Black Economic Progress in the Jim Crow South: Evidence from Rosenwald Schools*

Shariq Mohammed
University of Michigan

Paul Mohnen
University of Michigan

April 6, 2021

Abstract

This paper explores the role of human capital and labor market barriers in driving the racial gap in occupational standing in the Jim Crow South. To shed light on this question, we study the labor market impact of the Rosenwald Schools Initiative, a large-scale school construction program aimed at improving educational opportunities for blacks in the rural South during the early 20th century. We build a novel dataset linking Social Security application records to the 1920 and 1940 U.S. Censuses, allowing us to estimate the impact of Rosenwald schools on outcomes in adulthood for both men and women. Using a triple difference strategy, we show that exposure to Rosenwald schools led to greater educational attainment among rural black men and women, consistent with prior work. We find that these gains led to greater labor force participation among black women, and a higher propensity to work in white collar jobs among black men. However, we find no evidence that black men broke into white collar jobs that involved interacting with whites (e.g. sales jobs), in line with accounts that they tended to be excluded from those jobs. Our findings suggest that while human capital barriers played an important role, labor market discrimination limited the gains to education for blacks.

JEL: I24, J71, N32

*For helpful comments and discussions, we thank Dan Aaronson, Martha Bailey, Elizabeth Cascio, Dan Fetter, Price Fishback, Sasha Indarte, Gary Solon, Gavin Wright, seminar participants at the University of Arizona, University of Michigan, Wharton (BEPP), and conference participants at the American Economic Association, Economic History Association, and Social Science History Association Annual Meetings, and Equitable Growth. We thank Jessica Tinor for excellent research assistance. We gratefully acknowledge financial support from the Population Studies Center at the University of Michigan through a Small Grant Award, funded by the Marshall Weinberg Summer Research Prize Fund (G002832), and the use of the Population Studies Center's services and facilities (P2C HD041028). This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

1 Introduction

A vast literature has been dedicated to understanding the determinants of black economic progress over the 20th century, particularly the large gains that occurred after 1940. A central question has been the relative importance of improvements in black educational attainment, both in terms of quantity and quality, as opposed to reductions in institutional barriers favoring white workers over black workers (Smith and Welch, 1989; Donohue and Heckman, 1991). Beyond its historical significance, the answer to this question is relevant to present-day policies seeking to reduce the black-white wage gap.

One important setting in which this debate has played out is the early 20th century South, where the majority of blacks lived at the time. During this era, schools in the South were racially segregated under the "separate-but-equal" doctrine. As a result of public sector discrimination, black schools were severely underfunded, and many black children lacked access to schooling altogether. Based on time series evidence and decomposition exercises showing that measures of human capital can explain a substantial fraction of the black-white earnings gap among early 20th century cohorts, some studies have attributed a predominant role to racial gaps in schooling resources (Smith, 1984; Carruthers and Wanamaker, 2017). On the other hand, motivated in part by contemporary accounts that blacks faced substantial discrimination in the labor market (e.g. Myrdal, 1944), others have argued that truly equal schools would not have led to parity in the labor market (Wright, 1986; Margo, 1990).

This paper brings new evidence to this debate by exploiting a natural experiment which differentially improved access to quality education for black children in the South: the Rosenwald Schools Initiative. The Rosenwald Schools Initiative was a large-scale school construction program supported by the Rosenwald Fund, one of several private philanthropies advancing black education in the South (Donohue et al., 2002). Between 1913 and 1932, it helped fund the construction of nearly 5,000 schools in predominantly rural areas across 15 Southern states. By the end of the program in 1932, the total capacity of Rosenwald schools was around one third of the black school-age population in the rural South, making it the largest educational intervention of its kind at the time. Building on prior work showing that Rosenwald schools had a positive impact on black educational outcomes (Aaronson and Mazumder, 2011), we explore the extent to which these gains translated into better labor market outcomes, which is informative about the relative importance of human capital and labor market barriers.

Key to answering this question is the ability to observe individuals in childhood to determine their likely exposure to Rosenwald schools, and again in adulthood to measure their labor market outcomes. We build a new linked dataset which uses Social Security application records as an intermediate link between the 1920 and 1940 Censuses. Exploiting these administrative records has two main advantages. First, it allows us to link women across Censuses thanks to the availability of both married and birth names in Social Security application records. Second, key linking variables such as names and time and place of birth information are more accurately reported in administrative records than in Census records, which helps increase the number and quality of matches we are able to make.

We make links between Social Security application records and Census records using a supervised machine learning (ML) approach. Assessing whether two different records correspond to the same person based on limited and imperfect information such as names is a complex decision problem where human input is invaluable. The ML approach involves making manual linking decisions for a small random sample of records and "training" an algorithm to make similar decisions for millions of records. The combination of high-quality linking information in Social Security application records and our supervised ML approach allows us to achieve high match rates while fixing the error rate at 3 percent (relative to human trainers). Our dataset contains around half a million individuals born in the South between 1906 and 1918, which can be found in Social Security application records and linked to both the 1920 and 1940 Censuses.

To estimate the causal effect of Rosenwald schools, we leverage cross-county and time variation in the spread of Rosenwald schools across the South, as well as variation in the size of these schools. As a result, black children were differentially exposed to Rosenwald schools depending on the capacity of Rosenwald schools in the county they grew up in (if any), during the years in which they were of school-going age. More specifically, we compare 1940 outcomes of individuals who were differentially exposed to Rosenwald schools based on their age and county of residence in 1920. We adopt the triple difference framework from Aaronson and Mazumder (2011) which relies on the fact that Rosenwald schools explicitly targeted blacks in rural areas, allowing us to exploit whites and urban blacks as control groups for rural blacks, where rural status is measured in 1920. This allows us to difference out unobservable rural or black-specific factors that may be correlated with Rosenwald exposure and independently affect outcomes in 1940.

We begin by confirming that exposure to Rosenwald schools had a significant positive impact on the educational attainment of rural black men and women in 1940, consistent with Aaronson and Mazumder (2011)'s original results on black literacy and school attendance (based on pooled Census data), as well as their results on black male educational attainment (based on World War II enlistee records). We then examine whether greater educational attainment led to better labor market outcomes. In the tradition of Heckman and Payner (1989) and Margo (1990), our analysis of labor market standing centers around occupations rather than income. Valuable insights can be obtained by studying the nature of occupational gains and losses, insights which would otherwise be lost by focusing on summary income measures. Moreover, a major drawback of the income question in the 1940 Census is that it only asked about wage income and therefore paints an incomplete picture of total income, which comprised farm and business income for many workers.

Despite the significant gains in years of education, we find that blacks exposed to Rosenwald schools did not experience corresponding gains in average occupational standing in 1940. This masks some gains in white collar jobs, particularly among black men. However, we show that the gains in white collar jobs were concentrated in a few specific occupations that did not involve interacting with people (e.g. stenographers, typists), or mainly involved interacting with black customers (e.g. self-employed store owners). This finding is in line with accounts that blacks tended to be excluded from jobs that involved interacting with white co-workers or white customers, such as sales jobs (Goldin, 1990; Margo, 1990). As a result, many blacks exposed to Rosenwald schools instead ended up in low-skill service jobs for which they were overqualified (e.g. waiters, maids, cooks), offsetting the gains in white collar jobs. We also show that the gains in white collar jobs were particularly large for black men whose fathers were themselves white collar workers in 1920, highlighting a complementarity between access to education and family background.

At the extensive margin, we find that exposure to Rosenwald schools had a positive effect on black women's labor force participation in 1940. Correspondingly, and in line with Aaronson et al. (2014), we find that black women exposed to Rosenwald schools were less likely to be married or have children in 1940, suggesting that greater opportunities in the labor market induced black women to delay marriage and fertility. We also find that black men exposed to Rosenwald schools were less likely to be married in 1940, and provide suggestive evidence that this reflects thinner marriage markets due to greater female labor force participation in Rosenwald counties. Lastly, we show that our male findings are robust to using alternative Census linked datasets which do not rely on Social Security application records, restricting to individuals born in their 1920 county of residence (to address endogenous location choice concerns), and different ways of re-weighting the linked sample to make it representative of the population.

Overall, the fact that Rosenwald schools enabled some black men to obtain white collar jobs, in which they were heavily underrepresented, suggests that poor schooling opportunities hindered blacks' prospects in the labor market. On the other hand, the fact that they failed to break into white collar jobs that involved interacting with whites suggests that labor market discrimination limited the gains to education for blacks. Ultimately, we conclude that while barriers to human capital accumulation played a role, institutional barriers in the labor market fundamentally restricted the scope of black economic progress in the early 20th century South, echoing the central conclusion in Margo (1990).

The paper is organized as follows. Section 2 provides background information on the Rosenwald Schools Initiative. In Section 3, we outline the construction of our new linked dataset. Section 4 describes the triple difference strategy used to estimate the effects of exposure to Rosenwald schools on outcomes in 1940, which are presented in Section 5. Section 6 concludes.

2 The Rosenwald Rural Schools Initiative

In 1917, Chicago industrialist and philanthropist Julius Rosenwald established the Rosenwald Fund. One of the largest projects supported by the Fund was a school construction program called the Rosenwald Rural Schools Initiative. Its goal was to improve educational opportunities for black children in the rural South, which were severely lacking at the time.[1] The initiative originated from a partnership between Julius Rosenwald and Booker T. Washington, the principal of Tuskegee Institute in Alabama. Washington, who viewed the advancement of black education as the best path towards achieving black economic progress in the South, convinced Rosenwald to build 6 pilot schools in Alabama around 1913.[2] Between 1914 and 1919, Rosenwald helped fund the construction of another 700 schools across the South, primarily in Alabama, Louisiana, Tennessee, Kentucky, North Carolina, and Virginia. The program quickly expanded during the 1920s, resulting in the construction of nearly 5,000 schools across 15 states by the end of the program in 1932. The spread of Rosenwald schools over time is shown in Figure 1. The schools were erected at a total cost of $28.4 million (nominal terms), $4.7 million of which was contributed by the Rosenwald Fund (Embree and Waxman, 1949).

The Rosenwald Rural Schools Initiative had several key features. First, Rosenwald schools mainly provided primary education (grades 1 to 8), given that average educational attainment among rural blacks was very low at the time.[3] Second, the program was deliberately designed to provide physical resources, such as buildings, that could not easily be expropriated by predominantly white local educational authorities. Third, in addition to expanding access to education, the Rosenwald Fund placed a strong emphasis on providing high-quality education. Rosenwald schools were modernly-designed and had all the necessary amenities (e.g. lighting, ventilation, sanitation) and supplies (e.g. books, blackboards, desks) conducive to a good learning environment. Although schools varied in size, ranging from single-classroom schoolhouses to large buildings containing 22 classrooms, the Rosenwald Fund insisted that schools be built according to pre-specified blueprints. It also invested in complementary measures such as teacher salaries, the construction of teacher homes, and supported efforts to increase the length of school terms. Another measure aimed at improving the quality of instruction was the partnership with the Jeanes Fund, under which so-called "Jeanes supervisors" would travel from school to school to train

[1] The schooling system in the South was characterized by large racial gaps in expenditures-per-student, school term lengths, student-to-teacher ratios, and average teacher salaries (Margo, 1990).

[2] Washington believed that challenging Jim Crow segregation laws was unwise, and instead urged the black community to fight for better educational opportunities. The alternative view, held by other prominent black activists such as W. E. B. Du Bois, one of the founders of the National Association for the Advancement of Colored People (NAACP), was to settle for nothing less than equal rights. While Washington's incremental progress approach initially dominated, the latter view eventually prevailed, culminating in the civil rights movement.

[3] Average years of education among black men and women born in the South between 1900 and 1905 was only around 5 to 6 years.

teachers (Donohue et al., 2002).

A unique feature of the Rosenwald Initiative was its matching grant funding scheme: it required local communities to bear the majority of the cost of building a new school. On average, the Rosenwald Fund only covered around 20 percent of the total cost, with local blacks contributing 24 percent, local whites 5 percent, and local public authorities the remaining 51 percent.[4] The purpose of this scheme was to stimulate public interest in the provision of black education, foster cooperation between different parties within the local community, and only build schools in communities that had demonstrated a strong interest in the school's long-term success. Another key requirement was that the school become part of the public school system. While the Rosenwald Fund facilitated the initial construction of the school, the state was to be responsible for its continued operation by covering teacher salaries and other expenses.

Numerous studies have explored the legacy of Rosenwald schools, including the impact of Rosenwald schools on educational outcomes (Aaronson and Mazumder, 2011), health outcomes (Frisvold and Golberstein, 2013), fertility (Aaronson et al., 2014), incarceration (Eriksson, 2020), and migration (Aaronson et al., forthcoming). In this paper, we fill an important gap in this literature by studying the impact of Rosenwald schools on labor market outcomes. By assessing whether blacks exposed to Rosenwald schools were eventually able to obtain better jobs, we shed light on two important questions. First, as mentioned in the introduction, the magnitude and nature of labor market gains is informative about the relative importance of human capital vs. labor market barriers in explaining the racial gap in occupational standing in the early 20th century South. Second, our findings can be viewed as a test of whether the Rosenwald Schools Initiative ultimately succeeded in lifting blacks' economic opportunities, as Booker T. Washington had envisioned.

3 Data

3.1 SS-5 Records

The ideal dataset for our research question would link all individuals between the 1920 Census, in which we can observe individuals' county of residence in childhood and determine likely exposure to Rosenwald schools, and the 1940 Census, in which we can observe individuals' outcomes in adulthood. Instead of the usual approach of linking Census records together directly, we use Social Security application records as an intermediate link between the two Censuses. More specifically, we exploit the public version of the Social Security Numerical Identification (NUMIDENT) File, released by the National Archives and Records Administration. This file includes over 72 million

[4] The median school construction cost was $3,200 (nominal terms). Local communities often provided non-monetary contributions as well, in the form of land and labor.

entries of form SS-5 information (Application for a Social Security Card), corresponding to around 40 million unique individuals who died prior to 2007 and whose deaths were not state-reported. Each SS-5 record lists the number holder's full name, date of birth, sex, race and place of birth. Crucially, it also includes the number holder's parents' full names.

Using SS-5 records as an intermediate link between the 1920 and 1940 Censuses has two important advantages. First, it allows us to link women between Censuses. Women have largely been excluded from Census-to-Census linked datasets because they typically adopt their husband's last name when they get married. This makes it impossible to link women who appear under their birth name (or "maiden" name) as children but under their married name as adults, precluding a comprehensive analysis of the impact of childhood conditions on outcomes in adulthood for women. In contrast, SS-5 records for married women list both their married name (own last name field) and their birth name (father last name field).[5] Importantly, even though women participated in the labor market at lower rates than men historically, the SS-5 data contains the spouses of primary earners as surviving spouses received benefits after the 1939 Amendments to the Social Security Act. As a result, the proportion of men and women in SS-5 records is roughly equal.

The second advantage of exploiting SS-5 records is that key linking variables (names, time of birth, place of birth) are more accurately reported than in Census records. Information in SS-5 records is reported by an individual about him or herself, whereas Census enumerators often interviewed one member of a household about everyone living in it, and might have mis-transcribed the information provided to them (especially names). In addition, individuals have an incentive to report personal information accurately in administrative records since this information is directly tied to Social Security and other benefits, unlike in the Census. For example, middle names, which are crucial for disambiguation in name-based linking, are much more complete in SS-5 records. Only 19 percent of SS-5 entries have an empty middle name, and 62 percent of SS-5 entries have a middle name of at least 4 characters (relative to 69 percent and 2 percent in the 1940 Census respectively). Census records also notoriously suffer from age heaping, and place of birth information is not always accurately reported (especially if members of the household are born in different states or countries).

There are two main drawbacks to using SS-5 records as an intermediate link between Census records. First, SS-5 records do not have universal coverage. By definition, individuals who never applied for a Social Security number are naturally absent in the SS-5 data.[6] In addition, individuals

[5]Women have 1.8 entries in the SS-5 data on average as they tend to file a new form SS-5 upon (re-)marriage to update their name information. SS-5 records also include an "other name" field, which is often the birth name for married women. As a result, 94 percent of women have at least one last name which differs from their father's last name, which gives us a chance to find them in the Census regardless of their marital status at that moment in time.

[6]This could be problematic as some occupations, such as most self-employed workers, were not covered by Social Security until the 1950 Amendments. However, we note that the SS-5 data contains individuals who *ever* applied for a Social Security number, which will include individuals who eventually transitioned into covered employment.

who were still alive in 2007 or whose deaths were reported to the Social Security Administration by the states are excluded from the public version of the NUMIDENT. The NUMIDENT has good coverage of deaths that occurred from 1988 onwards and very low coverage prior to that, which to the best of our knowledge is due to the fact that most deaths prior to 1988 were reported by the states. Figure O.1 in the Online Appendix illustrates the coverage rate of SS-5 records by cohort for individuals born in the South, as implied by 1940 Census population counts by age. Coverage for the 1900-1910 cohorts is low, as many of these individuals died before 1988. Coverage then sharply increases as the state-reported death restriction loses bite, before steadily declining as individuals from later cohorts are more likely to be alive in 2007. Coverage by race is roughly even, and women are slightly overrepresented in earlier cohorts and underrepresented in later cohorts as they tend to live slightly longer than men.

Despite the uneven cohort coverage, SS-5 records have excellent geographic coverage. Panel A in Online Appendix Figure O.2 plots state of birth shares in SS-5 records against corresponding shares in the 1940 Census, separately by gender and race. Panel B plots state of death shares among the subset of SS-5 records which can be found in the NUMIDENT death files (for which we know place of death), against corresponding state of *residence* shares in the 1970 and 1980 Censuses (Ruggles et al., 2020), which proxies for place of death. In both panels, we focus on individuals born in the South between 1906 and 1918, which are the cohorts of interest in our analysis. In both graphs, the shares all lie very close to the 45 degree line for all demographic groups. This implies that SS-5 records do not disproportionately cover individuals who were born/died in specific states.

The second drawback of using SS-5 records as an intermediate link between Censuses is that requiring two links instead of one naturally reduces the size of the linked sample. This is partly mitigated by three factors: (1) the higher quality of SS-5 records, (2) linkages between SS-5 records and the 1920 Census are household-level links rather than individual-level links (which will result in higher match rates), and (3) the supervised ML linking approach we adopt. Nevertheless, in Section 5.6 we show that our findings for males are robust to using publicly available Census-to-Census linked datasets which do not rely on SS-5 records.

3.2 Linking SS-5 Records to Census Records using Machine Learning

Linking Method

Our dataset consists of a series of linkages between SS-5 and Census records. Each of these involves linking records mainly based on names, which is notoriously challenging due to common names and name misspellings. A common approach is to use deterministic algorithms, essentially matching on exact names or phonetic codes when certain conditions are met (e.g. Abramitzky et al., 2012). The appeal of these methods is that they are easy to implement, computationally cheap, and transparent. The downside is that they tend to be characterized by high rates of false matches, which can ultimately affect the downstream empirical analysis in unknown ways (Bailey et al., forthcoming).

In this paper, we adopt a supervised machine learning approach in which we first make manual linking decisions for a random sample of records, which is then used as input data to train an algorithm to make similar linking decisions for millions of records.[7] The motivation for this approach is that human input is invaluable in assessing whether to link two records. Rather than examining pairs of records, human trainers examine an *entire set* of potential candidates (which we will refer to as "potentials") for a particular record (which we will refer to as a "primary"), and decide whether or not to make a link and which link to make. This is crucial because there are often multiple plausible candidates, and choosing between them or deciding there is too much ambiguity to do so is not something that can easily be automated. In addition, human trainers are adept at identifying possible name misspellings, nicknames, or inverted first and middle names.

The resulting hand-linked data is then used to discipline a machine learning algorithm to mimic human decisions. In this paper, we use the two-stage model developed by Murray et al. (2020) specifically for one-to-one matching based on names. In the first stage, the probability that there exists a link in the set of potential candidates is modeled using a random forest model.[8] The random forest learns to label entire sets as containing a link or not based on the training data and *set-level* features provided by the modeler. With multiple potentials and multiple linking fields, the number of possible features is very large, but examples include average/maximum string distance scores, the difference between the highest and second-highest score, name commonality scores for the primary, and indicators for how many potentials match the primary along various dimensions. In the second stage, the probability that any link in the set is the true link conditional on there being a link in the set is modeled using a logit-style discrete choice model. In contrast to the first-stage model, the features in the second-stage model are *pair-level* features between the primary and each potential (e.g. string distance scores, indicators for matching age/race). The parameters of the logit model are estimated using maximum likelihood on the subset of primaries for which a link was found in the training data.

To generate the stage 1 and stage 2 probabilities, we use 10-fold cross-validation.[9] The resulting probabilities are then multiplied to obtain the unconditional match probability for every primary-potential pair. The model-based links are determined using a threshold rule: potentials

[7]The general approach described in the section comes from the Longitudinal, Intergenerational Family Electronic Micro-Database (LIFE-M) Project (`https://life-m.org/`).

[8]For more details on random forests, see Chapter 15 in Friedman et al. (2009).

