	63,474	0.4%	12.6%	1902	1802
New Orleans	LA	60,568	0.3%	13.0%	1902	1843
Bronx	NY	57,004	0.3%	13.3%	1899	1652
Minneapolis	MN	53,921	0.3%	13.6%	1912	1885
Jersey City	NJ	53,201	0.3%	13.9%		1889
Washington	DC	49,314	0.3%	14.2%	1899	1865
Providence	RI	44,884	0.3%	14.4%		1874
Rochester	NY	41,978	0.2%	14.7%		1912
Indianapolis	IN	39,280	0.2%	14.9%	1909	1831
Louisville	KY	39,104	0.2%	15.1%	1899	1871

Note: See Data Appendix for definition of each variable and description of processing.

Table 1.3: Effect of Carnegie Grant on Years of Schooling

	1	**2**	**3**	**4**	**5**	**6**
1(Carnegie Grant by Age 5)	0.572	0.179	0.076	0.090	0.134	0.101
	(0.034)	(0.019)	(0.018)	(0.009)	(0.010)	(0.017)
Child controls	Yes	Yes	Yes	Yes	Yes	Yes
Enumeration District (FEs)		Yes	Yes	Yes	Yes	Yes
Parent controls			Yes	Yes	Yes	Yes
County-by-Birth Year (FEs)				Yes	Yes	Yes
Census Microfilm Page (FEs)					Yes	
Household (FEs)						Yes
Observations	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550
R^2	0.20	0.26	0.29	0.32	0.55	0.73
Mean dep. var.	8.99	8.99	8.99	9.01	9.00	8.88

1. Data are the restricted 1900–1940 complete count census data, available through NBER.
2. Standard errors clustered at the childhood county of residence level.
3. Child controls are birthplace, birth order, and birth year. All are interacted with race (white, black, and Native American)-by-childhood census year (indicators for whether I see each 1940 adult in the 1900, 1910, 1920, or 1930 census).
4. Enumeration district fixed effects are enumeration district indicators (roughly 65,000 in each census year) interacted with race-by-childhood census year. Enumeration districts are occasionally not unique within a county. In these rare cases, the fixed effect combines individuals in potentially geographically discontinuous neighborhoods.
5. Parent controls are mother's birthplace, father's birthplace, mother's age, father's age, mother's occupation, father's occupation, mother's industry, and father's industry. All are interacted with race-by-childhood census year fixed effects. Missing values are coded as their own category.
6. County-by-birth year fixed effects include county fixed effects interacted with birth year fixed effects, interacted with race-by-childhood census year fixed effects.
7. Census microfilm page numbers are numbered in the data, but occasionally repeat within microfilm reel. In those cases, my Census microfilm page fixed effects will group those pages together.
8. There are 17.8 million observations in my analysis sample, but only 17.5 million have valid educational attainment information, so the model in column 1 contains 17.5 million records (after dropping singletons).

Table 1.4: Effect of Carnegie Grant on High School Graduation Rate

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.064	0.029	0.017	0.015	0.020	0.014
by Age 5)	(0.004)	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550
R^2	0.10	0.15	0.19	0.21	0.47	0.68
Mean dep. var.	0.28	0.28	0.28	0.29	0.28	0.26

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.5: Effect of Carnegie Grant on College Attendance

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.043	0.024	0.017	0.018	0.021	0.015
by Age 5)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550
R^2	0.03	0.08	0.12	0.14	0.42	0.64
Mean dep. var.	0.12	0.12	0.12	0.12	0.12	0.11

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.6: IV Effect of Carnegie Grant on Years of Schooling

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.75	0.33	0.14	0.17	0.27	0.24
by Age 5)	(0.05)	(0.05)	(0.04)	(0.02)	(0.02)	(0.04)
First stage:						
Coefficient	0.77	0.54	0.53	0.53	0.51	0.43
F-statistic	756	272	266	508	462	352
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
N	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.7: Effect of Carnegie Grant on Log Annual Wage Income

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.381	-0.022	-0.058	-0.045	-0.044	-0.003
	(0.032)	(0.013)	(0.011)	(0.008)	(0.009)	(0.018)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	16,766,377	16,749,418	16,748,204	16,618,602	14,182,462	5,027,116
R^2	0.07	0.11	0.12	0.15	0.41	0.61
Mean dep. var.	4.91	4.91	4.91	4.91	4.90	4.87

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.8: Effect of Carnegie Grant on Probability $\geq$ 50 Dollars of Non-Wage Income

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	-0.049	-0.001	0.004	0.004	0.005	0.007
	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,240,954	17,223,995	17,222,752	17,094,494	14,687,385	5,292,713
R^2	0.07	0.10	0.11	0.14	0.40	0.60
Mean dep. var.	0.31	0.31	0.31	0.31	0.32	0.33

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.9: Effect of Carnegie Grant on Log Imputed Total Income from 1950 Data

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.166	0.029	0.008	-0.004	-0.002	0.025
by Age 5)	(0.013)	(0.004)	(0.004)	(0.003)	(0.004)	(0.011)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,342,421	17,325,555	17,324,336	17,196,140	14,793,112	5,339,327
R^2	0.09	0.12	0.13	0.16	0.42	0.60
Mean dep. var.	7.21	7.21	7.21	7.21	7.24	7.25

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.10: Effect of Carnegie Grant on Log Imputed Wage Income from 1950 Data

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.425	0.006	-0.038	-0.037	-0.033	0.016
by Age 5)	(0.035)	(0.015)	(0.012)	(0.007)	(0.008)	(0.018)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,342,421	17,325,555	17,324,336	17,196,140	14,793,112	5,339,327
R^2	0.07	0.11	0.11	0.15	0.41	0.61
Mean dep. var.	5.85	5.85	5.85	5.86	5.85	5.81

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.11: Effect of Carnegie Grant on Log Imputed Non-Wage Income from 1950 Data

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	-0.296	0.023	0.048	0.041	0.040	0.039
by Age 5)	(0.025)	(0.013)	(0.011)	(0.007)	(0.008)	(0.019)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,342,421	17,325,555	17,324,336	17,196,140	14,793,112	5,339,327
R^2	0.07	0.10	0.11	0.14	0.40	0.60
Mean dep. var.	3.21	3.21	3.21	3.21	3.26	3.32

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.12: Effect of Carnegie Grant on Occupational Prestige

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.126	0.050	0.030	0.021	0.027	0.020
by Age 5)	(0.007)	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	16,127,337	16,110,400	16,109,139	15,978,774	13,517,806	4,746,742
R^2	0.08	0.12	0.14	0.17	0.43	0.63
Mean dep. var.	0.00	0.00	0.00	0.00	0.02	0.01

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.
3. Occupational prestige is standardized so that the average in the full matched sample is zero and the standard deviation is 1.

Table 1.13: Effect of Education on Various Outcomes (Instrumented with Compulsory Schooling)

	Weekly Wage	**Annual Wage**	**Annual Non-Wage**	**Annual Non-Wage**$\geq$ \$275	**Annual Total**	**Occ. Prestige**	**Occ. Safety**	**Housing Price**
Panel A: Men and Women								
Years of Education	-0.003	-0.09	0.33	0.03	0.07	0.09	0.025	0.089
	(0.016)	(0.07)	(0.06)	(0.01)	(0.06)	(0.02)	(0.011)	(0.019)
First stage F-statistic:	39	.	.	.	.	.	.	
N	3,680,223	5,066,060	.	.	.	.	3,021,522	
Average in Sample	5.43	6.68	2.47	0.27	7.60	0.00	0.14	10.70
Panel B: Men								
Years of Education	-0.014	-0.24	0.52	0.05	0.04	0.08	0.033	0.099
	(0.021)	(0.10)	(0.12)	(0.01)	(0.05)	(0.03)	(0.015)	(0.025)
First stage F-statistic:	21	.	.	.	.	.	.	.
N	2,166,387	2,502,089	.	.	.	.	.	1,467,960
Average in Sample	5.80	8.26	3.20	0.34	9.34	0.00	0.06	10.71
State of Birth (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year of Birth (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region*Birth Year (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age^4, census year, gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

1. See notes to Table 1.3.
2. Column 1 replicates the 2SLS estimates from Columns 4 and 6 of Table 1 in Stephens and Yang (2014) by subsetting to the sample of respondents with non-zero wage income.
3. In Columns 2-6, respondents with zero reported income are treated as zeroes and I log annual income measures.
4. In Column 7, the universe is the set of respondents who report owning a house.
5. Standard errors clustered at the state of birth-by-year of birth level.
6. Occupational prestige is standardized so that the average in the sample is zero and the standard deviation is 1.

Table 1.14: Effect of Carnegie Grant on Occupational Choice

	Prof. and Tech.	**Farm Owner**	**Manager**	**Clerical**	**Sales**	**Craftsmen**	**Operative**	**Service**	**Labor**
1(Carnegie Grant by Age 5)	0.005	0.007	-0.007	0.008	0.005	-0.006	-0.007	-0.001	-0.005
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Child controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Enumeration District (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Parent controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County-by-Birth Year (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Household (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
% in matched sibling sample	0.06	0.13	0.09	0.06	0.06	0.16	0.18	0.06	0.19
N	5,590,275	same	same	same	same	same	same	same	same
R^2	0.56	0.66	0.55	0.54	0.54	0.55	0.57	0.56	0.60

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.
3. Occupations defined using the major categories from the IPUMS occ1950 variable. I classify a small number of farm managers as managers, so that the farm owner category contains only farm owners and tenant-farmers. A complete mapping of granular occupations to these 9 broader occupations can be found here: `https://usa.ipums.org/usa-action/variables/occ1950#codes_section`
4. These regressions are subset to individuals reporting a valid occupation.

Table 1.15: Effect of Education on Occupational Choice (Instrumented with Compulsory Schooling)

	Prof. and Tech.	**Farm Owner**	**Manager**	**Clerical**	**Sales**	**Craftsmen**	**Operative**	**Service**	**Labor**
Panel A: Men and Women									
Years of Education	0.018	0.021	0.001	-0.001	0.005	-0.012	-0.042	0.004	0.005
	(0.006)	(0.004)	(0.005)	(0.008)	(0.005)	(0.006)	(0.009)	(0.005)	(0.004)
First stage F-statistic:	39	.	.	.	.	.	.	.	
N	4,374,405	.	.	.	.	.	.	.	.
Fraction in Sample	0.186	0.014	0.109	0.190	0.073	0.142	0.150	0.097	0.038
Panel B: Men									
Years of Education	0.007	0.033	-0.006	0.003	0.006	-0.011	-0.036	0.004	-0.001
	(0.009)	(0.008)	(0.009)	(0.007)	(0.006)	(0.011)	(0.011)	(0.006)	(0.006)
First stage F-statistic:	21	.	.	.	.	.	.	.	
N	2,442,498	.	.	.	.	.	.	.	.
Fraction in Sample	0.178	0.023	0.148	0.062	0.070	0.237	0.171	0.057	0.055
State of Birth (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year of Birth (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region*Birth Year (FEs)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age^4, census year, gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

1. See notes to Table 1.13.
2. Standard errors clustered at the state of birth-by-year of birth level.
3. Models are linear probability models where the outcome in each column is an indicator for whether a respondent was in that occupation. Sample restricted to those with valid occupations.

Effect of Occupation on Life Expectancy, Conditional on Reaching Age 65+

Occupation	Broad Category	Count in Matched Sample	Effect on Life Expectancy	1950 Excess Mortality	2017 Mortality Rate
Teachers	Professional	142,231	1.45	58	0.4
Farm owner	Farm owner	2,251,993	0.87	96	24.0
Prof. and tech. (general)	Professional	100,467	0.46	85	3.0
Electricians	Craftsmen	106,774	0.45	113	8.3
Janitors and sextons	Service	126,975	0.44	117	2.8
Insurance agents and brokers	Sales	107,309	0.15	96	0.5
Farm laborers, unpaid family	Laborers	797,318	0.13	96	16.4
Farm laborers, wage workers	Laborers	260,365	0.12	96	16.4
Mechanics (automobile)	Craftsmen	199,788	0.12	96	7.2
Manager, and proprietors	Managers	1,273,996	0.07	86	1.1
Bookkeepers	Clerical	212,975	0.03	94	0.6
No listed occupation	—	1,184,701	0.00		
Carpenters	Craftsmen	326,215	-0.00	85	7.5
Clerical (general)	Clerical	599,441	-0.08	83	0.6
Salesmen (general)	Sales	803,248	-0.09	86	1.6
Foremen	Craftsmen	261,173	-0.16	96	1.0
Mechanics (general)	Craftsmen	190,796	-0.16	78	8.1
Operatives (general)	Operatives	1,225,092	-0.32	97	9.4
Machinists	Craftsmen	225,744	-0.39	120	5.4
Deliverymen	Operatives	167,220	-0.42	107	26.8
Truck drivers	Operatives	614,081	-0.43	107	26.8
Laborers (general)	Laborers	1,827,286	-0.60	169	9.7
Miners and laborers	Operatives	327,067	-0.64	173	11.7
Painters annd construction	Craftsmen	182,390	-0.99	130	14.2
Total		13,514,642			

1. This table only includes occupations with more than 100,000 workers in my 1940 matched sample.
2. Broad Category defined using the major categories from the IPUMS occ1950 variable. A complete mapping of granular occupations to these 9 broader occupations can be found here: `https://usa.ipums.org/usa-action/variables/occ1950#codes_section`
3. Effect on Life Expectancy is the value of the occupation fixed effect from the regression in Equation 4.
4. 1950 Excess Mortality is defined in Guralnick (1963) as the ratio of the number of deaths from death certificates indicating a "usual occupation" equal to the given occupation relative to the expected number of deaths for people of that age. So for example, teachers have a 1950 Excess Mortality of 58, indicating that only 58 teachers died relative to the 100 who you would expect to have died given the age distribution of teachers who died.
5. 2017 data from the "Hours-based fatal injury rates by industry, occupation, and selected demographic characteristics, 2017" table of the Census of Fatal Occupational Injuries (CFOI), from: `https://www.bls.gov/iif/oshcfoi1.htm#2017`. 2017 Mortality Rate is number of on-the-job deaths per 100,000 full-time equivalent workers. I include it in the table when I can find a comparable occupation in the 2017 CFOI data for a given occupation from 1940.

Table 1.16: Projected Effect of Carnegie Grant on Life Expectancy, Conditional on Reaching Age 65

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.009	0.027	0.025	0.017	0.019	0.006
by Age 5)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,448,037	17,431,086	17,429,864	17,302,197	14,905,019	5,384,479
R^2	0.03	0.05	0.06	0.09	0.37	0.57
Mean dep. var.	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.
3. For each respondent, the outcome is the sum of the industry and occupation coefficients from Equation 4.

Appendix Tables

Table 1.17: Effect of Carnegie Grant on Educational Attainment, State Cluster

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.572	0.179	0.076	0.090	0.134	0.101
by Age 5)	(0.061)	(0.024)	(0.021)	(0.008)	(0.007)	(0.016)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550
R^2	0.20	0.26	0.29	0.32	0.55	0.73
Mean dep. var.	8.99	8.99	8.99	9.01	9.00	8.88

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood state of residence level.

