The Missing Link(s):

Women and Intergenerational Mobility*

Lukas Althoff[†] Harriet Brookes Gray[‡] Hugo Reichardt[§]

[Most recent version here]

First version: November 19, 2022.

This version: November 29, 2023.

Abstract

Throughout US history, mothers played an important role in educating their children. However, this role is obscured because data limitations have prevented researchers from including women in intergenerational studies. By leveraging data from Social Security applications, we build a large panel that tracks both men and women over time despite marital name changes. We measure intergenerational mobility as the share of variation in child outcomes explained by parental background (R^2), which we decompose into mothers' and fathers' separate contributions. We find that a mother's human capital is a crucial determinant of her child's outcomes, often surpassing the influence of fathers, especially for daughters and Black children. Generally, maternal human capital is more important for groups and places with limited access to education. Incorporating mothers' human capital suggests that, contrary to current evidence, intergenerational mobility increased over the 19th century.

*We thank Leah Boustan, Ellora Derenoncourt, Alice Evans, John Grigsby, Ilyana Kuziemko, and Pablo Valenzuela for insightful comments. Pedro Carvalho and Alex Shaffer provided excellent research assistance. This paper previously circulated under the title "Intergenerational Mobility and Assortative Mating."

[†]Stanford Institute for Economic Policy Research, Stanford University. lalthoff@stanford.edu

[‡]Department of Economics, Yale University. harriet.brookesgray@yale.edu

[§]Department of Economics, London School of Economics. h.a.reichardt@lse.ac.uk

1. INTRODUCTION

Mothers played a key role in educating their children, especially before the rise of public schools around 1900. Despite their prominent role in children's development, the historical contribution of mothers to their children's outcomes has been understudied due to data limitations. Existing evidence on how parental background has shaped child outcomes over US history focuses almost exclusively on the role of fathers' incomes. As a result, our understanding of intergenerational mobility remains incomplete.

In this paper, we study the influence of parental status, encompassing the roles of both mothers and fathers, on the outcomes of their children in the US from 1850 to 1940. To do so, we build a comprehensive and representative panel—one of the first to include women for this era. Historical administrative records allow us to track women's records over time despite name changes upon marriage. Our analysis first quantifies the intergenerational transmission of human capital and income. We then assess mothers' and fathers' separate contributions to shaping child outcomes. Our findings reveal that maternal human capital is a stronger predictor of child outcomes than paternal human capital. Disparities in school access across places and demographic groups explain much of the variation in the significance of maternal human capital. These results are consistent with mothers' central role as educators during the 19$^{\text{th}}$ century.

We first overcome the challenge of linking women's census records despite name changes by leveraging historical administrative data from Social Security Number applications. These applications provide both married and maiden names for millions of mothers and married female applicants. Using these data, we link the census records of 21 million women, resulting in a highly representative panel. We make this new dataset publicly available to further the reassessment of women's contribution to US history.

Second, to assess the importance of both mothers and fathers, we propose measuring intergenerational mobility as the share of variation in child outcomes explained by parental background (R^2). Unlike traditional mobility measures, such as the parent-child coefficient, the R^2 measure accommodates multiple parental inputs. We show that the R^2

has many desirable properties as a mobility measure and—in the special case of a single parental input—has a one-to-one relationship with the rank-rank coefficient. Another advantage of R^2 is that it can be separated into each parent's predictive power using statistical decomposition methods (Shapley, 1953; Owen, 1977). Lastly, using a range of parental variables to flexibly predict the child's outcome can alleviate the effect of measurement error in any one variable, a problem that is particularly acute in the historical context (Ward, 2023).

Third, we use a recently developed semiparametric latent variable method (Fan et al., 2017) to study rank-rank relationships between parents and children when only binary proxies of the variables of interest are observed. In the historical data, such binary proxies are common. For example, we use literacy as a binary proxy for human capital. Using the latent variable method, we recover rank-rank relationships between the latent outcomes of interest while imposing only mild assumptions on the distribution of the unobserved variables.

Our first key finding is that parental human capital significantly influenced the intergenerational transmission of income. While the separate importance of parental human capital and income is a central aspect of intergenerational mobility theory Becker et al. (2018), most prior empirical studies focused on income-to-income transmission alone. Our findings show that intergenerational mobility increased over this period after incorporating the role of parental human capital, contrasting findings based on parental income alone.

Consistent with our results, historians highlight the pivotal role of parental human capital. Until around 1900, public schooling was limited and home education was common. Mothers, primarily engaged in home production in this era, were key educators of their children (Dreilinger, 2021). "[T]he middle class mother was advised that she and she alone had the weighty mission of transforming her children into the model citizens of the day" (Margolis, 1984, p. 13).

Our second main findings shows that, in line with the qualitative historical evidence, mothers' human capital often more strongly predicts their children's outcomes than fathers'. We find that mothers on average account for 13% more of the variation in child human capital than fathers—for pre-1900 cohorts up to 43% more. This finding suggests that mothers play a critical role in the intergenerational transmission of human capital.

Our third key finding is that mothers' human capital is particularly predictive of daughters' and Black children's outcomes. For daughters, mothers account for one-third more of variation in outcomes than fathers—for Black children almost two-thirds more. This finding suggests that mothers may have an even greater role in shaping the outcomes of female and Black children. Thus, it is particularly important to incorporate the role of mothers when comparing mobility estimates across groups.

As a potential mechanism for the differential importance of maternal human capital, we explore variation in access to schools across race, place, and time. We find that in places where schools are less prevalent, mothers are particularly important predictors of child human capital. For children born before the rise of widespread public schools, school access explains half of the variation in mothers' predictive power relative to fathers. Similarly, we find that as school provision expands over time, the predictive power of maternal human capital decreases. These findings underscore the historical role of mothers in home education, particularly in settings where formal schooling options were limited or absent (Kaestle and Vinovskis, 1978; Margolis, 1984; Dreilinger, 2021).

We validate our main results using census data on children (ages 13–16) in parental households. We relate child outcomes to their parents' outcomes using census cross-sections, bypassing the need for record linkage. Results on transmission of human capital and formal schooling based on the cross-section mirror those based on our new panel. Furthermore, we validate our semiparametric latent variable model and show that, in contrast to ordinary least squares (OLS), it correctly identifies levels and trends in rank-rank mobility based on binary proxies (e.g., literacy) for unobserved underlying outcomes of interest (e.g., human capital).

A key contribution of this paper is to construct one of the most extensive and representative panels on intergenerational mobility to include women, building on the foundations of previous work. Craig et al. (2019) and Bailey et al. (2022) pioneered the effort to link women's records by expanding automated record linkage developed by Abramitzky et al. (2021b). However, the information they use to do so—historical birth, marriage, and death certificates—are available only for selected states and periods. Buckles et al. (2023) innovatively use crowd-sourced family trees, leading to significantly larger sample sizes. However, concerns about representativeness remain owing to the selectivity of users who contribute to these genealogies. In contrast to prior work, we leverage historical *administrative* data, allowing for both scale and representativeness. Espín-Sánchez et al. (2023) employ a small subset of the same administrative data. They develop parametric assumptions under which the role of women in intergenerational mobility can be inferred from the outcomes of male family members. Instead, we develop methods to estimate women's role in intergenerational mobility directly, understanding the mechanisms underlying their impact on children's human capital and economic outcomes.

This paper also deepens our insights into how mothers' background shaped Americans' life chances throughout history. Earlier studies often used father-son dynamics to measure intergenerational mobility (Abramitzky et al., 2021a; Ward, 2023). More recent work has extended this to father-daughter relationships based on new linked panels that include women or more modern survey and administrative data, revealing differences in mobility between sons and daughters (Craig et al., 2019; Jácome et al., 2021; Buckles et al., 2023; Chetty et al., 2014). Card et al. (2022) explore early-life human capital transmission from aggregate statistics on both parents. These prior studies do not assess mothers' importance in the intergenerational transmission of economic outcomes. Our paper emphasizes the separate roles of mothers and fathers in shaping their children's outcomes, uncovering that maternal human capital is a stronger predictor of child outcomes than father-based proxies, especially for female and Black children.

Incorporating mothers in studying the evolution of mobility over US history seems especially pressing given that in other contexts, evidence suggests that mothers are key determinants of child outcomes. For Norway, Black et al. (2005) show that transmission of additional parental education to their children can be detected only between mothers (not fathers) and sons (not daughters). García and Heckman (2023) show that programs to increase mothers' parenting skills increase intergenerational mobility. Leibowitz (1974) shows that mothers' education is a strong predictor of child IQ whereas fathers' education is not, which they argue is a result of mothers spending more time with their children than fathers.

Lastly, this paper is part of an ongoing reassessment of empirical evidence on intergenerational mobility in US history. Ward (2023) has illuminated the impact of measurement errors on mobility estimates. Jácome et al. (2021) demonstrate that excluding certain groups, notably Black daughters, skews perceptions of mobility trends. Eshaghnia et al. (2022) show that mobility measured in lifetime outcomes is magnitudes lower than mobility measured at a single point of a person's life cycle. Our empirical findings underline the significance of mothers and human capital, showing that a father-only focus inflates mobility rates and confounds comparisons of mobility across groups and over time.