[9]10-fold cross-validation involves splitting the training data into 10 equal-sized "folds," and using 9 folds to generate predicted probabilities for the remaining fold. This process is repeated 10 times, each time varying the fold under consideration, such that the resulting probabilities are not fitted on the data itself.

that have a match probability exceeding the threshold are called links. A key decision is how to set this threshold. In general, the lower the threshold the higher match rate but the lower the agreement rate with human trainers. Conversely, the higher the threshold the lower the match rate but the higher the agreement rate. This creates a natural trade-off between making more links and fewer false links. The advantage of using training data is that it allows us to control this trade-off. For each possible threshold on a grid between 0 and 1, we compute the share of trainer links that the model was able to reproduce ("recall rate") and the share of model links that agreed with the trainers ("precision rate"). Plotting these points in a two-dimensional space traces a precision-recall frontier, from which we can choose any point. For all our linkages, we choose the threshold associated with a 97 percent precision rate. Using the trained model and the cross-validated probability threshold, we then scale linking to the full sample of records. The match rate achieved by the model is roughly equal to the trainer match rate multiplied by the model recall rate.[10]

One valid concern with this type of supervised machine learning approach is that the quality of the model links is only as good as the quality of the human links. In other words, the precision rate targeted by the model is relative to human decisions, not the truth. We mitigate this in two ways: (1) by instructing trainers only to make links when they are very confident, and (2) by aggregating decisions from multiple trainers, as links that are independently chosen by different trainers are more likely to be objectively good links.[11]

Linking Sequence

We construct our linked dataset in three steps. This section briefly describes each step, but more extensive details can be found in Appendix B. We first link men in SS-5 records to individuals in the 1940 Census based on full names, age in 1940, and state/country of birth. Potentials in the Census are generated by blocking on sex, state/continent of birth, a +/-3 year window around the primary's age in 1940, and first or last name initial. We then select the top 25 candidates based on a name similarity score, which is the set trainers observe when making linking decisions. The linking process for women is analogous except that we try to link them twice: once using their birth name and once using their married name. To generate potential candidates, we additionally block on marital status (i.e. women who are linked using their birth/married name are linked to single/non-single women in the Census).

[10]Note that model performance hinges on the features chosen by the modeler. By designing features that enable the model to better approximate human decisions, we can push the precision-recall frontier outwards.

[11]Following the LIFE-M training process, two trainers first make linking decisions for the full random sample of primaries. The decisions of the two original trainers are then compared, and cases for which they disagree on are sent to three additional trainers (mixed in with a random sample of cases for which there was agreement). Only links for which the two original trainers agreed on, or links chosen by four out of five trainers are considered as links for the purpose of training the machine learning algorithm. Our current linking results are based on one to two trainers only, but the full process just described is ongoing.

Next, we link siblings in SS-5 records together based on parent names and children's year and place of birth. Note that SS-5 records contain the birth names of both parents, which makes it easier to confidently identify siblings. Potentials candidates are generated by blocking on first two initials of father last name, and a +/-10 year window around the primary's year of birth. We then select the top 20 candidates based on a parent name similarity score. Here, trainers are allowed to select multiple links as individuals can have multiple siblings. Nevertheless, we treat trainers' decisions as if they were examining each primary-potential pair separately, and use a random forest model to model these decisions. The resulting sibling linkages are used to create a crosswalk between SSNs and family IDs in the SS-5 data (see Appendix B for details).

Lastly, the reconstituted families from sibling linkages are linked to households in the 1920 Census based on parent names (except mother last name, which typically matches the father last name in the Census), and information on children (first and middle name, age in 1920, and place of birth). To generate potential candidates, we block on the union of children's state/continent of birth and father last name initial, and select the top 20 candidates based on a parent name similarity score. Unlike SS-5-1940 Census linking which is based on relatively little information, there is less ambiguity in household linking. This makes it easier for trainers to confidently choose between potential candidates (i.e. higher trainer match rate), and for the model to approximate trainers' decisions as they are more straightforward (i.e. higher recall rate).

We restrict our linked sample to black and white men and women who were born between 1906 and 1918 in one of 14 Southern states where Rosenwald schools were built.[12] We chose these cohorts as they were of school-going age during the Rosenwald era, and were also old enough in 1940 to measure their labor market outcomes.

Linked Sample

Table 1 presents our linking results. The first row in Panel A shows the starting number of individuals in SS-5 records that fall into our population of interest. We start with around 800,000 white men, 275,000 black men, 1.1 million white women, and 350,000 black women. The fifth row shows that we are ultimately able to link around 25 percent of white men, 22 percent of white women, 11 percent of black men, and 10 percent of black women to both the 1940 and 1920 Censuses at 97 percent precision. In the final row of Panel A, we also compute the effective coverage rate of our linked sample, defined as the number of links divided by the corresponding population in 1940 (bottom row of Table 1). Because many individuals are not contained in SS-5 records, the coverage rate of our linked sample is lower at around 6-7 percent for whites and 2.5 percent for

[12]The 14 states are: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, and Virginia. Missouri is excluded as only a handful of Rosenwald schools were built in that state at the very end of the program.

blacks. Our final sample contains around 500,000 individuals.

How do our match rates compare to existing linked datasets? Panel B shows match rates between the 1920 and 1940 Censuses from the IPUMS Multigenerational Longitudinal Panel (Helgertz et al., 2020b), which we will refer to as MLP. MLP also uses a supervised machine learning approach to generate links. There are two steps in the linking process. First, men are linked between adjacent Censuses as individuals. The main innovation of MLP is that it exploits rich contextual information contained in the Census as part of the linking process (e.g. county of residence, street address, members of household). The upside of doing this is that it reduces the ambiguity in the linking decisions, producing higher match rates. The downside is that the linked sample will naturally be skewed towards men who do not move between Censuses and have stable household structures. The consequences of this hidden selection depends on the question of interest. MLP also has a second stage which tries to link members living in the same household as men linked in the first stage. Among others, this makes it possible to link women, predominantly adult women living with their spouses and young daughters living with their fathers. The male MLP match rates in Panel B are similar to ours, at around 27 percent for white males, and 11 percent for black males.[13] However, the number of links and effective coverage rate is significantly higher in MLP because it links the entire U.S. population in 1920. The match rates for women are much lower at around 8 percent for white women and 4 percent for black women, and these links contain very few women that married between 1920 and 1940.

Panels C and D display match rates between the 1920 and 1940 Censuses from the Census Linking Project (Abramitzky et al., 2020), which we will refer to as CLP. CLP is based on the automated linking algorithm proposed in Abramitzky et al. (2012). Panel C shows results for their standard method. The match rate for white men is around 25 percent, and the match rate for black men is around 13 percent. Similar to MLP, they link all individuals in the 1920 Census, which implies that match rates and effective coverage rates are similar. Although match rates are similar, the accuracy of the links is likely significantly lower, as there is no training data to discipline the algorithm.[14] One way to increase the accuracy of links is to impose more stringent linking criteria. Panel D shows results for their conservative method, which requires names to be unique within a 5-year age band. The price to pay is a lower match rate (around 18 percent for white men and 7 percent for black men). Although the accuracy of these links is likely higher than using the standard method, it remains unknown. CLP does not link women.

Overall, our method yields match rates at the frontier of historical Census linking while targeting a high precision rate of 97 percent. In addition, we are able to link men and women at

[13]In interpreting these numbers, it is important to note that MLP targets a precision rate of 90 percent in the first stage and 97 percent in the second stage.

[14]Helgertz et al. (2020a) perform a manual check of the accuracy of links generated using this method, and report a precision rate of 82 percent.

similar rates by exploiting birth and married names. These match rates are achieved only using time invariant characteristics such as names, time and place of birth information.

Table 2 assesses the representativeness of our linked sample in terms of demographics and 1940 characteristics.[15] There are several noticeable differences between our linked sample and the population which mainly reflect two factors. First, due to the fact that men and earlier cohorts are underrepresented in SS-5 records, they are naturally underrepresented in our linked sample. Second, name-based linking tends to produce samples in which blacks are underrepresented and more educated individuals overrepresented, a feature that is shared by all linking methods. Column 4 shows that we can statistically reject the null hypotheses that the means in the population and linked sample are equal. Many of these differences stem from the fact that most outcomes are correlated with demographics, age in particular (e.g. marital status, labor force participation). In fact, Table O.3 in the Online Appendix shows that if we abstract away from these demographic differences by computing means by gender and race for the 1918 cohort, the linked sample appears very well balanced, with the notable exception of educational attainment.

To make our linked sample representative of the population, we re-weight on demographics (sex, cohort, race, state of birth) and educational attainment by generating inverse propensity score weights as described in Bailey et al. (2020).[16] The weighted means in column 3 of Table 2 are much closer to the population means, even for outcomes we did not target. Although we can still statistically reject equality of means for many outcomes, the magnitude of these differences is very small. Online Appendix Tables O.1 and O.2 show that our weighted linked sample is representative of the population within gender and race categories as well.

3.3 Rosenwald Schools

Information on Rosenwald schools comes from Aaronson and Mazumder (2011) and Aaronson et al. (2014), which was originally collected and processed from Fisk University's Rosenwald database. Key pieces of information include the budget year (proxy for the year of construction), the state and county in which the school was built, and the number of classrooms (proxy for the number of teachers). In total, the database includes 4,932 Rosenwald schools built across 888 counties.

[15] Online Appendix Figure O.3 shows that our linked sample has good coverage of 1920 counties of residence.

[16] More specifically, we stack the linked sample and a 10 percent random sample of the population in 1940, and estimate a linear probability model where the dependent variable is an indicator for being in the linked sample and the explanatory variables are fixed effects for sex-by-race-by-cohort, sex-by-race-by-state of birth, sex-by-cohort-by-state of birth, and sex-by-years of education. The resulting coefficients are used to create inverse propensity score weights.

4 Empirical Strategy

In the empirical analysis, we estimate the impact of individuals' exposure to Rosenwald schools, based on where they lived in 1920 and when they were born, on their outcomes in 1940. Following Aaronson and Mazumder (2011), exposure to Rosenwald schools is defined as the average number of Rosenwald teachers per rural black child aged 7-17 in the individual's county during the years he/she was aged 7-13 (assuming a class size of 45):[17]

$$\text{ROSE}_{bc} = \frac{1}{7} \times \sum_{t=b+7}^{b+13} \frac{\text{Rosenwald teachers}_{ct} \times 45}{\text{rural black school-age population (7-17)}_{ct}} \tag{1}$$

where b denotes birth cohort and c denotes county of residence in 1920. The cross-sectional and time variation in the average number of Rosenwald teachers per rural black school-age child is depicted in Online Appendix Figure O.4. Again following Aaronson and Mazumder (2011), we adopt a triple difference strategy which exploits the targeted nature of the Rosenwald Initiative: Rosenwald schools were attended by black children only and primarily served rural areas. This allows us to use whites and urban blacks as control groups for rural blacks. Consider an individual i born in year b in state s and living in county c in 1920. Our main regression specification, estimated separately by gender, has the form:

$$\begin{aligned} \text{y}_{ibsc} = {} & \text{county}_c + \text{cohort}_b + \text{state of birth}_s + \text{cohort}_b \times \text{state of birth}_s + \text{black}_i \\ & + \text{rural}_i + \text{black}_i \times \text{rural}_i + \text{cohort}_b \times \text{black}_i + \text{cohort}_b \times \text{rural}_i \\ & + \text{cohort}_b \times \text{black}_i \times \text{rural}_i + \gamma_0 \cdot \text{ROSE}_{bc} + \gamma_1 \cdot \text{ROSE}_{bc} \times \text{black}_i \\ & + \gamma_2 \cdot \text{ROSE}_{bc} \times \text{rural}_i + \gamma_3 \cdot \text{ROSE}_{bc} \times \text{black}_i \times \text{rural}_i + \Gamma \cdot \text{X}_i + \varepsilon_{ibsc} \end{aligned} \tag{2}$$

where y_{ibsc} is an outcome in 1940 and X_i are controls for 1920 household characteristics (father literacy, mother literacy, homeownership status, father occupational income score).[18] Observations are weighted using the inverse propensity score weights from Table 2, and standard errors are clustered at the 1920 county of residence level.[19]

The county fixed effects in equation (2) capture time invariant unobservable factors common to all black and white cohorts from a particular county. In particular, this guards against the fact that

[17]Exposure is based on ages 7-13 because Rosenwald schools were mostly primary schools, and the handful of Rosenwald high schools cannot be separately identified in the data.

[18]To estimate this regression, we restrict our sample to individuals who were not only born but also lived in one of the 14 Southern states in which Rosenwald schools were built (96 percent of the original sample). If any of the 1920 household controls are missing, we follow Aaronson and Mazumder (2011) and recode missing values to zero and include separate dummies for missing values to avoid dropping those observations.

[19]In Section 5.6, we show that our results are not sensitive to using alternative weights, or not weighting at all.

Rosenwald schools might have been built in counties that had certain characteristics that independently affected 1940 outcomes of blacks and whites alike. For example, Aaronson and Mazumder (2011) document that Rosenwald counties had larger black population shares and higher white literacy rates than non-Rosenwald counties. The cohort fixed effects, which flexibly vary by state of birth or race-by-rural status, capture unobservable factors common to all individuals in a specific cohort cell. In particular, this means we are not simply comparing rural blacks from later cohorts to rural blacks from earlier cohorts, which could confound the effect of Rosenwald schools with the effect of other cohort-specific factors (e.g. steady improvement in non-Rosenwald educational opportunities in rural areas over time).

The γ's are the coefficients of interest. One appealing feature of this empirical setup is that different combinations of these parameters represent the causal effect of total exposure to Rosenwald schools (relative to no exposure) under different identifying assumptions. For example, the sum of all the γ's captures the effect of Rosenwald schools on rural blacks. Because Rosenwald exposure could be correlated with rural or black-specific unobservable factors that might independently affect 1940 outcomes, we can difference out these factors by exploiting rural whites or urban blacks as controls groups. For example, areas with better economic conditions or with more progressive views towards black education might have been in a more favorable position to meet the requirements for the construction of a new Rosenwald school. To the extent that these factors equally affect rural blacks and rural whites, or urban blacks and rural blacks, we can eliminate them through differencing. The corresponding difference-in-difference estimators are $\gamma_1 + \gamma_3$ (black-white, rural) and $\gamma_2 + \gamma_3$ (black, rural-urban) respectively. Lastly, we can difference out both rural and black-specific unobservable factors using the triple-difference estimator γ_3 (black, rural-urban – white, rural-urban), which will be our preferred estimator in the analysis.

It is worth noting that urban blacks were not entirely untreated. Some Rosenwald schools might have been built in urban areas or areas that are labeled as urban according to the 1920 Census question. Carruthers and Wanamaker (2013) also show that expenditures on white schools increased in Rosenwald counties, hinting at spillover effects in terms of educational resources. To the extent that urban blacks and whites either directly or indirectly benefited from Rosenwald schools, our difference-in-difference and triple difference estimates will slightly understate the impact of Rosenwald schools on rural blacks.

Our regression specification is very similar to the one used in Aaronson and Mazumder (2011) to estimate the impact of exposure to Rosenwald schools on educational attainment among World War II enlistees. However, one key difference is that they only observe county at enlistment (and county of birth for a subset of enlistees), and impute rural status by classifying counties as entirely rural or predominantly urban (counties that fall into neither categories are excluded from the analysis). In contrast, thanks to our linkages to the 1920 Census we can observe the county individuals grew up in as well as whether they grew up in a rural or urban area, and therefore better measure exposure to Rosenwald schools for individuals in *all* counties.

One concern that even the triple difference estimator cannot address is the possibility of unobservable factors correlated with Rosenwald exposure that are specific to rural blacks. This concern is amplified by the fact that Rosenwald schools did not propagate across the South in a random fashion due to the Rosenwald Fund's matching grant scheme. As a result, rural black communities that received a Rosenwald school might have been characterized by certain unobservable attributes, such as a high demand for education or positive race relations, that put them on a different trajectory than communities that did not, regardless of whether the Rosenwald school was eventually built or not. To try to alleviate this concern, we will perform two exercises. First, we will show that in a placebo sample of individuals who completed their education prior to the Rosenwald era, future Rosenwald exposure does not have a positive effect on educational attainment, implying that the positive effects on educational attainment documented in the main analysis do not simply reflect a continuation of pre-existing cohort trends. Second, we will present alternative results that rely on the first pilot schools built in Alabama during the 1910s, whose location was mainly driven by proximity to the Tuskegee Institute rather than local factors. Despite differences in the sample and estimation, these findings will reveal qualitatively similar patterns as in our main analysis, suggesting that the non-random placement of Rosenwald schools is not driving our main findings.

Lastly, although we showed that our re-weighted linked sample is representative of the population in terms of demographics and 1940 outcomes (partly by design), one might be worried that individuals in our linked sample are selected in terms of exposure to Rosenwald schools in 1920. Table 3 tests this hypothesis by estimating equation (2) in a sample where we stack our linked sample and the population of interest in the 1920 Census (all individuals born in the South between 1906 and 1918), and where the dependent variable in an indicator being in the linked sample.[20] Observations in the linked sample are assigned their inverse propensity score weight from Table 2, while observations in the population are assigned a constant weight (the weights are normalized to sum to one in each sample). Effectively, this tests whether exposure to Rosenwald schools is predictive of being in the linked sample. The results for males are shown in columns 1-3 for different combinations of fixed effects, with or without 1920 household controls. Our preferred triple difference estimate in Panel D of column 3 is close to zero and not statistically significant. The corresponding coefficient for females in Panel D of column 6 is slightly larger but also not

[20] An alternative approach would be to identify individuals in the population who are included in our linked sample, and define the dependent variable accordingly. We are unable to do this as links between SS-5 records and the 1920 Census are at the *household level* rather than at the individual level. We have generated results using an intermediate approach where the sample is the population of interest and the dependent variable is an indicator equal to one if the individual's household is contained in our linked sample (unweighted regressions). These results are qualitatively similar to those presented here.

statistically significant. Therefore, this type of selection is unlikely to be of great concern.

5 Results

5.1 Educational Attainment

Using pooled Census data, Aaronson and Mazumder (2011) show that Rosenwald schools had a positive impact on black school attendance and literacy rates. They also document a positive effect on black male educational attainment using World War II enlistee records, concluding that the Rosenwald Initiative can account for around 40 percent of the black-white convergence in educational attainment among cohorts born in the South between 1910 and 1925. We begin our analysis by establishing that Rosenwald schools had a positive impact on educational attainment in 1940 among blacks in our linked sample, as we would expect given the findings above.

Table 4 shows the impact of Rosenwald schools on male educational attainment. The first three columns show estimates for different specifications where the dependent variable is years of education. The simple difference estimate in Panel B of column 1 for a baseline specification which includes county fixed effects, cohort-by-state of birth fixed effects, and race-by-rural fixed effects, implies that full exposure to Rosenwald schools increases years of education by 0.59 among rural black males. The difference-in-difference estimates in Panel C, which difference out rural-specific unobservable factors (black-white rural) or black-specific unobservable factors (black, rural-urban), imply an effect of 0.85 and 1.2 years respectively. The triple difference estimate in Panel D which nets out both rural and race-specific unobservable factors is larger in magnitude at 1.3 years. Column 2 allows the race-by-rural fixed effects to vary by cohort to address the possibility of race and rural-specific cohort trends in educational attainment (common across all counties) that may confound the effect of rising Rosenwald exposure in column 1. Interestingly, the resulting estimates are slightly larger. On the other hand, further controlling for 1920 household characteristics, such as whether one's parents are literate, reduces the magnitude of the estimates in column 3. Our preferred triple difference estimate in Panel D implies that full exposure to Rosenwald schools increases years of education by a statistically significant 1.4 years among rural black males. This effect is sizable: going from no exposure to an exposure of 0.29, the average exposure among the latest cohort we examine (1918), implies an effect whose magnitude is 8 percent of the 1940 average educational attainment among pre-Rosenwald cohorts (1900-1905) of black men born in the South.

Columns 4-6 illustrate that the gains are concentrated in the upper tail of the education distribution. In particular, the triple difference estimate in column 5 shows that complete exposure to Rosenwald schools increases the probability of completing high school (12 years or more) by a statistically significant 8.2 percentage points. Despite several differences in the data we use and the way we measure county of exposure and rural status, Aaronson and Mazumder (2011) report a triple difference estimate of 1.377 for educational attainment (Table 5, column 2), which is remarkably similar to our estimate of 1.4. Their triple difference estimate for high school completion is also very similar to ours (0.09 vs. 0.082), though their estimate for some high school (9-11 years) is fairly large whereas we find almost no effect (0.204 vs. 0.027).

Table 5 shows corresponding results for female educational attainment. Our preferred triple difference estimate in Panel D of column 3 shows that full exposure to Rosenwald schools increased years of education among rural black women by a statistically significant 1.2 years. Put differently, going from no exposure to a Rosenwald exposure of 0.29 translates into an effect whose magnitude is 6 percent of the 1940 average educational attainment among pre-Rosenwald cohorts (1900-1905) of black women born in the South. In contrast to rural black men, these gains are more evenly spread across some high school and high school completion. The triple difference estimate for some high school implies a statistically significant increase of 7.4 percentage points in response to full exposure to Rosenwald schools. The estimate for high school completion stands at 5.8 percentage points, but is not statistically significant. Although there are no estimates in the literature to compare these to, the finding that exposure to Rosenwald schools had a positive impact on female educational attainment is entirely consistent with Aaronson and Mazumder (2011)'s results on school attendance and literacy, which cover both sexes.

A primary concern for the interpretation of our results is the presence of cohort trends specific to rural blacks that are correlated with Rosenwald exposure. For example, if Rosenwald schools were built in counties where rural black educational attainment would have experienced relative improvements even in the absence of these schools, our triple difference estimates might overstate the impact of Rosenwald schools. We test for the presence of pre-existing trends in educational attainment using a placebo sample of men and women born in the South between 1886 and 1898 (20 years prior to our main sample). We link these individuals in SS-5 records to the 1900 and 1940 Censuses using the methodology described in Section 3.[21] We then estimate equation (2) using 1900 rather than 1920 information (county of residence, rural/urban status, household characteristics), and assigning Rosenwald exposure 20 years *later* (e.g. assign Rosenwald exposure for the 1910 cohort in county c to the 1890 cohort in county c). Since these individuals completed their primary education prior to the construction of the first Rosenwald schools, their educational attainment should be orthogonal to future Rosenwald exposure. This exercise is similar in spirit to the placebo test in Aaronson et al. (2014), in which they show that future Rosenwald exposure is

[21] To make links between SS-5 records and the 1900 Census, we directly recycled the training data and model used to make links between SS-5 records and the 1920 Census. We generate separate inverse propensity score weights for this alternative sample using the same weighting variables.

uncorrelated with women's fertility decisions prior to the Rosenwald Initiative.