Table 1.18: Effect of Carnegie Grant on Educational Attainment, only Siblings

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.621	0.171	0.099	0.098	0.121	0.101
by Age 5)	(0.034)	(0.019)	(0.018)	(0.014)	(0.015)	(0.017)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	5,716,477	5,713,985	5,713,550	5,569,796	5,421,502	5,375,550
R^2	0.17	0.27	0.30	0.35	0.66	0.73
Mean dep. var.	8.83	8.83	8.83	8.86	8.88	8.88

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.19: Effect of Carnegie Grant on Educational Attainment, Continuous Measure

	1	**2**	**3**	**4**	**5**	**6**
%Exposure	0.626	0.273	0.098	0.136	0.230	0.199
Age 5–20	(0.035)	(0.033)	(0.030)	(0.014)	(0.015)	(0.027)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,412,080	17,395,212	17,393,991	17,265,943	14,867,729	5,375,550
R^2	0.20	0.26	0.29	0.32	0.55	0.73
Mean dep. var.	8.99	8.99	8.99	9.01	9.00	8.88

1. See notes to Table 1.3.
2. Treatment is the fraction of years age 5–20 with access to a Carnegie grant.
3. Standard errors clustered at the childhood county of residence level.

Table 1.20: Effect of Carnegie Grant on Educational Attainment, Midwest

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.507	0.196	0.111	0.104	0.149	0.073
by Age 5)	(0.048)	(0.024)	(0.022)	(0.013)	(0.013)	(0.023)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	6,153,830	6,148,172	6,147,000	6,109,797	5,463,382	2,071,528
R^2	0.12	0.19	0.23	0.26	0.49	0.71
Mean dep. var.	9.54	9.54	9.54	9.55	9.49	9.28

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.21: Effect of Carnegie Grant on Educational Attainment, Whites

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.559	0.171	0.071	0.088	0.133	0.104
by Age 5)	(0.033)	(0.019)	(0.018)	(0.010)	(0.010)	(0.017)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	15,878,910	15,878,220	15,877,839	15,814,796	13,874,495	5,064,397
R^2	0.15	0.21	0.25	0.28	0.52	0.72
Mean dep. var.	9.28	9.28	9.28	9.28	9.23	9.06

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Effect of Carnegie Grant on Educational Attainment, Only Pre-Exposure

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.619	0.125	0.084	0.076	0.116	0.102
by Age 5)	(0.043)	(0.021)	(0.019)	(0.013)	(0.015)	(0.025)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	13,825,663	13,813,393	13,812,192	13,693,528	11,998,357	4,451,156
R^2	0.19	0.25	0.28	0.31	0.54	0.73
Mean dep. var.	8.76	8.76	8.76	8.77	8.79	8.69

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.22: Effect of Carnegie Grant on Educational Attainment, High School Control

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.453	0.167	0.070	0.089	0.130	0.097
by Age 5)	(0.030)	(0.018)	(0.017)	(0.009)	(0.010)	(0.017)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Age at HS Ent.	Yes	Yes	Yes			
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr*				Yes	Yes	Yes
Age at HS Ent.						
Census Page					Yes	
Household						Yes
Observations	17,411,951	17,395,059	17,393,831	17,265,883	14,867,646	5,375,276
R^2	0.20	0.26	0.29	0.32	0.55	0.73
Mean dep. var.	8.99	8.99	8.99	9.01	9.00	8.88

1. See notes to Table 1.3.
2. Age at HS Entry indicates the age of each child when a high school opens in their place of residence.
3. Standard errors clustered at the childhood county of residence level.

Table 1.23: Effect of College Libraries on Educational Attainment

	1	2	3	4	5	6
1(College Lib.	0.635	0.108	0.023	-0.005	0.036	0.019
by Age 5)	(0.027)	(0.029)	(0.030)	(0.026)	(0.026)	(0.048)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	17,343,845	17,327,003	17,325,785	17,197,694	14,799,176	5,349,331
R^2	0.20	0.26	0.29	0.32	0.55	0.73
Mean dep. var.	8.99	9.00	9.00	9.01	9.00	8.88

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.

Table 1.24: Effect of Carnegie Grant on Educational Attainment, Plus 20 Years

	1	2	3	4	5	6
1(Carn. Grant	0.489	0.061	-0.012	-0.006	0.014	0.053
by Age 5)	(0.030)	(0.014)	(0.015)	(0.013)	(0.019)	(0.038)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	13,117,519	13,104,271	13,103,075	12,984,218	11,222,540	4,126,383
R^2	0.19	0.25	0.28	0.32	0.54	0.73
Mean dep. var.	8.71	8.71	8.71	8.72	8.74	8.65

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.
3. Unlike in the baseline specification, where I use the entire matched sample, this placebo test subsets to children living in places that did not receive a Carnegie grant and children who received a Carnegie grant when they were 21 years old or older. This ensures that none of the children in this sample had access to a Carnegie grant before they made their human capital investment decisions.

Table 1.25: Effect of Carnegie Grant on Educational Attainment, Lagged 20 Years

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.434	0.055	-0.016	0.029	0.038	0.016
by Age 5)	(0.035)	(0.012)	(0.014)	(0.016)	(0.016)	(0.021)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr				Yes	Yes	Yes
Census Page					Yes	
Household						Yes
Observations	15,812,214	15,795,879	15,794,656	15,666,064	13,242,318	4,758,184
R^2	0.20	0.27	0.30	0.33	0.56	0.74
Mean dep. var.	9.01	9.01	9.01	9.02	9.00	8.87

1. See notes to Table 1.3.
2. Standard errors clustered at the childhood county of residence level.
3. Unlike in the baseline specification, where I use the entire matched sample, this placebo test subsets to children living in places that did not receive a Carnegie grant and children who received a Carnegie grant when before the age of 5. This ensures that all of the children in this sample had access to a Carnegie grant before they made their human capital investment decisions or never had access to a Carnegie grant.

Table 1.26: Effect of Carnegie Grant on Educational Attainment, Perfect Matches

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.659	0.103	0.072	0.077	0.119	0.060
by Age 5)	(0.037)	(0.015)	(0.014)	(0.013)	(0.016)	(0.027)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Age at HS Ent.	Yes	Yes	Yes			
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr*				Yes	Yes	Yes
Age at HS Ent.						
Census Page					Yes	
Household						Yes
Observations	9,058,899	9,041,858	9,040,595	8,904,648	6,467,682	2,248,509
R^2	0.17	0.26	0.31	0.34	0.61	0.78
Mean dep. var.	9.67	9.67	9.67	9.70	9.74	9.61

1. See notes to Table 1.3 and 1.22.
2. Standard errors clustered at the childhood county of residence level.
3. This sample consists of matches that are accurate according to the two match accuracy measures discussed in the Matching Appendix.

Table 1.27: Effect of Carnegie Grant on Educational Attainment, All Matches

	1	**2**	**3**	**4**	**5**	**6**
1(Carn. Grant	0.530	0.152	0.063	0.070	0.116	0.087
/by Age 5)	(0.031)	(0.017)	(0.015)	(0.008)	(0.009)	(0.012)
Child covs.	Yes	Yes	Yes	Yes	Yes	Yes
Age at HS Ent.	Yes	Yes	Yes			
Enum. dist.		Yes	Yes	Yes	Yes	Yes
Parent covs.			Yes	Yes	Yes	Yes
Cnty*brthyr*				Yes	Yes	Yes
Age at HS Ent.						
Census Page					Yes	
Household						Yes
Observations	21,755,034	21,738,814	21,737,626	21,620,669	19,503,367	7,850,353
R^2	0.19	0.25	0.28	0.30	0.51	0.70
Mean dep. var.	9.01	9.01	9.01	9.01	9.01	8.83

1. See notes to Table 1.3 and 1.22.
2. Standard errors clustered at the childhood county of residence level.
3. This sample consists of all adults in 1940 who match to at least one early childhood census record, independent of match quality. See the Matching Appendix for more details.

CHAPTER 2

THE BLACK-WHITE LIFETIME EARNINGS GAP

2.1 Introduction

A large literature analyzes black-white earnings gaps. For example, Bayer and Charles (2018) show that the black-white gap in median income shrank from roughly 1 log point in 1940 to 0.66 log points in 1950 before stagnating from 1950 to 2014.[1] And earlier work by Card and Krueger (1993) states that "The narrowing of the black-white earnings gap between 1960 and the mid-1970's represents one of the most significant episodes of relative progress for African Americans in U.S. history." Cross-sectional earnings gaps represent changes in the relative price of available white and black labor in a given year. But economists and policymakers often point to these trends as measures of the relative life outcomes of black and white men. Income gaps in one year are an incomplete measure of the economic outlook for black and white men in the United States because many children died before they could be surveyed as working adults. A more complete way to think about inequality between groups is to start at birth and measure the lifetime outcomes for these groups. From this starting point, mortality is quite important.

This idea is not new. Many researchers propose measures of inequality that rely on a birth cohort's adult outcomes. But in the same papers where researchers propose looking at outcomes by birth cohort, they instead calculate inequality cross-sectionally. In one of the first papers that used census microdata to analyze black-white inequality, Smith and Welch (1989) studied cross-sectional black-white gaps in educational attainment and earnings for adults in the census, claiming that "[a]mong men born in this century, there has been a substantial narrowing of racial difference in years of school completed." But Smith and Welch ignored the large number of black men with low levels of education who died before they could be enumerated in decennial censuses.

When social scientists estimate economic parameters using survey data, death acts as a form

1. See Figure 9A in Bayer and Charles (2018).

of sample selection. If unequal access to clean water killed a randomly selected 10% of black workers in the United States each year, and sharp improvements in environmental policy suddenly gave everyone access to clean water, standard measures of income inequality would not change because the average income of living black workers would not change. But the increase in life expectancy would obviously have large effects on welfare.[2] This is an exaggerated version of what actually happened in the early 1900s as black mortality rates declined. In this paper, I measure outcomes conditional on birth instead of looking at the set of people who happened to live until the age of 30.

A set of papers considers the economic value of health improvements. Murphy and Topel (2006) use a structural model of health and labor decisions to estimate that increases in life expectancy since 1900 were worth trillions of dollars to society. In related work, Murphy and Topel (2005) show that increases in life expectancy since 1968 were worth more for black men than they were for white men. And Becker et al. (2005) and Gallardo-Albarran (2019) measure cross-country inequality by combining GDP and life expectancy to construct a measure of "full income." Becker and his co-authors shows that cross-country inequality shrank significantly between 1960 and 2000 when they measure inequality using this combined 'full income' metric. Gallardo-Albarran uses a similar method to show that GDP-based measures of welfare understate cross-country improvements in living standards in the early 1900s.

In this paper, I provide the first long-run estimates of lifetime earnings by race in the United States. Researchers have used administrative datasets to calculate lifetime earnings in Germany,

2. Other economists have recognized that death is a form of sample selection when measuring welfare. For example, Kanbur and Mukherjee (2007) construct new measures of poverty that account for the fact that "There is a glaring paradox in all commonly used measures of poverty. The death of a poor person, because of poverty, reduces poverty according to these measures." And Currie (2011) argues that "A possible drawback to using data on births [to measure welfare effects of pollution abatement] is that pollution could affect the probability of a conception or of a live birth. If we suppose that pollution abatement would lead to fewer fetal deaths, and more births, and that the marginal fetus lost due to pollution is more vulnerable and less healthy than others, then focusing on births will tend to understate the beneficial effects of abatement by increasing the number of less healthy infants whose birth weight is recorded."

Norway, and the United States (Bonke et al. 2015; Bhuller et al. 2017; Tamborini et al. 2015; Guvenen et al 2017). But researchers have only separated out lifetime earnings estimates by race using synthetic cohorts from 1970-2000 (See Table 4-2 of Jena, Philipson, and Sun (2010). After I calculate lifetime earnings by race, I construct a 'full income' measure of welfare based on income, life expectancy, and a structural model of optimal consumption. I use this model to discuss black-white inequality in the 20th century.

In the following sections, I describe the census microdata that I use to estimate life expectancy and earnings measures. I use this data to calculate and compare cross-sectional black-white earnings gaps and lifetime earnings by birth cohort. I then put an economic value on the black-white life expectancy gap for each birth cohort using research from the literature on how to calculate the value of a statistical life (VSL). I propose a simple structural model of optimal consumption where newborns with different types face known exogenous sequences of mortality risk and income each year. I calibrate this model using Census data and show that declining mortality risk for black children between 1900 and 1940 halved the black-white welfare gap. Convergence then slowed, and a significant black-white welfare gap still exists; this gap is driven by persistent black-white gaps in earnings and mortality rates.

2.2 Data

2.2.1 Census Microdata

I use the 1910-2010 public-use decennial census and 2006-2014 American Community Survey microdata from IPUMS to estimate earnings gaps by race and birth cohort (Ruggles et al. 2020).[3] These data represent cross-sectional nationally representative samples of the U.S. population every ten years from 1910-2000 and annually from 2006-2014. For all the analyses described below, I

3. I do not use data from the 2001-2005 ACS because the 2001-2005 ACS do not include institutionalized individuals in their sampling frames. I also do not use the 2015 and 2016 ACS data because the Census Bureau modified the race variable in those years, complicating comparisons between pre-2015 data and 2015-2016 data.

subset to black and white men born in the U.S. between 1900 and 1970. I use sample weights throughout so that my estimates are nationally representative of the American-born black and white populations.

I use the census microdata to calculate three separate measures of annual earnings: labor earnings, post-tax labor earnings, and total earnings. I start with labor earnings from 1940-2014 and total earnings from 1950-2014. In 1940, only labor earnings are available.[4] In all years, I assume that all males have zero earnings before the age of 16, because these younger men were often not in the sampling frame of the decennial census's income questions.[5] Because no earnings measures are available before 1940, I must first calculate earnings for males before 1940 so that I can estimate lifetime earnings for the 1900-1970 birth cohorts. I begin by estimating non-labor earnings for each male in the 1940 census using data from 1950. I regress 1950 non-labor earnings at the individual level on indicators for race, age, occupation, industry, state of birth, and state of residence.[6] I then use these regression coefficients to predict non-labor earnings for each male in 1940.[7]

4. Also, in 1940 1.5% of respondents were institutionalized and these respondents were not asked to report labor income. I assume that all of these respondents have zero earnings, but that does not affect any of this paper's conclusions.

5. Males under the age of 16 had very low measurable earnings in the early 1900s. According to Table 2 of the U.S. Census of Manufacturers report on the Earnings of Wage-Earners, less than 1% of total earnings were paid to workers under the age of 16. While this evidence is from only one (large) sector of the economy, anything close to a 1% change in the total earnings of any group in my sample will not affect, in any way, the conclusions of this paper. Source: `https://catalog.hathitrust.org/Record/008433603`

6. There are roughly 300 distinct occupations and 20 distinct industries in the census microdata. I also separately regress an indicator for whether or not each person worked on those indicators and included the predicted value from that regression in the wage regression.