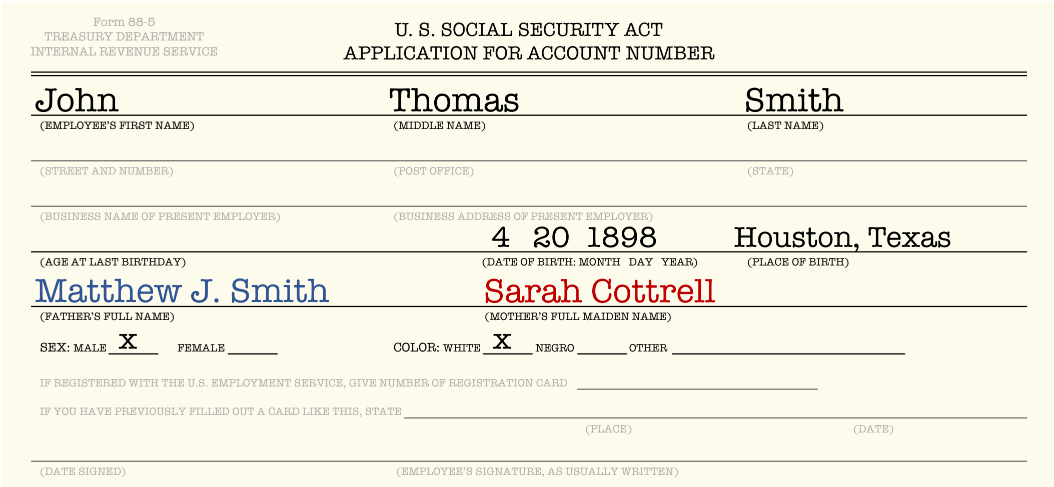
2. A New Panel that Includes Women (1850–1940)

The main empirical challenge in including women to study the long-run evolution of intergenerational mobility is the lack of suitable panel data. In this section, we describe how we overcome this hurdle by combining census records with historical administrative data that contain the married and maiden names of millions of women. Using these data, we link adult men and women in historical censuses (1850-1940) to their childhood census records. The resulting panel data is unique in its coverage and representativeness, particularly because it includes women.

2.1 Historical Administrative Data (Social Security Administration)

Our historical administrative data comprise 41 million Social Security Number (SSN) applications, covering the near-universe of applicants. For data privacy reasons, only applicants who died before 2008 are included. The data contain applicant's name, age, race, place of birth, and the maiden names of their parents (see Figure 1). Based on these data, we can derive the married and maiden names of millions of women including all

FIGURE 1: Social Security Application Form


Form 88-5
TREASURY DEPARTMENT
INTERNAL REVENUE SERVICE

U. S. SOCIAL SECURITY ACT
APPLICATION FOR ACCOUNT NUMBER

John (EMPLOYEE'S FIRST NAME) Thomas (MIDDLE NAME) Smith (LAST NAME)

(STREET AND NUMBER) (POST OFFICE) (STATE)

(BUSINESS NAME OF PRESENT EMPLOYER) (BUSINESS ADDRESS OF PRESENT EMPLOYER)

(AGE AT LAST BIRTHDAY) 4 20 1898 (DATE OF BIRTH: MONTH DAY YEAR) Houston, Texas (PLACE OF BIRTH)

Matthew J. Smith (FATHER'S FULL NAME) Sarah Cottrell (MOTHER'S FULL MAIDEN NAME)

SEX: MALE X FEMALE ____ COLOR: WHITE X NEGRO ____ OTHER ____

IF REGISTERED WITH THE U.S. EMPLOYMENT SERVICE, GIVE NUMBER OF REGISTRATION CARD ____

IF YOU HAVE PREVIOUSLY FILLED OUT A CARD LIKE THIS, STATE ____ (PLACE) (DATE)

(DATE SIGNED) (EMPLOYEE'S SIGNATURE, AS USUALLY WRITTEN)

Notes: This figure sketches a filled-in Social Security application form. Besides the applicants' name, address, employer, year and state of birth, and race, the application includes the father's name and the mother's maiden name. We access a digitized version of these data.

applicants' mothers as well as the smaller group of female applicants who were married at the time of application. We sourced a digitized version of these data from the National Archives and Records Administration (NARA).

Representativeness. Initially, SSN applicants were not representative of the US population, as the SSN system was launched in 1935 to register employed individuals, excluding self-employed and certain other occupations (Puckett, 2009). However, its scope rapidly expanded; for example, Executive Order 9397 in 1943 and the IRS's adoption of SSNs for tax reporting in 1962 increased its coverage. Throughout, the share of female applicants has been close to 50 percent. The representativeness of our sample is further improved by parents who enter our sample irrespective of whether they applied for an SSN.

Coverage. The data has extensive coverage of men and women born in the 1880s or after. The majority of Americans born in or after 1915 were assigned an SSN and therefore enter our data as applicants—a fact we establish by comparing each cohort's number of births and SSNs (CDC, 2023; SSA, 2023). The share of Americans with an SSN rises from 64 percent for those born in 1915 to 80 percent for those born in 1920, 90 percent for 1935, and close to 100 percent starting with those born in 1950. Parents extend this coverage from around 1915 to approximately 1890, given that the median age at the first birth of a child was 25 among parents of the 1915 cohort.[1] A detailed examination on linked samples will address remaining selection issues.

2.2 Census Data

We use the full-count census data for all available decades between 1850 and 1940 (Ruggles et al., 2020). These data include each person's full names, state and year of birth, sex, race, marital status, and other information. The data also identify family interrelationships for individuals in the same household. For those who live with their parents or spouses, we therefore also observe parental or spousal information.

2.3 Linking Method

We use a multi-stage linking process to maximize the utility of SSN application data, building on existing methods of automated record linkage (Abramitzky et al., 2021b). This procedure consists of three stages: linking SSN applicants to census records, linking applicants' parents to census records, and tracking these records over time.

First stage: Applicant SSN $\leftrightarrow$ census. We start by linking each SSN applicant to their corresponding census record, using a rich set of criteria such as full names of the applicants *and* their parents, year and state of birth, race, and sex. The criteria are then progressively relaxed to the literature standard, which involves only first and last name with spelling variations allowed, state of birth, and year of birth within a 5-year band. For married female applicants, we search for potential census matches using both their maiden and married names. A link is established if a unique match is found; if dual matches occur, we discard the individual.

Second stage: Parent SSN $\leftrightarrow$ census. After linking SSN applicants to their census records, we focus on connecting their parents to the census. Since specific birth details for mothers are not available in the SSN applications, we cannot directly link them like we do for the applicants. However, if a child's SSN application is successfully matched to

[1] According to the 1920 census, the average age for first-time parents is 28 for fathers and 24 for mothers.

a census record, and that record shows the child residing with their parents, we can link the parents to that specific census household. For parents who are not SSN applicants themselves, we create a synthetic identifier similar to an SSN.

Third stage: Census $\leftrightarrow$ census. Having assigned unique SSNs or synthetic identifiers to millions of individuals in the census records, we can link these records over time. We cover all possible pairs of census decades from 1850 to 1940.

In principle, it would be possible to establish additional links across census records by using standard or machine learning methods. These methods would be particularly useful for men and never-married women, where the issue of name changes does not apply. However, we choose not to use these methods for two reasons. First, our dataset's unique value lies in its ability to trace women from childhood to adulthood, capturing name changes upon marriage—a feature not addressable by standard linking or machine learning methods. Second, using different methods for different subgroups would compromise the comparability and representativeness of our sample, as not all groups would be linked based on a consistent set of criteria.

2.4 Our New Panel

In the first two stages, our process assigns SSNs to 36 million census records—16 million applicants and 20 million parents. The implied linking rate is 40 percent for applicants, surpassing the more typical 25 percent of prior studies thanks to our use of detailed information, notably parent names. In the third stage, we link 112 million census records over time, tracking each of the 36 million individuals through more than three census decade pairs on average.

A standout feature of the panel is the inclusion of 12 million women for whom we observe pre- and post-marriage data. The sample sizes are largest for women born between the 1890s and the 1920s, with each birth decade containing 1.5 to 3 million women. These data allows us to overcome critical data limitations to study the role of women in intergenerational mobility throughout US history.

Our panel is highly representative of the overall US population across several metrics, including gender and race (see Figure 2). Women comprise 46 percent of our linked sample in 1940. The sample mirrors the US-born and foreign-born shares of the population. While Black Americans are slightly underrepresented, our panel exceeds the representativeness of other samples in this dimension as well. Socioeconomic factors like income, home ownership, years of education, and literacy also align well with the broader population. Our sample over-represents married individuals, possibly because, if they are known to us, we use the names of a person's children or spouses in the linking procedure, improving linking rates of those who have children, a spouse, or both. We reweight our sample to more closely resemble the US population's characteristics throughout this paper.[2] Our reweighted sample is close to perfectly representative of the full population, even in characteristics not directly targeted by the reweighting method. The panel maintains its representative quality even in the earliest census decades (see Appendix Figure C.7).

Moreover, our panel offers broad coverage. It captures 7–20 percent of the US population from 1910–1940 and 1–5 percent from 1850–1900 (see Figure 3). This extensive reach makes our sample highly valuable for longitudinal studies.

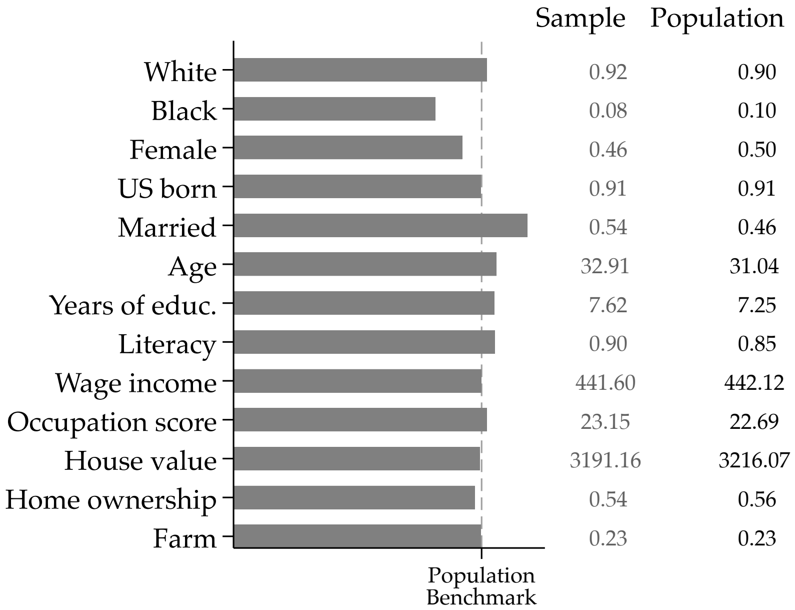
2.5 Validation via Census Cross-Section

Studying intergenerational mobility in childhood outcomes does not require census linking, allowing us to leverage the full census population of older children in parental households. Results based on this sample provide a valuable benchmark for results derived from our panel.

Specifically, we use census cross-sections to analyze parent and child outcomes for families where children reside in their parents' household. We focus on the early life outcomes of literacy and school attendance, limiting our observation to children aged 13–16. Within this age range, the likelihood of a child living apart from their parents is small, minimizing selection into the sample.