The estimates for the placebo test are shown in Table 6. Due to a combination of low coverage in SS-5 records and lower life expectancy for these earlier cohorts, the sample sizes for this exercise are relatively small. We focus on the triple difference estimates in Panel D since this is our preferred estimate in the main analysis. If anything, future Rosenwald exposure seems to be negatively correlated with past cohort trends in educational attainment, though most of the estimates are statistically insignificant. This finding is consistent with the notion that Rosenwald schools were built in places where schooling opportunities for blacks were particularly lacking and where white literacy was higher.[22] Margo (1986) notes that educational attainment among blacks born in the late nineteenth century was most likely lower than what was recorded in the 1940 Census. Following Collins and Margo (2006), in columns 4 and 8 we use literacy, defined as 3 years of education or more, as an alternative measure of educational attainment that is less prone to measurement error. For both males and females, the triple difference estimates are positive, but small and statistically insignificant. All in all, we do not find any evidence that educational attainment was already trending upwards in Rosenwald counties. If anything, we find suggestive evidence of the opposite, which implies that our estimates might be slightly understating the impact of Rosenwald schools on educational attainment.

5.2 Labor Force Participation

Table 7 shows the impact of Rosenwald schools on male labor force participation. We decompose labor force status into 6 categories: private/non-emergency employment, public emergency employment, unemployment, and three types of non-labor force participation (housework, school, other). The triple difference estimate in column 3 of Panel D suggests that full exposure to Rosenwald schools reduced the probability of rural black men to be employed in public emergency jobs in 1940 by 4.7 percentage points, significant at the 10 percent level. This type of employment mainly included public relief jobs under New Deal programs such as the Works Progress Administration, the Civilian Conservation Corps, and the National Youth Administration. These jobs were not particularly desirable as they were characterized by limited work hours and offered low subsistence-level wages (Fishback, 2017). Therefore, a reduction in public emergency jobs most likely represents an improvement in labor market status among rural black men. This decline is offset by increases in private employment, school attendance, and "other" non-labor force participation, none of which are statistically significant. The slight increase in non-labor force par-

[22] Areas with higher white literacy rates might have been more receptive to the idea of advancing black education, perhaps recognizing that it served their economic interests. The Rosenwald Fund might have also followed Booker T. Washington's non-confrontational philosophy when it came to black-white relations, and deliberately avoided areas with low levels of white literacy.

ticipation seems somewhat puzzling, but a closer look suggests this could be the product of census enumeration errors that failed to distinguish between employees and occupants of institutional facilities (e.g. prisons, asylums, homes for the elderly or handicapped).[23]

The corresponding results for females in Table 8 show that exposure to Rosenwald schools had a positive effect on black female labor force participation. According to the triple difference estimates in Panel D, complete exposure to Rosenwald schools raised private/non-emergency employment in 1940 by a statistically significant 10 percentage points among rural black women, offset by an equally large decline in home production. To get a sense of magnitude, our estimate implies that going from no exposure to a Rosenwald exposure to 0.29 (the average exposure among the 1918 cohort) leads to an increase in labor force participation equivalent to 6 percent of the average labor force participation rate in 1940 among pre-Rosenwald cohorts (1900-1905) of black women born in the South.

To put this finding into perspective, we make two observations. First, the labor force participation rate was much lower among black women than among black men in 1940 (46 percent vs. 92 percent among our cohorts of interest). Second, while the black female labor force participation rate did not change much over the first half of the 20th century, the white female labor force participation rate more than doubled from 14.5 percent in 1900 to 31.8 percent in 1950 (Boustan and Collins, 2014). Many scholars have attributed this rise in large part to a combination of rising high school completion among white women and rising demand for such workers, most notably in clerical occupations. In a similar way, our findings suggest that education was an important determinant of black women's labor force participation. However, as we will show in the next section, few black women who entered the labor force as a result of exposure to Rosenwald schools were able to obtain clerical jobs, and many instead had to settle for low-skill service jobs.

5.3 Occupational Standing

We now turn to the main question of interest: to what extent did the gains in educational attainment as a result of exposure to Rosenwald schools translate into gains in occupational standing? We classify occupations into five major categories: farmers, unskilled jobs, skilled/semi-skilled jobs,

[23]The Census Bureau's original code for "other" non-labor force participation in the 1940 Census included institutional inmates whose work status was entered as out of the labor force by default (Jenkins, 1983). However, recent work by Eriksson (2020), based on a more sophisticated way of identifying prison inmates, documents that Rosenwald exposure *reduced* the probability of incarceration among rural black men. To reconcile these facts, one possible explanation is that employees of institutional facilities (e.g. janitors, cooks) were mis-categorized as inmates. To this point, Jenkins (1983) states that "enumerators' entries for institutional inmates were not reliable. Many enumerators failed to distinguish between employees of institutions and their families, on the one hand, and residents of institutions, on the other hand. Institutions were often not identified and frequently quasi-households, such as monasteries, convents, and other homes, were identified as institutions. It was also found that there was inconsistency in the exclusion of inmates of various types of institutions from the labor force" (p. 97).

white collar jobs, and "other" jobs. Farmers include farm owners/tenants and farm managers. Unskilled jobs include farm laborers, laborers, private household workers (e.g. cooks, maids), and service workers (e.g. waiters, barbers). Skilled/semi-skilled jobs include operatives (e.g. bus drivers, meat cutters) and craftsmen (e.g. carpenters, mechanics). White collar jobs include sales workers (e.g. advertising agents, insurance brokers), clerical workers (e.g. cashiers, bank tellers), professionals (e.g. accountants, engineers), and managers. "Other" jobs is a residual category covering non-occupational responses such as housekeeping or taking care of one's parents.

Black and white workers held vastly different jobs on average in 1940. This is illustrated in Panel C of Online Appendix Tables O.1 and O.2. Among our cohorts of interest, non-farm white male workers were roughly evenly distributed across white collar jobs (23 percent), unskilled jobs (27 percent), and skilled/semi-skilled jobs (35 percent). In contrast, black workers were disproportionately concentrated in unskilled jobs (63 percent), with only a minority in skilled/semi-skilled jobs (17 percent) and white collar jobs (4 percent). The picture among female workers is similar. 41 percent of white female workers held white collar jobs, with another 21 percent in skilled/semi-skilled jobs and 18 percent in unskilled jobs. In contrast, 74 percent of black female workers had unskilled jobs, with only 7 percent each in white collar and skilled/semi-skilled jobs. As a result, the average job held by a white man paid 87 percent more than the average job held by a black man among our cohorts of interest, and the corresponding number for women was 113 percent.[24] These gaps reflect two realities: (1) blacks lacked the necessary skills to obtain some jobs, and (2) qualified blacks were excluded from some jobs on the basis of their race. In this section, we disentangle the role of these two factors by characterizing the occupational gains among blacks who received a boost to their education as a result of Rosenwald schools.

Table 9 shows the impact of Rosenwald exposure on male occupational standing. In columns 1 and 2, we assess whether individuals exposed to Rosenwald schools had better jobs on average in 1940, as measured by race-specific occupational income scores. The occupational income scores are based on the average wage income in 1940 (column 1) or the average total income in 1950 (column 2) among male workers aged 18 to 65 within each 3-digit occupation, separately by race. Whether we consider the simple difference, difference-in-difference, or triple difference estimates, we find that exposure to Rosenwald schools did not have a statistically significant effect on the occupational standing of rural black men in 1940.

Since income-based measures of occupational standing can potentially cloud reallocation across occupations and are just one way to rank occupations, columns 3-7 directly look at the effect of Rosenwald exposure on individuals' propensity to hold different occupations in 1940.[25] Interest-

[24]These numbers are based on race-specific occupational income scores from the 1950 Census.

[25]We assign occupations to individuals based on their occupational response, regardless of whether they were employed or unemployed at the time of the Census.

ingly, we find that the estimates for average occupational standing mask large offsetting effects across occupation groups. In particular, the triple difference estimates in Panel D imply that exposure to Rosenwald schools had a positive effect on rural black men's probability of holding white collar and unskilled jobs, and a negative effect on their probability of being farmers and holding skilled/semi-skilled jobs. In terms of magnitude, complete exposure to Rosenwald schools raised employment in white collar and unskilled jobs by 6.4 and 12.2 percentage points respectively, and reduced employment in farming and skilled/semi-skilled jobs by 5.2 and 13.4 percentage points respectively (all statistically significant at the 5 percent level or higher). These effects are fairly large relative to pre-Rosenwald cohort means, shown at the bottom of Table 9.

Panel A of Figure 2 decomposes the triple difference estimates into finer occupation groups (x-axis). The positive effect on white collar employment is driven by increases in clerical and managerial jobs, while the positive effect on unskilled employment is driven by an increase in service jobs. The decline in skilled/semi-skilled jobs is concentrated in operative jobs. Table O.4 in the Online Appendix digs further and presents triple difference estimates for a selected set of 3-digit occupations that are driving the occupational results (broken down by 3-digit industry for some occupations). Rural black men primarily made gains in two types of clerical jobs, stenographers/typists/secretaries and messengers/office boys. The gains in managerial occupations were concentrated in the miscellaneous manager category. The majority of this effect comes from self-employed managers that most likely owned a store (e.g. food stores, gas stations). The gains in service jobs were concentrated in the miscellaneous service worker category, primarily in the restaurant and hospitality industries. The losses in skilled/semi-skilled jobs were concentrated in plumbers/pipe fitters, and miscellaneous operatives in a variety of manufacturing industries.

Table 10 shows the impact of exposure to Rosenwald schools on female occupational standing. Due to lower labor force participation, only around 40 percent of women in our linked sample reported an occupation. Similar to their male counterparts, we find no evidence that Rosenwald schools had any significant effect on occupational income scores of black women in 1940. Again, this masks a positive effect on employment in white collar and unskilled jobs, offset by a negative effect on employment in skilled/semi-skilled jobs, though only the latter effect is statistically significant. In terms of white collar jobs, Online Appendix Table O.5 shows that rural black women made small but significant gains in office machine operators and library attendants. The triple difference estimates for stenographers/typists/secretaries and miscellaneous clerical workers are larger in magnitude but imprecisely estimated. There is also a positive but insignificant effect on the probability of being teachers, which was the most common occupation among educated black women at the time. The positive effect of exposure to Rosenwald schools on employment in unskilled jobs and the negative effect on employment in skilled/semi-skilled jobs were almost entirely concentrated in two job categories: miscellaneous private household workers (e.g. maids, cooks), and miscellaneous operatives.

In his analysis of black labor market, Margo (1990) notes that "a majority of black professionals in the South were found in just two occupations, teaching and preaching" (p. 91). It is apparent from our findings that blacks exposed to Rosenwald schools made little to no gains in professional occupations that required post-secondary education, such as lawyers, engineers or doctors. This finding is perhaps not so surprising given that Rosenwald schools mainly provided primary education, and Southern blacks had very limited access to high school and college education.[26] In addition to limited post-secondary schooling opportunities, limited family resources—which Margo (1990) refers to as "intergenerational drag"—was likely another important factor restricting upward mobility among Southern blacks. In Table 11, we explore this point by documenting heterogeneity in the occupation results by father occupation category in 1920 (farmer, unskilled, skilled/semi-skilled, white collar). The triple difference estimates in column 4 show that the positive impact of Rosenwald exposure on white collar employment was most pronounced among black men whose fathers were themselves white collar workers in 1920, suggesting that family background played an important role in shaping the gains from schooling.

While a short supply of highly-educated workers can explain why blacks were heavily underrepresented at the top of the occupational ladder, many white collar jobs did not require more than a high school education. A prime example are sales jobs, such as advertising agents and stock salesmen, in which the median education level among white male workers was no more than 12 years. Given the finding that exposure to Rosenwald schools significantly raised high school completion among rural black men, why did they not experience any gains in these jobs? We argue that this is consistent with the notion that blacks were effectively excluded from certain occupations due to racial discrimination.[27] White employers often refused to hire black workers for certain positions, either because it was deemed inappropriate for black and white workers to work alongside each other, or because white customers preferred not to interact with black workers (Spero and Harris, 1931; Goldin, 1990; Margo, 1990).[28] Another telling piece of evidence comes from the 1940 Women's Bureau Survey. In response to a question regarding their hiring practices with respect to

[26] Anderson (1988) summarizes the state of black higher education in the early 20th century: "Thus a convergence of circumstances—the lack of federal and state support for the development of black higher education, the opposition of industrial philanthropy, and the impoverishment of missionary and black religious philanthropy—combined to retard the development of black higher education during the first two decades of the 20th century" (p. 248).

[27] Myrdal (1944) describes black men's plight as follows: "As a wage earner the Negro is excluded from many trades. Where he works he is commonly held down to the status of laborer and is excluded from skilled work. But there are always possibilities for him to enter these jobs, and he is always struggling to do so. In the occupations traditionally associated with upper or middle class status, the exclusion policy is usually much more complete and 'settled.' This is because it is fortified by 'social' considerations, as well as by economic ones" (p. 304).

[28] For example, Margo (1990) notes that "whites simply refused to work for a black foreman. Black access to apprenticeship and training programs in the skilled blue-collar trades was jealously restricted by prejudiced employees, employers, and trade unions. White employers did not hire blacks in retail sales or office work because white customers or clients would be offended" (p. 95).

colored workers, almost 50 percent of the surveyed firms reported that they either had explicit or implicit policies prohibiting the hiring of black clerical workers (Goldin, 1990). These accounts suggest that racial prejudice among employers, employees, and customers combined to prevent blacks from getting white collar jobs, either by denying them the opportunity to acquire the necessary experience or by barring them outright. Consistent with our findings, the few white collar jobs that blacks were able to get either did not necessarily involve interacting with whites (e.g. stenographers, typists), were low-status jobs and therefore "acceptable" for blacks (e.g. messengers, office boys), or mainly involved interacting with other blacks (e.g. self-employed storekeepers).

We formalize these ideas in Table 12, where we separately estimate the effect of Rosenwald exposure on different categories of white collar jobs, depending on whether they typically required a post-secondary education and whether they involved interacting with whites. Education requirements are assigned based on median years of education among Southern white male workers aged 18-65 in 1940. To determine whether jobs required interacting with whites, we first identify jobs that involved "dealing with people" based on task content descriptions from the 1956 Dictionary of Occupational Titles (U.S. Department of Labor, 1956). The resulting classification of jobs is shown in Online Appendix Table O.6.[29] Jobs that involved interacting with people are assumed to have involved interacting with whites, with three exceptions—teachers, clergymen, and self-employed managers—which mainly involved interacting with blacks (Myrdal, 1944; Margo, 1990). There are two takeaways from Table 12. First, blacks made no gains in white collar jobs that required post-secondary education, as our occupation results already hinted at. The only exception is the positive but insignificant effect on the probability of being a teacher among black women. Second, in jobs that required no more than a high school education, blacks primarily made gains in jobs that did not involve interacting with whites, in line with the notion that they tended to be excluded from jobs that did.

One puzzling finding from our occupational results is that exposure to Rosenwald schools had a negative effect on the probability of holding operative jobs and a positive effect on the probability of holding service jobs in 1940, both among black men and black women. On the surface, this might seem odd as operative jobs paid more on average than service jobs and private household jobs.[30] However, it is important to keep in mind that the types of operative jobs that blacks were

[29]Our classification is necessarily imperfect as there are many more job titles in the Dictionary of Occupational Titles (DOT) than there are in the Census, and the mapping between them is not obvious. In column 3, we indicate the closest DOT occupation(s) we could find for every 3-digit Census occupation (column 1). When we could not find a closely related occupation or when there were too many (empty entry), we manually imputed whether jobs involved interacting with people based on job descriptions in the 1949 DOT (U.S. Department of Labor, 1949), as well as our own judgement. In some cases, there were several related job titles that mapped into different categories. We either leaned towards one category based on occupation *string* frequencies within 3-digit Census occupations among black workers (e.g. blacks were more likely to be stenographers/typist than secretaries), or otherwise defaulted to the occupation that involved interacting with people (e.g. funeral attendant vs. embalmer).

[30]Among Southern black men, the average occupational income score for operative jobs was 6 percent higher than

able to obtain at the time were often the most undesirable, dangerous, and physically-demanding ones (Spero and Harris, 1931; Bailer, 1943). Even at the Ford Motor Company, which offered some of the better-paying manufacturing jobs to blacks, black workers were disproportionately assigned to hot, dangerous foundry jobs while earning the same as their white peers (Foote et al., 2003). Therefore, one possibility is that education altered blacks' preferences away from such jobs in favor of safer, less physically demanding service jobs.

Another possibility is that educated blacks simply earned more as service workers than as operatives. It seems plausible that a high school education was not valued in operative jobs, but might have been a desirable attribute for service workers who would interact with white customers (e.g. waiters, porters) or be in the employ of white families (e.g. maids). We probe this point for male workers in Online Appendix Table O.7, where the dependent variables are indicators for being in a particular race-specific 1940 wage income quartile within an occupation group (service workers or operatives). The results imply that the negative effect of Rosenwald exposure on operative jobs among rural black men was concentrated in the bottom two quartiles, while there was a statistically significant increase in the top quartile of service jobs. Online Appendix Table O.8 does the same for female workers, where service jobs are replaced with private household jobs. The increase in private household jobs was entirely driven by an increase in the top quartile, while there was a statistically significant decline in the second and fourth quartiles of operative jobs. While these results are consistent with the notion that some individuals "upgraded" from the bottom quartiles of operative jobs to the top quartiles of service jobs, it is important to keep in mind these coefficients only measure net changes across the occupational distribution. The estimates in Panel C of Table 11 provide further supporting evidence. They show that the positive effects on unskilled jobs and the negative effects on semi-skilled jobs in response to Rosenwald exposure were concentrated among black men and women whose fathers held semi-skilled jobs in 1920. These individuals would have known what operative jobs entailed—and what it took to enter these professions—and still opted for service jobs, which strongly suggests that the latter were more desirable.

Lastly, another possible explanation is that black boys and girls were simply trained and conditioned to enter the service economy while in school. Many black schools in the early 20th century South placed a strong emphasis on "industrial" or "vocational" training, which focused on developing skills like gardening, woodworking, cooking, laundering, and housekeeping (Anderson, 1988). The proponents of industrial education included Booker T. Washington and Northern philanthropists, who believed that a "classical" education would be of little use to blacks in the labor market given the institutional constraints at the time, as well as white Southerners, who saw it as a

for service jobs. Among Southern black women, the average income score for operative jobs was 58 percent higher than for private household jobs. Part of the larger gap for women might reflect selection: 57 percent of black women worked in private household jobs while only 8 percent of them worked in operative jobs.

way of maintaining the existing hierarchy of jobs.

5.4 Wage Income

While we have mainly focused on occupations so far, income is another key indicator of labor market standing. The advantage of income is that it varies even among individuals who report the same occupation. However, one important drawback is that, while the 1940 Census was the first decennial census to include a question on income, it only asked about wage and salary income. Other important components of total income, chiefly farm and business income, are unobserved. Online Appendix Table O.9 documents income reporting patterns by occupation group in the 1940 Census. The first column shows the share of individuals reporting no wage income. The second column shows the share of individuals reporting less than 6 dollars in weekly wage income, which is half of the federal weekly minimum wage in 1940 (Goldin and Margo, 1992). The share of individuals reporting very low income partly captures the extent to which workers may have had other sources of compensation, possibly in-kind (e.g. food, lodging). The third column shows the share of individuals who reported earning more than 50 dollars in non-wage income during the previous year. The fourth column shows the share of individuals who are self-employed, for which wage and salary income is probably a poor measure of total income.

The summary statistics in Panel A for Southern black male workers illustrate that wage and salary income likely provides an incomplete picture of total income for several occupation groups. For instance, 86 percent of farmer reported no wage income, and correspondingly 97 percent were self-employed and 75 percent reported earning more than 50 dollars in non-wage (farm) income. Similarly, although managers comprised a much smaller fraction of the black male workforce (1.2 percent), 62 percent of them reported no wage income, with 71 percent being self-employed and 58 percent reporting more than 50 dollars in non-wage (business) income. The patterns are broadly similar for white workers. These statistics imply that one must be careful when interpreting effects on wages in 1940. This is especially relevant in our context, since we found that black men exposed to Rosenwald schools were less likely to be farmers and more likely to be self-employed managers.

Motivated by these income reporting patterns, the first four columns in Table 13 first show the impact of Rosenwald exposure on individuals' propensity to report no wage income, very low wage income, non-wage income, and being self-employed in 1940. The estimates in Panel A imply that rural black men were more likely to report no wage income and non-wage income, and less likely to report very low wage income and being self-employed, though none of these coefficients are statistically significant. The estimates in Panel B imply that rural black women were significantly less likely to report no wage income (consistent with greater labor force participation), and slightly more likely to report non-wage income (not statistically significant).

With these caveats in mind, columns 5 and 6 in Table 13 show the impact of Rosenwald schools on log annual wage income and log weekly wage income (conditional on reporting at least 1 week worked last year) in 1940. Complete exposure to Rosenwald schools raises annual wage income by 7.6 percent among rural black men and by 9.4 percent among rural black women, neither of which are statistically significant. The corresponding coefficients for weekly wage income are 3.7 percent and 11.2 percent respectively, also not statistically significant. Overall, while highly imprecise, these estimates suggest that exposure to Rosenwald schools may have had a positive effect on wage income. However, as we have argued, these results must be interpreted with extreme caution given that wage income only provides a partial picture of total income, and many workers reported no wage income which raises selection concerns.