7. Census earnings data is topcoded, and that topcoding often effects 0-2% of all observations. In the tables I present in this paper, I do not correct for this topcoding, but it has no effect on any of my results. To confirm that this is the case, I replicate the main results from this paper by assuming that the income distribution in each year closely matched the income distribution in 2000, when topcoding had a negligible effect on reported income. For example, in 1950, which has the most topcoding of any census year, roughly 2% of labor income valaues are topcoded. So, I see that in 2000 the 98th percentile of the labor income distribution was roughly \$190,000. The average of income values in the top 2% of the 2000 income distribution is roughly \$300,000, so I multiply all topcoded income values in 1950 by $\frac{300000}{190000} = 1.6$. In 1960-1990, the topcoding 'ratio' is around

At this point, I have labor and total income measures for 1940 onward. I calculate post-tax earnings by merging federal individual personal income tax rates and personal exemption amounts from the appropriate year onto each individual in the 1940-2014 census and ACS data. I apply these tax rates and personal exemptions to each person's nominal labor earnings to estimate each person's post-tax earnings. I rely on federal income tax rates and not state or local income tax rates for two reasons. First, there is no standardized time series of state income tax rates in the early or mid 1950s.[8] Second, federal income taxes are the largest source of income tax revenue. For example, in 2015, federal, state, and local personal income tax revenue totaled 1.5 trillion dollars, 336 billion dollars, and 32 billion dollars respectively.[9]. The disparity between federal and non-federal income tax revenue was even larger in the 1900s.

I inflate all earnings variables to 2015 dollars using the CPI-U. To estimate the earnings of men pre-1940, I use the same imputation procedure that I used to impute 1940 non-labor earnings. I regress 1940 labor, total, and post-tax earnings for people in my sample on indicators for race, age, occupation, industry, state of birth, and state of residence. I use these regression coefficients to predict the amount of earned income for all men in decennial years 1910, 1920, and 1930.

I then collapse the census microdata into estimates of the average income for each age*year*race cell. I linearly interpolate these income averages for each age*race cell in years between adjacent census years.[10] For example, I take the average income of white 16 year olds in 1940 and 1950, and linearly interpolate average incomes of white 16 year olds for the years 1941-1949. I also linearly interpolate population counts for each age*race cell across missing years. Because pre-1940 earnings information was based on the relationship between 1940 demographic information and

1.15.

8. One working paper attempts to construct a federal and state income tax calculator, but the calculator is incomplete and not yet available to researchers (Bakija 2017).

9. See `https://www.taxpolicycenter.org/statistics/revenue-government-level`

10. Each age*race cell has roughly ten snapshots of census and ACS data that I use to interpolate income values. My results are virtually identical when I interpolate income data using more flexible cubic splines.

earnings, I lastly multiply all pre-1940 earnings cells by the ratio of average earnings per person in that year to average earnings per person in 1940.[11] In a small number of sparsely populated cells describing respondents at young ages, the imputed earnings are very slightly negative. In those cases, I zero out the earnings in that cell.

2.2.2 Life Expectancy Data

I calculate the number of black and white men born in the U.S. in each year between 1900 and 1970 using CDC summary tables and Census microdata. The CDC publishes U.S. birth counts separately for blacks and whites from 1959-present and for whites and non-whites from 1910-1959.[12] The CDC also produces the probability of surviving through age $t \in \{1, 5, 10, 15, ..., 100\}$ for black and white men in decennial census years from 1900 to the present.[13]

11. Source: Series D722-727 on page 44 of the Labor section in the U.S. Census Bureau's 1970 Historical Statistics of the United States, Colonial Times to 1970. See `https://www2.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970/hist_stats_colonial-1970p1-chD.pdf` for more details about how the series was constructed.

12. I rely on birth counts separated by the race of the child when available (not the race of the mother). See Table 1-1: Live Births, Birth Rates, and Fertility Rates, by Race: United States 1909-2000. See `https://www.cdc.gov/nchs/data/statab/t001x01.pdf`. The CDC data described here rely heavily on birth and death certificates.

13. See Table 20 of Arias (2011). The CDC constructs these survival probabilities using census data, medicare records, and birth and death certificates. CDC mortality rates for 1900 and 1910 rely on a smaller number of 10 death registration states that collected complete vital statistics information. And mortality data for 1920 rely on 34 states with high-quality data. Black births were underreported in the early 1900s because of the CDC's reliance on death registration states and because of poor birth certificate quality. This in turn biases black infant mortality rates. Eriksson, Niemesh, and Thomasson (2018) show that black infant mortality rates are biased upward relative to white infant mortality rates because black newborns are less likely to be enumerated by government agencies. For example, in 1940 94% of white births were registered on birth certificates, while only 82% of black births were registered on birth certificates, which is consistent with earlier work showing that black births were undercounted on birth certificates by roughly 20% in 1880-1940 (Coale and Rivers, 1973). Eriksson and coauthors argue that the black infant mortality rate for the 1915 birth cohort was 11.1%. Meanwhile, the CDC data (Arias 2015, Table 20) implies a black infant mortality rate of roughly 14.9%. While this difference is quantitatively large, my estimates of lifetime earnings gaps are largely driven by death rates before the age of 20. And this 3.8pp mortality rate underestimation is not quantitatively important for my welfare estimates when compared to the probability of death before the age of 20, which was roughly 29% for the 1915

I use census microdata to construct a measure of birth counts to validate the CDC mortality rates. In Figures 2.1 and 2.2, I plot the number of white and black males born each year from 1900 to 1970. The CDC datapoints are the raw birth counts from the CDC. In years prior to 1959, I assume that the number of non-white non-black births equals 55,000, which is the number of non-white non-black births in 1959.[14] The imputed datapoints take the CDC survivorship rates, linearly interpolate them to fill in years and ages without available data, and apply them to the counts of people in the census and ACS microdata in the appropriate birth cohort, race, and age group. This produces implied birth counts from each age*birth cohort*race cell in the census microdata. Figures 2.1 and 2.2 show the median number of implied births from cells aged 5-60 in the census.[15] The survivorship rates provided by the CDC, applied to Census counts closely match the CDC birth counts. Although the CDC birth counts do tend to be roughly 5% larger for both blacks and whites.

As readers can see in Figures 2.3 and 2.4, blacks had significantly shorter lifespans than whites in the early 1900s. In Figure 2.3, I plot black and white life expectancy for males by birth cohort using each year's age-specific mortality rates. The life expectancy numbers in this figure come directly from the CDC and rely on a variety of statistical procedures to identify true life expectancy (Arias, 2011). White males born in 1900 had an average life expectancy at birth of 47 years and black males born in 1900 had an average life expectancy at birth of 33 years. Note the sharp drop in life expectancy for the 1918 white and black birth cohorts. This drop is due to the Spanish Flu, which single-handedly decreased life expectancy by 11.8 years in 1918 (Noymer and Garenne 2000).[16] Life expectancy increased dramatically between the 1900 and 1940 birth cohorts, and the

---

birth cohort of blacks.

14. I rely on this assumption in years before 1959 when the CDC did not separate out births of black children from births of all non-white children.

15. I subset to these age ranges because in ages 0-4 and 61+, census population counts have various rounding, undercounting, and small sample size issues in some years.

16. This 11.8 year drop in life expectancy in Figure 2.3 is based on a cumulation of the age-specific decreases in life expectancy for Americans in 1918. In other words, males born in 1918 did not have a realized lifespan 11.8 years lower than males born in 1917. But if one had given ever

black-white gap in male life expectancy shrank from 14.1 years to 10.6 years. In Figure 2.4, I plot the probability of surviving through age 30 by race and birth cohort using actual mortality rates experienced by these birth cohorts, through age 30. We can see that the largest convergence in life expectancy took place between the 1900 and 1920 birth cohorts.

2.3 Cross-Sectional Earnings Gaps

In Table 2.1, I report the average earnings of living blacks and white 30 year-old men every ten years, beginning in 1930 and ending in 2020.[17] I inflate the earnings values into 2014 dollars using the CPI-U. I do not subset to wage-earners, so men with no earnings are included in the average. The 1930 through 2020 years of data in Table 2.1 correspond to the 1900 through 1990 birth cohorts. White earnings roughly double over this time period while black earnings roughly triple. The ratio of white to black average earnings at age 30 increases from 1930 to 1940 before declining between 1940 and 1950 and declining slowly in the following years. Post-tax labor earnings (column 2 of each panel) have a slightly smaller white/black earnings ratio for the simple reason that income tax rates in the U.S. are progressive and whites had higher average incomes than blacks. Total earnings, which includes government transfers, investment, and other income, display similar trends to labor earnings. Table 2.1 shows that white males born in 1900 who were living at age 30 earned an average of 2 times the amount that living black males born in 1900 earned at age 30.

In Figure 2.5, I plot the ratio of white to black labor earnings for living men in different age groups to better visualize the evolution of cross-sectional white earnings relative to black earnings over time.[18] The figure separates out this ratio for living men in each birth cohort and in three age

male born in 1918 the age-specific mortality rates of cohorts in 1918 as they proceeded through life, the projected life expectancy of males born in 1918 would have been 11.8 years lower than the projected life expectancy of males born a year earlier.

17. I impute earnings for the pre-1940 and post-1914 years using the methods described in the data section.

18. The graphs for post-tax labor earnings and total earnings are virtually indistinguishable and

bins: men in their 30s, 40s, and 50s. To produce each line, I average the ratio of white to black earnings for each age in that group. Recall that I impute data in non-decennial census years before 2000. Also, for the 1975 birth cohorts and the men in their 50s, the white/black earnings ratio is projected forward in years that have no available data (2015-2029). In Figure 2.5, the white/black average earnings ratio is roughly 2 for the 1900 birth cohort in each age bin. The gap steadily drops to around 1.7 in the 1945 birth cohort before slowly increasing back up to 1.8-2.1 depending on the age bin. If we used this type of standard inequality metric as a measure of the evolution of black-white inequality for children born in different years, we might assume that there was no progress in the expected adult labor market outcomes of white men relative to black men between 1900 and 1970. But as I discussed above, this is misleading.

Lastly, in Figure 2.6 I show how the ratio of white to black labor earnings evolves over the life-cycle for four different birth cohorts from 1900 to 1975. For more recent cohorts, I impute data for later ages. The plot is noisy, because the sample sizes for a specific birth cohort*age*race cell are not particularly large (especially for blacks), but the general trend is that the relative earnings of whites relative to blacks increases from 1.5 at young ages to closer to 2 at higher ages.

2.4 Lifetime Earnings Gaps

I construct estimates of lifetime earnings for each birth cohort and race using the decennial census and ACS microdata described in the data section, imputing earnings for unavailable years and ages. In addition, to estimate lifetime earnings for more recent birth cohorts, I assume that real earnings will grow for each birth cohort*race group at a 1.5% rate from 2014 through 2060. I also assume that no one earns any income after age 89. I calculate two measures of lifetime earnings: undiscounted and discounted. For the discounted lifetime earnings, I discount each annual income datapoint back to the birth year for each cohort (assuming an annual discount rate of 0.96). I then sum the average earnings from age 16-89 for each race in each birth cohort,

show all the same trends.

weighted by the interpolated CDC mortality rates. This is a measure of the 'expected' earnings for each birth cohort, if the men in that birth cohort only knew their race, expected annual income at each age conditional on race, and expected annual mortality rates conditional on race, with no other information about ability or environment.

In the previous section, I calculated cross-sectional earning gaps, which are standard and reported widely in the literature on black-white inequality. I now provide the first estimates of lifetime earnings for birth cohorts in the U.S. In Table 2.2, I report the average undiscounted lifetime labor earnings, lifetime post-tax labor earnings, and lifetime total earnings for black and white birth cohorts born every five years from 1900 to 1970. In Table 2.3, I report discounted lifetime earnings ($\beta = 0.96$). Lifetime earnings roughly tripled for white men and increased by a factor of six for black men when we compare the 1900 and 1940 birth cohort. But lifetime earnings then increased only slightly from the 1940 to 1970 birth cohorts. The 'No Death' column reports the lifetime earnings each newborn male could expect to receive in the given cohort if they died at age 90 and received the average earnings for the living males in their birth cohort*race cell at each age. As expected, the hypothetical lifetime earnings in a world with no death are roughly 45-50% larger for whites than their actual average lifetime earnings. For blacks, the hypothetical lifetime earnings in a world with no mortality until age 90 are 2.3-2.6 times larger than the actual lifetime earnings.

I also include, in Tables 2.2 and 2.3, a column titled "+VSL" that column measures the economic value of the difference in life expectancy for each birth cohort of black and white men. I calculate this value as follows: I begin with a $2.2 million estimate of the value of a statistical life for the 1970 birth cohort.[19] I follow the U.S. government in assuming that the value of a statistical

19. Ashenfelter and Greenstone (2004) use changes in speed limits to estimate the causal effect of a change in mortality risk on time saved. By putting a value on each hour of travel time saved equal to the average hourly wage, they estimate that the value of a statistical life is $1.5 million in 1997 dollars, which is $2.22 million in 2015 dollars. Recent papers by Ioannidis et al. (2017) and Doucouliagos, Stanley, and Giles (2012) show that standard hedonic measures of the VSL suffer from "severe publication selection bias." The authors correct for this publication bias by conducting a meta-analysis of the VSL parameter after restricting to studies with a sufficient amount of power. They estimate a VSL of $1.47 million in 2000 dollars, which is virtually identical to the

life year is constant across ages and races (Aldy and Viscusi, 2008). Because the lifetime income and life expectancy of the average male born in the U.S. grew tremendously between 1900 and the present, I scale my estimate of the VSL by the changes in lifetime income and the changes in life expectancy each year. To do this, I first assume that the $1.5 million estimate comes from the 1970 birth cohort. White males born in 1970 had a life expectancy of 68 years. So, the 1970 birth cohort of white men valued each life-year at around $32,000 2015 dollars, which is 0.074-times their discounted lifetime earnings of $432,000. So, I use 0.074*(the average lifetime earnings of white men) as a measure of the value of a statistical life year for each birth cohort. This means that the statistical value of a life year increases as the average lifetime income of men grows. Here, I am assuming that a 1% change in lifetime income increases the statistical value of a life-year by 1%. This is not unrealistic, since recent estimates of the income elasticity of the value of a statistical life year range from 0.7 to 1.1 (Viscusi and Masterman, 2017).

In Tables 2.4 and 2.5, I report the ratio of white to black undiscounted and discounted lifetime earnings, for all three measures of earnings (labor, post-tax labor, and total earnings). I also report in the 'No Death' column what the ratio of white to black lifetime earnings would be in a world where everyone lived through age 89 and died at age 90. White males born in 1900 earned roughly 3.4 times as much over their lifetime as black males born in 1900 did. This ratio fell to 2.1 for the 1920 birth cohorts and has since fallen only slightly further, to 2.0 for the 1970 birth cohort. The last column in each panel of Table 2.3 presents the ratio of white to black expected consumption from the model I present in the next section.

It is clear that whites born in the early 1900s had significantly higher lifetime earnings than equivalent cohorts of black men. But white men could also expect to live much longer than black men in all birth cohorts. One simple way to combine both of these facts into one welfare metric is to add the value of the additional life-years that whites could expect to receive at birth to their lifetime income. The thought experiment here is that if whites and blacks could all expect to live

---

estimate of the social value of a statistical life from Ashenfelter and Greenstone (2004).

for the same number of years, but still expected to receive the lifetime earnings of their respective race*birth cohort, then the white birth cohort should be indifferent between receiving X additional years of expected life or receiving a lump sum payment at birth equal to the value of a life-year from the VSL literature times X. In Table 2.2, I presented this value of the black-white gap in life expectancy for each birth cohort in the '+VSL' column.