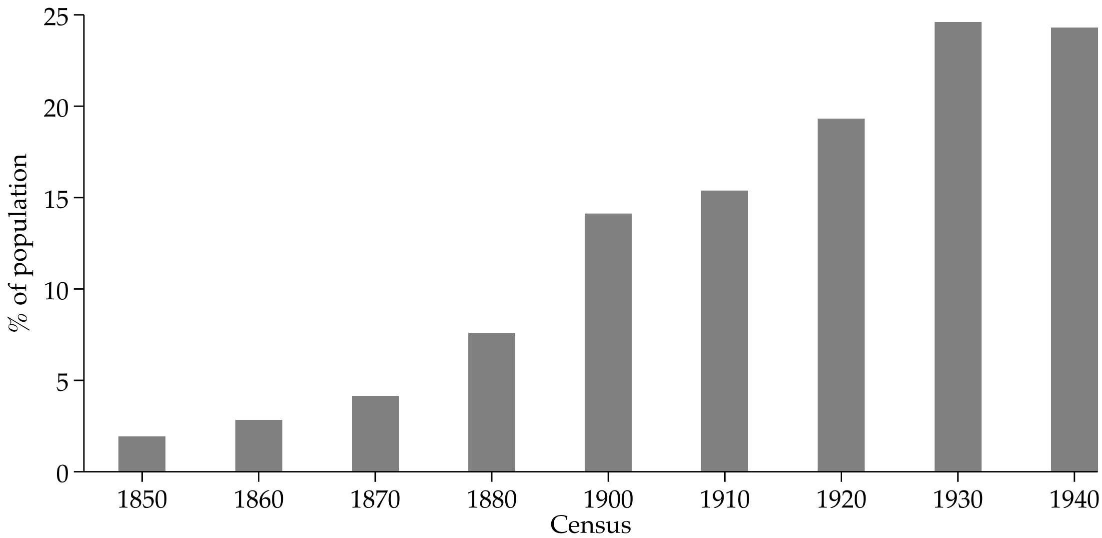
[2] We use a flexible non-parametric method to construct inverse propensity weights (see Appendix C.2).

FIGURE 2: Sample Balance Prior to Weighting (1940)



Notes: This figure shows the representativeness of characteristics among individuals in the 1940 census who we successfully assign an SSN compared to the full population in the 1940 census. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Because of the large sample sizes, even economically small differences are statistically significant.

FIGURE 3: Fraction of US Population Linked in Our New Panel



Notes: This figure shows the fraction of the full population of men and women that we successfully assign a Social Security Number (SSN). This includes parents of SSN applicants who did not apply for an SSN themselves and who we assign synthetic identifiers.

2.6 Economic Outcomes

To understand the role of mothers and fathers in shaping child outcomes, we require separate measures of each parent's outcomes. We therefore focus on human capital outcomes. In contrast to other economic outcomes, such as occupations or income, proxies for human capital reflect the status of both men and women. Furthermore, parental human capital is an important input to children's outcomes.

Historical data often only provide coarse proxies for underlying outcomes of interest. For example, instead of human capital, in the historical census we only observe a person's literacy as a binary proxy thereof. To recover rank-rank relationships in human capital and other outcomes despite this data limitation, we leverage a cutting-edge semiparametric latent variable method (see Section 3.4).

To measure parental background, we additionally consider household-level measures such as income. We incorporate household-level alongside individual-level information only when considering the overall importance of parental background, not when we aim to distinguish mothers' and fathers' separate contributions.

For children, we consider outcomes during both child- and adulthood. During childhood (ages 13–16), we measure literacy (as a proxy for human capital), school attendance, and total years of schooling completed. During adulthood (ages 20–54), we measure literacy, years of education, and occupational income scores. Again, to accurately measure the status of both sons and daughters, we use individual- and household-level information depending on the dimension of status.

3. Measuring Intergenerational Mobility with Multiple Inputs

In this section, we propose a statistical model of intergenerational mobility that accounts for the contributions of both fathers' and mothers' human capital to their children's socioeconomic outcomes. First, we propose using the R^2 of a regression of child outcomes on parental human capital as a mobility measure that integrates the roles of both parents. Second, we use a simple decomposition method that allows to assess the separate contribution of mothers and fathers to the overall R^2. Third, we use a state-of-the-art semiparametric latent variable method to identify intergenerational mobility in ranks when only a binary proxy of the outcome of interest is observed (e.g., literacy as a proxy for human capital).

3.1 A Simple Model of Intergenerational Mobility

We build on standard statistical models of intergenerational mobility where a child's socioeconomic status is a linear function of parental status:

$$\text{rank}\left(y_i^{\text{child}}\right) = \alpha + \beta_1 \text{rank}\left(y_i^{\text{mother}}\right) + \beta_2 \text{rank}\left(y_i^{\text{father}}\right) + \varepsilon_i, \tag{1}$$

where y_i, y_i^{mother}, and y_i^{father} are the outcomes of child i, their mother, and their father, respectively. We focus on ranked outcomes, such that we only consider mobility based on relative positions in the distribution. There are several advantages of this approach in our setting. First, rank-rank correlations are not affected by changes in the marginal distribution of outcomes which, given the long time horizon of our study, enhances the interpretability of the coefficients. Second, using ranked outcomes ensures that the marginal distributions of mother's and father's outcomes are identical, so that their relative contributions can be effectively compared. Third, we focus particularly on "human capital", a concept that is best understood and measured in relative, rather than absolute, terms.

This statistical model differs from most previous research by allowing for multiple parental inputs—most importantly to explicitly incorporate mothers alongside fathers as contributors to a child's outcomes. Note that this model can be extended to accommodate many different inputs including interactions between maternal and paternal effects.

3.2 R^2 as a Measure of Mobility with Multiple Inputs

We propose using the R^2 as an intuitive measure of intergenerational mobility that can account for multiple inputs. We can thereby capture the importance of both mothers and fathers:

$$R^2 = \frac{\sum_{i=1}^{N}\left(\widehat{y}_i^{child} - \bar{y}^{child}\right)^2}{\sum_{i=1}^{N}\left(y_i^{child} - \bar{y}^{child}\right)^2} = \frac{\text{Variance in child outcomes explained by parents}}{\text{Variance in child outcomes}},$$

where $\widehat{y}_i^{child}$ is the predicted outcome of child i from equation (1) and $\bar{y}^{child}$ is the average child outcome.

We argue that predictability as measured by the R^2—measuring the share of variation in child outcomes explained by parental background—is an intuitive measure of intergenerational mobility. In a perfectly mobile society, child outcomes cannot be predicted by parental background ($R^2 = 0$).

The R^2 has a direct relationship with traditional mobility measures—parent-child coefficients or, most commonly, father-son coefficients ($\widehat{\beta}$).[3] In a Appendix B.1, we show:

$$R^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}\left(y^{\text{child}}\right)}{\text{Var}\left(y^{\text{father}}\right)}. \tag{2}$$

Furthermore, rank-rank coefficients—among the most popular measures of mobility—have a one-to-one mapping to the R^2. In this and other cases where the variance of child and parent outcomes are identical, it follows from equation (2) that $R^2 = \beta^2$. For log-log coefficients, the equivalence holds absent changes in income inequality.

The advantage of R^2 is that it can provide an intuitive and easily interpretable measure of mobility even when considering multiple parental inputs. We use this advantage to include both mothers' and fathers' outcomes. Furthermore, it allows to include multiple dimension of parental socioeconomic status. Another advantage is that the R^2 can be decomposed into the contribution of individual inputs, as described in the next section.

[3]This parent-child coefficient is β as estimated via $y_i^{\text{child}} = \alpha + \beta y_i^{\text{parent}} + \varepsilon_i$.

3.3 Measuring Individual Inputs' Contribution to R^2

To assess the contribution of individual parent inputs in shaping child outcomes, we propose decomposing the overall R^2 using a statistical method developed by Shapley (1953); Owen (1977).

Using this decomposition method, we compute the contribution ϕ_j of each input $\boldsymbol{x}_j$ (comprising one or multiple regressors) to the overall variation in child outcomes explained by all inputs:

$$\phi_j = \sum_{T \subseteq V - \{\boldsymbol{x}_j\}} \frac{1}{k!} \left[R^2(T \cup \{\boldsymbol{x}_j\}) - R^2(T) \right],$$

where $R^2(T)$ represents the R^2 of regressing the dependent variable (i.e., y_i^{child}) on a set of variables $T \subseteq V$ (e.g., $V = \{y_i^{\text{mother}}, y_i^{\text{father}}\}$), and k is the number of variables in V (i.e., $k = |V|$). Intuitively, ϕ_j represents the weighted sum of marginal contributions that a parent makes to the variation in child outcomes explained by different combinations of parental inputs. In Appendix B.2, we describe the decomposition method in more detail and provide a closed-form expression for ϕ_j for equation (1) in terms of the estimated coefficients and the rank-rank correlation between mother's and father's outcomes.

The Shapley-Owen decomposition offers several unique advantages, being the only that satisfies three formal conditions defined by Young (1985); Huettner and Sunder (2011) that can be summarized as follows:

1. *Additivity.* Individual contributions to the R^2 add up to the total R^2.

2. *Equal treatment.* Regressors that are equally predictive receive equal values.

3. *Monotonicity.* More predictive regressors receive larger values.

While the Shapley-Owen decomposition method is popular in the machine learning literature (Redell, 2019; Lundberg and Lee, 2017), it has not been widely used in economics (recent exceptions from public economics and trade are Fourrey, 2023; Redding and Weinstein, 2023).

3.4 Measuring Mobility with Latent Inputs

Historical census data typically offer limited direct information about key economic outcomes. However, they include several binary indicators that serve as proxies for these unobserved continuous outcomes. In this section, we propose a method to recover rank-rank relationships in continuous latent outcomes under such constraints.