The findings in this section relate to a recent study by Carruthers and Wanamaker (2017). They show that a large fraction of the racial gap in wage income in the 1940 South can be explained by observable differences in education, as measured by years of education and county-level proxies of school quality (e.g. expenditures per pupil, teachers per pupil, average teacher salaries). Besides differences in the empirical strategy and underlying data, our analysis differs from theirs by mainly focusing on occupation which allows us to include individuals who did not report wage income in the 1940 Census. We argue that focusing on income gaps leads to an incomplete picture of economic standing as it sidesteps the fact that blacks and whites were concentrated in different segments of the labor market. While some of our results suggest that better schools might have helped blacks enter some jobs in which they were underrepresented, and perhaps earn higher wages within the set of jobs that were available to them, but many jobs were simply out of reach due to labor market discrimination. As such, our findings are consistent with the view in Wright (1986) and Margo (1990) that institutional barriers fundamentally limited black upward mobility in the early 20th century South. For example, Wright (1986) remarks that "(j)ob discrimination in the better-paying positions was far more important than wage differentials for the same job. Blacks could get the going wage in the unskilled market, but there was a virtual upper limit to their possible progress above that level" (p. 185).

5.5 Marital Status & Fertility

Table 14 explores the impact of Rosenwald schools on marital status and fertility. The first three columns in Panel B show that exposure to Rosenwald schools had a negative impact on the probability of being married and the probability of having children in 1940 among rural black women. Complete exposure to Rosenwald schools reduced the probability of being married by 8.7 percentage points and the probability of having any children by 10 percentage points, both statistically significant at the 5 percent level. These findings are consistent with the positive effect (of similar magnitude) on female labor force participation, and likely reflect a higher opportunity cost of starting a family and raising children.[31]

Interestingly, the first column in Panel A shows that exposure to Rosenwald schools also has a significant negative effect on the probability of being married in 1940 among rural black men. Given the corresponding finding for women, one possible explanation is that this reflects a decline in the number of potential female partners in the marriage market. Columns 6 and 7 provide some suggestive evidence to support this hypothesis. They show that own exposure to Rosenwald schools is positively correlated with spouse educational attainment and labor force participation among rural black men that have a spouse, consistent with the notion that spouses were themselves exposed to Rosenwald schools. This aligns with the notion that marriage markets were more local in the early 20th century, and that greater female labor force participation in response to Rosenwald schools might have reduced the number of potential partners for men, at least temporarily.

5.6 Robustness Checks

Alternative Linked Datasets

A key concern is that our results are partly driven by the nature of our linked sample, either because it relies on SS-5 records which only cover a specific subset of the population or due to the supervised ML approach used to make links between SS-5 records and Census records. In Online Appendix Table O.10, we show that our main findings are not sensitive to using alternative linked datasets. Panels A and B show results for males and females using the MLP dataset, and Panels C and D show results for males using two versions of the CLP dataset, which we described in Section 3.2. For comparability, we re-weight these datasets on demographics and educational attainment in 1940 using the same procedure we used to re-weigh our SS-5-based linked dataset.

We first examine the male results. Even though some estimates differ slightly in magnitude, the findings are qualitatively very similar. Among rural black men, exposure to Rosenwald schools has a positive effect on educational attainment, unskilled employment, and white collar employment, and a negative effect on skilled/semi-skilled employment and marital status. Two notable differences is that according to these alternative datasets there is no negative effect on farm employment, and a statistically significant negative effect on the probability of having any children (our triple difference estimate was also negative but half as large and not statistically significant). Although

[31] These results are broadly in line with the findings in Aaronson et al. (2014). One notable difference is that they also document a negative effect of Rosenwald exposure on number of children conditional on having at least one child. Since their analysis covers the period 1910-1930, one possible explanation is that some factors affecting the decision of how many children to have and how much to invest in them have changed by 1940. The theoretical framework in Galor (2012) illustrates that the intensive margin effect of a reduction in the cost of investing in children's education can be non-negative if the elasticity of investment in education is less than one in absolute value.

the point estimates differ, the associated 95 percent confidence intervals overlap to a large extent.

The female results in Panel B using the MLP dataset are also qualitatively similar, though more muted overall. For example, the effect of Rosenwald exposure on labor force participation, marital status and fertility among rural black women are smaller than our estimates and not statistically significant. However, we again emphasize that MLP contains very few women who got married between 1920 and 1940. Given that we focus on women aged 21 to 33 in 1940, this is an important subset of the female population. Capturing this subpopulation thanks to the availability of birth and married names is a key advantage of exploiting SS-5 records.

Differences across datasets are driven by a combination of two factors: (1) MLP and CLP link all individuals in the 1920 Census, irrespective of whether they can be found in SS-5 records or not, and (2) selection into the linked sample differs across datasets due to differences in linking methodology and targeted precision. The first factor, combined with the fact that the findings are broadly similar across datasets, implies that our results do not hinge on the selection of individuals into SS-5 records, which is reassuring. The second factor suggests that smaller coefficients obtained using alternative datasets could be due to attenuation bias since these datasets target a lower precision rate (or do not target any precision rate). This is supported by two pieces of evidence. First, the estimates using the standard Abramitzky et al. (2012) method in Panel C are smaller than the corresponding estimates in Panel D based on their conservative method, which almost surely produces fewer false links. Second, columns 2 and 3 in Online Appendix Tables O.12 and O.13 show that our point estimates tend to be smaller in magnitude when we build alternative linked samples targeting a precision rate of 95 or 90 percent (instead of 97 percent), consistent with the notion that classical measurement error induced by false links lead to attenuation bias (Bailey et al., forthcoming).

Alabama State Border Design

As discussed in Section 4, an important threat to our identification strategy is the possibility of rural-black-specific cohort trends that are correlated with the construction of Rosenwald schools. Following Aaronson and Mazumder (2011), we present alternative results based on the earliest Rosenwald schools built in Alabama. Unlike schools that were built during the program's expansion, the placement of these early schools was arguably driven by proximity to the Tuskegee Institute rather than local circumstances. To this end, we build a separate linked sample which covers individuals born between 1900 and 1906 in Alabama and four bordering states (Florida, Georgia, Louisiana, Mississippi, Tennessee). We link these individuals to the 1910 and 1940 Censuses using the same methodology used to generate our core linked sample. We further restrict the sample to individuals living in counties on either side of the Alabama state border in 1910 to make the treatment and control groups as similar as possible.

We adopt a similar regression specification as in Aaronson and Mazumder (2011) which uses a binary Rosenwald exposure measure (TUSKCOUNTY), corresponding to whether the county contained one of the Rosenwald schools built under Tuskegee, interacted with a race indicator. Because a handful of counties in the four neighboring states also received one of these early schools, we interact the binary exposure measure with an indicator which is equal to one for non-Alabama counties. Online Appendix Figure O.5 displays the counties along the Alabama state border, and which category they fall into. The specification is given by:

$$\begin{aligned} \mathrm{y}_{ibsc} &= \mathrm{county}_c + \mathrm{cohort}_b + \mathrm{state\ of\ birth}_s + \mathrm{black}_i + \mathrm{rural}_i + \mathrm{black}_i \times \mathrm{rural}_i \\ &+ \gamma_1 \cdot \mathrm{TUSKCOUNTY}_c \times \mathrm{black}_i \\ &+ \gamma_2 \cdot \mathrm{TUSKCOUNTY}_c \times \mathrm{black}_i \times \mathrm{non\text{-}AL}_c + \Gamma \cdot X_i + \varepsilon_{ibsc} \end{aligned} \quad (3)$$

We control for cohort fixed effects, state of birth fixed effects, race-by-rural fixed effects, 1910 household characteristics (same as baseline specification), and 1910 county of residence fixed effects (which absorb the baseline effect of TUSKCOUNTY). The coefficient of interest is γ_1, which captures the effect of early Rosenwald schools in Alabama on blacks relative to whites.

Online Appendix Table O.11 shows the resulting estimates, separately for men and women in Panels A and B. Although the coefficients are imprecisely estimated due to small sample sizes (around 500 men and 2,100 women), being in a Rosenwald county has a positive effect on the educational attainment of black men and women (around half a year), a positive effect on black female labor force participation, a positive effect on white collar employment among black men, and a negative effect on black women's probability of being married or having children in 1940. However, only the coefficient for female fertility is statistically significant. Other patterns in the table, such as the occupational patterns among black women, or the effects on marital status and fertility among black men, are qualitatively different than our main findings. In addition to differences in the estimating equation, one possible reason for why the patterns do not perfectly align with our main results is that here we focus on earlier cohorts born in very specific counties, for which the impact of Rosenwald schools might have been different for a variety of reasons (e.g. differences in pre-existing educational opportunities or downstream labor market opportunities). Although merely suggestive, these results broadly point in the same direction as our main findings.

Additional Robustness Checks

Our empirical strategy assigns exposure to Rosenwald schools based on individuals' county of residence and rural status in 1920. The advantage of doing so is that location in 1920 likely provides a good approximation of actual exposure to Rosenwald schools, by virtue of being the mid-point of the Rosenwald era. However, one concern is that some motivated parents may have

have endogenously chosen to move to Rosenwald counties to be able to send their children to these schools, which would affect the interpretation of our results. To address this concern, column 4 in Online Appendix Tables O.12 and O.13 restricts the sample to individuals who were born in their 1920 county of residence, which includes around two thirds of our linked sample.[32] Reassuringly, the triple difference estimates are very similar among this sub-population, implying that selective migration is unlikely to be an important factor.

Our main specification included controls for basic 1920 household characteristics (father literacy status, mother literacy status, homeownership status, and father occupational income score). Column 5 in Online Appendix Tables O.12 and O.13 shows that our main findings are not sensitive to additionally controlling for farm status, family size fixed effects, father and mother age and place of birth fixed effects, mother labor force status, and father 3-digit occupation fixed effects. Columns 6 and 7 show that our main findings are also not sensitive to alternative weighting schemes. In column 6, instead of the inverse propensity score weights described in Section 3.2, we use simple inverse probability weights which re-weight each sex-race-cohort-state of birth cell by the inverse of the probability of being in our linked sample in 1940. In particular, these weights deliberately do not adjust for educational attainment, which unlike demographic characteristics is an endogenous outcome. Column 7 presents unweighted results. The triple difference estimates are very similar regardless of the weighting strategy.

6 Conclusion

To what extent was the racial gap in occupational standing in the early 20th century South driven by differences in human capital versus labor market discrimination? Disentangling the relative contribution of these two factors is challenging from an empirical perspective, which has led to diverging views in the literature. This paper sheds new light on this question by exploring the labor market impact of the Rosenwald Schools Initiative, a large school building program which led to the construction of nearly 5,000 schools in the rural South between 1913 and 1932. Consequently, Rosenwald schools differentially expanded access to high-quality education for Southern blacks born at the beginning of the 20th century. Building on prior work showing that rural blacks exposed to Rosenwald schools experienced significant gains in educational attainment, we explore to what extent these gains translated into better labor market opportunities.

To answer this question, we build a new dataset linking Social Security application records to the 1920 and 1940 Censuses, which allows us to estimate the impact of childhood exposure to

[32]County of birth is available for around 92 percent of individuals in SS-5 records via a 12-character place of birth string. Not all place of birth strings can be mapped to counties due to the 12-character limit and the fact that some strings can refer to different places in the United States. We thank Evan Taylor for kindly sharing his crosswalk.

Rosenwald schools on labor market outcomes in adulthood for both men and women. We find that exposure to Rosenwald schools increased labor force participation among rural black women, and employment in white collar jobs among rural black men. Given that black men were heavily underrepresented in white collar jobs at the time, this suggests that human capital barriers contributed to the black-white gap in occupational standing. However, we also find that blacks failed to break into white collar jobs that involved interacting with whites, such as sales jobs, consistent with accounts that they tended to be excluded from those jobs. As a result, many educated blacks ended up in low-skill service jobs for which they were overqualified, such that exposure to Rosenwald schools did not lead to a meaningful improvement in average occupational standing. This suggests that labor market barriers limited the gains to education for blacks, consistent with the view in Wright (1986) and Margo (1990) that deep-rooted institutional barriers fundamentally hindered the ability of blacks to obtain good jobs and reach middle-class status.

Black economic prospects would only significantly improve after World War II and the civil rights movement. As such, we cannot rule out the possibility that some labor market gains may have materialized after 1940, as institutional barriers gradually broke down. The possibility of post-war gains will be testable once the 1950 full count Census is released to the public.

References

Aaronson, Daniel and Bhashkar Mazumder, "The Impact of Rosenwald Schools on Black Achievement," *Journal of Political Economy*, 2011, *119* (5), 821–888.

_ , _ , Seth G. Sanders, and Evan Taylor, "Estimating the Effect of School Quality on Mortality in the Presence of Migration: Evidence from the Jim Crow South," *Journal of Labor Economics*, forthcoming.

_ , Fabian Lange, and Bhashkar Mazumder, "Fertility Transitions Along the Extensive and Intensive Margins," *American Economic Review*, 2014, *104* (11), 3701–24.

Abramitzky, Ran, Leah Boustan, and Myera Rashid, *Census Linking Project: Version 1.0 [dataset]* 2020.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 2012, *102* (5), 1832–56.

Anderson, James D., *The Education of Blacks in the South, 1860–1935*, The University of North Carolina Press, 1988.

Bailer, Lloyd H., "The Negro Automobile Worker," *Journal of Political Economy*, 1943, *51* (5), 415–428.

Bailey, Martha J., Connor Cole, and Catherine Massey, "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2020, *53* (2), 80–93.

_ , _ , Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons From U.S. Historical Data," *Journal of Economic Literature*, forthcoming.

Boustan, Leah Platt and William J. Collins, "The Origin and Persistence of Black-White Differences in Women's Labor Force Participation," in "Human Capital in History: The American Record," University of Chicago Press, 2014, pp. 205–240.

Carruthers, Celeste K. and Marianne H. Wanamaker, "Closing the Gap? The Effect of Private Philanthropy on the Provision of African-American Schooling in the US South," *Journal of Public Economics*, 2013, *101*, 53–67.

_ and _ , "Separate and Unequal in the Labor Market: Human Capital and the Jim Crow Wage Gap," *Journal of Labor Economics*, 2017, *35* (3), 655–696.

Collins, William J. and Robert A. Margo, "Historical Perspectives on Racial Differences in Schooling in the United States," *Handbook of the Economics of Education*, 2006, *1*, 107–154.

Embree, Edwin Rogers and Julia Waxman, *Investment in People: The Story of the Julius Rosenwald Fund*, Harper, 1949.

Eriksson, Katherine, "Education and Incarceration in the Jim Crow South: Evidence from Rosenwald Schools," *Journal of Human Resources*, 2020, *55* (1), 43–75.

Fishback, Price, "How Successful Was the New Deal? The Microeconomic Impact of New Deal Spending and Lending Policies in the 1930s," *Journal of Economic Literature*, 2017, *55* (4), 1435–85.

Foote, Christopher L., Warren C. Whatley, and Gavin Wright, "Arbitraging a Discriminatory Labor Market: Black Workers at the Ford Motor Company, 1918–1947," *Journal of Labor Economics*, 2003, *21* (3), 493–532.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

Frisvold, David and Ezra Golberstein, "The Effect of School Quality on Black-White Health Differences: Evidence from Segregated Southern Schools," *Demography*, 2013, *50* (6), 1989–2012.

Galor, Oded, "The Demographic Transition: Causes and Consequences," *Cliometrica*, 2012, *6* (1), 1–28.

Goldin, Claudia, *Understanding the Gender Gap: An Economic History of American Women*, Oxford University Press, 1990.

_ **and Robert A. Margo**, "The Great Compression: The Wage Structure in the United States at Mid-Century," *The Quarterly Journal of Economics*, 1992, *107* (1), 1–34.

Heckman, James J. and Brook S. Payner, "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *The American Economic Review*, 1989, *79* (1), 138–177.

Helgertz, Jonas, Joseph R. Price, Jacob Wellington, Kelly Thompson, Steven Ruggles, and Catherine R. Fitch, "A New Strategy for Linking Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel," IPUMS Working Papers Working Paper No. 2020-03, Minnesota Population Center 2020.

_ **, Steven Ruggles, John Robert Warren, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Joseph P. Price, Evan Roberts, and Matthew Sobek**, *IPUMS Multigenerational Longitudinal Panel: Version 1.0 [dataset]*, Minneapolis, MN: IPUMS, 2020.

III, John J. Donohue and James J. Heckman, "Continuous Versus Episodic Change: The Impact of Civil Rights Policy on the Economic Status of Blacks," *Journal of Economic Literature*, 1991, *29* (4), 1603–1643.

_ **,** _ **, and Petra E. Todd**, "The Schooling of Southern Blacks: The Roles of Legal Activism and Private Philanthropy, 1910–1960," *The Quarterly Journal of Economics*, 2002, *117* (1), 225–268.

Jenkins, Robert, "Procedural History of the 1940 Census of Population and Housing," Technical Report, The Center for Demography and Ecology, University of Wisconsin Madison 1983.

Margo, Robert A., "Race, Educational Attainment, and the 1940 Census," *Journal of Economic History*, 1986, *46* (1), 189–198.

_ , *Race and Schooling in the South, 1880-1950: An Economic History*, University of Chicago Press, 1990.

Murray, Jared, Martha J. Bailey, and Connor Cole, "The Highlander Probability Model: Power and Precision from Imposing Constraints in One-to-One Matching," Working paper 2020.

Myrdal, Gunnar, *An American Dilemma: The Negro Problem and Modern Democracy*, Harper & Brothers Publishers, 1944.

Ruggles, Steven, 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.

Smith, James P., "Race and Human Capital," *The American Economic Review*, 1984, *74* (4), 685–698.

_ **and Finis R. Welch**, "Black Economic Progress after Myrdal," *Journal of Economic Literature*, 1989, *27* (2), 519–564.

Spero, Sterling Denhard and Abram Lincoln Harris, *The Black Worker: The Negro and the Labor Movement*, Columbia University Press, 1931.

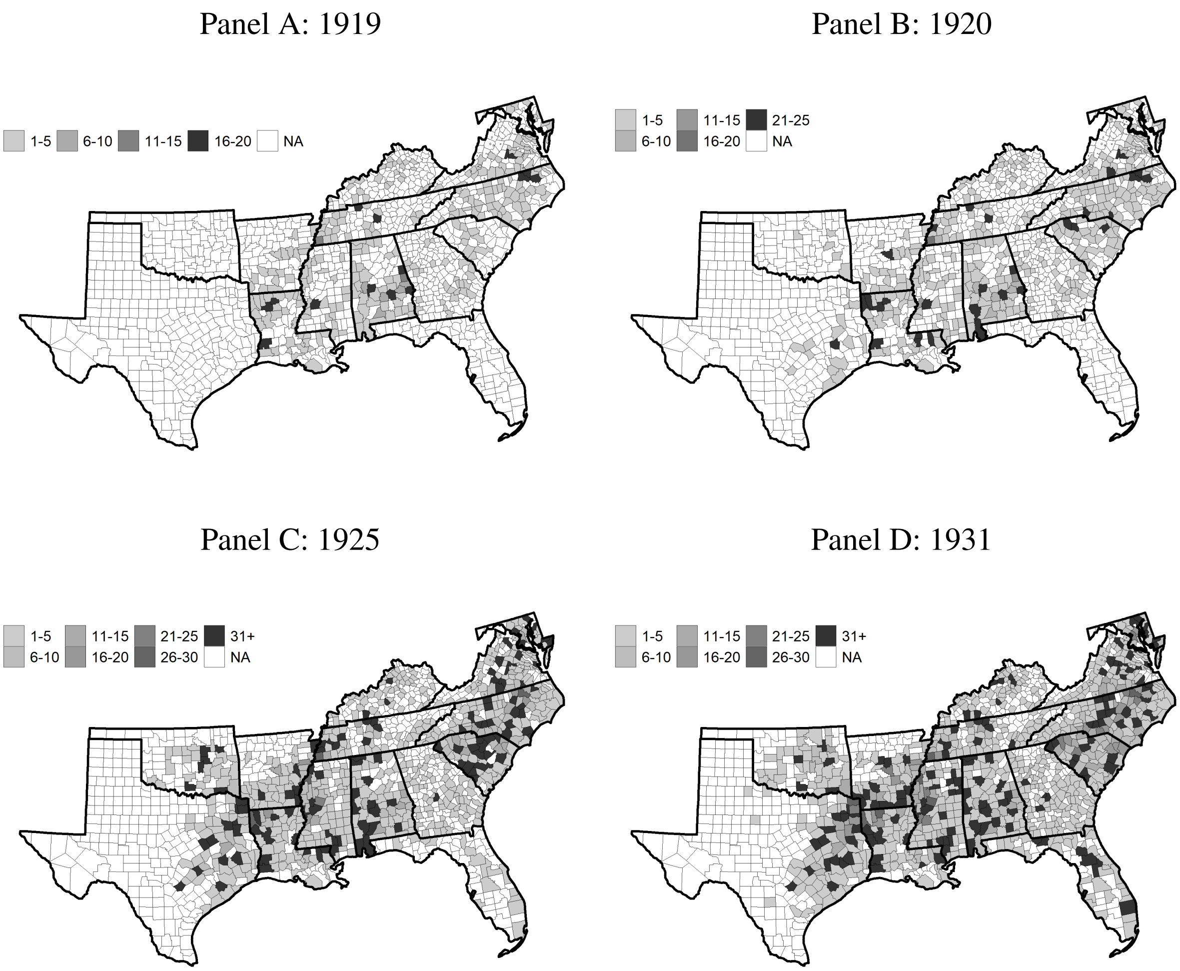
U.S. Department of Labor, *Dictionary of Occupational Titles*, second ed., Vol. I, United States Government Printing Office, 1949.

_ , *Estimates of Worker Trait Requirements for 4,000 Jobs as defined in the Dictionary of Occupational Titles*, United States Government Printing Office, 1956.

Wright, Gavin, *Old South, New South: Revolutions in the Southern Economy since the Civil War*, Basic Books, 1986.