In the third column of each panel in Table 2.3, I take the ratio of white lifetime earnings + the value of the extra years of life white cohorts received to black lifetime earnings. In Figure 2.7, I plot these six measures of relative white and black welfare annually. This method for combining lifetime earnings and life expectancy gaps into one welfare measure is useful because it is in dollars, but it is somewhat ad hoc. I must assume a fixed value of each year of life for a given birth cohort, independent of how long they live and independent of their annual income profile; in other words, there are no decreasing returns to life expectancy and there is no interaction between the return to additional income and the return to lower mortality risk. So, below I calibrate a simple model of consumption in a world with mortality risk to the census data.

2.5 Model

I consider a simple model of optimal consumption in a world with annual mortality risk. Time is discrete with T periods, $1, \ldots, T$. Each person born in year 0 makes a plan to consume c_a at each age a if they are alive, which provides them with utility at age a of $u(c_a)$. p_t is the independent and exogenous probability that a person dies at age t, so the probability of surviving through age a is $s_a = \prod_{j=1}^{a} (1 - p_j)$. Earnings at each age a, $y_a \geq 0$, are exogenous and expected lifetime utility, given a sequence of consumption $\{c_a\}$, is

$$U = \sum_{a=1}^{T} \beta^a s_a u(c_a)$$

where β is the standard discount rate.

At birth, people can sign contracts to borrow or loan money intertemporally at a constant annual interest rate r. So the budget constraint is: $\sum_{a=1}^{T} \frac{s_a c_a}{(1+r)^a} = \sum_{a=1}^{T} \frac{s_a y_a}{(1+r)^a}$. In other words, a person can sign a contract to borrow against future expected income to finance consumption, but she can only borrow against income that the lender expects her to receive given the commonly known probability of dying each period and the commonly known income process. This budget constraint assumes a complete annuities market. Readers can think of this market as one where each newborn agrees to send their income each year to the bank, and in return, they receive a predetermined amount of money from the bank each year that they are alive. The amount of money the bank sends each person is governed by the consumer's budget constraint, which states that the lifetime net present value of consumption must be no larger than the lifetime net present value of income. This is equivalent to a zero profit constraint for the bank that is borrowing and loaning money, so this is a world where many small banks compete in this annuity market with exogenous interest rate r.

At birth, each newborn chooses a potential sequence of consumption $\{c_1 \dots c_T\}$ to maximize expected lifetime utility subject to the budget constraint. Consumption here is the total number of dollars the newborn will consume in a given period if they are alive. So expected consumption in period t is $s_t c_t$. The Lagrangian for this maximization problem is:

$$\mathscr{L} = \sum_{a=1}^{T} \beta^a s_a u(c_a) + \lambda \sum_{a=1}^{T} \left(\frac{s_a(y_a - c_a)}{(1+r)^a} \right)$$

Differentiating with respect to c_t gives us a sequence of first order constraints that yield the standard Euler equation: $\beta(1+r) = \frac{u`(c_t)}{u`(c_{t+1})}$. For simplicity, assume that $u(c) = log(c)$. So the Euler equation simplifies to $c_{t+1} = \beta(1+r)c_t = [\beta(1+r)]^{t-1} c_1$. The logic follows identically if utility is CRRA, but it is more difficult to solve for a closed-form value of consumption each period.

Going back to the budget constraint, and plugging in the Euler equation from the previous step, we can see that

$$\sum_{a=1}^{T} \frac{\beta^{a-1} s_a c_1}{1+r} = \sum_{a=1}^{T} \frac{s_a y_a}{(1+r)^a}$$

So,

$$c_1 = \frac{\sum_{a=1}^{T} \frac{s_a y_a}{(1+r)^{a-1}}}{\sum_{a=1}^{T} s_a \beta^{a-1}}$$

This means that $\forall t$,

$$c_t = \frac{[\beta(1+r)]^{t-1} \sum_{a=1}^{T} \frac{s_a y_a}{(1+r)^{a-1}}}{\sum_{a=1}^{T} s_a \beta^{a-1}}$$

We can now calculate, at birth, the amount each person expects to consume at each age as a function of known parameters. So, we can deterministically calculate the utility a newborn can expect to receive given (1) constant parameters β and r, and (2) a sequence of probabilities of dying at each age $\{p_a\}$ and income at age a, $\{y_a\}$. Notice that T is the lowest age a such that $p_a = 1$. Because the decennial census sometimes topcodes age at 90, I will assume that everyone dies by the age of 90, and therefore no one consumes anything after age 89.

I can now compare the expected utility of blacks and whites over time and within a birth cohort by calculating levels of utility separately for black and white newborns with known annual incomes (from census microdata) and probabilities of dying (from the CDC).

2.6 Comparative Statics

We can easily see that consumption in all periods is increasing in income, but the effect of mortality risk, p_a, on consumption in each period is more complicated. With many applications of the chain rule, we arrive at this partial derivative of consumption in period t conditional on being alive (c_t) with respect to the probability of dying (p_s) in a given period $s > 1$:

$$\frac{\partial c_t}{\partial p_s} = \frac{[\beta(1+r)]^{t-1}}{(1-p_s)} \left(\frac{\left(\sum_{a=1}^{s-1} \frac{s_a y_a}{(1+r)^{a-1}}\right)\left(\sum_{a=s}^{T} [s_a \beta^{a-1}]\right) - \left(\sum_{a=s}^{T} \frac{s_a y_a}{(1+r)^{a-1}}\right)\left(\sum_{a=1}^{s-1} [s_a \beta^{a-1}]\right)}{\left[\sum_{a=1}^{T} s_a \beta^{a-1}\right]^2} \right)$$

This partial derivative can be either positive or negative depending on the distribution of earnings and mortality over a lifespan. While this might seem unrealistic, recall that c_t does not represent realized consumption at age t. Instead, it represents consumption at age t if the representative agent lives to age t.

Using the partial derivative $\frac{\partial c_t}{\partial p_s}$, we can once again use several applications of the chain rule to calculate how utility is affected by an increased probability of dying in a given period.

$$\frac{\partial U}{\partial p_s} =$$

$$\frac{1}{(1-p_s)}\sum_{t=1}^{T}\beta^t s_t\left(\frac{\left(\sum_{a=1}^{s-1}\frac{s_a y_a}{(1+r)^{a-1}}\right)\left(\sum_{a=s}^{T}[s_a\beta^{a-1}]\right)-\left(\sum_{a=s}^{T}\frac{s_a y_a}{(1+r)^{a-1}}\right)\left(\sum_{a=1}^{s-1}[s_a\beta^{a-1}]\right)}{\left(\sum_{a=1}^{T}\frac{s_a y_a}{(1+r)^{a-1}}\right)\left(\sum_{a=1}^{T}s_a\beta^{a-1}\right)}-log(c_t)\right)$$

Note that this sum is made up of two terms. The first term, a complicated ratio of terms, is bound below by -1 and above by 1. So a sufficient, but not necessary condition for $\frac{\partial U}{\partial p_s} < 0$ is if, $\forall t$, $log(c_t) > 1$. The reason this partial derivative can sometimes be positive is because if consumption is sufficiently close to zero, utility is negative, so an increase in the probability of dying actually increasing expected utility by moving it closer to zero. But for any plausible sequence of income, including all sequences of income we see in census microdata for race*birth cohort cells, this partial derivative will be negative; and an increase in mortality risk reduces utility.

Lastly, one value of this model is that it allows mortality risk to affect people differently as a function of their expected annual income. Equivalently, shocks to income will have different effects on people with higher or lower mortality risk. To measure this, we can calculate the cross partial derivative of c_t with respect to p_s and y_s, and we see that:

$$\frac{\partial c_t}{\partial y_s \partial p_s} = \frac{\beta^{t-1}(1+r)^{t-s}s_s}{(1-p_s)\left(\sum_{a=1}^{T}s_a\beta^{t-1}\right)^2} * \left[-\sum_{j=1}^{s-1}s_j\beta^{j-1}\right]$$

This is always negative, which means that at higher values of income, an increase in mortality risk in a given period has an increasingly negative effect on the level of expected consumption in all periods.

2.7 Calibration

I can now calibrate this model separately to any sequence of annual survivorship probabilities and earnings. I take each black and white birth cohort from 1900 to 1970 and calculate expected consumption and utility for that cohort. In Table 2.4, Table 2.5, and Figure 2.7, I show the ratio of dollars of lifetime expected consumption for white and black birth cohorts each year. These closely match the general time trends of my earnings-based inequality measures, with the largest convergence in expected lifetime consumption between the 1900 and 1920 birth cohorts. But this measure of relative consumption calibrated to total earnings (in the last column of Table 2.4) shows continual gradual progress in black-white inequality between the 1920 and 1970 birth cohorts.

In Figure 2.8, I plot the level of utility for blacks and whites in each birth cohort. Normally levels of utility are unitless and not particularly understandable. But in this model, levels of utility represent the net present value of future expected log consumption. As we can see in Figure 2.8, levels of utility have weakly monotonically increased from the 1900 to 1970 birth cohorts for blacks and whites. In Figure 2.9, I plot the gap between black and white utility levels. The largest convergence in black and white utility occurred between the 1910 and 1920 birth cohorts. Since then, the gap is roughly constant. Black males born in 1970 have earnings and mortality risk that net them welfare equivalent to the welfare received by white males born in 1930.

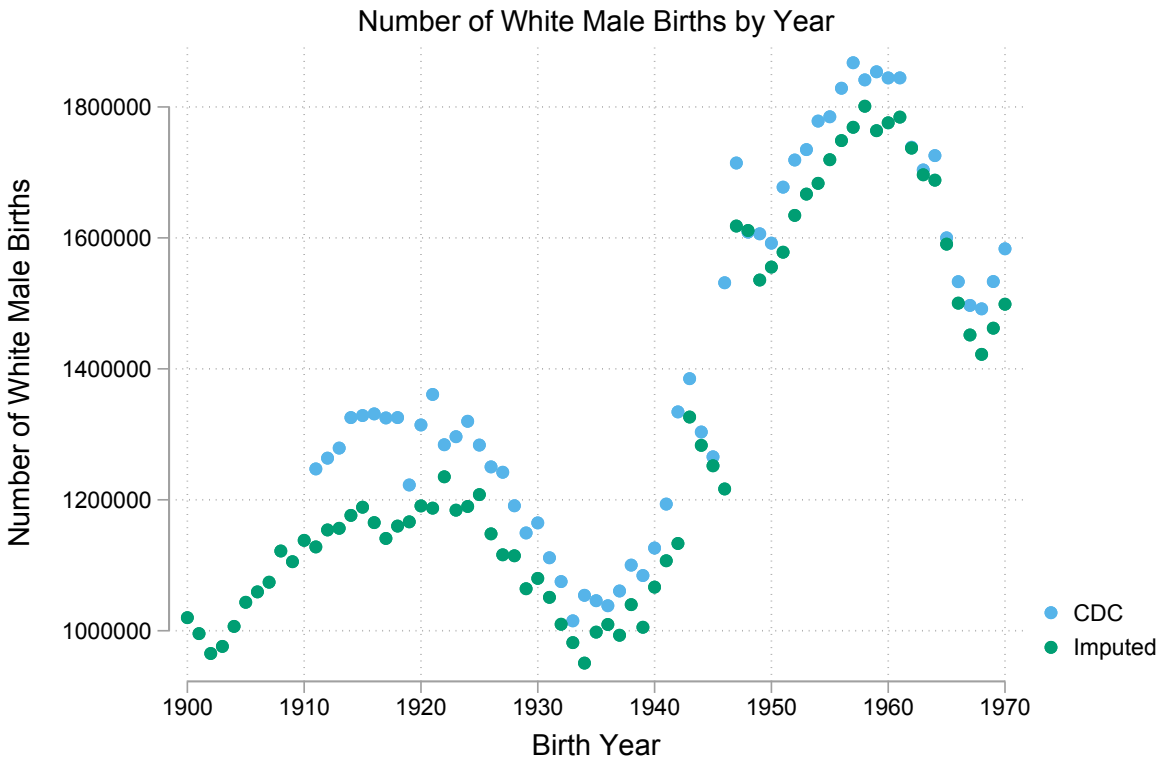
2.8 Conclusion

48% of black males born in 1900 died before the age of 30 as compared to only 26% of white males. These mortality rates declined to 22% and 15% by the 1920 birth cohort, and this sudden convergence had large effects on relative welfare. Black-white cross-sectional earnings gaps are often used as a measure of inequality, but these gaps ignore the large (albeit incomplete) convergence of black and white mortality rates from 1900 to the present. I calculate two 'full income' measures of black-white inequality that combine lifecycle earnings and mortality rates to

measure welfare.

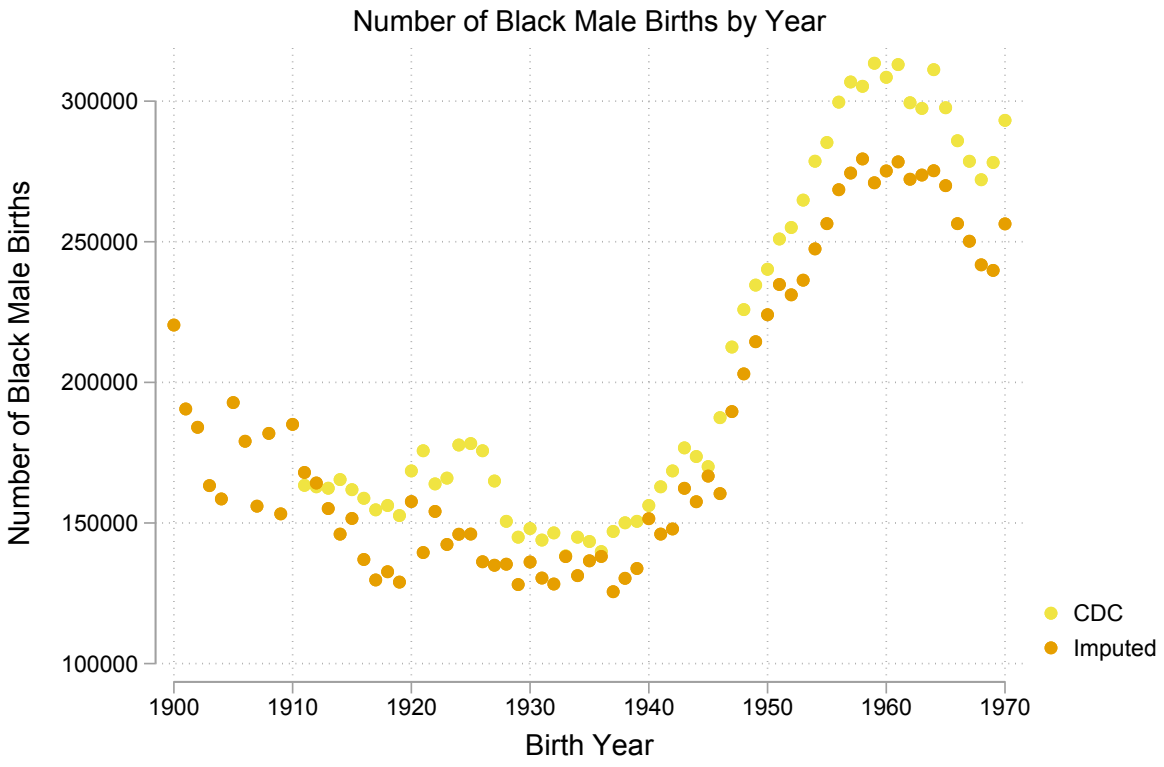
In the 1940 census, the average working-age white male reported earnings twice as large as the average working-age black male. But because blacks had significantly higher mortality rates at each age, white males born in 1900 could expect to have lifetime labor earnings 3.4 times the lifetime labor earnings of black males born in 1900. This ratio fell to 2.1 in the 1920 birth cohorts and has since fallen only slightly further, to 2.0 for the 1970 birth cohort. A simple structural model of consumption in a world with mortality risk similarly shows that almost all of the convergence in black-white welfare over the past century came from the rapid but incomplete convergence in black and white mortality rates between 1900 and 1920 birth cohorts. Black-white welfare gaps have declined only slightly since the 1920 birth cohort.