Most importantly, while human capital is not directly observable, we observe a binary proxy, literacy. Specifically, we assume that literacy, $\text{Lit}(h)$, is a weakly increasing function of human capital. That is, we assume that a person is literate if their human capital is above a threshold level $\bar{h}$:

$$\text{Lit}(h) = \begin{cases} 1 & \text{if } h > \bar{h} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We assume that human capital of parents and outcomes of children are drawn from a joint Gaussian copula distribution, i.e., that there exists a set of unknown monotonically increasing transformations $f_{\text{c}}, f_{\text{m}}, f_{\text{f}}$ such that $\left\{f_{\text{c}}\left(y_i^{\text{child}}\right), f_{\text{m}}\left(h_i^{\text{mother}}\right), f_{\text{f}}\left(h_i^{\text{father}}\right)\right\}^T \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}\right)$ with $\text{diag}(\boldsymbol{\Sigma}) = \mathbb{1}$. The Gaussian copula distribution is commonly used in the statistics literature due to its flexibility and good performance in practice (e.g. Liu et al., 2009, 2012; Zue and Zou, 2012). It is a family of probability distributions that includes the normal distribution, but allows for a much wider range of distributions.[4]

Under the Gaussian copula assumption and equation (3), we can identify the parameters in equation (1), even if only literacy is observed, not human capital directly. To do so, we use a statistical method derived by Fan et al. (2017) to estimate the covariance matrix $\boldsymbol{\Sigma}$ of an underlying Gaussian distribution in the presence of binary variables that are obtained by dichotimizing a latent variable satisfying the Gaussian copula distribution. The method in Fan et al. (2017) allows for a combination of binary and continuous variables. It can be extended to non-binary ordinal variables (Dey and Zipunnikov, 2022).

[4]For instance, since it includes any monotonic transformation of normally distributed random variables, it allows for skewed and multi-modal distributions.

Intuitively, the method leverages the information that parent-child correlations in literacy contain about the correlations in underlying human capital. More specifically, Fan et al. (2017) show that the Kendall's rank correlation coefficient is an invertible function of the elements of $\mathbf{\Sigma}$, the parameters of interest. We refer to Fan et al. (2017) for a more detailed and formal description of the estimator and its properties, notably $\sqrt{n}$-consistency. Because the marginal distribution of ranked variables are uniform between 0 and 100 by definition, $\mathbf{\Sigma}$ is sufficient to identify equation (1). After obtaining an estimate of the covariance matrix $\hat{\mathbf{\Sigma}}$, we obtain estimates of the parameters in equation (1) by simulating from $\mathcal{N}(0, \mathbf{\Sigma})$, transforming the variables into ranks, and estimating the relevant rank-rank regression.

We validate this method and show that, in contrast to ordinary least squares (OLS), it correctly identifies levels and trends in rank-rank mobility based on binary proxies for unobserved underlying outcomes of interest. First, we compare the R^2 from a rank-rank regression of parents' and children's years of education (the benchmark) with the R^2 obtained using our method for arbitrarily coarsened data (e.g., a dummy variable indicating if a person achieved more than 5 years of education). Our method yields results almost identical to the rank-rank benchmark. Second, we conduct a simulation where literacy serves as a binary proxy for human capital. We simulate human capital ranks, convert them into literacy dummies based on historical literacy rates, and compare the R^2 values from regressions using these dummies. Again, our method perfectly captures trends and levels in R^2 across years, whereas OLS yields far lower levels of R^2 and, importantly, trends in R^2 are strongly impacted by changing literacy rates across years.

We apply this method not only to measuring rank-rank mobility in human capital (through literacy), but also to measuring educational rank-rank mobility (through school attendance at a given age). Because we anticipate this method to be useful for future research facing similar data limitations, we have developed a Stata command for easy implementation by others.

4. Income Mobility and the Role of Parental Human Capital

In this section, we first demonstrate the theoretical, historical, and empirical motivations for including parental human capital in assessing intergenerational mobility. Next, we show that accounting for parental human capital not only increases the observed intergenerational persistence but also alters conclusions regarding mobility trends throughout US history. Lastly, we discuss a historical literature underscoring the vital influence of mothers in transmitting human capital to their children—a relationship we quantitatively assess in the subsequent parts of this paper.

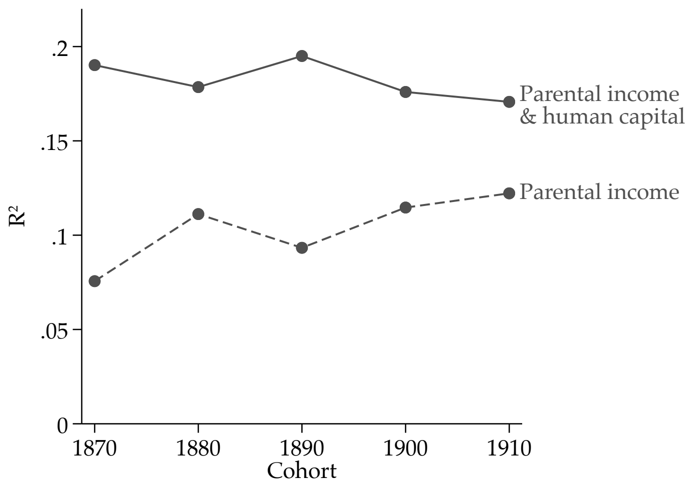
4.1 Income Mobility Accounting for Parental Human Capital

The theory of intergenerational mobility indicates that parental human capital, in addition to income, is a critical determinant of children's outcomes (Becker et al., 2018). Parental human capital may not only increase their capacity for monetary investments in their children's human capital but may also shape their children's human capital directly. However, many empirical studies focus on the relationship between parental and child income, not taking into account other aspects of parental background.

In addition to the theoretical rationale for including parental human capital in historically assessing income mobility, there are significant empirical reasons. In historical US data, direct income measurement is absent; instead, it is approximated through workers' occupations. Researchers have calculated average incomes for occupations using contemporary data, applying these averages to individuals in earlier census records. The lack of more detailed data has forced researchers to largely ignore within-occupation income variations and shifts in the relative status of occupations over time. Factoring in human capital can substantially enhance the assessment of parental background.

We empirically assess the importance of parental human capital in shaping their children's income. We compute the share of variation in a child's income accounted for by both parental income and parental human capital over time. As a literature benchmark, we compare these estimates to the contribution that parental income alone would make to the predictability of child incomes.

FIGURE 4: Share of Variation in Incomes Explained by Parental Background



Notes: This figure shows the share of the variance in a child's household income rank explained by (1) parents' household income ranks and their (latent) human capital ranks (R^2) and (2) parents' household income ranks alone. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score. Results are based on our new panel and sample weights are applied.

We find that parental income and human capital together account for large fractions of variation in children's incomes (see Figure 4). Most importantly, the broader measure of parental background that includes both income and human capital also suggests that intergenerational mobility increased over time—opposite to the conclusion one would derive from measures that ignore parental human capital.

We show that the predictive power of parental background varies across children of different sex and race (see Appendix Figure A.2). Sons generally exhibit lower intergenerational mobility compared to daughters. Specifically, white sons show the least mobility, with 13 to 19 percent of variation in household incomes linked to parental background. Black sons are more mobile than white sons, followed by White daughters, and most notably, Black daughters. Black daughters are not only the most mobile group, they are also the only group whose mobility increases over time. It is crucial to recognize that (1) high within-group mobility does not imply high mobility within the general population and that (2) high mobility does not necessarily equate to high *upward* mobility.

4.2 The Historical Role of Parental Human Capital

Emphasizing the role of parental human capital in shaping child outcomes is particularly important in the historical context. Prior to the establishment of public schools in the late 19th and early 20th centuries, parental home-education was central for children's human capital development. Even children who were enrolled in school in the late 19th century attended school less than four months a year on average (Dreilinger, 2021).

In the 19th century, women bore most of the responsibility to educate children in the home—a time marked by women's specialization in home production and the scarcity of public schools. Initially, in the early agrarian phase of US history, both men and women engaged in home-based industries. However, the first industrial revolution (around 1790–1830) ushered in factory work, a transition more pronounced among men, leading women to increasingly focus on domestic production. Consequently, women became the primary educators of children (Kaestle and Vinovskis, 1978; Margolis, 1984).

Mothers' pivotal role gained recognition from contemporary intellectuals, who advocated for the professionalization of women's role as educators. "The mother forms the character of the future man," Catharine Beecher, a famous American educator, wrote (Beecher, 1842). "The mother may, in the unconscious child before her, behold some future Washington or Franklin, and the lessons of knowledge and virtue, with which she is enlightening the infant mind, may gladden and bless many hearts," the Ladies' Magazine wrote (cited in Kuhn, 1947).

During this period, a substantial body of guidance was developed to equip women for this crucial responsibility. Beecher wrote: "Educate a woman, and the interests of a whole family are secured." Some even viewed home education as superior to formal school education. One hour in the "family school" may "do more towards teaching the young what they ought to know, than is now done by our whole array of processes and instruments of instruction" within schools and colleges, William Alcott, another American educator, wrote (cited in Kuhn, 1947).

5. The Role of Mothers in Shaping Child Outcomes in US History

We measure intergenerational mobility as the share of variance in child outcomes that is attributable to parental background. We decompose this predictive power into contributions from mothers and fathers. Our findings show that a mother's human capital more strongly predicts her child's outcomes than a father's. This difference is particularly pronounced for female and Black children. We corroborate these findings by using cross-sectional data on children living with their parents.

5.1 Parental Human Capital and Child Outcomes

Integrating our baseline model (equation 1) and our method for latent inputs (section 3.4), we evaluate the impact of parental human capital on children's human capital and income. Our sample comprises children whose outcomes are recorded in the census at age 21 or older.

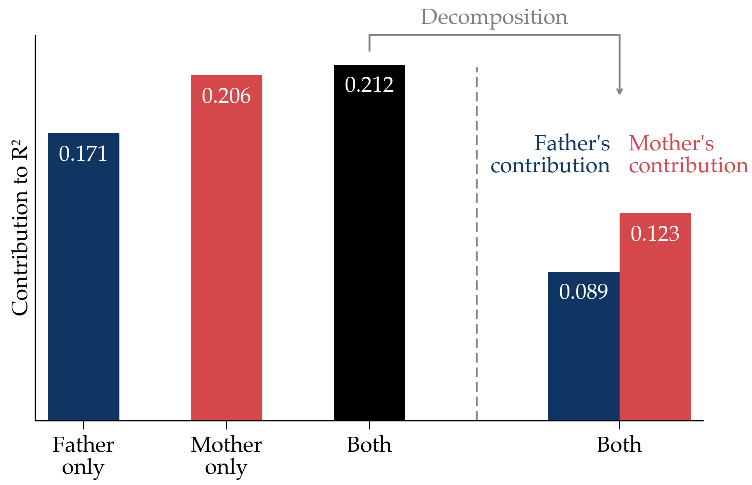
Figure 5 demonstrates the Shapley-Owen decomposition method. For children born in the 1880s, mothers' human capital alone accounts for 20.6% of the variation in child human capital. Fathers and mothers together predict 21.2% of the variation. Notably, using the Shapley-Owen decomposition, we find that mothers account for more than half (58%) of the joint predictive power.