A Figures & Tables

Figure 1: Number of Rosenwald Schools by County, 1919-1931

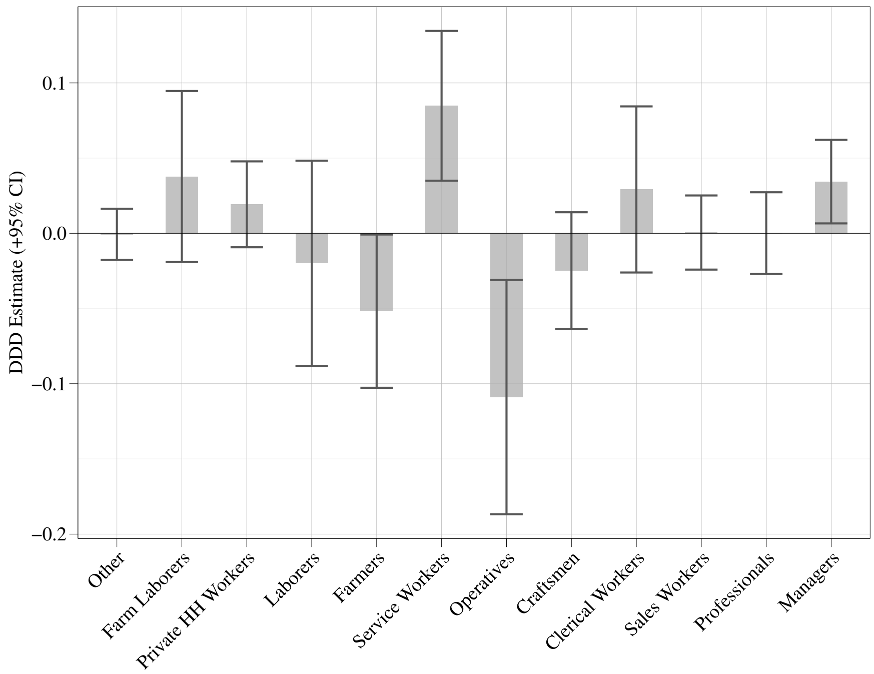


Notes: This figure shows the cumulative number of Rosenwald schools by county, separately by year.

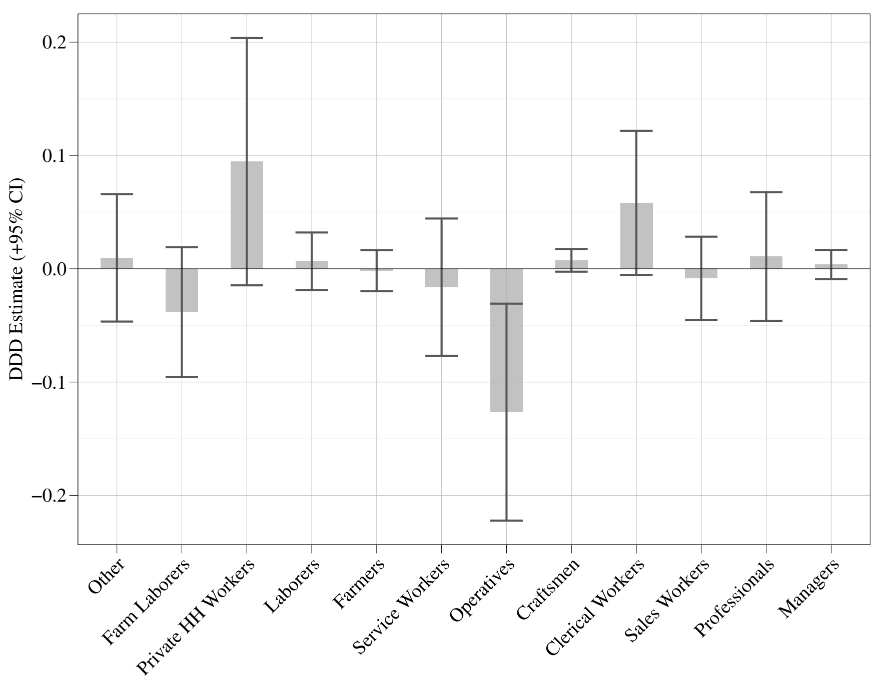
Source: Fisk Rosenwald Database.

Figure 2: The Impact of Rosenwald Schools on Occupation Groups

Panel A: Males



Panel B: Females



Notes: Each bar corresponds to the triple difference estimate from equation (2), where the dependent variable is an indicator for being in a particular occupation group in 1940 (*x*-axis). The error bars represent the corresponding 95% confidence intervals.

Table 1: Match Rates by Gender & Race

	White males	Black males	White females	Black females
Panel A: SS-5 Linked Dataset				
Individuals in SS-5 records born in South (1906-1918)	802,640	275,315	1,119,724	348,729
Links to 1940 Census at 97% precision	362,603	74,300	469,172	88,161
	0.452	0.27	0.419	0.253
Links to 1940 & 1920 Censuses at 97% precision	198,874	30,260	247,362	34,379
	0.248	0.11	0.221	0.099
Effective coverage rate (incl. SS-5 coverage)	0.06	0.026	0.073	0.026
Panel B: IPUMS Multigenerational Longitudinal Panel (MLP)				
Individuals in 1920 Census born in South (1906-1918)	3,322,027	1,254,049	3,215,680	1,253,143
Links to 1930 & 1940 Censuses	901,639	132,813	239,779	44,092
	0.271	0.106	0.075	0.035
Effective coverage rate	0.273	0.114	0.071	0.033
Panel C: Census Linking Project (CLP), Abramitzky Boustan Eriksson (2012) standard method				
Individuals in 1920 Census born in South (1906-1918)	3,322,027	1,254,049	—	—
Links to 1940 Census	824,599	164,893	—	—
	0.248	0.131	—	—
Effective coverage rate	0.25	0.141	—	—
Panel D: Census Linking Project (CLP), Abramitzky Boustan Eriksson (2012) conservative method				
Individuals in 1920 Census born in South (1906-1918)	3,322,027	1,254,049	—	—
Links to 1940 Census	584,877	89,769	—	—
	0.176	0.072	—	—
Effective coverage rate	0.177	0.077	—	—
Individuals in 1940 Census born in South (1906-1918)	3,299,484	1,165,698	3,368,993	1,346,122

Notes: The targeted precision rate in Panel A is relative to hand links in the training data. For details on the targeted precision rate in Panel B, see Helgertz et al. (2020a). There is no concept of precision in Panels C and D. The effective coverage rate is defined as the final number of links divided by the population of interest in the 1940 Census (bottom row).

Table 2: Representativeness of Linked Sample, Demographics & 1940 Characteristics

	Mean			Difference (two-sided *t*-test)	
	Population (10%)	Linked sample (unweighted)	Linked sample (weighted)	Linked sample (unweighted)	Linked sample (weighted)
	(1)	(2)	(3)	(4)	(5)
Female	0.513	0.551	0.511	0.038	-0.002
				[0]	[0.051]
Black	0.273	0.127	0.27	-0.146	-0.002
				[0]	[0.002]
Born in 1906-1910	0.361	0.122	0.365	-0.239	0.004
				[0]	[0]
Born in 1911-1914	0.316	0.289	0.313	-0.027	-0.003
				[0]	[0.001]
Born in 1915-1918	0.323	0.589	0.322	0.266	-0.001
				[0]	[0.101]
Born in AL	0.083	0.068	0.083	-0.014	0
				[0]	[0.741]
Born in AR	0.06	0.059	0.06	-0.001	0
				[0.011]	[0.612]
Born in FL	0.026	0.031	0.026	0.004	0
				[0]	[0.618]
Born in GA	0.095	0.077	0.095	-0.018	0
				[0]	[0.355]
Born in KY	0.079	0.092	0.079	0.014	0
				[0]	[0.442]
Born in LA	0.061	0.058	0.061	-0.002	0
				[0]	[0.365]
Born in MD	0.038	0.062	0.038	0.024	0
				[0]	[0.547]
Born in MS	0.058	0.041	0.058	-0.016	0
				[0]	[0.927]
Born in NC	0.094	0.086	0.094	-0.008	0
				[0]	[0.667]
Born in OK	0.062	0.081	0.062	0.019	0
				[0]	[0.941]
Born in SC	0.058	0.043	0.057	-0.015	0
				[0]	[0.317]
Born in TN	0.076	0.075	0.076	-0.001	0
				[0.001]	[0.434]
(table continues on next page)					
N	912,955	510,875			

Table 2 (cont.): Representativeness of Linked Sample, Demographics & 1940 Characteristics

	Mean			Difference (two-sided *t*-test)	
	Population (10%) (1)	Linked sample (unweighted) (2)	Linked sample (weighted) (3)	Linked sample (unweighted) (4)	Linked sample (weighted) (5)
Born in TX	0.139	0.147	0.139	0.008	0.001
				[0]	[0.292]
Born in VA	0.073	0.081	0.073	0.008	0
				[0]	[0.427]
Live in Northeast	0.032	0.024	0.029	-0.008	-0.003
				[0]	[0]
Live in Midwest	0.062	0.055	0.059	-0.008	-0.003
				[0]	[0]
Live in South	0.866	0.881	0.875	0.015	0.009
				[0]	[0]
Live in West	0.039	0.041	0.037	0.002	-0.002
				[0]	[0]
Live in urban area	0.441	0.445	0.413	0.004	-0.028
				[0]	[0]
Married	0.697	0.595	0.679	-0.102	-0.018
				[0]	[0]
Years of education	8.495	9.727	8.459	1.231	-0.036
				[0]	[0]
In labor force	0.625	0.598	0.617	-0.027	-0.008
				[0]	[0]
Occupation: white collar	0.141	0.197	0.148	0.056	0.007
				[0]	[0]
Occupation: farmer	0.07	0.055	0.082	-0.016	0.011
				[0]	[0]
Occupation: skilled/semi-skilled	0.172	0.165	0.162	-0.006	-0.009
				[0]	[0]
Occupation: unskilled	0.247	0.183	0.23	-0.064	-0.017
				[0]	[0]
Annual wage income ($1940)	367.247	370.008	364.953	2.761	-2.294
				[0.011]	[0.035]
First name commonality score	0.139	0.132	0.127	-0.007	-0.012
				[0]	[0]
Last name commonality score	0.087	0.038	0.043	-0.049	-0.044
				[0]	[0]
N	912,955	510,875			

Notes: Means in columns 3 and 5 are weighted using inverse propensity score weights, where the weighting variables are sex-by-race-by-cohort fixed effects, sex-by-race-by-state of birth fixed effects, sex-by-cohort-by-state of birth fixed effects, and sex-by-years of education fixed effects. *p*-value for two-sided *t*-test of equality of means in brackets (columns 4-5).

Table 3: Selection into Linked Sample, 1920 Exposure to Rosenwald Schools

	Dependent variable: 1[1920 Census household in linked sample]					
	Southern born males (1906-1918)			Southern born females (1906-1918)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Regression coefficients						
ROSE (γ_0)	0.05***	0.044**	0.045**	0.045***	0.014	0.014
	[0.017]	[0.018]	[0.018]	[0.01]	[0.011]	[0.011]
ROSE $\times$ black (γ_1)	0.05**	-0.012	-0.012	0.042**	-0.002	-0.002
	[0.02]	[0.022]	[0.022]	[0.019]	[0.02]	[0.02]
ROSE $\times$ rural (γ_2)	-0.051***	-0.037**	-0.038**	-0.059***	-0.013	-0.013
	[0.014]	[0.016]	[0.016]	[0.009]	[0.01]	[0.01]
Panel B: Simple difference						
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	0.029*	0.009	0.002	0.044***	0.028**	0.023*
	[0.015]	[0.016]	[0.016]	[0.011]	[0.012]	[0.012]
Panel C: Difference-in-difference						
Black-white, rural ($\gamma_1 + \gamma_3$)	0.031**	0.001	-0.004	0.058***	0.028**	0.021*
	[0.013]	[0.014]	[0.014]	[0.01]	[0.011]	[0.011]
Black, rural-urban ($\gamma_2 + \gamma_3$)	-0.07***	-0.024	-0.03	-0.043**	0.016	0.01
	[0.022]	[0.025]	[0.024]	[0.018]	[0.019]	[0.019]
Panel D: Triple difference						
Black, rural-urban – white, rural-urban (γ_3)	-0.019	0.013	0.008	0.016	0.029	0.023
	[0.022]	[0.023]	[0.023]	[0.02]	[0.021]	[0.021]
1920 county of residence fixed effects	✓	✓	✓	✓	✓	✓
Cohort $\times$ state of birth fixed effects	✓	✓	✓	✓	✓	✓
Black $\times$ rural fixed effects	✓			✓		
Cohort $\times$ black $\times$ rural fixed effects		✓	✓		✓	✓
1920 household controls			✓			✓
R^2	0.014	0.014	0.02	0.01	0.011	0.016
N	4,431,583	4,431,583	4,431,583	4,375,362	4,375,362	4,375,362

Notes: The samples in these regressions include individuals in our linked sample and individuals in the population of interest in 1920 (blacks and whites born in the South between 1906 and 1918). The dependent variable is an indicator equal to one for observations in the linked sample, and zero for observations in the population. 1920 household controls include father occupational income scores, and indicators for homeownership, father literacy, and mother literacy. Observations in the linked sample are assigned their inverse propensity score weight from Table 2, while observations in the population are assigned a constant weight (the weights are normalized to sum to one in each sample). Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 4: The Impact of Rosenwald Schools on Male Educational Attainment

	Dependent variable: Outcome in 1940					
	Years of education			Some high school (9-11 years)	High school or more (12+ years)	Some high school or more (9+ years)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Regression coefficients						
ROSE (γ_0)	-0.141	-0.116	-0.029	0.024	0.001	0.026
	[0.182]	[0.192]	[0.163]	[0.018]	[0.019]	[0.02]
ROSE $\times$ black (γ_1)	-0.458	-0.534	-0.617**	0.022	-0.067*	-0.045
	[0.318]	[0.353]	[0.311]	[0.029]	[0.036]	[0.034]
ROSE $\times$ rural (γ_2)	-0.125	-0.187	-0.234	-0.034**	-0.003	-0.037**
	[0.173]	[0.182]	[0.153]	[0.015]	[0.018]	[0.018]
Panel B: Simple difference						
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	0.586***	0.733***	0.528***	0.04**	0.013	0.053**
	[0.214]	[0.235]	[0.198]	[0.02]	[0.018]	[0.027]
Panel C: Difference-in-difference						
Black-white, rural ($\gamma_1 + \gamma_3$)	0.852***	1.035***	0.79***	0.049***	0.015	0.064**
	[0.19]	[0.221]	[0.176]	[0.018]	[0.017]	[0.026]
Black, rural-urban ($\gamma_2 + \gamma_3$)	1.185***	1.382***	1.174***	-0.006	0.079**	0.073*
	[0.328]	[0.348]	[0.31]	[0.036]	[0.034]	[0.039]
Panel D: Triple difference						
Black, rural-urban – white, rural-urban (γ_3)	1.31***	1.569***	1.407***	0.027	0.082**	0.109***
	[0.353]	[0.381]	[0.338]	[0.036]	[0.036]	[0.041]
1920 county of residence fixed effects	✓	✓	✓	✓	✓	✓
Cohort $\times$ state of birth fixed effects	✓	✓	✓	✓	✓	✓
Black $\times$ rural fixed effects	✓					
Cohort $\times$ black $\times$ rural fixed effects		✓	✓	✓	✓	✓
1920 household controls			✓	✓	✓	✓
Pre-Rosenwald cohort mean (1900-1905)	5.27	5.27	5.27	0.068	0.053	0.121
R^2	0.244	0.245	0.317	0.057	0.184	0.238
N	216,663	216,663	216,663	216,663	216,663	216,663

Notes: 1920 household controls include father occupational income scores, and indicators for homeownership, father literacy, and mother literacy. Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 5: The Impact of Rosenwald Schools on Female Educational Attainment

	Dependent variable: Outcome in 1940					
	Years of education			Some high school (9-11 years)	High school or more (12+ years)	Some high school or more (9+ years)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Regression coefficients						
ROSE (γ_0)	0.055	0.183	0.194	0.022	0.027	0.049*
	[0.139]	[0.14]	[0.128]	[0.023]	[0.017]	[0.028]
ROSE × black (γ_1)	0.066	-0.44	-0.469*	-0.014	-0.026	-0.04
	[0.269]	[0.313]	[0.281]	[0.037]	[0.041]	[0.043]
ROSE × rural (γ_2)	-0.151	-0.326**	-0.276**	-0.036	-0.015	-0.051**
	[0.132]	[0.149]	[0.137]	[0.023]	[0.018]	[0.026]
Panel B: Simple difference						
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	0.715***	0.839***	0.604***	0.047**	0.044***	0.09***
	[0.179]	[0.207]	[0.205]	[0.019]	[0.017]	[0.027]
Panel C: Difference-in-difference						
Black-white, rural ($\gamma_1 + \gamma_3$)	0.811***	0.982***	0.687***	0.06***	0.032*	0.092***
	[0.174]	[0.207]	[0.193]	[0.018]	[0.018]	[0.026]
Black, rural-urban ($\gamma_2 + \gamma_3$)	0.594**	1.096***	0.88***	0.038	0.043	0.082*
	[0.3]	[0.369]	[0.32]	[0.035]	[0.038]	[0.046]
Panel D: Triple difference						
Black, rural-urban – white, rural-urban (γ_3)	0.745**	1.422***	1.156***	0.074*	0.058	0.132***
	[0.302]	[0.357]	[0.316]	[0.042]	[0.042]	[0.048]
1920 county of residence fixed effects	✓	✓	✓	✓	✓	✓
Cohort × state of birth fixed effects	✓	✓	✓	✓	✓	✓
Black × rural fixed effects	✓					
Cohort × black × rural fixed effects		✓	✓	✓	✓	✓
1920 household controls			✓	✓	✓	✓
Pre-Rosenwald cohort mean (1900-1905)	6.02	6.02	6.02	0.098	0.075	0.173
R^2	0.187	0.188	0.264	0.034	0.176	0.213
N	266,150	266,150	266,150	266,150	266,150	266,150

Notes: 1920 household controls include father occupational income scores, and indicators for homeownership, father literacy, and mother literacy. Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 6: Placebo Test: Educational Attainment Among 1886-1898 Cohorts

	Dependent variable: Outcome in 1940							
	Males				Females			
	Years of education	Some high school (9-11 years)	High school or more (12+ years)	Literate (3+ years)	Years of education	Some high school (9-11 years)	High school or more (12+ years)	Literate (3+ years)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Regression coefficients								
ROSE (γ_0)	0.021	-0.049**	0.001	0.008	-0.441	-0.028	-0.058	0.029
	[0.396]	[0.025]	[0.048]	[0.023]	[0.311]	[0.028]	[0.041]	[0.034]
ROSE $\times$ black (γ_1)	0.723	0	0.118	-0.016	0.278	0.08*	0.036	-0.045
	[0.617]	[0.048]	[0.077]	[0.038]	[0.665]	[0.041]	[0.065]	[0.071]
ROSE $\times$ rural (γ_2)	-0.073	0.07**	-0.033	-0.006	0.602*	0.059**	0.066	-0.021
	[0.41]	[0.03]	[0.05]	[0.023]	[0.337]	[0.03]	[0.041]	[0.034]
Panel B: Simple difference								
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	-0.334	-0.004	-0.038	0.01	0.179	0.002	0.023	0.025
	[0.398]	[0.022]	[0.03]	[0.038]	[0.369]	[0.026]	[0.028]	[0.039]
Panel C: Difference-in-difference								
Black-white, rural ($\gamma_1 + \gamma_3$)	-0.282	-0.025	-0.006	0.009	0.018	-0.029	0.016	0.017
	[0.402]	[0.021]	[0.03]	[0.037]	[0.361]	[0.025]	[0.028]	[0.037]
Black, rural-urban ($\gamma_2 + \gamma_3$)	-1.078	0.044	-0.157**	0.018	0.342	-0.05	0.046	0.04
	[0.676]	[0.046]	[0.065]	[0.054]	[0.795]	[0.052]	[0.067]	[0.065]
Panel D: Triple difference								
Black, rural-urban – white, rural-urban (γ_3)	-1.005	-0.025	-0.124	0.025	-0.26	-0.109**	-0.021	0.062
	[0.716]	[0.051]	[0.082]	[0.054]	[0.753]	[0.049]	[0.07]	[0.081]
R^2	0.258	0.069	0.135	0.154	0.257	0.067	0.157	0.134
N	57,782	57,782	57,782	57,782	47,219	47,219	47,219	47,219

Notes: All regressions include 1900 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1900 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 7: The Impact of Rosenwald Schools on Male Labor Force Participation

	Dependent variable: Outcome in 1940						
	In labor force				Not in labor force		
	All	Employed		Unemployed	Housework	School	Other
		Private/ non-emergency	Public emergency				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Regression coefficients							
ROSE (γ_0)	0.002	-0.001	0.002	0.001	0.004	0.002	-0.008
	[0.01]	[0.017]	[0.012]	[0.012]	[0.003]	[0.006]	[0.01]
ROSE × black (γ_1)	0.01	-0.022	0.049**	-0.017	0.002	0	-0.012
	[0.019]	[0.027]	[0.022]	[0.015]	[0.004]	[0.01]	[0.017]
ROSE × rural (γ_2)	0.008	-0.005	0.003	0.009	0	-0.003	-0.005
	[0.009]	[0.016]	[0.011]	[0.01]	[0.003]	[0.005]	[0.009]
Panel B: Simple difference							
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	-0.012	-0.006	0.008	-0.014	0.001	0.007	0.005
	[0.015]	[0.021]	[0.012]	[0.011]	[0.005]	[0.006]	[0.012]
Panel C: Difference-in-difference							
Black-white, rural ($\gamma_1 + \gamma_3$)	-0.023*	-0.001	0.002	-0.024**	-0.003	0.008	0.018
	[0.013]	[0.018]	[0.01]	[0.01]	[0.004]	[0.005]	[0.011]
Black, rural-urban ($\gamma_2 + \gamma_3$)	-0.024	0.016	-0.043**	0.002	-0.006	0.006	0.024
	[0.023]	[0.031]	[0.021]	[0.018]	[0.006]	[0.012]	[0.021]
Panel D: Triple difference							
Black, rural-urban – white, rural-urban (γ_3)	-0.032	0.021	-0.047*	-0.007	-0.005	0.008	0.029
	[0.023]	[0.032]	[0.025]	[0.019]	[0.006]	[0.011]	[0.02]
Pre-Rosenwald cohort mean (1900-1905)	0.926	0.793	0.057	0.076	0.006	0.002	0.067
R^2	0.05	0.052	0.041	0.034	0.035	0.041	0.045
N	220,181	220,181	220,181	220,181	220,181	220,181	220,181