Figure 2.1:



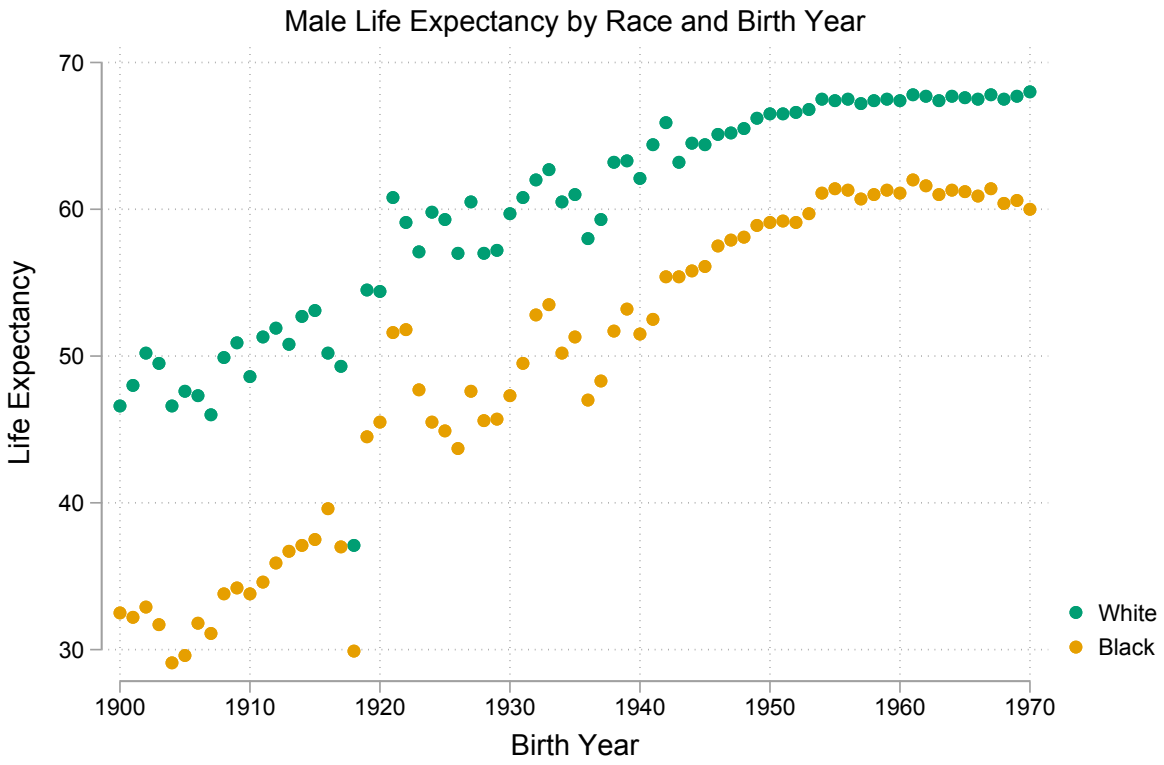
Note: CDC birthcounts based on published CDC data which relies in birth certificates. Imputed birthcounts apply age-specific mortality rates to census birthcounts. The imputed data point for each birth year equals the median imputed birthcount from age 5-60 census age groups.

Figure 2.2:



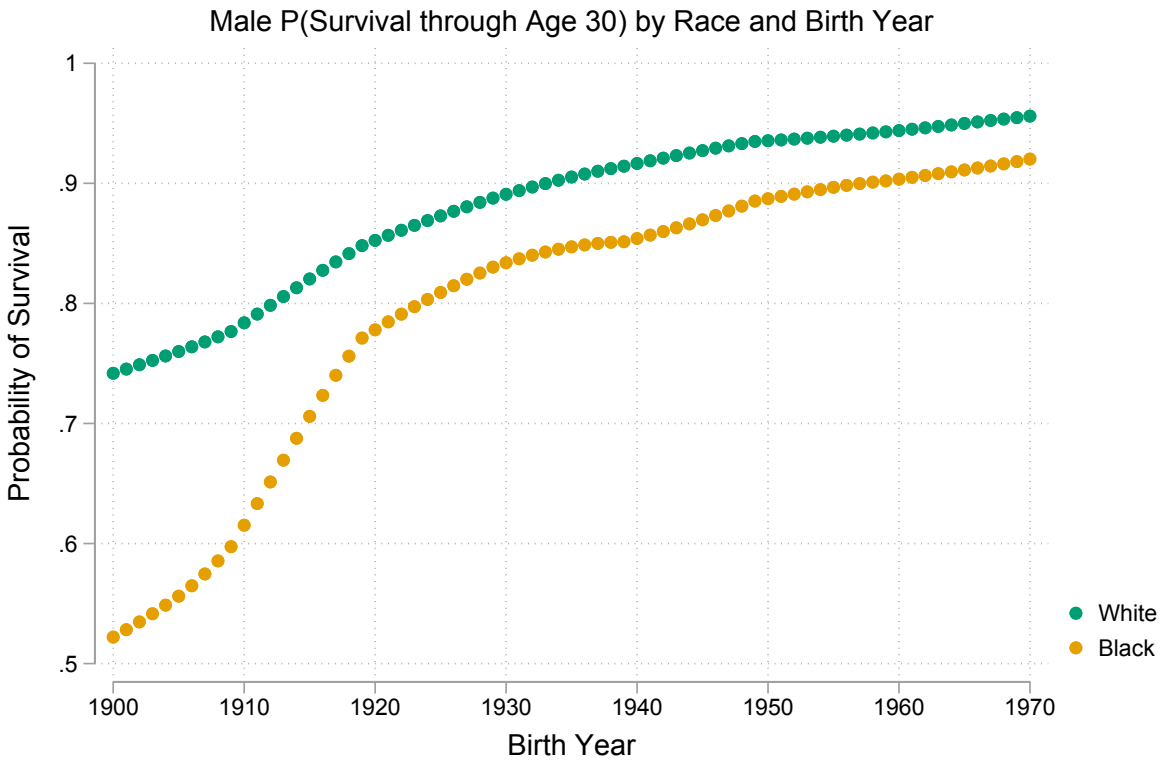
Note: See note to Figure 2.1.

Figure 2.3:



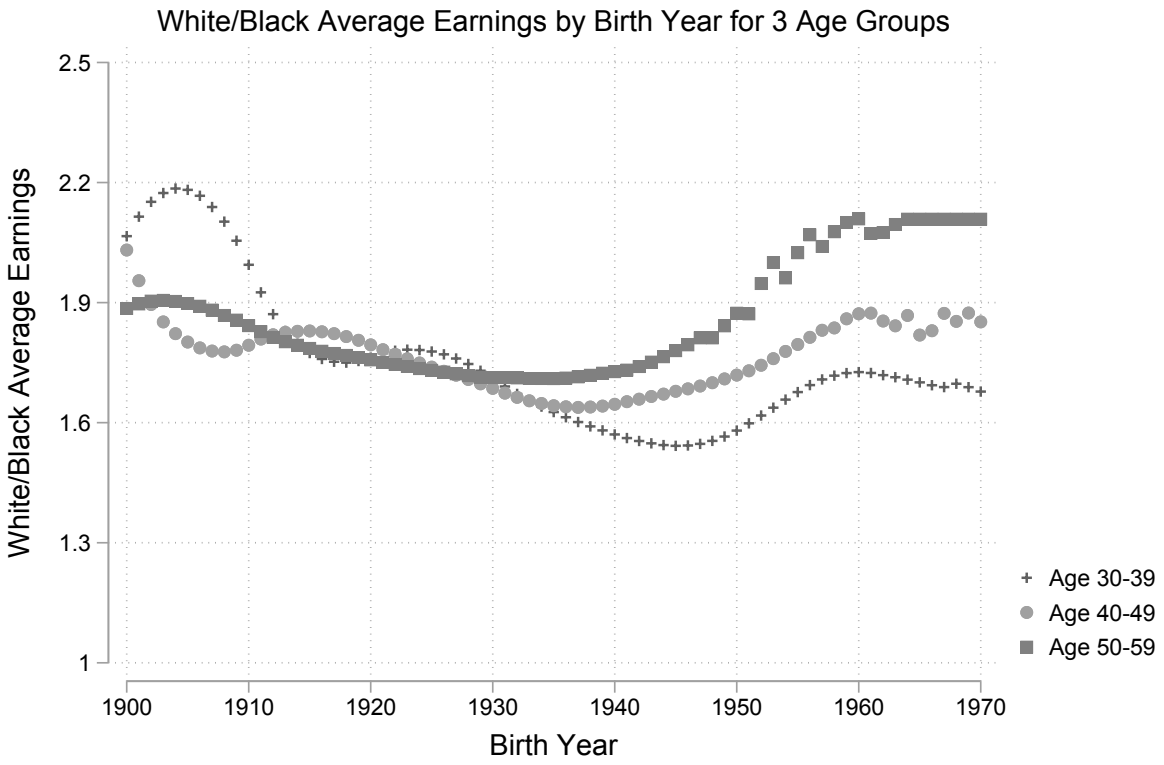
Note: Life expectancy by race taken directly from the CDC (Arias 2011). The life expectancy measures are calculated using that year's mortality rates for each cohort. So the Spanish Flu, which caused large drops in life expectancy for all ages in 1918, had a noticeable effect on life expectancy of black and white males in 1918.

Figure 2.4:



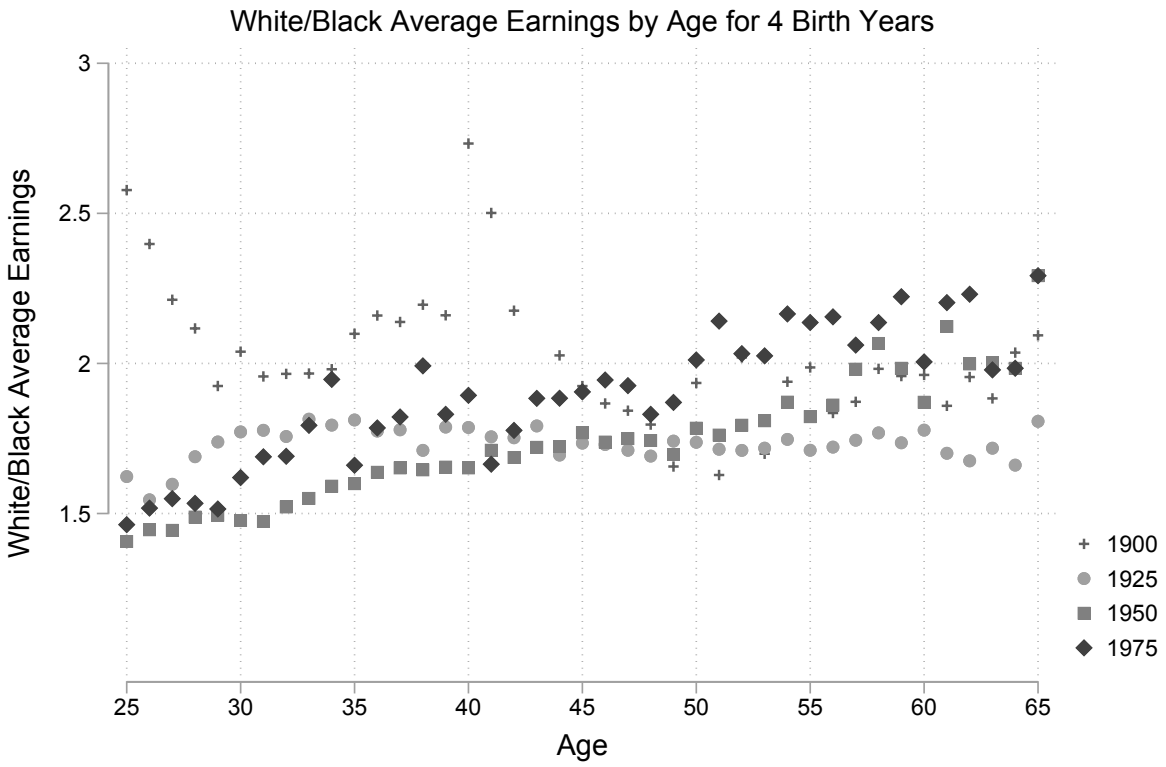
Note: Survivorshop rate taken directly from the CDC (Arias 2011).

Figure 2.5:



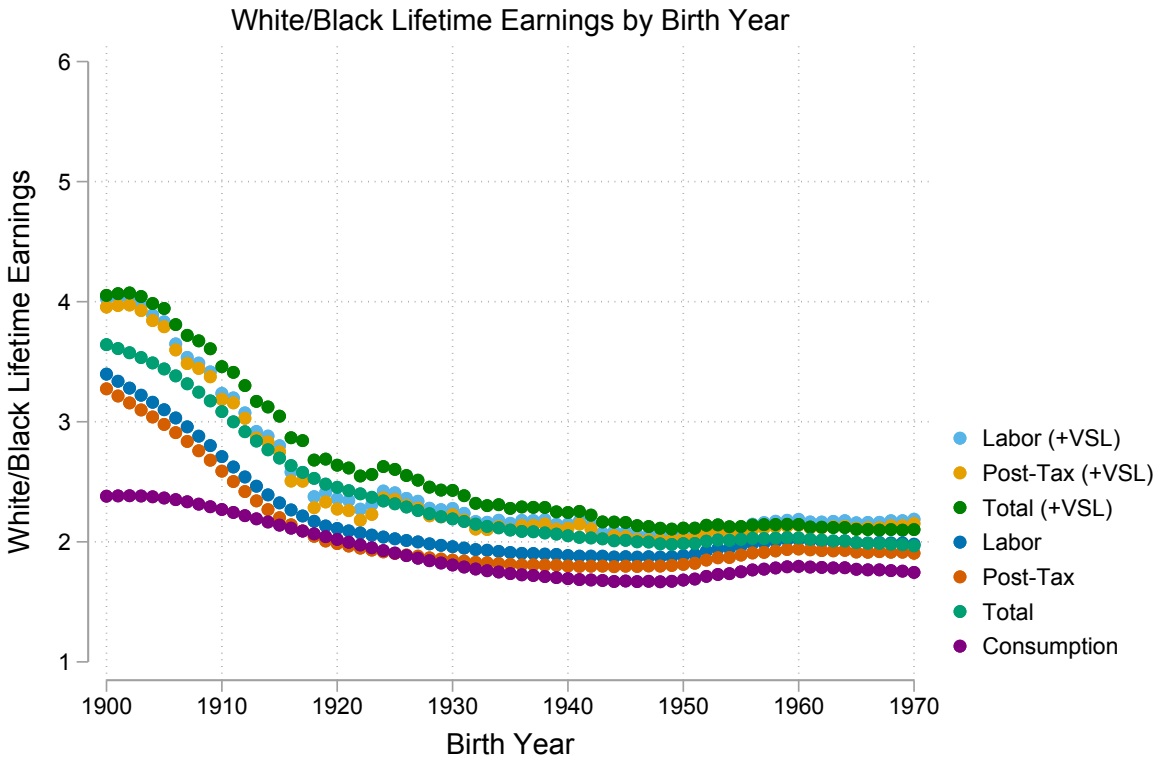
Note: See note to Table 2.2. Earnings are linearly interpolated by race*age*birth year cell in non-census years and then averaged within 10-year age bins for each birth cohort to produce these scatter plots at an annual level. I assume that earnings after 2014 are equal to earnings in 2014 (within race*age cells). This is only relevant for the oldest age group in this figure.