Next, we examine how the transmission varies by child gender and race. We analyze two outcomes: children's human capital and their formal schooling. Panel A of Figure 6 shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks. We recover human capital rank-rank transmission using information on literacy and the semiparametric latent variable method introduced in section 3.4. Panel B shows this relationship for ranks in years of school.

The distinction between human capital and formal schooling is particularly crucial historically, when parental education was the primary source of learning.

FIGURE 5: Illustrating our Decomposition Method

Intergenerational Transmission of Human Capital

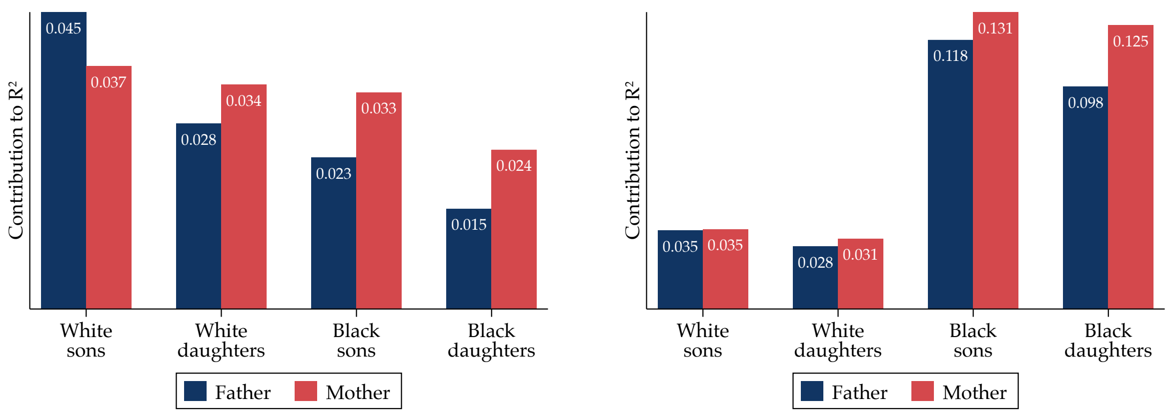


Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. We decompose the overall R^2 using the Shapley-Owen method to quantify each parent's contribution. Results are based on our new panel, specifically children born in the 1880s; sample weights are applied.

FIGURE 6: Parental Human Capital & Child Outcomes (1920s cohort)

(A) Rank in (Latent) Human Capital

(B) Rank in Years of School



Notes: Panel A shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. Panel B shows the results based on ranks in years of school. We decompose the overall R^2 using the Shapley-Owen method to quantify each parent's contribution. Results are based on our new panel and sample weights are applied.

Our first key finding is that mothers' contributions are generally larger than fathers'. This finding is particularly pronounced among female and Black children. In earlier cohorts, mothers' contributions are even larger, exceeding fathers' across all groups including white sons (see Appendix Figure A.3). Generally, maternal human capital was most predictive in the earliest cohorts (mid-1800s). The pronounced impact of mothers on daughters and Black children aligns with their historical lack of access to educational resources (Kober and Rentner, 2020). For daughters, it could also suggest the presence of gender-specific role model effects (e.g., Bettinger and Long, 2005).

Second, we find that white children experienced higher mobility in formal schooling than Black children. Among white children, only 6 to 7% of variation in school education can be explained by parental education. In contrast, among Black children, parents account for 22 to 25% of variation in school education. This finding may reflect the fact that school access among white children had massively expanded in the early 1900s (see Appendix Figure A.6). In contrast, most Black children lived in the Jim Crow South with restricted school access, shorter school years, and poor school quality.

The third key finding is that mobility in human capital differs far less across race than mobility in formal schooling. Among Black children, 4 to 6 percent of variation in human capital are explained by parental human capital—for white children it is 6 to 8 percent. This finding is consistent with the notion that formal school education can partly be substituted via parental education.

5.2 Validating our Results in the Census Cross-Section

To validate our panel-based findings, we analyze the census cross-section of children aged 13–16 living with their parents. We use literacy, school attendance, and completed education as proxies for children's human capital; for parents, we consider literacy and years of education.

Our cross-sectional analysis closely mirrors our panel data results: Mothers' predictive power is higher than fathers', especially for female and Black children (see Appendix Figure A.4). Across different measures of human capital and schooling, similar trends emerge. Compared to our panel-based findings, measuring human capital during childhood in the cross-section suggests stronger intergenerational persistence. This finding reflects intra-generational mobility in human capital. For example, in this period, many individuals became literate only during adulthood.

6. Mothers' Key Role in Child Education before Widespread School Access

So far, we have established that mothers' human capital is more predictive of their child's human capital than fathers'. This section examines examines mothers' role in home education before the advent of widespread schooling as an explanation for mothers' disproportionate influence in human capital transmission, as suggested by historical literature. We analyze human capital transmission and mothers' relative impact on children born between the 1850s and 1920s, correlating these findings with local school access.

6.1 Public Schools and the Rise of Educational Mobility

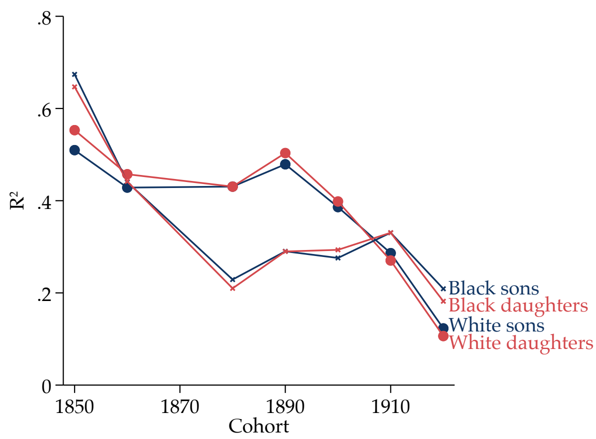
Our comparison of intergenerational human capital transmission over time reveals increasing educational mobility in American children. In the 1850s, parental background accounted for 50 to 70% of variation in child human capital, while for those born in the 1920s, this figure dropped to 10 to 20% (see Panel A of Figure 7).

The trend, however, varies significantly by race. After slavery ended, Black children saw a rapid increase in mobility, which then plateaued. White children's mobility remained stable until the early 1900s, followed by a surge, marking the first time since the Civil War that white Americans surpassed Black children in educational mobility.

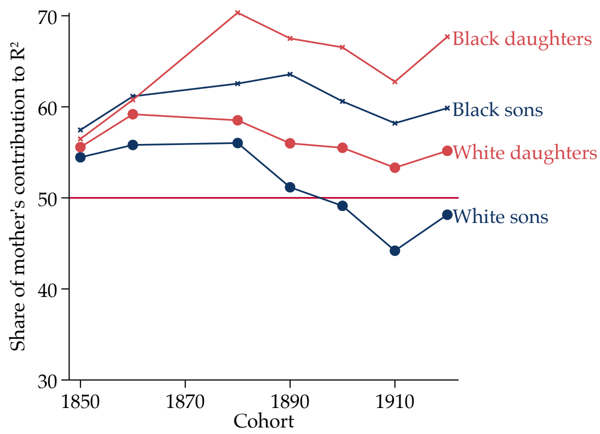
Mothers, more than fathers, have consistently influenced child human capital (see Panel B of Figure 7). This impact is most pronounced for daughters and Black children, where maternal influence is notably stronger. Over time, mothers' relative influence on white children, especially sons, has diminished, whereas for Black daughters, it has grown.

FIGURE 7: Transmission of (Latent) Human Capital Ranks Across Cohorts

(A) Transmission by Group



(B) Mothers' Relative Contribution



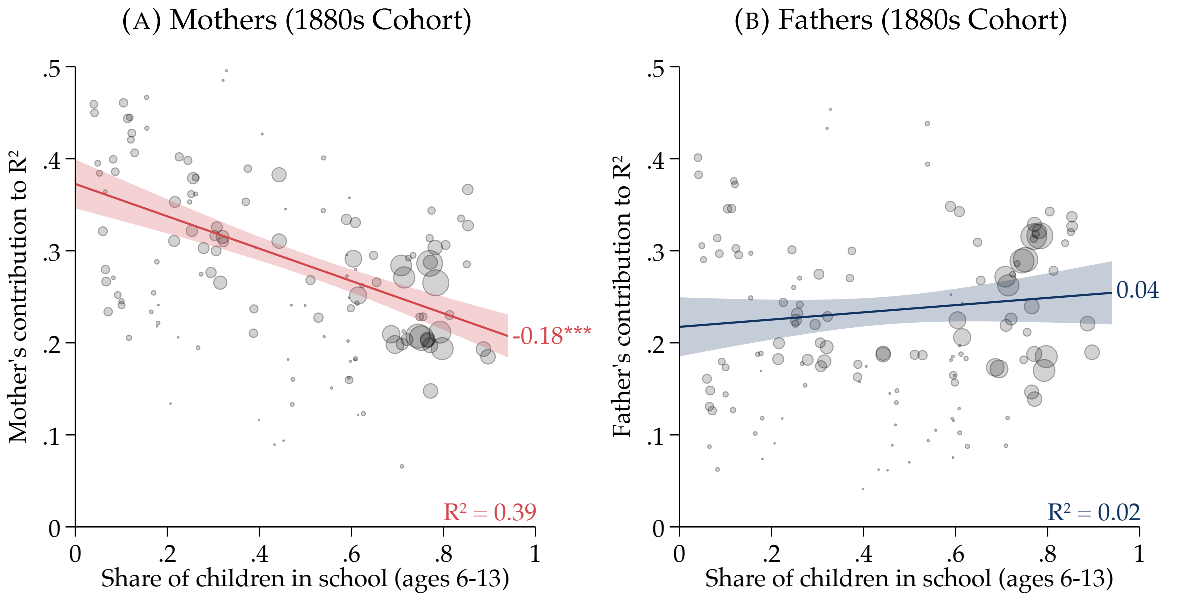
Notes: Panel A shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2) across cohorts and groups. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. Panel B shows mothers' relative contribution to the overall R^2 using the Shapley-Owen method. Results are based on the census cross-section of children ages 13–16 in their parents' household.