Notes: All regressions include 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 8: The Impact of Rosenwald Schools on Female Labor Force Participation

	Dependent variable: Outcome in 1940						
	In labor force				Not in labor force		
	All	Employed		Unemployed	Housework	School	Other
		Private/ non-emergency	Public emergency				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Regression coefficients							
ROSE (γ_0)	0.048*	0.038*	0.001	0.009*	-0.05**	-0.001	0.003
	[0.026]	[0.022]	[0.004]	[0.005]	[0.025]	[0.003]	[0.007]
ROSE × black (γ_1)	-0.101***	-0.105***	0.008	-0.005	0.088***	-0.001	0.014
	[0.025]	[0.028]	[0.008]	[0.009]	[0.025]	[0.006]	[0.014]
ROSE × rural (γ_2)	-0.02	-0.013	-0.003	-0.004	0.019	0.002	0
	[0.023]	[0.02]	[0.003]	[0.005]	[0.022]	[0.003]	[0.007]
Panel B: Simple difference							
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	0.025	0.021	-0.002	0.006	-0.03	0.008	-0.003
	[0.022]	[0.021]	[0.004]	[0.008]	[0.023]	[0.005]	[0.011]
Panel C: Difference-in-difference							
Black-white, rural ($\gamma_1 + \gamma_3$)	-0.003	-0.004	0	0.001	0.002	0.007	-0.006
	[0.021]	[0.02]	[0.004]	[0.007]	[0.021]	[0.005]	[0.01]
Black, rural-urban ($\gamma_2 + \gamma_3$)	0.078***	0.088***	-0.011	0.002	-0.067**	0.01	-0.02
	[0.027]	[0.028]	[0.008]	[0.013]	[0.029]	[0.007]	[0.017]
Panel D: Triple difference							
Black, rural-urban – white, rural-urban (γ_3)	0.098***	0.101***	-0.008	0.006	-0.086***	0.008	-0.02
	[0.031]	[0.033]	[0.008]	[0.013]	[0.031]	[0.008]	[0.016]
Pre-Rosenwald cohort mean (1900-1905)	0.458	0.404	0.014	0.039	0.465	0.003	0.074
R^2	0.072	0.065	0.013	0.018	0.084	0.023	0.042
N	270,520	270,520	270,520	270,520	270,520	270,520	270,520

Notes: All regressions include 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 9: The Impact of Rosenwald Schools on Male Occupational Standing

	Dependent variable: Outcome in 1940						
	Occupational income score		Occupation category				
	1940 basis (log)	1950 basis (log)	Other	Farmer	Unskilled	Skilled/ semi-skilled	White collar
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Regression coefficients							
ROSE (γ_0)	0.011	0.012	-0.003	0.009	0.006	-0.023	0.011
	[0.025]	[0.017]	[0.004]	[0.018]	[0.02]	[0.029]	[0.023]
ROSE × black (γ_1)	-0.03	-0.022	-0.003	0.032*	-0.057*	0.076*	-0.048*
	[0.026]	[0.025]	[0.007]	[0.017]	[0.032]	[0.041]	[0.027]
ROSE × rural (γ_2)	-0.017	-0.006	0.002	0.028*	-0.051***	0.045*	-0.024
	[0.023]	[0.016]	[0.004]	[0.016]	[0.019]	[0.027]	[0.021]
Panel B: Simple difference							
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	-0.01	-0.021	-0.004	0.017	0.02	-0.036	0.003
	[0.029]	[0.027]	[0.007]	[0.023]	[0.029]	[0.028]	[0.018]
Panel C: Difference-in-difference							
Black-white, rural ($\gamma_1 + \gamma_3$)	-0.004	-0.027	-0.003	-0.02	0.065**	-0.058**	0.016
	[0.026]	[0.024]	[0.005]	[0.019]	[0.026]	[0.023]	[0.014]
Black, rural-urban ($\gamma_2 + \gamma_3$)	0.01	-0.011	0.002	-0.023	0.071	-0.089*	0.04
	[0.039]	[0.035]	[0.009]	[0.028]	[0.045]	[0.046]	[0.032]
Panel D: Triple difference							
Black, rural-urban – white, rural-urban (γ_3)	0.026	-0.005	-0.001	-0.052**	0.122***	-0.134***	0.064**
	[0.035]	[0.031]	[0.009]	[0.026]	[0.042]	[0.044]	[0.03]
Pre-Rosenwald cohort mean (1900-1905)	1.51	2.12	0.003	0.175	0.571	0.203	0.049
R^2	0.461	0.492	0.023	0.127	0.156	0.08	0.173
N	203,885	202,904	205,415	205,415	205,415	205,415	205,415

Notes: All regressions include 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 10: The Impact of Rosenwald Schools on Female Occupational Standing

	Dependent variable: Outcome in 1940						
	Occupational income score		Occupation category				
	1940 basis (log)	1950 basis (log)	Other	Farmer	Unskilled	Skilled/ semi-skilled	White collar
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Regression coefficients							
ROSE (γ_0)	-0.003	-0.002	0.033*	-0.015**	-0.001	-0.005	-0.012
	[0.019]	[0.021]	[0.02]	[0.006]	[0.021]	[0.019]	[0.024]
ROSE $\times$ black (γ_1)	-0.015	0.003	0.008	0.004	0.002	0.04	-0.054
	[0.032]	[0.032]	[0.02]	[0.004]	[0.032]	[0.048]	[0.044]
ROSE $\times$ rural (γ_2)	0.008	0.017	-0.033*	0.004	0.002	0.037*	-0.009
	[0.017]	[0.017]	[0.018]	[0.005]	[0.019]	[0.021]	[0.023]
Panel B: Simple difference							
Black rural ($\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$)	0.004	0.011	0.018	-0.009	0.049	-0.048*	-0.01
	[0.024]	[0.029]	[0.024]	[0.009]	[0.03]	[0.025]	[0.024]
Panel C: Difference-in-difference							
Black-white, rural ($\gamma_1 + \gamma_3$)	-0.001	-0.003	0.018	0.002	0.049	-0.079***	0.011
	[0.022]	[0.025]	[0.02]	[0.008]	[0.03]	[0.028]	[0.023]
Black, rural-urban ($\gamma_2 + \gamma_3$)	0.022	0.011	-0.024	0.002	0.048	-0.082*	0.056
	[0.036]	[0.041]	[0.028]	[0.009]	[0.042]	[0.044]	[0.039]
Panel D: Triple difference							
Black, rural-urban – white, rural-urban (γ_3)	0.014	-0.006	0.01	-0.002	0.047	-0.119**	0.064
	[0.037]	[0.042]	[0.029]	[0.009]	[0.042]	[0.049]	[0.044]
Pre-Rosenwald cohort mean (1900-1905)	1.46	2.05	0.102	0.025	0.725	0.089	0.059
R^2	0.631	0.54	0.111	0.065	0.293	0.116	0.239
N	107,477	90,451	108,892	108,892	108,892	108,892	108,892

Notes: All regressions include 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in brackets, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 11: Occupation Results: Heterogeneity by Father Occupation Category in 1920

	Dependent variable: Male outcome in 1940				Dependent variable: Female outcome in 1940			
	Farmer	Unskilled job	Skilled/ semi-skilled job	White collar job	Farmer	Unskilled job	Skilled/ semi-skilled job	White collar job
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Father farmer (owner/manager) in 1920								
ROSE × black × rural	0.244*	-0.194	-0.113	0.165	0.002	0.459**	-0.082	-0.120
	(0.136)	(0.216)	(0.189)	(0.153)	(0.022)	(0.217)	(0.165)	(0.236)
R^2	0.099	0.128	0.092	0.080	0.087	0.277	0.134	0.156
N	104,429	104,429	104,429	104,429	51,048	51,048	51,048	51,048
Panel B: Father unskilled job in 1920								
ROSE × black × rural	-0.027	0.129	-0.126	0.042	0.002	-0.026	-0.106	0.093
	(0.040)	(0.078)	(0.081)	(0.051)	(0.015)	(0.077)	(0.091)	(0.092)
R^2	0.242	0.275	0.213	0.198	0.224	0.378	0.251	0.251
N	21,906	21,906	21,906	21,906	12,011	12,011	12,011	12,011
Panel C: Father skilled/semi-skilled job in 1920								
ROSE × black × rural	-0.009	0.263***	-0.313***	0.040	0.007	0.186*	-0.249***	0.035
	(0.055)	(0.102)	(0.121)	(0.060)	(0.018)	(0.112)	(0.087)	(0.109)
R^2	0.265	0.226	0.150	0.168	0.303	0.332	0.235	0.245
N	28,355	28,355	28,355	28,355	16,076	16,076	16,076	16,076
Panel D: Father white collar job in 1920								
ROSE × black × rural	-0.207***	0.042	-0.233	0.435***	-0.040	0.113	-0.047	0.113
	(0.075)	(0.206)	(0.207)	(0.159)	(0.034)	(0.247)	(0.177)	(0.256)
R^2	0.221	0.226	0.162	0.193	0.173	0.288	0.177	0.227
N	24,941	24,941	24,941	24,941	15,084	15,084	15,084	15,084

Notes: Sample in each panel restricted to individuals whose fathers were in a particular occupation category in 1920 (indicated in panel title). All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 12: White Collar Jobs: Breakdown by Whether Job Typically Requires More Than High School Education & Whether Job Involves Dealing With People

	Dependent variable: White collar job in 1940			
	Job requires high school or less		Job requires more than high school	
	Job involves dealing with people (except self-employed managers)	Job does not involve dealing with people (including self-employed managers)	Job involves dealing with people (except teachers & clergymen)	Job does not involve dealing with people (including teachers & clergymen)
	(1)	(2)	(3)	(4)
Panel A: Males				
ROSE × black × rural	0.010	0.060**	0.002	-0.008
	(0.022)	(0.029)	(0.004)	(0.009)
R^2	0.069	0.084	0.036	0.029
N	205,415	205,415	205,415	205,415
Panel B: Females				
ROSE × black × rural	0.004	0.037	0.000	0.023
	(0.016)	(0.036)	(0.008)	(0.024)
R^2	0.058	0.157	0.023	0.055
N	108,892	108,892	108,892	108,892

Notes: All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 13: The Impact of Rosenwald Schools on Wage Income

	Dependent variable: Outcome in 1940					
	1[no income]	1[weekly wage income < \$6]	1[non-wage income ⩾ \$50]	1[self-employed]	Wage income (log)	
					Annual	Weekly
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Males						
ROSE × black × rural	0.046	-0.021	0.023	-0.022	0.076	0.037
	(0.036)	(0.029)	(0.032)	(0.029)	(0.075)	(0.062)
R^2	0.076	0.070	0.078	0.101	0.245	0.244
N	220,181	220,181	220,181	220,181	154,850	153,291
Panel B: Females						
ROSE × black × rural	-0.098***	-0.005	0.023	-0.003	0.094	0.112
	(0.033)	(0.026)	(0.025)	(0.009)	(0.105)	(0.084)
R^2	0.067	0.084	0.020	0.019	0.233	0.312
N	270,520	270,520	270,520	270,520	81,302	80,198

Notes: Sample restricted to individuals with positive wage income in column 5, and additionally 1+ week worked last year in column 6. All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table 14: The Impact of Rosenwald Schools on Marital Status, Fertility, Child School Attendance & Spouse Outcomes

	Dependent variable: Own outcome in 1940					Dependent variable: Spouse outcome in 1940		
	Married	Any children	Number of children	Number of children	Any child (5-17) in school	Years of education	In labor force	Occupational income score (log)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Males								
ROSE × black × rural	-0.133***	-0.034	0.002	0.243	0.042	0.921**	0.106*	-0.018
	(0.046)	(0.056)	(0.129)	(0.252)	(0.235)	(0.457)	(0.060)	(0.103)
R^2	0.143	0.133	0.182	0.232	0.239	0.323	0.124	0.710
N	220,181	220,181	220,181	59,501	15,615	102,312	103,913	19,377
Panel B: Females								
ROSE × black × rural	-0.087**	-0.101**	-0.210*	0.024	0.049	0.645	-0.009	0.040
	(0.040)	(0.042)	(0.114)	(0.170)	(0.088)	(0.500)	(0.014)	(0.036)
R^2	0.068	0.089	0.145	0.173	0.169	0.319	0.084	0.542
N	270,520	270,520	270,520	131,233	67,207	171,640	174,366	171,146
Sample	All	All	All	Any child	Any child (5-17)	Spouse	Spouse	Spouse

Notes: Sample restrictions indicated in bottom row. All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

B Data Appendix

B.1 Linking SS-5 Records to the 1940 Census as Individuals

The first step in constructing our linked sample involves linking SS-5 records to the 1940 Census. We start by describing the way we link males. For each primary SS-5 record, we search for potential candidates in the 1940 Census that match on sex, first or last name initial, state/continent of birth, and a +/-3 year window around the primary's age in 1940.[33] We then compute a Jaro-Winkler name similarity score, and keep the top 25 potentials.

To generate training data, we randomly draw 1,000 primaries and make manual linking decisions as follows. For each primary and the corresponding 25 potentials from the previous step, we display full names, age in 1940, country of birth (if born outside the U.S.), and race.[34] The trainer either chooses a single best candidate from the list or declines to make a link. Failing to make a link occurs either when there is no plausible candidate in the list, or when there are multiple plausible candidates. Note that clear-cut links are not always characterized by an exact match on first and last name due to misspellings and nicknames. In our training data, we are able to link 53 percent of males to the 1940 Census.

To train the two-stage model from Murray et al. (2020) to make linking decisions for males, we first split the training data into a "training set" and a "test set," using a 70 percent/30 percent split. The training set is used as input data to train the algorithm, and the test set can be used to assess the out-of-sample performance of the model (which usually aligns quite well with in-sample performance). Probabilities are generated using 10-fold cross-validation as described in Section 3.2.[35] The male model is able to achieve a recall rate of 77 percent at 97 percent precision.

The process of linking females is analogous, except for a few key differences. First, we generate two version of each female record, one with the married name as the last name and one with the birth name as the last name. Birth and married names are determined by comparing the last name field and the father last name field. Linking women both ways is important because we do not know under which last name they appear in the 1940 Census. Second, we additionally block on marital status when generating potential candidates. That is, women under their birth name are linked to single women in the Census. Conversely, women under their married name are linked to non-single women in the Census (i.e. married, divorced, widowed). Third, we must decide what to do when women are linked using their married name *and* birth name, which can happen when

[33] We block on continent of birth for individuals born outside the U.S. because countries are more likely to be mis-specified than states (especially for European countries where borders changed over time). Blocking on first and last name initial ensures that potentials candidates can be found even if one of these initials is misspecified.

[34] We display race and country of birth as optional pieces of information to take into account, knowing that they are not always accurately reported.

[35] The full set of model features we used to train this model (and all other models) is available upon request.

both names are common. We use the following conflict resolution rule: if the difference in model probabilities is 0.1 or greater, keep the link with the highest probability; discard both otherwise.

To generate training data for females, we manually link a random sample of 500 SSNs (corresponding to around 1,500 primaries). For SSNs that have conflicting links, we manually resolve cases by either choosing one link over the other or deciding there is too much ambiguity. This allows us to compute SSN-level match rates, and assess the accuracy of the conflict resolution rule. Ultimately, we are able to manually link 59 percent of female SSNs to the 1940 Census. The SSN-level recall rate in the female model is 77 percent at 97 percent precision.[36]

Using the trained male and female models and the corresponding cross-validated probability thresholds, we generate links for the full sample of SS-5 records. We mention two final details. SS-5 records with unknown genders are linked twice: once using the male model and once using the female model. We resolve any male vs. female conflicts within SSN using the same conflict resolution rule used for female married vs. birth name conflicts. Lastly, while the model links each primary to a single potential in the Census, it is possible for the same Census potential to be linked to two different primaries due to common names. To resolve those cases, we prioritize links that have a higher precision rate, and discard the remaining conflicts.

B.2 Linking Siblings in SS-5 Records

The second step in constructing our linked sample involves linking siblings together in the SS-5 records based on parent names. The purpose of these linkages is to reconstruct the entire family structure (children + parents) for the final linking step, in which we link reconstructed families in SS-5 records to households in the 1920 Census. We generate sets of potential siblings for each primary as follows. We start with all unique SSNs in the SS-5 records (if SSNs appear multiple times, we keep the most complete information from each field). For each primary SSN, we generate potential siblings from the remaining SSNs by first blocking on father first two last name initials, and a +/-10 year window around the primary's year of birth.[37] For all potentials within the block, we compute a Jaro-Winkler score between parent names fields, and keep the top 21 potentials. Lastly, we drop redundant primary-primary pairs (which will have a maximum score of 1 by definition).

To generate training data, we randomly draw 500 primaries and make manual linking decisions as follows. For each primary and the corresponding 20 potentials from the previous step, we

[36]The female model is estimated at the primary level to mirror the fact that trainers make independent linking decisions for each primary, even if they belong to the same SSN.

[37]This blocking scheme is necessary for computational reasons, as we do not block on place of birth (siblings can be born in different states/countries). The +/-10 year window does not imply that all siblings must be born at most 10 years apart, as will become clear below (two siblings born more than 10 years apart can be linked together via a common sibling which is born within 10 years of both siblings).

display the year of birth, state/country of birth, race, and parent full names.[38] Here, trainers can choose multiple links as individuals can have multiple siblings. Note that SS-5 records list the mother's full birth name, which greatly reduces ambiguity. In our training data, we are able to find at least 1 sibling for around 40 percent of primaries, and the average number of siblings conditional on having at least one sibling is 1.6.

Similar to SS-5-1940 Census linking, we split the training data into a training set (70 percent) and test set (30 percent).[39] We then train a random forest model on the training set and use 10-fold cross-validation to generate match probabilities and determine the threshold associated with a 97 percent precision rate. The model is able to achieve a 90 percent recall rate at 97 percent precision.

Using the trained random forest model and the cross-validated probability threshold, we generate siblings links in the full sample of SS-5 records. The final step involves creating a crosswalk between SSNs and family IDs from the collection of sibling links, allowing us to organize SS-5 records into families (similar to household IDs in the Census). The simplest way to think about this problem is that SS-5 records are nodes and siblings links edges in a network. Family IDs are then clusters of nodes formed from the edges between them. We use the concept of connected components to "draw" an initial set of family IDs. In other words, any two nodes within a family can be connected to each other via an arbitrary chain of edges, but two nodes in different families cannot. This means that siblings that are born 10 years apart can still belong to the same family if they were both linked to a common sibling.

Lastly, we assign a quality tag to each family ID based on the degree of connectivity between nodes (i.e. if every sibling within the family is connected to all other siblings, then we can be very confident in the quality of this family ID), and trim/split a small number of family IDs that seem dubious (e.g. implausibly large number of children).[40] In the end, we are able to reconstitute around 5.9 million non-trivial families (i.e. at least 2 siblings) in the SS-5 records, out of a starting number of 40 million unique SSNs. Around 35 percent of SSNs are linked to at least one other SSN in the SS-5 data.

B.3 Linking SS-5 Records to the 1920 Census as Households

The last step in constructing our linked sample involves linking reconstituted families in SS-5 records to households in the 1920 Census, when children are still living with their parents as dependents. We generate potentials households for each family by blocking on father last name

[38] In practice, the main consideration is parent name similarity, but year of birth differences and whether the siblings are born in the same state or the same country is also informative.

[39] In contrast to one-to-one linking where each set constitutes a distinct linking decision, in one-to-many linking we consider each primary-potential pairs as a separate linking decision, even though these decisions were made simultaneously to some extent. As a result, the effective size of the training data is roughly $N = 10{,}000$ (500×20).

[40] The full details on how we generate family IDs and associated quality tags is available upon request.

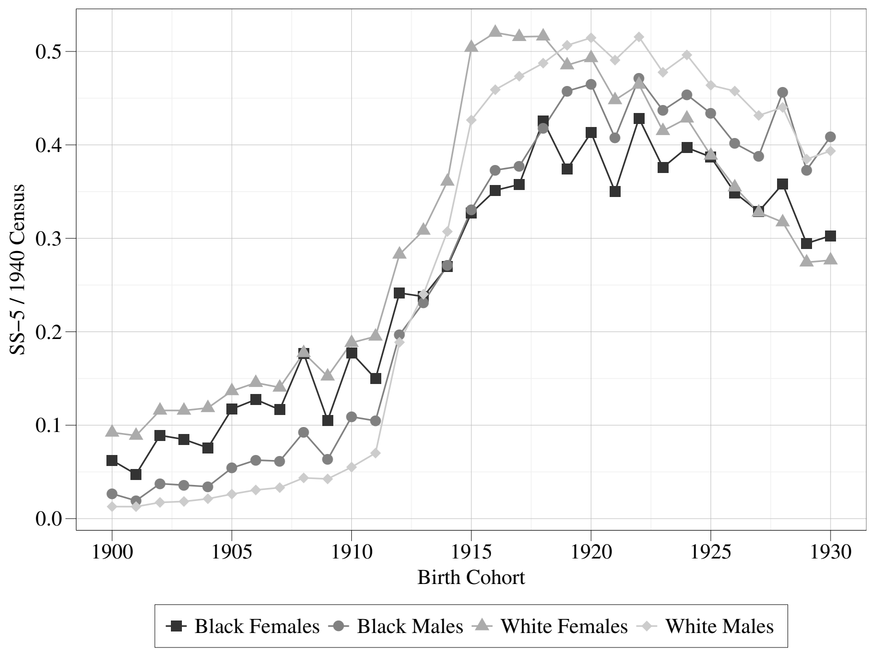
initial and the union of children's state/continent of birth. Blocking on the union of children's place of birth implies that the only requirement we impose is that at least one of the potential children's place of birth matches one of the primary sibling's place of birth. Note that SS-5 records do not contain parents' place of birth, which is why we block on children's place of birth. Within each block, we select the top 20 potential households based on a Jaro-Winkler score between parent names.