Figure 2.6:



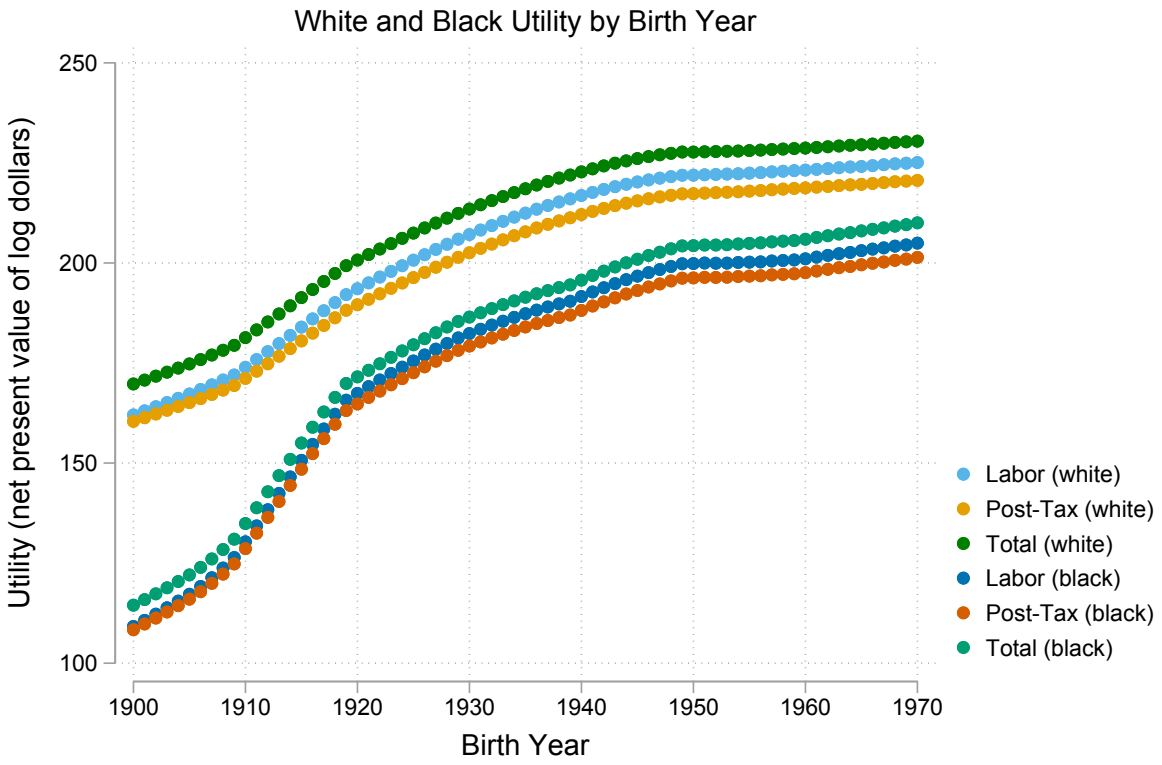
Note: See note to Figure 2.5.

Figure 2.7:



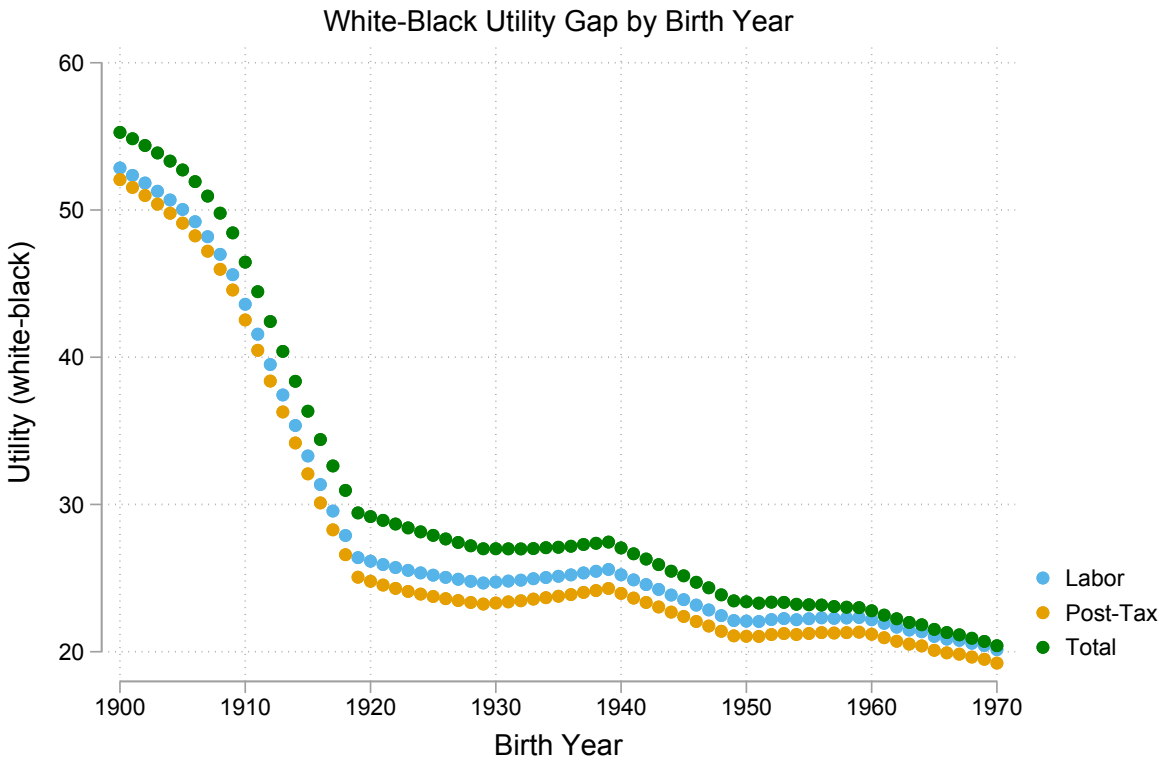
Note: See note to Table 2.2. This is the ratio of undiscounted lifetime earnings. The +VSL lines correspond to the thought experiment I describe in the text, where I take the ratio of undiscounted white lifetime earnings + the value of the additional life-years each white male can expect to receive to undiscountd black lifetime earnings. The consumption line corresponds to the lifetime expected consumption of white males divided by the lifetime expected consumption of black males born in each birth cohort. Here, consumption is calculated assuming $\beta = 1$, but this ratio is quite similar when the consumption line is calculated using $\beta = 0.96$.

Figure 2.8:



Note: Each line is the level of utility for whites or blacks, extracted from the structural model I calibrate in the paper. I assume a discount rate of $\beta = 0.96$

Figure 2.9:



Note: See note to Figure 2.8. This is simply the level difference in white and black utility from Figure 2.8. I include it to emphasize the large convergence in white and black utility between the 1900 and 1920 birth cohorts, followed by general stagnation.

Table 2.1: Average Earnings at Age 30

	White Earnings			Black Earnings			White/Black Earnings		
Year	Labor	Post-tax	Total	Labor	Post-tax	Total	Labor	Post-tax	Total
1930	16,033	15,786	23,765	7,701	7,664	11,753	2.08	2.06	2.02
1940	16,019	15,805	23,169	6,611	6,588	10,688	2.42	2.40	2.17
1950	23,862	20,465	29,471	13,430	12,009	15,295	1.78	1.70	1.93
1960	36,993	30,471	42,820	20,916	18,088	22,732	1.77	1.68	1.88
1970	52,122	42,127	57,423	33,883	28,577	35,481	1.54	1.47	1.62
1980	46,917	38,443	52,454	31,770	27,045	34,241	1.48	1.42	1.53
1990	45,179	37,418	49,895	26,154	22,311	27,737	1.73	1.68	1.80
2000	47,061	38,613	51,566	29,963	25,242	32,172	1.57	1.53	1.60
2010	38,740	32,687	42,294	21,751	18,868	24,163	1.78	1.73	1.75
2020	38,535	32,463	41,337	24,061	20,689	25,418	1.60	1.57	1.63

Note: Earnings data from default IPUMS census and ACS microdata samples. Units are dollars. Labor earnings are only available in the Census data from 1940-2014 and total earnings only available from 1950-2014. I do not use ACS data from 2015-2016 because the race variable is not easily comparable to 1900-2014 census and ACS data. Data subset to American-born black and white men who report being age 30 in the census and ACS. Each cell is the average earnings across all living white or black 30-year old men, including men who earn zero dollars. Labor earnings rely on the INCWAGE IPUMS variable, Total earnings rely on the INCTOT IPUMS variable. Post-tax earnings apply federal income tax rates to the INCWAGE variable. Earnings values in 2020 are equal to 2014 earnings values. 1930 labor and total income values and 1940 total income values are imputed as I describe in the text of the paper. For example, I regress 1940 labor earnings at the individual level on indicators for race, age, occupation, industry, state of birth, and state of residence. I then use these regression coefficients to predict labor earnings for each male in 1930.

Table 2.2: Lifetime Earnings (undiscounted)

	Whites							Blacks					
	Labor		Post-Tax		Total			Labor		Post-Tax		Total	
Birth Cohort	Earnings	No Death	Earnings	No Death	Earnings	No Death	+VSL Gap	Earnings	No Death	Earnings	No Death	Earnings	No Death
1900	669	1,067	595	936	1,099	2,077	124	197	523	182	475	302	959
1905	779	1,199	673	1,025	1,257	2,301	184	251	608	226	541	365	1,066
1910	921	1,354	774	1,130	1,473	2,590	179	340	721	299	630	477	1,240
1915	1,118	1,543	919	1,262	1,771	2,914	229	481	862	417	743	656	1,460
1920	1,349	1,762	1,089	1,418	2,101	3,219	158	639	1,006	548	859	856	1,664
1925	1,574	1,977	1,258	1,578	2,406	3,489	297	777	1,154	661	979	1,038	1,885
1930	1,785	2,175	1,423	1,732	2,681	3,725	291	911	1,290	770	1,088	1,223	2,104
1935	1,976	2,351	1,573	1,871	2,929	3,943	252	1,034	1,424	868	1,194	1,396	2,314
1940	2,141	2,502	1,705	1,992	3,157	4,161	298	1,135	1,524	948	1,271	1,539	2,472
1945	2,239	2,584	1,789	2,064	3,307	4,307	244	1,193	1,552	995	1,293	1,643	2,561
1950	2,237	2,563	1,800	2,060	3,330	4,333	217	1,187	1,507	993	1,260	1,678	2,574
1955	2,248	2,571	1,818	2,078	3,369	4,408	177	1,143	1,432	962	1,204	1,667	2,551
1960	2,294	2,615	1,859	2,117	3,461	4,538	190	1,136	1,404	958	1,183	1,702	2,590
1965	2,371	2,688	1,922	2,177	3,606	4,724	199	1,191	1,450	1,004	1,221	1,808	2,718
1970	2,487	2,807	2,016	2,273	3,801	4,971	261	1,256	1,512	1,058	1,273	1,933	2,878

Note: Earnings data from default IPUMS census and ACS microdata samples. Units are thousands of dollars. Labor earnings are only available in the Census data from 1940-2014 and total earnings only available from 1950-2014. I do not use ACS data from 2015-2016 because the race variable is not easily comparable to 1900-2014 census and ACS data. Data subset to American-born black and white men who report being age 30 in the census and ACS. Each cell represents the average lifetime earnings for white and black males born in the given birth cohort. Earnings are calculated by summing average earnings for each birth cohort and race across all ages, weighted by the age-specific CDC mortality rate, and assuming that all men live only until age 89. Labor earnings rely on the INCWAGE IPUMS variable, Total earnings rely on the INCTOT IPUMS variable. Post-tax earnings apply federal income tax rates to the INCWAGE variable. Earnings values after 2014 are equal to 2014 earnings values, assuming an additional 1.5% annual growth rate. 1910, 1920, and 1930 labor and total income values and 1940 total income values are imputed as I describe in the text of the paper. For example, I regress 1940 labor earnings at the individual level on indicators for race, age, occupation, industry, state of birth, and state of residence. I then use these regression coefficients to predict labor earnings for each male in 1930. In non-surveyed years from 1916-2014, earnings are linearly interpolated at the age*race*birth year level before being summed to construct lifetime earnings measures. Pre-1940 earnings are adjusted to account for annual variation in wages. This has no effect on my results. The 'No Death' columns assume that all newborns live until age 89. The '+VSL Gap' column estimates the value of the black-white gap, calculated as described in the text.

Table 2.3: Lifetime Earnings ($\beta = 0.96$)

	Whites							Blacks					
	Labor		Post-Tax		Total			Labor		Post-Tax		Total	
Birth Cohort	Earnings	No Death	Earnings	No Death	Earnings	No Death	+VSL Gap	Earnings	No Death	Earnings	No Death	Earnings	No Death
1900	122	177	114	163	183	279	130	34	79	33	74	52	123
1905	136	193	124	174	200	299	186	42	89	39	82	59	132
1910	156	214	137	187	225	325	175	55	105	50	94	74	149
1915	187	244	159	206	264	362	221	80	130	71	114	101	175
1920	229	284	189	234	313	409	154	112	161	98	140	134	207
1925	276	332	225	270	364	458	301	143	195	123	167	168	246
1930	324	379	262	306	414	504	304	173	227	148	194	204	284
1935	367	420	295	337	459	546	269	201	257	170	217	235	319
1940	402	453	323	363	497	581	322	226	283	190	238	263	348
1945	420	467	339	376	518	598	264	243	296	204	248	283	364
1950	418	461	340	374	517	592	234	242	288	204	242	284	359
1955	416	457	341	374	514	588	189	229	270	195	229	273	342
1960	419	458	344	376	516	590	199	223	259	190	220	268	333
1965	422	459	346	376	521	593	204	227	261	193	221	276	339
1970	432	468	354	383	533	605	261	235	267	199	226	288	351

Note: See note to Table 2.2. The only difference here is that the discount rate is set to $\beta = 0.96$.