Historians highlight mothers' important role in educating their children in the 19th century (Kaestle and Vinovskis, 1978; Margolis, 1984; Dreilinger, 2021). While the spread of school access around 1900 was rapid, it was highly unequal. Specifically, Black children and girls were slower to gain access than boys. "When public schools did open up to girls, they were sometimes taught a different curriculum from boys and had fewer opportunities for secondary or higher education" (Kober and Rentner, 2020). Similarly, schools for Black children had drastically lower quality than schools for white children (Card and Krueger, 1992; Althoff and Reichardt, 2023).

In line with the hypothesis that mothers' importance reflects their role in home schooling, our analysis indicates a strong correlation between the disparity in mothers' and fathers' contributions to child human capital and the child's school access (see Figure 8). Mothers are more predictive of child outcomes in areas with limited school access. In 1880, this correlation explained 39 percent of the variation in mothers' contribution to variation in child human capital. Conversely, fathers' contribution showed no correlation with school access.

Expressing mothers' relative to fathers' contribution, school access is even more nega-

FIGURE 8: Mothers' Human Capital as Substitute for Local Schools



Notes: This figure shows the relationship between local school access and parental contributions to child human capital. We compute the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2) across cohorts and groups. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. Panels A and B show mothers' and fathers' contributions to the overall R^2 using the Shapley-Owen method. Each dot represents a child group from the 1880s, categorized by race, sex, and state. Sample size weights are applied. School access is determined by the race and sex-specific share of children aged 6–13 in school.

tively correlated, explaining 51 percent of the variation across groups (Panel A, Appendix Figure A.5). As school access spread over time, this correlation remained strong but accounting for only 6% of the variation of mothers' importance for children born in the 1920s (Panel B, Figure A.5).

Our evidence on the reduced impact of parental human capital as public education access improves is also consistent with Biasi (2023), who shows that equalizing school resources can reduce geographic disparities in intergenerational mobility.

7. CONCLUSION

This paper revisits intergenerational mobility in the US from 1850 to 1940, emphasizing the role of human capital, especially the mother's. We develop a novel, comprehensive panel that includes women in early U.S. history, introduce the R^2 metric for assessing mobility with multiple parental inputs, leverage advanced statistical techniques to analyze intergenerational transmission under data constraints, and dissect the impact of maternal and paternal backgrounds. Our findings highlight the significant influence of maternal human capital on children's outcomes, particularly for daughters and Black children. We propose that limited school access might explain why this impact varies based on race, location, and across time.

There are several promising avenues for future research. We expanded the parental status measurement to separately encompass maternal and paternal roles. For example, to measure intergenerational mobility, we considered both parents' human capital and parents' household income. Future research could integrate broader parental background measures like wealth or social norms. Given the importance of the location in which a person grows up—as documented in previous research (e.g., Chetty and Hendren, 2018; Althoff and Reichardt, 2023)—future research could also use the R^2 mobility metric to factor in neighborhood quality alongside parental background. Another promising avenue for future work would be to assess changes in maternal transmission of economic outcomes over the 20th century, especially amid rising female labor participation (Goldin, 1977, 1990, 2006) and single-motherhood (Althoff, 2023).

Our new panel dataset serves as a foundation for future work on the role of women in shaping US history. Future researchers may find this dataset helpful to reevaluate questions that require panel data but have been studied exclusively for men, as well as to consider new questions that focus specifically on the role of women.

REFERENCES

ABRAMITZKY, R., L. BOUSTAN, E. JACOME, AND S. PEREZ (2021a): "Intergenerational Mobility of Immigrants in the United States over Two Centuries," *American Economic Review*, 111, 580–608.

ABRAMITZKY, R., L. P. BOUSTAN, K. ERIKSSON, J. J. FEIGENBAUM, AND S. PÉREZ (2021b): "Automated Linking of Historical Data," *Journal of Economic Literature*, 59, 865–918.

ALTHOFF, L. (2023): "Two Steps Forward, One Step Back: Racial Income Gaps among Women since 1950," Working Paper.

ALTHOFF, L. AND H. REICHARDT (2023): "Jim Crow and Black Economic Progress After Slavery," Working Paper.

BAILEY, M. J., P. Z. LIN, S. MOHAMMED, P. MOHNEN, J. MURRAY, M. ZHANG, AND A. PRETTYMAN (2022): "LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database," dataset: `https://doi.org/10.3886/E155186V2`.

BECKER, G. S., S. D. KOMINERS, K. M. MURPHY, AND J. L. SPENKUCH (2018): "A Theory of Intergenerational Mobility," *Journal of Political Economy*, 126.

BEECHER, C. E. (1842): *Treatise on Domestic Economy*, Boston: T. H. Webb, & Co.

BETTINGER, E. AND B. LONG (2005): "Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students," *American Economic Review*, 95, 152–157.

BIASI, B. (2023): "School Finance Equalization Increases Intergenerational Mobility," *Journal of Labor Economics*, 41, 1–38.

BLACK, S. E., P. J. DEVEREUX, AND K. G. SALVANES (2005): "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital," *American Economic Review*, 95, 437–449.

BUCKLES, K., J. PRICE, AND H. WILBERT (2023): "Family Trees and Falling Apples: Intergenerational Mobility Estimates from US Genealogy Data," Working Paper.

CARD, D., C. DOMNISORU, AND L. TAYLOR (2022): "The Intergenerational Transmission of Human Capital: Evidence from the Golden Age of Upward Mobility," *Journal of Labor Economics*, 40, S1–S493.

CARD, D. AND A. B. KRUEGER (1992): "School Quality and Black-White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics*, 107, 151–200.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2023): "Live Births, Birth Rates, and Fertility Rates, by Race of Child: United States, 1909-80," dataset: `https://www.cdc.gov/nchs/data/statab/t1x0197.pdf`.

CHETTY, R. AND N. HENDREN (2018): "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 133, 1107–1162.

CHETTY, R., N. HENDREN, P. KLINE, E. SAEZ, AND N. TURNER (2014): "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility," *American Economic Review Papers and Proceedings*, 104, 141–147.

CRAIG, J., K. A. ERIKSSON, AND G. T. NIEMESH (2019): "Marriage and the Intergenerational Mobility of Women: Evidence from Marriage Certificates 1850-1910," Working Paper.

DEY, D. AND V. ZIPUNNIKOV (2022): "Semiparametric Gaussian Copula Regression modeling for Mixed Data Types (SGCRM)," Working Paper.

DREILINGER, D. (2021): *The Secret History of Home Economics: How Trailblazing Women Harnessed the Power of Home and Changed the Way We Live*, W.W. Norton & Company.

ESHAGHNIA, S., J. J. HECKMAN, R. LANDERSØ, AND R. QURESHI (2022): "Intergenerational Transmission of Family Influence," Working Paper 30412, National Bureau of Economic Research.

ESPÍN-SÁNCHEZ, J.-A., J. P. FERRIE, AND C. VICKERS (2023): "Women and the Econometrics of Family Trees," Working Paper 31598, National Bureau of Economic Research, Cambridge, MA.

FAN, J., H. LIU, Y. NING, AND H. ZOU (2017): "High dimensional semiparametric latent graphical model for mixed data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79, 405–421.

FOURREY, K. (2023): "A Regression-Based Shapley Decomposition for Inequality Measures," *Annals of Economics and Statistics*, 39–62.

GARCÍA, J. L. AND J. J. HECKMAN (2023): "Parenting Promotes Social Mobility Within and Across Generations," *Annual Review of Economics*, 15, 349–388.

GOLDIN, C. (1977): "Female labor force participation: The origin of black and white differences, 1870 and 1880," *Journal of Economic History*, 87–108.

——— (1990): *Understanding the gender gap: An economic history of American women*, Oxford University Press.

——— (2006): "The quiet revolution that transformed women's employment, education, and family," *American economic review*, 96, 1–21.

HUETTNER, F. AND M. SUNDER (2011): "Decomposing R^2 with the Owen value," Working paper.

JÁCOME, E., I. KUZIEMKO, AND S. NAIDU (2021): "Mobility for All: Representative Intergenerational Mobility Estimates over the 20th Century," Working Paper 29289, National Bureau of Economic Research.

KAESTLE, C. F. AND M. A. VINOVSKIS (1978): "From Apron Strings to ABCs: Parents, Children, and Schooling in Nineteenth-Century Massachusetts," *American Journal of Sociology*, 84, S39–S80, supplement: Turning Points: Historical and Sociological Essays on the Family.

KOBER, N. AND D. S. RENTNER (2020): "History and Evolution of Public Education in the US," Online report: `https://files.eric.ed.gov/fulltext/ED606970.pdf`.

KUHN, A. L. (1947): *The Mother's Role in Childhood Education: New England Concepts 1830-1860*, Yale University Press.

LEIBOWITZ, A. (1974): "Home Investments in Children," in *Economics of the Family: Marriage, Children, and Human Capital*, ed. by T. W. Schultz, University of Chicago Press, 432–456.

LIU, H., F. HAN, M. YUAN, J. LAFFERTY, AND L. WASSERMAN (2012): "High-dimensional semiparametric Gaussian copula graphical models," *Annals of Statistics*, 40, 2293–2326.

LIU, H., J. LAFFERTY, AND L. WASSERMAN (2009): "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research*, 10.

LUNDBERG, S. M. AND S.-I. LEE (2017): "A Unified Approach to Interpreting Model Predictions," Working Paper.

MARGOLIS, M. L. (1984): *Mothers and Such: Views of American Women and Why They Changed*, Berkeley and Los Angeles: University of California Press.