We generate training data for a random sample of 250 primary families. We display the following information to trainers: father full name, mother first and middle name, and children's first and middle name, age in 1920, and state/country of birth code (to avoid overcrowding the screen). We also display the modal race among children. We only display primary siblings that would have been alive in 1920, and hide potential children in the Census if there is little chance that they match any of the primary siblings (based on age and place of birth considerations only). We are able to manually link 47 percent of families to a household in the 1920 Census in the training data. We train the two-stage model from Murray et al. (2020) as described in SS-5-1940 linking.[41] The model is able to reproduce around 85 percent of links made by trainers at 97 percent precision. Similar to SS-5-1940 linking, it is possible for the same Census household to be linked to two different families in the SS-5 data after scaling up the model. We use a similar strategy to resolve those conflicts.

[41] We also generated training data to link families to the 1910 Census (250 cases, 34 percent match rate). The ML algorithm was trained using the pooled 1920 and 1910 training data.

Online Appendix

Figure O.1: Cohort Coverage in SS-5 Records for Individuals Born in South, 1900-1930



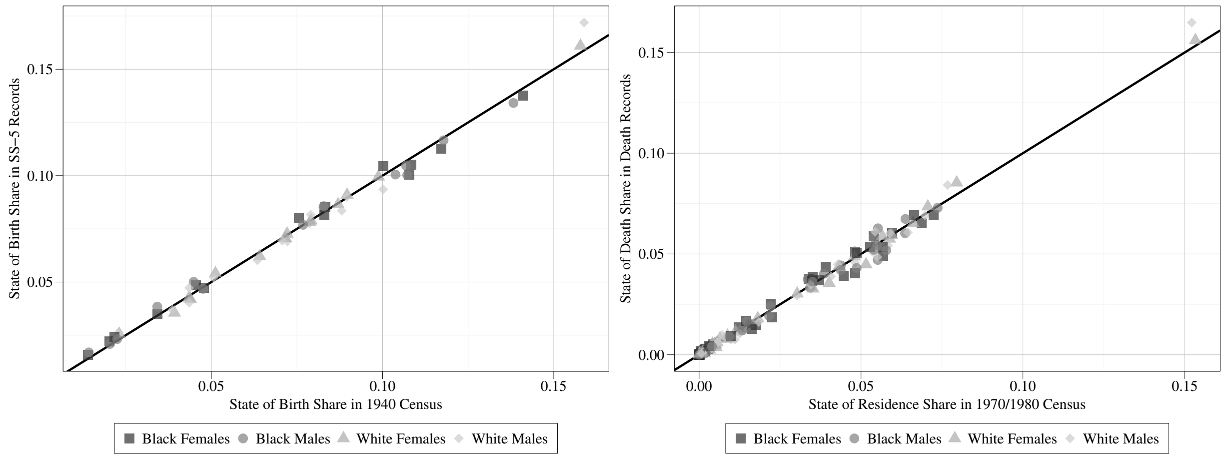
Notes: This figure plots the implied coverage rate in SS-5 records for individuals born in the South, defined as number of individuals in SS-5 records divided by the corresponding number of individuals in the 1940 Census, separately by cohort (x-axis) and gender-race cell (legend).

Source: NUMIDENT SS-5 records, 1940 U.S. Census.

Figure O.2: State of Birth & State of Death Coverage in SS-5 Records for Individuals Born in South between 1906 and 1918

Panel A: State of Birth

Panel B: State of Death



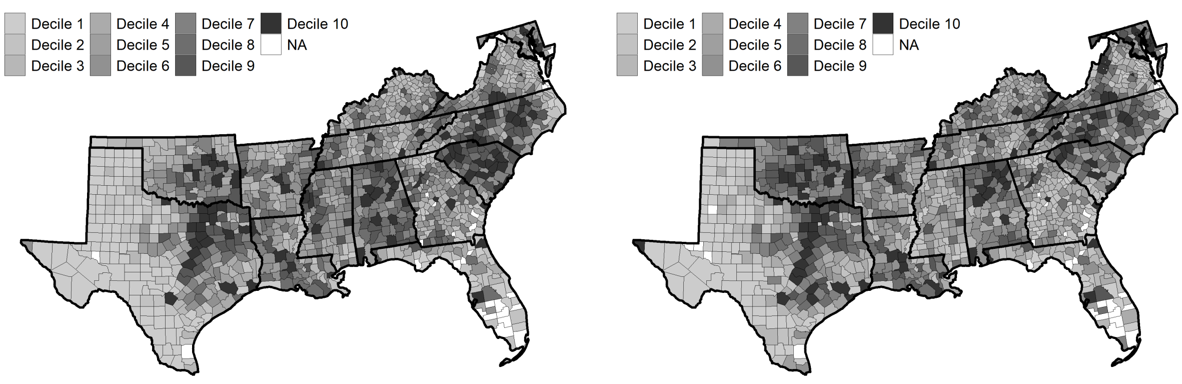
Notes: Panel A plots state of birth shares in SS-5 records against the corresponding shares in the 1940 Census among individuals born in the South between 1906 and 1918, separately gender-race cell (legend). For the same cohorts, Panel B plots state of death shares in SS-5 records (among the subset of SSNs that can be found in NUMIDENT death records) against state of residence shares in the 1970 and 1980 Censuses, separately gender-race cell (legend).

Source: NUMIDENT SS-5 and death records, 1940, 1970 and 1980 U.S. Censuses.

Figure O.3: 1920 County of Residence Shares in Linked Sample vs. 1920 Census

Panel A: 1920 Census

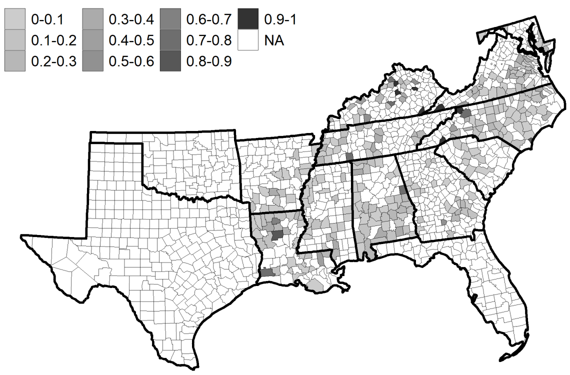
Panel B: Linked Sample



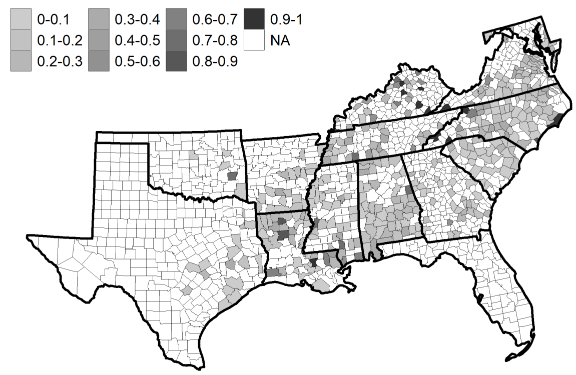
Notes: This figure plots 1920 county of residence shares in the 1920 Census (Panel A) and 1920 county of residence shares in our linked sample (Panel B), among individuals born in the South between 1906 and 1918.

Figure O.4: Rosenwald Exposure by County, 1919-1931

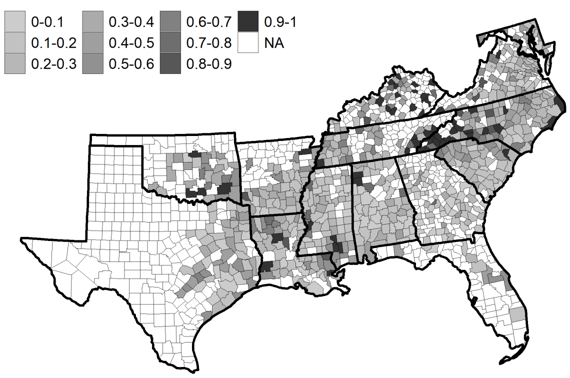
Panel A: 1919



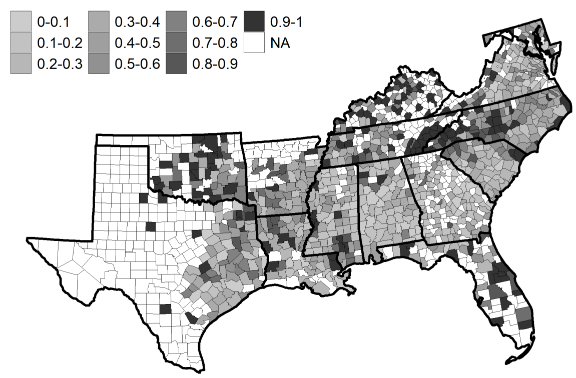
Panel B: 1920



Panel C: 1925

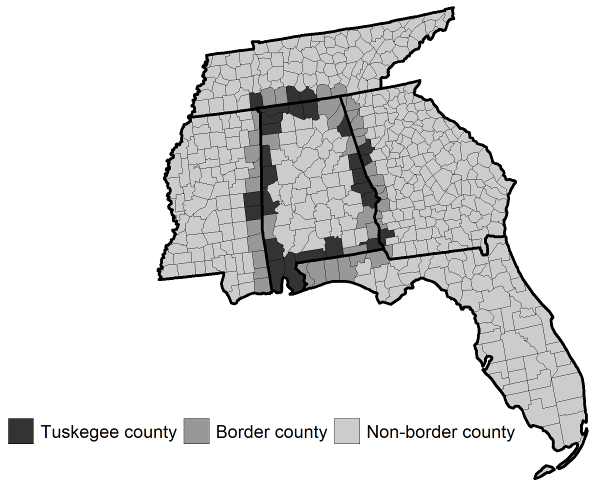


Panel D: 1931



Notes: This figure plots our Rosenwald exposure measure by county, drawn from Aaronson and Mazumder (2011), separately by year.

Figure O.5: Alabama State Border Design



Notes: This figure shows the counties used in the Alabama state border design described in Section 5.6.

Table O.1: Means by Race in Population vs. Linked Sample: Males

	White males			Black males		
	Population	Linked sample (unweighted)	Linked sample (weighted)	Population	Linked sample (unweighted)	Linked sample (weighted)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Demographics						
Born in 1906-1910	0.36	0.06	0.37	0.36	0.11	0.36
Born in 1911-1914	0.31	0.28	0.31	0.32	0.3	0.31
Born in 1915-1918	0.32	0.66	0.33	0.32	0.59	0.32
Born in AL	0.07	0.06	0.07	0.11	0.1	0.11
Born in AR	0.06	0.06	0.06	0.05	0.05	0.05
Born in FL	0.02	0.03	0.02	0.03	0.04	0.03
Born in GA	0.08	0.07	0.08	0.14	0.12	0.14
Born in KY	0.1	0.1	0.1	0.02	0.02	0.02
Born in LA	0.05	0.05	0.05	0.08	0.08	0.09
Born in MD	0.04	0.07	0.04	0.02	0.03	0.03
Born in MS	0.04	0.03	0.04	0.11	0.1	0.11
Born in NC	0.09	0.08	0.09	0.1	0.11	0.1
Born in OK	0.08	0.09	0.07	0.01	0.02	0.01
Born in SC	0.04	0.04	0.04	0.11	0.1	0.11
Born in TN	0.09	0.08	0.09	0.05	0.05	0.05
Born in TX	0.16	0.16	0.16	0.08	0.09	0.09
Born in VA	0.07	0.08	0.07	0.08	0.09	0.07
Panel B: 1920 characteristics						
Rural status	—	0.74	0.78	—	0.85	0.88
Rosenwald exposure (ROSE)	—	0.21	0.14	—	0.2	0.14
Father literacy status	—	0.94	0.92	—	0.75	0.71
Mother literacy status	—	0.96	0.94	—	0.82	0.78
Homeownership status	—	0.5	0.52	—	0.29	0.28
Panel C: 1940 characteristics						
Years of education	9	9.96	8.82	5.86	6.91	5.8
Some high school	0.22	0.22	0.22	0.11	0.15	0.11
High school or more	0.3	0.42	0.28	0.08	0.15	0.08
Head of household	0.56	0.42	0.56	0.52	0.44	0.56
Married	0.64	0.49	0.63	0.65	0.55	0.66
Any children	0.41	0.26	0.41	0.34	0.31	0.41
In labor force	0.93	0.93	0.94	0.92	0.92	0.93
Occupation: other	0.01	0.01	0.01	0.01	0.01	0.01
Occupation: farmer	0.15	0.13	0.17	0.15	0.16	0.2
Occupation: unskilled	0.27	0.26	0.27	0.63	0.6	0.58
Occupation: skilled/semi-skilled	0.35	0.33	0.33	0.17	0.17	0.16
Occupation: white collar	0.23	0.27	0.23	0.04	0.06	0.05
Occupational income score (1950)	26.87	27	26.77	14.4	14.16	13.79
Annual wage income ($1940)	641.4	601.7	633.7	344.1	329.8	338
Live in state of birth	0.74	0.79	0.77	0.7	0.77	0.77
N	3,283,836	190,891		1,162,163	29,290	

Notes: Means in columns 3 and 6 weighted using inverse propensity score weights.

Table O.2: Means by Race in Population vs. Linked Sample: Females

	White females			Black females		
	Population	Linked sample (unweighted)	Linked sample (weighted)	Population	Linked sample (unweighted)	Linked sample (weighted)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Demographics						
Born in 1906-1910	0.36	0.17	0.36	0.35	0.16	0.35
Born in 1911-1914	0.32	0.29	0.31	0.32	0.29	0.32
Born in 1915-1918	0.32	0.55	0.33	0.32	0.55	0.33
Born in AL	0.07	0.07	0.07	0.11	0.1	0.11
Born in AR	0.06	0.06	0.06	0.05	0.04	0.05
Born in FL	0.02	0.03	0.02	0.03	0.04	0.04
Born in GA	0.08	0.07	0.08	0.14	0.12	0.14
Born in KY	0.1	0.1	0.1	0.02	0.03	0.02
Born in LA	0.05	0.06	0.05	0.09	0.08	0.09
Born in MD	0.04	0.06	0.04	0.02	0.03	0.02
Born in MS	0.04	0.04	0.04	0.11	0.09	0.11
Born in NC	0.09	0.08	0.09	0.1	0.12	0.1
Born in OK	0.08	0.08	0.07	0.01	0.02	0.01
Born in SC	0.04	0.04	0.04	0.11	0.1	0.11
Born in TN	0.09	0.08	0.09	0.05	0.05	0.05
Born in TX	0.16	0.15	0.16	0.08	0.08	0.08
Born in VA	0.07	0.07	0.07	0.08	0.1	0.07
Panel B: 1920 characteristics						
Rural status	—	0.76	0.78	—	0.83	0.85
Rosenwald exposure (ROSE)	—	0.18	0.14	—	0.19	0.14
Father literacy status	—	0.93	0.92	—	0.75	0.73
Mother literacy status	—	0.95	0.94	—	0.82	0.79
Homeownership status	—	0.5	0.51	—	0.29	0.29
Panel C: 1940 characteristics						
Years of education	9.56	10.03	9.4	6.83	7.95	7.11
Some high school	0.23	0.23	0.23	0.15	0.19	0.16
High school or more	0.37	0.43	0.35	0.12	0.21	0.14
Head of household	0.69	0.61	0.68	0.61	0.52	0.59
Married	0.76	0.68	0.74	0.72	0.62	0.66
Any children	0.56	0.49	0.57	0.47	0.45	0.5
In labor force	0.29	0.32	0.27	0.46	0.43	0.41
Occupation: other	0.19	0.17	0.2	0.11	0.12	0.13
Occupation: farmer	0.01	0.01	0.01	0.01	0.01	0.02
Occupation: unskilled	0.18	0.17	0.18	0.74	0.66	0.68
Occupation: skilled/semi-skilled	0.21	0.18	0.19	0.07	0.08	0.08
Occupation: white collar	0.41	0.47	0.42	0.07	0.12	0.1
Occupational income score (1950)	29.75	30.35	29.73	14	15.09	14.5
Annual wage income ($1940)	187	199.4	170.7	120.7	129.4	119
Live in state of birth	0.74	0.78	0.78	0.69	0.74	0.74
N	3,355,482	237,431		1,328,069	33,089	

Notes: Means in columns 3 and 6 weighted using inverse propensity score weights.

Table O.3: Means by Sex × Race in Population vs. Linked Sample (Unweighted): 1918 Cohort

	White males		Black males		White females		Black females	
	Pop.	Linked sample	Pop.	Linked sample	Pop.	Linked sample	Pop.	Linked sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Demographics								
Born in AL	0.07	0.06	0.1	0.09	0.07	0.07	0.1	0.09
Born in AR	0.06	0.06	0.05	0.05	0.06	0.07	0.05	0.04
Born in FL	0.02	0.03	0.03	0.04	0.03	0.03	0.03	0.04
Born in GA	0.08	0.07	0.14	0.12	0.08	0.07	0.14	0.12
Born in KY	0.1	0.1	0.02	0.02	0.1	0.1	0.02	0.03
Born in LA	0.05	0.05	0.08	0.09	0.05	0.06	0.08	0.08
Born in MD	0.04	0.07	0.02	0.03	0.04	0.07	0.02	0.03
Born in MS	0.04	0.03	0.1	0.09	0.04	0.03	0.1	0.09
Born in NC	0.09	0.08	0.11	0.12	0.09	0.08	0.11	0.12
Born in OK	0.08	0.09	0.02	0.02	0.08	0.09	0.02	0.02
Born in SC	0.04	0.04	0.12	0.1	0.04	0.04	0.12	0.1
Born in TN	0.09	0.08	0.04	0.05	0.09	0.08	0.05	0.05
Born in TX	0.15	0.15	0.08	0.09	0.15	0.14	0.08	0.08
Born in VA	0.07	0.08	0.08	0.1	0.07	0.07	0.08	0.1
Panel B: 1940 characteristics								
Years of education	9.14	9.86	6.13	6.97	9.8	10.13	7.25	8.1
Some high school	0.23	0.23	0.14	0.17	0.24	0.24	0.19	0.22
High school or more	0.33	0.41	0.09	0.15	0.41	0.46	0.15	0.22
Head of household	0.21	0.2	0.23	0.23	0.46	0.44	0.41	0.37
Married	0.27	0.26	0.33	0.32	0.57	0.53	0.57	0.5
Any children	0.11	0.1	0.14	0.15	0.34	0.32	0.36	0.35
In labor force	0.86	0.87	0.88	0.88	0.34	0.37	0.43	0.41
Occupation: other	0.03	0.03	0.03	0.03	0.18	0.16	0.14	0.13
Occupation: farmer	0.1	0.1	0.11	0.11	0.01	0.01	0.01	0.01
Occupation: unskilled	0.38	0.34	0.7	0.67	0.22	0.19	0.74	0.69
Occupation: skilled/semi-skilled	0.32	0.32	0.13	0.15	0.19	0.19	0.05	0.07
Occupation: white collar	0.16	0.21	0.03	0.04	0.4	0.46	0.05	0.09
Occupational income score (1950)	23.82	24.73	12.97	13.23	28.93	29.64	13.2	14.11
Annual wage income ($1940)	361.2	394.3	222.5	235.5	160.9	183.5	87.2	100.8
Live in state of birth	0.79	0.81	0.79	0.82	0.78	0.8	0.76	0.77
N	279,901	34,481	95,114	4,857	279,555	33,590	98,591	4,917

Notes: Means in linked sample are unweighted.

Table O.4: The Impact of Rosenwald Schools on Male Selected 3-Digit Occupations

	Triple difference estimate (γ_3) (1)	Mean annual wage income in 1940 (black workers, '00$) (2)	Median years of education in 1940 (white workers) (3)	Black-white ratio of emp. shares in 1940 (high school or more) (4)
Panel A: Managers				
Managers, officials, and proprietors (n.e.c.)	0.028**	9	11	0.2
	(0.014)			
Food stores, except dairy products	0.006*	8.1	9	0.3
	(0.003)			
Gasoline service stations	0.006*	7.1	10	0.2
	(0.003)			
Misc. entertainment and recreation services	0.003***	8.2	10	0.5
	(0.001)			
Printing, publishing, and allied industries	0.001**	13.8	12	0.3
	(0.001)			
Buyers and shippers, farm products	0.001*	8.4	9	0.1
	(0.001)			
Panel B: Clerical workers				
Messengers and office boys	0.008**	9.7	11	3.8
	(0.003)			
Stenographers, typists, and secretaries	0.006**	9.3	12	0.3
	(0.003)			
Panel C: Service workers				
Service workers, except private HH (n.e.c.)	0.026**	4.8	8	12.1
	(0.012)			
Eating and drinking places	0.014**	4.3	8	9.4
	(0.007)			
Hotels and lodging places	0.011	5	9	18.7
	(0.010)			
Waiters and waitresses	0.017	6	10	7.8
	(0.011)			
Attendants, professional and personal service (n.e.c.)	0.002*	5.1	10	2.7
	(0.001)			
Panel D: Craftsmen				
Plumbers and pipe fitters	-0.012***	6.9	8	0.5
	(0.004)			
Panel E: Operatives				
Operative and kindred workers (n.e.c.)	-0.105***	7.2	8	1.1
	(0.030)			
Yarn, thread, and fabric mills	-0.033**	5.8	6	0.2
	(0.016)			
Laundering, cleaning, and dyeing services	-0.020**	6.2	9	6.1
	(0.010)			
Apparel and accessories	-0.016**	6.7	8	2.2
	(0.007)			
Blast furnaces, steel works, and rolling mills	-0.007**	9.9	8	1.7
	(0.003)			

Notes: Column 1 shows the triple difference estimate from equation (2), where the dependent variable is an indicator for being in a specific 3-digit occupation (row) or 3-digit occupation × 3-digit industry (indented row) in 1940. Robust standard errors in parentheses, clustered at the county level.. *** 1%, ** 5%, * 10% significance. Summary statistics in columns 2-4 based on male workers aged 18-65 in 1940 (with a high school education or more in column 4).