Table 2.4: White/Black Lifetime Earnings (undiscounted)

	Labor				Post-Tax				Total			
Birth Cohort	Earnings	No Death	+VSL Gap	Cnsm.	Earnings	No Death	+VSL Gap	Cnsm.	Earnings	No Death	+VSL Gap	Cnsm.
1900	3.40	2.04	4.03	2.30	3.27	1.97	3.96	2.23	3.64	2.17	4.05	2.38
1905	3.10	1.97	3.83	2.20	2.98	1.90	3.79	2.14	3.44	2.16	3.94	2.37
1910	2.71	1.88	3.24	2.05	2.59	1.80	3.19	1.98	3.08	2.09	3.46	2.27
1915	2.32	1.79	2.80	1.88	2.20	1.70	2.75	1.79	2.70	2.00	3.05	2.14
1920	2.11	1.75	2.36	1.76	1.99	1.65	2.27	1.66	2.45	1.93	2.64	2.02
1925	2.03	1.71	2.41	1.69	1.90	1.61	2.35	1.59	2.32	1.85	2.60	1.91
1930	1.96	1.69	2.28	1.64	1.85	1.59	2.22	1.55	2.19	1.77	2.43	1.81
1935	1.91	1.65	2.15	1.60	1.81	1.57	2.10	1.52	2.10	1.70	2.28	1.74
1940	1.89	1.64	2.15	1.58	1.80	1.57	2.11	1.50	2.05	1.68	2.24	1.69
1945	1.88	1.67	2.08	1.57	1.80	1.60	2.04	1.51	2.01	1.68	2.16	1.67
1950	1.89	1.70	2.07	1.60	1.81	1.64	2.03	1.54	1.99	1.68	2.11	1.68
1955	1.97	1.80	2.12	1.69	1.89	1.73	2.07	1.63	2.02	1.73	2.13	1.75
1960	2.02	1.86	2.19	1.76	1.94	1.79	2.14	1.70	2.03	1.75	2.14	1.80
1965	1.99	1.85	2.16	1.75	1.91	1.78	2.11	1.69	1.99	1.74	2.10	1.77
1970	1.98	1.86	2.19	1.74	1.91	1.79	2.15	1.68	1.97	1.73	2.10	1.75

Note: See note to Table 2.2. This is the ratio of white to black lifetime undiscounted earnings.

Table 2.5: White/Black Lifetime Earnings ($\beta = 0.96$)

	Labor				Post-Tax				Total			
Birth Cohort	Earnings	No Death	+VSL Gap	Cnsm.	Earnings	No Death	+VSL Gap	Cnsm.	Earnings	No Death	+VSL Gap	Cnsm.
1900	3.54	2.24	7.31	2.57	3.46	2.20	7.41	2.50	3.55	2.27	6.07	2.66
1905	3.27	2.17	7.73	2.44	3.20	2.13	7.98	2.36	3.39	2.27	6.54	2.62
1910	2.83	2.04	6.01	2.25	2.75	1.99	6.26	2.16	3.04	2.19	5.41	2.48
1915	2.33	1.88	5.09	2.05	2.24	1.80	5.34	1.95	2.62	2.07	4.81	2.33
1920	2.04	1.76	3.42	1.92	1.94	1.67	3.51	1.81	2.33	1.97	3.47	2.20
1925	1.94	1.71	4.05	1.84	1.83	1.61	4.28	1.73	2.16	1.87	3.95	2.07
1930	1.87	1.67	3.63	1.78	1.77	1.58	3.82	1.68	2.03	1.78	3.53	1.96
1935	1.83	1.63	3.16	1.74	1.73	1.55	3.31	1.65	1.95	1.71	3.10	1.88
1940	1.78	1.60	3.21	1.70	1.70	1.53	3.40	1.62	1.89	1.67	3.12	1.83
1945	1.73	1.58	2.82	1.69	1.66	1.52	2.95	1.62	1.83	1.64	2.76	1.80
1950	1.73	1.60	2.70	1.71	1.67	1.54	2.81	1.64	1.82	1.65	2.64	1.79
1955	1.81	1.70	2.64	1.79	1.75	1.64	2.72	1.72	1.88	1.72	2.57	1.85
1960	1.88	1.77	2.77	1.85	1.81	1.71	2.86	1.78	1.92	1.77	2.67	1.89
1965	1.86	1.76	2.76	1.83	1.79	1.70	2.85	1.76	1.89	1.75	2.63	1.85
1970	1.84	1.75	2.95	1.81	1.78	1.69	3.09	1.75	1.85	1.73	2.76	1.82

Note: See note to Table 2.2. This is the ratio of white to black lifetime earnings with annual discount rate $\beta = 0.96$.

CHAPTER 3

MATCHING CENSUS DATA

3.1 Introduction

Below, I describe the matching process that I use to match children and young adults from the 1900–1930 decennial censuses to adults in the 1940 decennial census. I use this matched data in Chapter 1 of my dissertation. I follow the standard matching algorithm from Abramitzky, Boustan, and Eriksson (2012), first independently matching each childhood census to the 1940 census using an iterative matching procedure. After each step, I collect all matches and remove them from the data–I then attempt to match the remaining records. If any childhood record matched to multiple adult records from 1940, I remove that childhood record and those multiple adult records from the search for matches because it is impossible to ascertain which of the adult records should match to the childhood record.

Step 1: I begin by matching 0–25 year old males in the 1900 census to adult men aged 20–65 in the 1940 census exactly on:

- The NYSIIS[1] of the first name
- The NYSIIS of the last name
- Exact age
- State or country of birth
- Race

I then iterate on age, allowing for differences of one year and then two years in either direction. In Step 1, I search for matches in five iterations because of this age mismatch allowance: I search

1. NYSIIS codes are New York State Identification and Intelligence System Phonetic codes: a phonetic coding system for names.

for exact matches on age, matches where the adult record's reported age is one year before the childhood record, matches where the adult record's reported age is one year after the childhood record, matches where the adult record's reported age is two years before the childhood record, and matches where the adult record's reported age is two years after the childhood record.

Step 2: After this first set of matches, I repeat Step 1, but instead of looking for NYSIIS codes corresponding to each respondent's name, I match exactly using standardized first and last names, where I standardize nicknames and remove any remaining middle initials. Lastly, I repeat Step 1 using the raw name strings in the census.

Step 3: I then repeat Steps 1 and 2 without the restriction that records match exactly on race. This allows for changes in reported race between the childhood and adult census year.

Step 4: I then perform three more searches for matches: looking for exact matches on (1) NYSIIS first/last name, then (2) cleaned name, and then (3) raw name, along with birth state/country, and race (no age restriction, to allow for transcription errors in the age variable).

Step 5: Lastly, I repeat Step 4, relaxing the restriction that the childhood record match to the adult record on race.

I then append together the four sets of matches from 1900 to 1940, 1910 to 1940, 1920 to 1940, and 1930 to 1940. For each of the adult men in 1940, I keep the two highest-quality matches, based on the ordering defined in Steps 1–5 above. As a tiebreaker, I always choose the earliest record that matched to an adult in 1940. So for example, if an adult male in 1940 matches exactly on NYSIIS first/last name, age, state of birth, and race to a record in each of 1900, 1910, 1920, and 1930, I will keep the 1900 and 1910 records as potential matches.

For each adult record that matched to at least one childhood record, I calculate two measures of match accuracy:

1. Parental Name Accuracy: For adult records in 1940 that matched to two childhood records (A and B), I compare the reported parents of A and B in the early censuses. If the first two letters of A's mother and B's mother are the same or the first two letters of A's father and B's

father are the same, I call this a high-quality match, because it is very unlikely to happen by chance unless A and B represent the same person.

2. Birthplace Accuracy: In 1940, a subset of adults were asked to report the birthplace of their mother and father. So, for all adult records in 1940 that match to at least one childhood record, I compare the reported birthplace of the adult in 1940 with the childhood record that matched to the adult record. If the reported birthplaces of the records' mother or father match across censuses, I call this a high-quality match.

Below, I present a table showing the fraction of potential matches from the first quality check, for records that matched to two childhood census records. These 21,364,440 records represent 10,682,220 adult males who each matched to two childhood census records (from two separate decennial census years between 1900–1930) such that both childhood census records have at least one parent in the household with a non-missing name. For each of the 36 iterations discussed above, I report the fraction of these matched childhood-adult record tuples that have at least one parent (mother or father) whose first two letters of their first names are the same across the two childhood records in the tuple. I also present a table describing the second measure of match accuracy.

Table 3.1: Parental Name Accuracy Measure by Loop Iteration

Iteration	# Low-Quality	% Low-Quality	# High-Quality	% High-Quality	Total
1	2,647,659	26	7,529,692	74	10,177,351
2	855,068	35	1,601,099	65	2,456,167
3	637,356	28	1,622,583	72	2,259,939
4	388,385	51	368,776	49	757,161
5	260,936	52	244,288	48	505,224
6	355,820	41	514,657	59	870,477
7	108,645	53	94,549	47	203,194
8	79,033	47	87,882	53	166,915
9	55,407	70	23,626	30	79,033
10	39,788	71	16,122	29	55,910
11	194,585	52	177,261	48	371,846
12	61,608	63	36,915	37	98,523
13	47,120	59	32,473	41	79,593
14	33,660	74	11,796	26	45,456
15	24,723	75	8,215	25	32,938
16	148,035	64	82,435	36	230,470
17	107,070	72	41,018	28	148,088
18	80,409	71	32,711	29	113,120
19	79,217	78	21,792	22	101,009
20	56,350	78	15,611	22	71,961
21	12,621	76	4,063	24	16,684
22	7,786	81	1,799	19	9,585
23	5,757	80	1,438	20	7,195
24	5,442	85	985	15	6,427
25	4,345	86	726	14	5,071
26	4,187	83	884	17	5,071
27	3,672	84	696	16	4,368
28	3,013	84	590	16	3,603
29	3,210	86	516	14	3,726
30	2,454	86	399	14	2,853
31	1,013,371	67	508,752	33	1,522,123
32	315,203	79	82,932	21	398,135
33	236,460	83	49,605	17	286,065
34	166,011	80	40,809	20	206,820
35	34,586	87	5,086	13	39,672
36	20,004	88	2,663	12	22,667
Total	8,098,996	38	13,265,444	62	21,364,440

Table 3.2: Birthplace Accuracy Measure by Loop Iteration

Iteration	# Low-Quality	% Low-Quality	# High-Quality	% High-Quality	Total
1	643,931	11	5,203,196	89	5,847,127
2	231,149	20	926,495	80	1,157,644
3	175,153	16	954,147	84	1,129,300
4	112,087	30	262,979	70	375,066
5	77,548	31	176,493	69	254,041
6	76,934	13	505,112	87	582,046
7	26,405	24	82,484	76	108,889
8	18,962	20	77,911	80	96,873
9	15,287	34	30,031	66	45,318
10	11,327	34	21,748	66	33,075
11	44,061	16	224,098	84	268,159
12	14,549	26	41,389	74	55,938
13	10,936	22	37,786	78	48,722
14	9,222	33	18,816	67	28,038
15	7,095	33	14,109	67	21,204
16	50,073	28	126,433	72	176,506
17	29,779	33	59,342	67	89,121
18	21,548	32	45,054	68	66,602
19	20,336	35	37,542	65	57,878
20	14,963	36	26,791	64	41,754
21	4,347	29	10,877	71	15,224
22	2,358	33	4,778	67	7,136
23	1,688	32	3,649	68	5,337
24	1,531	33	3,055	67	4,586
25	1,260	34	2,448	66	3,708
26	935	27	2,473	73	3,408
27	868	30	2,043	70	2,911
28	663	27	1,790	73	2,453
29	679	27	1,859	73	2,538
30	635	30	1,517	70	2,152
31	423,746	39	655,240	61	1,078,986
32	120,088	40	181,679	60	301,767
33	79,616	36	139,186	64	218,802
34	58,668	35	108,112	65	166,780
35	10,928	35	20,452	65	31,380
36	5,104	28	13,043	72	18,147
Total	2,324,459	19	10,024,157	81	12,348,616

I now present the same table for the first measure of accuracy, but instead of examining all potential matches, I focus on pairs of childhood records in two decennial censuses from 1900–1930 that each matched to the same adult record in 1940 where each of the two childhood records independently matched to the adult record in the same iteration of the loop described above.

Table 3.3: Parental Name Accuracy Measure by Loop Iteration for Matches from the Same Loop

Iteration	# Low-Quality	% Low-Quality	# High-Quality	% High-Quality	Total
1	919,616	14	5,664,724	86	6,584,340
2	156,606	17	770,426	83	927,032
3	95,132	13	639,040	87	734,172
4	36,646	18	170,702	82	207,348
5	19,410	16	98,548	84	117,958
6	77,994	21	297,204	79	375,198
7	10,232	25	30,668	75	40,900
8	5,160	19	22,292	81	27,452
9	2,572	26	7,308	74	9,880
10	1,512	25	4,506	75	6,018
11	46,146	34	90,660	66	136,806
12	6,804	38	11,206	62	18,010
13	3,886	33	7,736	67	11,622
14	2,090	35	3,862	65	5,952
15	1,178	32	2,554	68	3,732
16	15,758	35	29,294	65	45,052
17	5,674	31	12,520	69	18,194
18	3,616	29	8,982	71	12,598
19	3,270	29	8,024	71	11,294
20	1,992	26	5,714	74	7,706
21	710	48	782	52	1,492
22	188	47	216	53	404
23	96	35	182	65	278
24	72	39	112	61	184
25	64	46	74	54	138
26	192	47	216	53	408
27	122	46	144	54	266
28	82	39	130	61	212
29	100	49	106	51	206
30	74	50	74	50	148
31	288,988	45	356,608	55	645,596
32	39,188	55	31,798	45	70,986
33	35,844	66	18,262	34	54,106
34	25,976	51	24,964	49	50,940
35	1,910	67	934	33	2,844
36	1,698	70	730	30	2,428
Total	1,810,598	18	8,321,302	82	10,131,900

To interpret this table, consider the first row, which shows that 6,584,340 potential childhood records (for 3,292,170 adult records) matched in the first iteration of the loop described above. 86% of these tuples have two childhood records whose fathers or mothers have the same first two letters of their first names in each of the childhood census records. If false positives in this set of matches are distributed independently across potential matches for all potential matches within a loop iteration, and if this measure of accuracy correctly identifies valid matches, then $0.86^{0.5} =$ 93% of matches in the first iteration of the loop are accurate. I use the tabulation from Table 3.3 to calculate accuracy measures for each iteration of the looped matching algorithm described above. Accuracy ranges from 93% for the first iteration of the loop down to $0.30^{0.5} = 55\%$ for the last (36th) iteration of the loop.