OWEN, G. (1977): "Values of games with a priori unions," in *Essays in Mathematical Economics and Game Theory*, ed. by R. Heim and O. Moeschlin, New York: Springer.

PUCKETT, C. (2009): "The Story of the Social Security Number," *Social Security Bulletin*, 69.

REDDING, S. J. AND D. E. WEINSTEIN (2023): "Accounting for Trade Patterns," Working paper.

REDELL, N. (2019): "Shapley Decomposition of R-Squared in Machine Learning Models," Working Paper.

RUGGLES, S., S. FLOOD, R. GOEKEN, J. GROVER, E. MEYER, J. PACAS, AND M. SOBEK (2020): "IPUMS USA: Version 10.0," dataset: `https://doi.org/10.18128/D010.V10.0`.

SHAPLEY, L. (1953): "A value for n-person games," in *Contributions to the Theory of Games*, ed. by H. Kuhn and A. Tucker, Princeton University Press, vol. 2.

SOCIAL SECURITY ADMINISTRATION (2023): "Number of Social Security card holders born in the U. S. by year of birth and sex," dataset: `https://www.ssa.gov/oact/babynames/numberUSbirths.html`.

WARD, Z. (2023): "Intergenerational Mobility in American History: Accounting for Race and Measurement Error," *American Economic Review*, 113, 3213–3248.

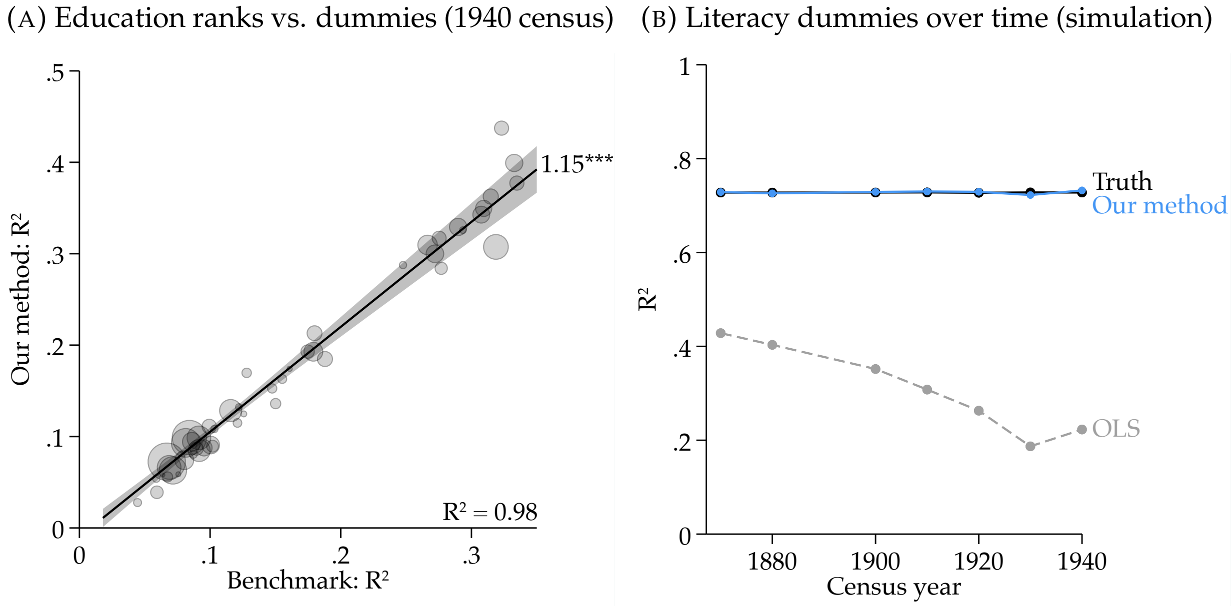
YOUNG, H. P. (1985): "Monotonic solutions of cooperative games," *International Journal of Game Theory*, 14, 65–72.

ZUE, L. AND H. ZOU (2012): "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571.

APPENDIX

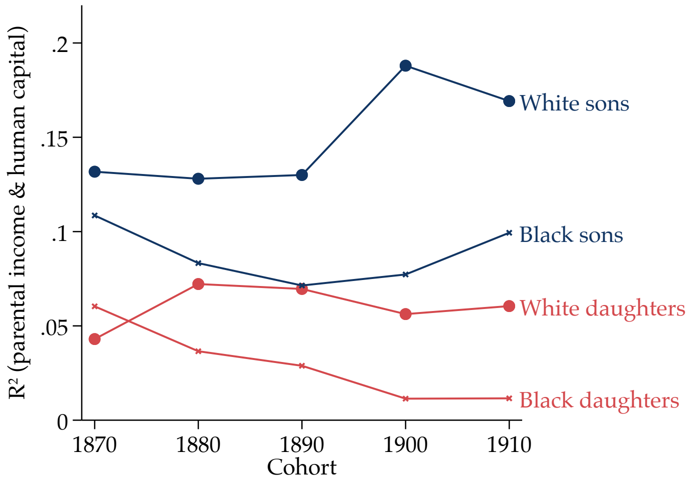
A. Appendix Figures

Figure A.1: Validation of the Semiparametric Latent Variable Method



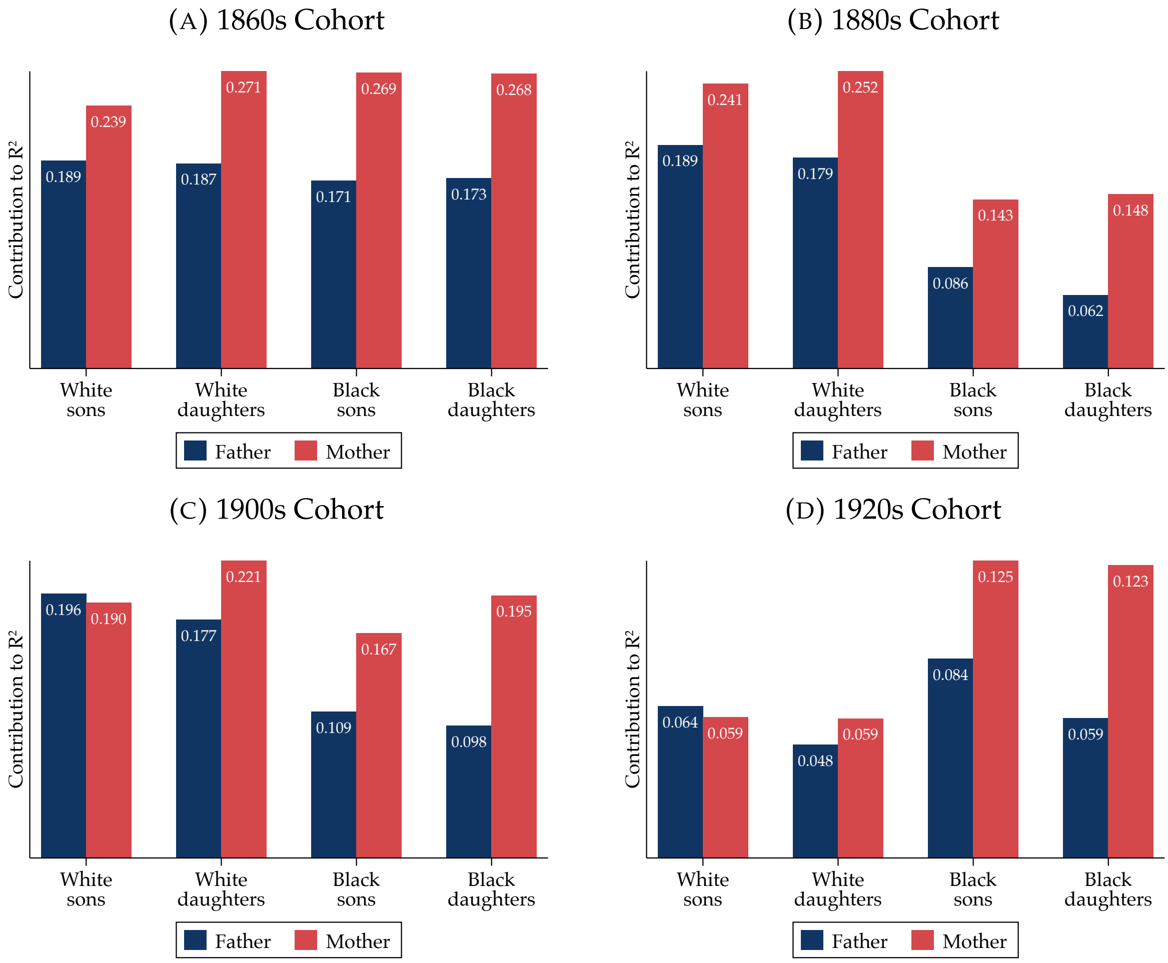
Notes: This figure demonstrates the effectiveness of our semiparametric latent variable method in identifying rank-rank relationships from binary proxies. Panel A contrasts the R^2 values from rank-rank regressions using actual and coarsened educational data from the 1940 census. We coarsen the data by arbitrarily categorizing individuals based on their educational attainment: more than 11 years for children, 9 for mothers, and 7 for fathers. Each dot represents a US state, weighted by sample size and focusing on children aged 13—21 living with parents. Panel B illustrates a simulation where literacy serves as a binary proxy for human capital. We simulate human capital ranks, convert them into literacy dummies based on historical literacy rates, and compare the R^2 values from regressions using these dummies. The "Truth" line represents the R^2 from human capital rank regression, "Our method" from our latent variable method using literacy dummies, and "OLS" from a standard OLS regression with the same dummies.