Table O.5: The Impact of Rosenwald Schools on Female Selected 3-Digit Occupations

	Triple difference estimate (γ_3) (1)	Mean annual wage income in 1940 (black workers, '00$) (2)	Median years of education in 1940 (white workers) (3)	Black-white ratio of emp. shares in 1940 (high school or more) (4)
Panel A: Clerical workers				
Stenographers, typists, and secretaries	0.019	9.3	12	0.3
	(0.021)			
Clerical and kindred workers (n.e.c.)	0.016	12	12	0.5
	(0.014)			
Office machine operators	0.007**	11	12	0.3
	(0.003)			
Attendants and assistants, library	0.004*	5.4	12	1
	(0.002)			
Panel B: Professionals				
Teachers (n.e.c.)	0.019	8.5	16	2.5
	(0.024)			
Panel C: Private household workers				
Private household workers (n.e.c.)	0.116**	4.2	8	27.4
	(0.053)			
Panel D: Operatives				
Operative and kindred workers (n.e.c.)	-0.097**	7.2	7	1.1
	(0.048)			
Yarn, thread, and fabric mills	-0.040	5.8	6	0.2
	(0.025)			
Apparel and accessories	-0.029	6.7	7	2.2
	(0.024)			
Miscellaneous manufacturing industries	-0.011**	6.7	8	1.1
	(0.005)			
Laundry and dry cleaning operatives	-0.019	6.9	6	5
	(0.015)			

Notes: Column 1 shows the triple difference estimate from equation (2), where the dependent variable is an indicator for being in a specific 3-digit occupation (row) or 3-digit occupation $\times$ 3-digit industry (indented row) in 1940. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance. Summary statistics in columns 2 and 3 based on male workers aged 18-65 in 1940. The black-white ratio of employment shares in column 4 is based on female workers aged 18-65 with a high school education or more in 1940.

Table O.6: White Collar Jobs: Median Years of Education & Interaction Task Content

1950 Census job title	Median years of education in 1940 (white workers)	Dictionary of Occupational Titles (DOT) job title	Job requires dealing with people
(1)	(2)	(3)	(4)
Panel A: Professionals			
Actors and actresses	12	Actor	Yes
Airplane pilots and navigators	13	Airplane pilot, commercial	No
Architects	16	Architect, marine	No
Artists and art teachers	12	Commercial artist, illustrating	No
Athletes	12		Yes†
Authors	14	Editorial writer	No
Chemists	16	Chemist, pharmaceutical	No
Chiropractors	15	Chiropractor	Yes
Clergymen	16		Yes†
Professors and instructors, subject not specified	17	Professor, college or university	Yes
Dancers and dancing teachers	11	Dancer	No
Dentists	16	Dental hygienist	Yes
Designers	12	Hat designer	No
Draftsmen	12		No†
Editors and reporters	14	Reporter	Yes
Engineers, chemical	16	Chemical engineer, reseach and development	No
Engineers, civil	16		No†
Engineers, electrical	16		No†
Engineers, industrial	16		No†
Engineers, mechanical	16		No†
Engineers, metallurgical, metallurgists	16	Metallurgist extractive	No
Entertainers (n.e.c.)	9		Yes†
Farm and home management advisors	16	Home-lighting advisor	No
Funeral directors and embalmers	12	Funeral attendant	Yes
		Embalmer	No
Lawyers and judges	16	Lawyer, civil	Yes
Librarians	13	Librarian assistant	Yes
		Librarian	No
Musicians and music teachers	12		Yes†
Nurses, professional	10	Nurse, general duty	Yes
Optometrists	14	Optometrist	Yes
Osteopaths	16	Osteopath	Yes
Pharmacists	14	Pharmacist	No
Photographers	12	Photographer, street	Yes
		Photographer, commercial	No
Physicians and surgeons	17	Orthopedic surgeon	Yes
		Surgeon I	No
Radio operators	12	Radio operator	No
Religious workers	12		Yes†
Social and welfare workers, except group	14		Yes†

(table continues on next page)

Table O.6 (cont.): White Collar Jobs: Median Years of Education & Interaction Task Content

1950 Census job title (1)	Median years of education in 1940 (white workers) (2)	Dictionary of Occupational Titles (DOT) job title (3)	Job requires dealing with people (4)
Panel A: Professionals (cont.)			
Sports instructors and officials	13		Yes†
Surveyors	9	Surveyor	No
Teachers (n.e.c.)	16	Teacher, high school	Yes
Technicians, testing	12		No†
Technicians (n.e.c.)	12		No†
Therapists and healers (n.e.c.)	12	Physical therapist	Yes
Veterinarians	16	Veterinarian	No
Professional, technical and kindred workers (n.e.c.)	12		No†
Panel B: Managers			
Buyers and department heads, store	12	Buyer II	No
Buyers and shippers, farm products	9		No†
Conductors, railroad	8	Conductor, road freight	No
Credit men	12	Credit cashier	Yes
Floormen and floor managers, store	12		No†
Inspectors, public administration	12	All-around inspector	No
Managers and superintendents, building	10	Superintendant, maintenance	No
Officers, pilots, pursers and engineers, ship	8	Purser	Yes
		Ship pilot	No
Officials and administrators (n.e.c.), public administration	12		Yes†
Officials, lodge, society, union, etc.	12		Yes†
Postmasters	12		Yes†
Purchasing agents and buyers (n.e.c.)	12	Purchasing agent	Yes
Managers, officials, and proprietors (n.e.c.)	11	Manager, sales	Yes
Panel C: Clerical occupations			
Agents (n.e.c.)	12	Purchasing agent	Yes
Attendants and assistants, library	12	Librarian assistant	Yes
Attendants, physicians and dentists office	11		Yes†
Baggagemen, transportation	9	Baggage porter	Yes
Bookkeepers	12	Bookkeeper II	No
Collectors, bill and account	11	Collector II	Yes
Express messengers and railway mail clerks	12	Corporate messenger	Yes
		Messenger, bank	Yes
		Mail clerk	No
Mail carriers	11		Yes†
Messengers and office boys	11	Corporate messenger	Yes
		Mail boy	No
Office machine operators	12	Office machine serviceman	No
Shipping and receiving clerks	10	Shipping clerk I	No
Stenographers, typists, and secretaries	12	Stenographer	No
		Typist	No
		Secretary	Yes

(table continues on next page)

Table O.6 (cont.): White Collar Jobs: Median Years of Education & Interaction Task Content

1950 Census job title	Median years of education in 1940 (white workers)	Dictionary of Occupational Titles (DOT) job title	Job requires dealing with people
(1)	(2)	(3)	(4)
Panel C: Clerical occupations (cont.)			
Telegraph messengers	9		Yes†
Telegraph operators	10	Telegraph operator	No
Telephone operators	9	Telephone operator I	No
Ticket, station, and express agents	12	Ticket taker	Yes
Clerical and kindred workers (n.e.c.)	12	File clerk I	No
		Receptionist I	Yes
Panel D: Sales occupations			
Advertising agents and salesmen	12	Salesperson, general	Yes
Auctioneers	10		Yes†
Demonstrators	10	Demonstrator	Yes
Hucksters and peddlers	7	Peddler I	Yes
Insurance agents and brokers	12		Yes†
Newsboys	10		Yes†
Real estate agents and brokers	12		Yes†
Stock and bond salesmen	12		Yes†
Salesmen and sales clerks (n.e.c.)	12		Yes†

Notes: Median years of education in column 2 based on white male workers aged 18-65 in 1940. Column 4 specifies whether the D.O.T. job title in column 3 requires "dealing with people," based on task content descriptions in U.S. Department of Labor (1956). If we could not find a closely related DOT job title for a particular 1950 Census job title (column 3 empty), we inferred whether the job requires dealing with people based on job descriptions in U.S. Department of Labor (1949) as well as our own judgement (indicated by † symbol). Column 3 either specifies the closest DOT job title for the corresponding 1950 Census job title in column 1 (e.g. "Dancer" for "Dancers and dancing teachers"), one example of a closely related DOT job title if there are several that map into the same category in column 4 (e.g. "Chemist, pharmaceutical" for "Chemists"), or two examples if there are several that map into different categories in column 4 (e.g. "Corporate messenger" and "Mail boy" for "Messengers and office boys").

Table O.7: Male Service Workers vs. Operatives by Income Quartile

	Dependent variable: 1[occupation group × wage income quartile in 1940]							
	Service workers				Operatives			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Black quartiles (1940$):	$\leqslant 330$	(330,520]	(520,792]	> 792	$\leqslant 355$	(355,572]	(572,840]	> 840
White quartiles (1940$):	$\leqslant 440$	(440,780]	(780,1200]	> 1200	$\leqslant 500$	(500,780]	(780,1200]	> 1200
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ROSE × black × rural	0.038*	0.016	0.029	0.023*	-0.054**	-0.040*	-0.035	-0.019
	(0.020)	(0.017)	(0.019)	(0.014)	(0.025)	(0.021)	(0.023)	(0.021)
R^2	0.039	0.041	0.044	0.043	0.045	0.054	0.043	0.050
N	153,375	153,375	153,375	153,375	153,375	153,375	153,375	153,375

Notes: All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table O.8: Female Private Household Workers vs. Operatives by Income Quartile

	Dependent variable: 1[occupation group × wage income quartile in 1940]							
	Private household workers				Operatives			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Black quartiles (1940$):	$\leqslant 104$	(104,188]	(188,312]	> 312	$\leqslant 200$	(200,360]	(360,540]	> 540
White quartiles (1940$):	$\leqslant 100$	(100,185]	(185,344]	> 344	$\leqslant 300$	(300,500]	(500,675]	> 675
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ROSE × black × rural	-0.000	-0.004	-0.018	0.105***	-0.017	-0.056***	-0.032	-0.041*
	(0.035)	(0.037)	(0.039)	(0.038)	(0.022)	(0.019)	(0.021)	(0.024)
R^2	0.140	0.119	0.126	0.136	0.051	0.060	0.069	0.075
N	77,431	77,431	77,431	77,431	77,431	77,431	77,431	77,431

Notes: All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table O.9: Income Reporting by Occupation Group in 1940, Southern Born Males

	Wage income categories (%)		Non-wage income	Share self-	Employment
	Missing	< \$6 (weekly)	⩾ \$50 (%)	employed (%)	share (%)
	(1)	(2)	(3)	(4)	(5)
Panel A: Southern born black males (18-65)					
Other	88.8	3.1	11.9	4.3	0
Farm laborers	28.3	43.8	16.3	2.2	17.4
Private household workers	12.1	30.7	17.3	4.9	2.8
Laborers	10.2	10.9	10.5	2.2	27
Farmers	85.9	9.8	74.6	96.8	21
Service workers	9.6	10.6	14.2	3.7	10.4
Operatives	8.5	8.1	9.6	3.5	12
Craftsmen	19.7	6.4	19.9	17.1	4.4
Clerical workers	7.4	4.9	11.9	1.9	0.9
Sales workers	27.6	10.5	26.5	22	0.7
Professionals	25.7	8.3	29.8	15.5	2
Managers	62.4	5.1	58	70.8	1.2
Panel B: Southern born white males (18-65)					
Other	85.6	2.8	14.2	6.1	0
Farm laborers	38.8	23.6	23.8	3.4	9.4
Private household workers	17.6	18.9	22.2	7.9	0.2
Laborers	11.9	10.6	15.7	4.5	10.2
Farmers	84.8	7.1	76.4	96	21.5
Service workers	14.4	5.1	21.3	10.7	3.5
Operatives	8.8	3.2	13.8	5.1	16.4
Craftsmen	13.4	4.1	19.6	11.3	14.3
Clerical workers	6.4	2.1	15.6	1.9	5.6
Sales workers	15.9	3	22.3	8.7	5.6
Professionals	26.2	2.1	37	24.2	4.6
Managers	47.2	1.3	52.6	52.7	8.7

Notes: South includes 14 Rosenwald states (AL, AR, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, and VA).

Table O.10: Alternative Linked Datasets

	Dependent variable: Outcome in 1940							
	Years of education	In labor force	Farmer	Unskilled job	Skilled/ semi-skilled job	White collar job	Married	Any children
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Males, Multigenerational Longitudinal Panel (MLP)								
ROSE × black × rural	0.998***	-0.038***	-0.006	0.099***	-0.145***	0.044**	-0.085**	-0.071**
	(0.287)	(0.014)	(0.019)	(0.024)	(0.033)	(0.022)	(0.042)	(0.031)
R^2	0.358	0.033	0.139	0.158	0.082	0.184	0.164	0.162
N	981,799	997,145	933,676	933,676	933,676	933,676	997,145	997,145
Panel B: Females, Multigenerational Longitudinal Panel (MLP)								
ROSE × black × rural	0.757**	0.063	-0.005	0.116**	-0.152**	0.034	-0.046	-0.025
	(0.348)	(0.045)	(0.007)	(0.056)	(0.070)	(0.043)	(0.039)	(0.037)
R^2	0.298	0.081	0.043	0.335	0.157	0.294	0.058	0.080
N	267,960	272,503	155,982	155,982	155,982	155,982	272,503	272,503
Panel C: Males, Census Linking Project (CLP), standard Abramitzky Boustan Eriksson (2012) method								
ROSE × black × rural	0.875**	-0.021	-0.015	0.053***	-0.083***	0.042**	-0.117***	-0.072***
	(0.378)	(0.013)	(0.014)	(0.020)	(0.024)	(0.019)	(0.028)	(0.026)
R^2	0.282	0.021	0.083	0.153	0.061	0.141	0.147	0.125
N	933,613	949,570	898,765	898,765	898,765	898,765	949,570	949,570
Panel D: Males, Census Linking Project (CLP), conservative Abramitzky Boustan Eriksson (2012) method								
ROSE × black × rural	1.039***	-0.025	-0.001	0.076***	-0.117***	0.045**	-0.134***	-0.072**
	(0.377)	(0.017)	(0.016)	(0.024)	(0.027)	(0.022)	(0.032)	(0.034)
R^2	0.304	0.025	0.100	0.150	0.063	0.156	0.156	0.135
N	635,744	646,228	612,838	612,838	612,838	612,838	646,228	646,228

Notes: Each coefficient corresponds to the triple difference estimate from equation (2), where the dependent variable is indicated in the column, using the linked sample indicated in the panel title (see text for details). All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table O.11: Alabama State Border Design

	Dependent variable: Outcome in 1940							
	Years of education	In labor force	Farmer	Unskilled job	Skilled/ semi-skilled job	White collar job	Married	Any children
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Males born in 1900-1906 linked to 1910 & 1940 Censuses								
TUSKCOUNTY × black	0.539	0.002	-0.079	-0.037	-0.030	0.146	0.052	0.090
	(0.957)	(0.078)	(0.101)	(0.097)	(0.110)	(0.092)	(0.108)	(0.117)
R^2	0.361	0.221	0.256	0.248	0.211	0.241	0.204	0.222
N	515	531	504	504	504	504	531	531
Panel B: Females born in 1900-1906 linked to 1910 & 1940 Censuses								
TUSKCOUNTY × black	0.423	0.102	0.068*	-0.127	0.131*	0.063	-0.093	-0.182**
	(0.441)	(0.072)	(0.039)	(0.111)	(0.068)	(0.084)	(0.056)	(0.079)
R^2	0.301	0.121	0.243	0.379	0.196	0.307	0.096	0.094
N	2,173	2,243	643	643	643	643	2,243	2,243

Notes: Sample restricted to individuals in 1910-1940 linked sample living in Alabama state border counties in 1910 (see Online Appendix Figure O.5). All regressions include a Tuskegee county indicator interacted with a black indicator and a non-Alabama county indicator (TUSKCOUNTY × black × non-AL), 1910 county of residence fixed effects, cohort fixed effect, state of birth fixed effects, race-by-rural fixed effects, and 1910 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Observations weighted using inverse propensity score weights. Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table O.12: Robustness Checks: Male Results

	Baseline results	95% precision	90% precision	Reside in county of birth	Additional 1920 HH controls	Inverse probability weights	Unweighted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Dependent variable: years of education in 1940							
ROSE × black × rural	1.407***	1.278***	1.023***	1.314***	1.411***	1.380***	1.468***
	(0.338)	(0.345)	(0.341)	(0.381)	(0.314)	(0.273)	(0.266)
R^2	0.317	0.307	0.281	0.344	0.351	0.325	0.283
N	216,663	232,426	270,462	143,992	216,663	216,663	216,663
Panel B: Dependent variable: unskilled job in 1940							
ROSE × black × rural	0.122***	0.123***	0.136***	0.127***	0.106***	0.109***	0.091**
	(0.042)	(0.040)	(0.034)	(0.048)	(0.040)	(0.038)	(0.035)
R^2	0.156	0.150	0.137	0.168	0.168	0.168	0.133
N	205,415	220,399	256,335	136,212	205,415	205,415	205,415
Panel C: Dependent variable: skilled/semi-skilled job in 1940							
ROSE × black × rural	-0.134***	-0.117***	-0.120***	-0.155***	-0.124***	-0.137***	-0.132***
	(0.044)	(0.043)	(0.036)	(0.054)	(0.044)	(0.042)	(0.040)
R^2	0.080	0.076	0.066	0.090	0.100	0.067	0.053
N	205,415	220,399	256,335	136,212	205,415	205,415	205,415
Panel D: Dependent variable: white collar job in 1940							
ROSE × black × rural	0.064**	0.046	0.036	0.054	0.072**	0.073***	0.082***
	(0.030)	(0.028)	(0.028)	(0.037)	(0.031)	(0.028)	(0.025)
R^2	0.173	0.169	0.156	0.195	0.207	0.192	0.164
N	205,415	220,399	256,335	136,212	205,415	205,415	205,415

Notes: Each coefficient corresponds to the triple difference estimate from equation (2), where the dependent variable is indicated in the panel title, for different linked samples (columns 2-3), sample restrictions (column 4), control variables (column 5), and weighting schemes (columns 6-7). All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Additional 1920 household controls in column 5 include father 3-digit occupation fixed effects, an indicator for farm status, father and mother age-by-place of birth fixed effects, an indicator for mother labor force participation, and fixed effects for family size. Observations weighted using inverse propensity score weights (except in columns 6-7). Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.

Table O.13: Robustness Checks: Female Results

	Baseline results	95% precision	90% precision	Reside in county of birth	Additional 1920 HH controls	Inverse probability weights	Unweighted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Dependent variable: years of education in 1940							
ROSE × black × rural	1.156***	1.200***	1.179***	1.162***	1.188***	1.024***	1.157***
	(0.316)	(0.296)	(0.306)	(0.402)	(0.287)	(0.307)	(0.301)
R^2	0.264	0.254	0.232	0.285	0.295	0.259	0.236
N	266,150	297,277	358,058	177,068	266,150	266,150	266,150
Panel B: Dependent variable: in labor force in 1940							
ROSE × black × rural	0.098***	0.094***	0.076***	0.077**	0.097***	0.093***	0.082**
	(0.031)	(0.032)	(0.027)	(0.033)	(0.031)	(0.031)	(0.032)
R^2	0.072	0.069	0.063	0.077	0.078	0.069	0.063
N	270,520	302,184	364,004	180,056	270,520	270,520	270,520
Panel C: Dependent variable: married in 1940							
ROSE × black × rural	-0.087**	-0.091**	-0.057	-0.086	-0.084**	-0.100**	-0.105**
	(0.040)	(0.039)	(0.036)	(0.052)	(0.040)	(0.041)	(0.043)
R^2	0.068	0.065	0.058	0.074	0.076	0.075	0.068
N	270,520	302,184	364,004	180,056	270,520	270,520	270,520
Panel D: Dependent variable: any children in 1940							
ROSE × black × rural	-0.101**	-0.114***	-0.085**	-0.113**	-0.102**	-0.099**	-0.101**
	(0.042)	(0.043)	(0.038)	(0.048)	(0.042)	(0.041)	(0.041)
R^2	0.089	0.085	0.078	0.096	0.097	0.095	0.089
N	270,520	302,184	364,004	180,056	270,520	270,520	270,520

Notes: Each coefficient corresponds to the triple difference estimate from equation (2), where the dependent variable is indicated in the panel title, for different linked samples (columns 2-3), sample restrictions (column 4), control variables (column 5), and weighting schemes (columns 6-7). All regressions include the baseline Rosenwald exposure measure (ROSE), its interaction with a black indicator (ROSE × black), its interaction with a rural indicator (ROSE × rural), 1920 county of residence fixed effects, cohort-by-state of birth fixed effects, cohort-by-race-by-rural fixed effects, and 1920 household controls (father occupational income score, homeownership status, father literacy status, mother literacy status). Additional 1920 household controls in column 5 include father 3-digit occupation fixed effects, an indicator for farm status, father and mother age-by-place of birth fixed effects, an indicator for mother labor force participation, and fixed effects for family size. Observations weighted using inverse propensity score weights (except in columns 6-7). Robust standard errors in parentheses, clustered at the county level. *** 1%, ** 5%, * 10% significance.