Lastly, I select one childhood census record for each adult census record in 1940 that matched to two childhood census records. I prioritize childhood records in three steps:

1. I first prioritize 0–15 year old childhood census records over 16–25 childhood census records

2. , I then prioritize the earliest childhood census record if a pair of census records matched according to the parental name accuracy measure described above

3. I then prioritize matches from the first 20 iterations of the loop described above, because those matches are on average more accuracy than matches from loop iterations 21–36

4. Lastly, I prioritize the earliest match (prioritizing matches from 1900 over matches from 1910 and so on)

Below, I show which matches I keep for the final analysis sample using this prioritization of matches where there were multiple childhood records to choose from.

Table 3.4: Potential Matches for the Final Analysis Sample

Iteration	# Drop	% Drop	# Keep	% Keep	Total
1	5,846,785	38	9,406,353	62	15,253,138
2	1,349,945	34	2,612,177	66	3,962,122
3	1,403,608	39	2,231,742	61	3,635,350
4	393,326	30	912,405	70	1,305,731
5	365,490	38	605,397	62	970,887
6	513,599	34	994,743	66	1,508,342
7	112,056	31	246,907	69	358,963
8	109,368	36	194,999	64	304,367
9	38,142	27	103,068	73	141,210
10	43,492	37	74,093	63	117,585
11	232,207	29	568,947	71	801,154
12	56,420	28	148,073	72	204,493
13	56,214	33	115,342	67	171,556
14	22,442	24	70,060	76	92,502
15	26,630	34	52,860	66	79,490
16	135,609	27	366,944	73	502,553
17	81,372	29	195,423	71	276,795
18	78,107	35	147,692	65	225,799
19	48,899	27	131,797	73	180,696
20	55,674	37	95,431	63	151,105
21	13,261	29	32,964	71	46,225
22	7,622	34	14,719	66	22,341
23	6,433	37	10,947	63	17,380
24	4,868	36	8,477	64	13,345
25	5,016	40	7,414	60	12,430
26	4,437	35	8,135	65	12,572
27	3,701	35	6,910	65	10,611
28	3,345	36	6,037	64	9,382
29	3,045	35	5,548	65	8,593
30	2,910	38	4,807	62	7,717
31	988,681	36	1,770,324	64	2,759,005
32	267,271	38	437,845	62	705,116
33	195,571	36	344,787	64	540,358
34	140,320	35	255,775	65	396,095
35	27,268	38	45,025	62	72,293
36	15,366	37	26,449	63	41,815
Total	12,658,500	36	22,260,616	64	34,919,116

From the 22,260,616 childhood records at this, I drop an additional three records that do not have a valid state of residence in the childhood census data. And I drop 4,442,495 matches that were low-quality according to either of the two accuracy measures defined above.[2] This leaves me with 17,818,118 records in my final analysis sample out of 39,175,147 of the 20–65 year old men in 1940 who I attempted to match to 0–25 year old males in the 1900–1930 decennial censuses. This represents a match rate of 45.5%. And based on the accuracy estimates from Table 3.3, 80% of these matches are true matches and 20% of these matches are false positives. This is towards the frontier of the methods analyzed in Abramitsky et al. (2020), although my calculation of accuracy measures is new in the literature, making comparisons difficult.

2. In Appendix Table 1.26 I show that my main educational attainment results are unchanged when I subset to matches that are accurate, according to one of the two measures described above. And in Appendix Table 1.27 I show that including these 4.4 million matches that are low-quality has only a small effect on my main results.

References

Rolf Aaberge, Magne Mogstad, and Vito Peragine. "Measuring long-term inequality of opportunity." *Journal of Public Economics*, Vol. 95, pp. 194-204, 2011.

Daniel Aaronson and Bhashkar Mazumder. "The Impact of Rosenwald Schools on Black Achievement." *Journal of Political Economy*, Vol. 119, No. 5, pp. 821-888, 2011.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. "Europe's Tired, Poor, and Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review*, Vol. 102, No. 5, pp. 1832-56, 2012.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy*, Vol. 122, No. 3, pp. 467-506, 2014.

Ran Abramitzky, Leah Platt Boustan, Katherine Eriksson, James J. Feigenbaum, and Santiago Pérez. "Automated Linking of Historical Data." *Journal of Economic Literature* , forthcoming.

Joseph E. Aldy. "Birds of a Feather: Estimating the Value of Statistical Life from Dual-Earner Families." *Journal of Risk and Uncertainty*, Vol. 58, pp. 187-205, 2019.

Joseph E. Aldy and W. Kip Viscusi "Adjusting the Value of a Statistical Life for Age and Cohort Effects". *Review of Economics and Statistics*, Vol. 90, pp. 573-581, 2008.

Joshua D. Angrist and Alan B. Krueger. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, Vol. 106, No. 4, pp. 979-1014, 1991.

Elizabeth Arias. "United States Life Tables, 2011." *National Vital Statistics Reports, 2015*, Vol. 64, No. 11, pp. 1-62, September 22, 2015.

Orley Ashenfelter and Michael Greenstone. "Using Mandated Speed Limits to Measure the Value of a Statistical Life." *Journal of Political Economy*, Vol. 112, pp. 226-267, 2004.

Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. "Top Incomes in the Long Run of History." *Journal of Economic Literature*, Vol. 49, No. 1, pp. 3-71, 2011.

Martha J. Bailey and William J. Collins. "The Wage Gains of African-American Women in the 1940s." *Journal of Economic History*, Vol. 66, No. 3, pp. 737-777, 2006.

Martha Bailey, Brian Jacob, Michael Kevane, and William Sundstrom. "Carnegie's Legacy and the Growth of American Cities: Did Public Libraries Have Measurable Effects?" Unpublished Data, 2011.

Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. "How Well do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of Economic Literature*, Vol. 58, No. 4, pp. 997-1044, 2020.

Jon Bakija. “Documentation for a Comprehensive Historical U.S Federal and State Income Tax Calculator Program.” Working Paper, June 2017.

Nathaniel Baum-Snow and Derek Neal. “Mismeasurement of Usual Hours Worked In the Census and ACS.” *Economics Letters*, Vol. 102, Is. 1, pp. 39-41, January 2009.

Patrick Bayer and Kerwin Kofi Charles. “Divergent Paths: A New Perspective on Earnings Differences between Black and White Men since 1940.” *Quarterly Journal of Economics*, Vol. 133, No. 3, pp. 1459-501, 2018.

Gary S. Becker, Tomas J. Philipson, Rodrigo R. Soares. “The Quantity and Quality of Life and the Evolution of World Inequality.” *American Economic Review*, Vol. 95, No. 1, pp. 277-291, March 2005.

Enrico Berkes and Peter Nencka. “Knowledge access: The effects of Carnegie libraries on innovation.” Working Paper, 2019.

Rachana Bhatt. “The Impact of Public Library Use on Reading, Television, and Academic Outcomes.” *Journal of Urban Economics*, Vol. 68, No. 2, pp. 148-166, 2010.

Dan Black, Yu-Chieh Hsu, Seth Sanders, and Lowell Taylor. “Combining Forward and Backward Estimates of Mortality.” Working Paper, 2014.

Dan Black, Yu-Chieh Hsu, Seth Sanders, Lynne Schofield, and Lowell Taylor. “Black-White Mortality Differentials at Older Ages: Inferential Difficulties and New Evidence.” Working Paper, 2016.

George Sylvan Bobinski. “Carnegie Libraries: Their History and Impact on American Public Library Development.” American Library Association, 1969.

Timm Bonke, Giacomo Corneo, and Holger Luthen. “Lifetime Earnings Inequality in Germany.” *Journal of Labor Economics*, Vol. 33, No. 1, pp. 171-208, 2015.

David Card and Alan B. Krueger. “Trends in Relative Black-White Earnings Revisited.” *American Economic Review: Papers and Proceedings*, Vol. 83, No. 2, pp. 85-91, 1993.

Andrew Carnegie. “The gospel of wealth and other timely essays.” Century, 1901.

Census Bureau. “Instructions to Enumerators, Population and Agriculture, 1940.” *1940 Census Publications*, Form PA-1.

Census Bureau. “Agriculture 1950 - A Graphic Summary.” *1950 Census Publications*, Vol. 5, No. 6, pp. 69-102, 1950.

Raj Chetty and Nathaniel Hendren. “The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects.” *Quarterly Journal of Economics*, Vol. 133, No. 3, pp. 1107-62, 2018.

Ansley J. Coale and W. Rives Norfleet Jr. "A Statistical Reconstruction of the Black Population of the United States 1880-1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility." *Population Index*, Vol. 39, No. 1, pp. 3-36, January 1973.

Sergio Correia. "A Feasible Estimator for Linear Models with Multi-Way Fixed Effects." Working Paper, 2016.

Hawthorne Daniel. "Public Libraries for Everyone: The Growth and Development of Library Services in the United States, Especially Since the Passage of the Library Services Act." Doubleday, 1961.

Chris Doucouliagos, T.D. Stanley, and Margaret Giles. "Are estimates of the value of a statistical life exaggerated?." *Journal of Health Economics*, Vol. 31, pp. 197-206, 2012.

Esther Duflo. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy." *American Economic Review*, Vol. 91, No. 4, pp. 795-813, 2001.

Katherine Eriksson, Gregory T. Niemesh, and Melissa Thomasson. "Revising Infant Mortality Rates for the Early 20th Century United States." *Demography*, Vol. 55, No. 6, pp. 2001-24, 2018.

James Feigenbaum. "Intergenerational Mobility during the Great Depression." *Working Paper*, 2015.

Joseph P. Ferrie. "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." *Historical Methods*, Vol. 29, No. 4, pp. 141-156, 1996.

Jennie M. Flexner. "Circulation Work in Public Libraries." American Library Association, 1927.

Daniel Gallardo-Albarran "Missed opportunities? Human welfare in Western Europe and the United States, 1913–1950." *Explorations in Economic History*, Vol. 72, pp. 57-73, April 2019.

A.P. Garbin and Frederick L. Bates. "Occupational Prestige and Its Correlates: A Re-examination." *Social Forces*, Vol. 44, No. 3, pp. 295-302, 1966.

Gregory Gilpin, Ezra Karger, and Peter Nencka. "The Returns to Public Library Investment." Working Paper, 2020.

Claudia Goldin. "How America Graduated from High School: 1910 to 1960." NBER Working Paper, June, 1994.

Claudia Goldin. "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century." *Journal of Economic History*, Vol. 58, No. 2, pp. 345-374, 1998.

Claudia Goldin. "Egalitarianism and the Returns to Education during the Great Transformation of American Education." *Journal of Political Economy*, Vol. 107, No. S6, pp. S65-S94, 1999.

Lillian Guralnick. “Mortality by Occupation and Cause of Death among Men 20 to 64 Years of Age: United States, 1950.” *Vital Statistics Special Reports*, Vol. 53, No. 3, September 1963

Fatih Guvenen, Greg Kaplan, Jae Song, and Justin Weidner. “Lifetime Incomes in the United States over Six Decades.” NBER Working Paper, April 2017.

Mark E. Hill and Ira Rosenwaike. “The Social Security Administration’s Death Master File: the completeness of death reporting at older ages.” *Social Security Bulletin*, Vol. 64, No. 1, pp. 45-51, 2001.

Robert W. Hodge, Paul M. Siegel and Peter H. Rossi. “Occupational Prestige in the United States, 1925-63.” *American Journal of Sociology*, Vol. 70, No. 3, pp. 286-302, 1964.

Anupam B. Jena, Tomas J. Philipson, and Eric C. Sun. “Health and Wealth Disparities in the United States.” The AEI Press, Washington D.C. 2010.

D. Gale Johnson. “Agriculture and the Wealth of Nations.” *American Economic Review*, Vol. 87, No. 2, pp. 1-12, 1997.

Ravi Kanbur and Diganta Mukherjee. “Premature Mortality and Poverty Measurement.” *Bulletin of Economic Research*, Vol. 59, Is. 4, pp. 339-359, 2007.

Michael J. Kevane and William A. Sundstrom. “The Development of Public Libraries in the United States, 1870–1930: A Quantitative Assessment.” *Information & Culture*, Vol. 49, No. 2, pp. 117-144, 2014.

Michael J. Kevane and William A. Sundstrom. “Public Libraries and Political Participation: 1870-1940.” Working Paper, 2016.

Eric Klinenberg. “Palaces for the people: How social infrastructure can help fight inequality, polarization, and the decline of civic life.” Broadway Books, 2018.

Trevon D. Logan and John M. Parman. “The National Rise in Residential Segregation.” *Journal of Economic History*, Vol. 77, No. 1, pp. 127-170, 2017.

Diana L. Magnuson and Miriam L. King. “Comparability of the Public Use Microdata Samples: Enumeration Procedures.” *Historical Methods*, Vol. 28, No. 1, pp. 27-32, 1995.

Robert Sidney Martin. “Carnegie Denied: Communities Rejecting Carnegie Library Construction Grants, 1898-1925.” Greenwood Press, 1993.

Kevin M. Murphy and Robert H. Topel. “Black-White Differences in the Economic Value of Improving Health.” *Perspectives in Biology and Medicine*, Vol. 48, No. 1, pp. 176-194, 2005.

Kevin M. Murphy and Robert H. Topel “The Value of Health and Longevity.” *Journal of Political Economy*, Vol. 114, No. 5, pp. 871-904, 2006.

Christopher A. Neilson and Seth D. Zimmerman. "The Effect of School Construction on Test Scores, School Enrollment, and Home Prices." *Journal of Public Economics*, Vol. 120, pp. 18-31, 2014.

Amir B Ferreira Neto. "Do Public Libraries Impact Local Labor Markets? Evidence from Appalachia." Working Paper, 2019.

Andrew Noymer and Michel Garenne. "The 1918 influenza epidemic's effects on sex differentials in mortality in the United States." *Population and Development Review*, Vol. 26, No. 3, pp. 565-581, 2000.

Philip Oreopoulos and Kjell G. Salvanes. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives*, Vol. 25, No. 1, pp. 159-184, 2011.

Maurizio Pisati. "SPMAP: Stata module to visualize spatial data." *Statistical Software Package*, 2018.

Samuel H. Preston and Michael R. Haines. "New Estimates of Child Mortality during the Late Nineteenth Century." Fatal Years: Child Mortality in Late Nineteenth-Century America, Chapter 2, pp. 50-87, 1991.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. "IPUMS USA: Version 10.0 [dataset]." Minneapolis, MN: IPUMS, 2020. `https://doi.org/10.18128/D010.V10.0`

Martin Saavedra and Tate Twinam. "A Machine Learning Approach to Improving Occupational Income Scores." *Explorations in Economic History*, Vol. 75, 2020.

Paul Mathew Siegel. "Prestige in the American Occupational Structure." Ph.D. dissertation, University of Chicago, 2018.

James P. Smith and Finis R. Welch. "Black Economic Progress After Myrdal." *Journal of Economic Literature*, Vol. 27, No. 2, pp. 519-564, June 1989.

Melvin Stephens Jr and Dou-Yan Yang. "Compulsory Education and the Benefits of Schooling." *American Economic Review*, Vol. 104, No. 6, pp. 1777-92, 2011.

Christopher R. Tamborini, Chang Hwan Kim, and Arthur Sakamoto. "Education and Lifetime Earnings in the United States." *Demography*, Vol. 52, Is. 4, pp. 1383-1407, August 2015.

Petra Todd. "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated." *Handbook of Development Economics*, Vol. 4, pp. 3848-94, 2008.