FIGURE A.2: Group-Specific Parental Background-to-Income Mobility



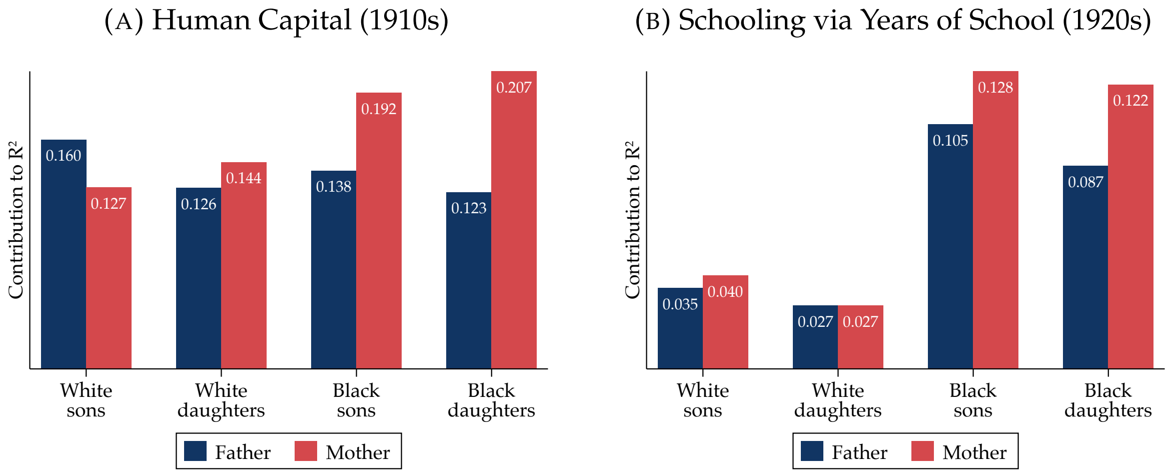
Notes: This Figure shows the share of the variance in a child's household income rank explained by parents' household income ranks and their (latent) human capital ranks (R^2) across cohorts and groups. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score. Results are based on our new panel and sample weights are applied.

FIGURE A.3: Parental & Child Human Capital



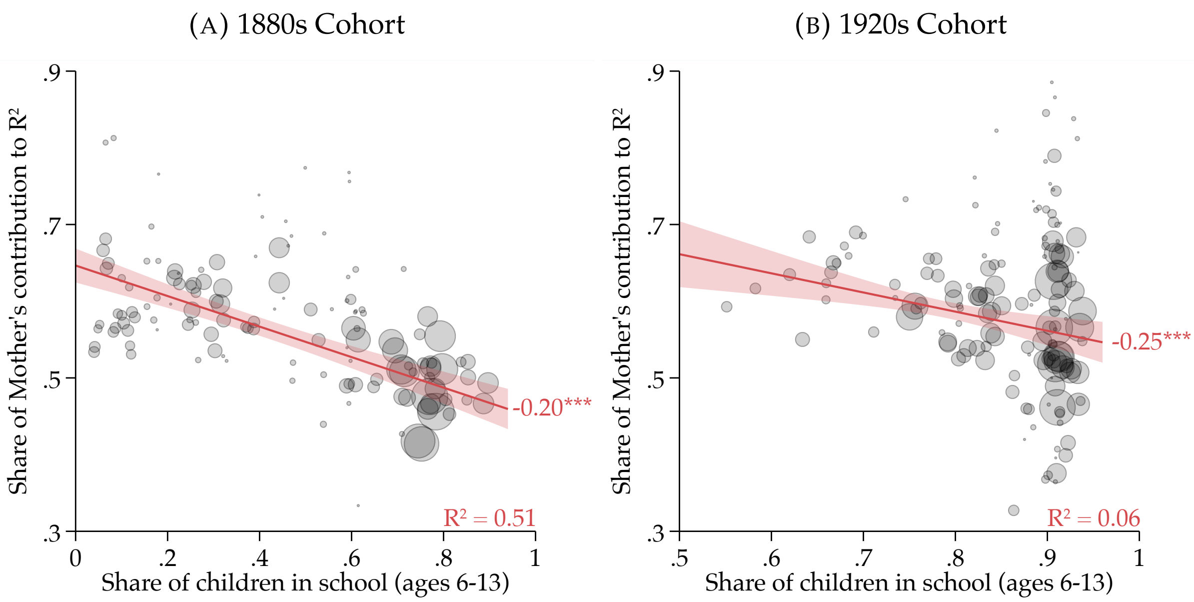
Notes: This figure shows the share of variation in child human capital explained by paternal and maternal human capital across cohorts. We use literacy to measure the rank-based transmission of human capital based on the method we introduce in section 3.4. Results are based on the census cross-section of children ages 13–16 in their parents' household.

FIGURE A.4: Validation of Results via Census Cross-Section



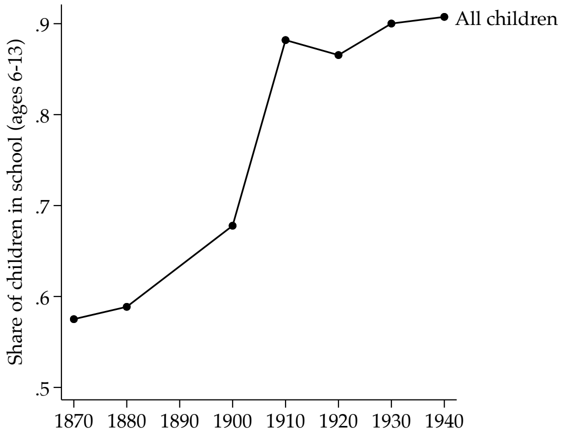
Notes: This figure shows the share of variation in child human capital explained by paternal and maternal human capital for children at ages 13-16 in the 1930 census (1910s cohort); and variation in child schooling explained by paternal and maternal schooling for children at ages 13-16 in the 1940 census (1920s cohort). We use literacy and years of education (the former only observed up until the 1930 census, the latter only observed in the 1940 census) to measure the rank-based transmission of human capital and schooling based on the method we introduce in section 3.4.

FIGURE A.5: Mothers' Human Capital as Substitute for Local Schools



Notes: This figure shows the relationship between local school access and mothers' *relative* contributions to child human capital (as a share of total variation explained). Literacy is used as the measure for rank-based transmission of human capital (section 3.4). Each dot represents a child group from the 1880s, categorized by race, sex, and state. Sample size weights are applied. School access is determined by the race and sex-specific share of children aged 6–13 in school. Results are based on the census cross-section of children ages 13–16 in their parents' household.

FIGURE A.6: Increasing Access to Schools



Notes: This figure shows the share of children aged 6–13 who attend school across time.

B. METHODS APPENDIX

B.1 Equivalence Between R^2 and Coefficients

B.1.1 One input

In a linear regression with a single explanatory variable, $y_i = \alpha + \beta x_i + \varepsilon_i$, the coefficient β and the R^2 are defined as follows:

$$\beta = \text{cor}(x, y) \cdot \sqrt{\frac{\text{Var}(y)}{\text{Var}(x)}} \tag{4}$$

$$R^2 = \text{cor}(x, y)^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}(x)}{\text{Var}(y)}, \tag{5}$$

where $\text{cor}(x, y)$ is the correlation between y and x and $\text{Var}(y)$ is the variance of y.

Rank-rank coefficients. Rank-rank coefficients are a popular measure of mobility. By construction, quantile-ranked outcomes share the same distribution. Therefore, if both y and x are outcomes in quantile-ranks, we have $\text{Var}(y) = \text{Var}(x)$ so that $R^2 = \widehat{\beta}^2$.

Intergenerational elasticity coefficients. Intergenerational elasticities are another common measure of mobility. Such elasticities are estimated in a regression of $\log(y)$ and $\log(x)$ where y and x are a child and a parent's outcome, respectively. Such an elasticity is equal to $\sqrt{R^2}$ if and only if $\text{Var}(\log(y)) = \text{Var}(\log(x))$. A sufficient condition for these variances to equate is that the marginal distribution of children's outcomes are a shifted version of that of the parents, i.e. $y \sim bx$ for some $b > 0$.

B.1.2 Two inputs

In a linear regression with two explanatory variables, $y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, the R^2 will in part depend on the correlation between $x_{i,1}$ and $x_{i,2}$—i.e., the level of parental assortative mating:[5]

$$R^2 = \widehat{\beta}_1^2 \frac{Var(x_1)}{Var(y)} + \widehat{\beta}_2^2 \frac{Var(x_2)}{Var(y)} + 2\widehat{\beta}_1\widehat{\beta}_2 \frac{Cov(x_1, x_2)}{Var(y)}. \tag{6}$$

Rank-rank coefficients. Again, using that by construction, quantile-ranked outcomes share the same distribution, we have $\text{Var}(y) = \text{Var}(x_1) = \text{Var}(x_2)$ so that $R^2 = \widehat{\beta}_1^2 + \widehat{\beta}_2^2 + 2\widehat{\beta}_1\widehat{\beta}_2\widehat{\rho}_{1,2}$, where $\widehat{\rho}_{1,2}$ is the correlation between x_1 and x_2.

B.2 Shapley-Owen Decomposition of the R^2

The Shapley-Owen decomposition of R^2 (Shapley, 1953; Owen, 1977) provides a way to quantify the contribution of each independent variable to a model. The method was introduced in cooperative game theory as a method for fairly distributing gains to players. It has been used more recently as a way to interpret black-box model predictions in machine learning (Redell, 2019; Lundberg and Lee, 2017), as well as in some economics research on inequality (Azevedo et al., 2012; Fourrey, 2023).

For a given set of k vectors of regressors $V = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k\}$, we create sub-models for each possible permutation of vectors of regressors.

The marginal contribution of each vector of regressor $\boldsymbol{x}_j \in V$ is:

[5] We use that $R^2 \equiv \frac{Var(y) - Var(\varepsilon)}{Var(y)}$ and

$$\begin{aligned} Var(y) &= Var(\beta_1 x_1 + \beta_2 x_2 + \varepsilon) \\ \frac{Var(y) - Var(\varepsilon)}{Var(y)} &= \beta_1^2 \frac{Var(x_1)}{Var(y)} + \beta_2^2 \frac{Var(x_2)}{Var(y)} + 2\beta_1\beta_2 \frac{Cov(x_1, x_2)}{Var(y)} \end{aligned}$$

$$\Delta_j = \sum_{T \subseteq V - \{\boldsymbol{x}_j\}} \left[R^2(T \cup \{\boldsymbol{x}_j\}) - R^2(T) \right]$$

where $R^2(T)$ represents the R^2 of regressing the dependent variable on a set of variables $T \subseteq V$ (e.g., $V = \{y_i^{\text{mother}}, y_i^{\text{father}}\}$). The marginal contribution gives us the sum of the contributions that the vector of regressors $\boldsymbol{x}_j$ makes to the R^2 of each sub-model. Then, the Shapley-value ϕ_j for the vector of regressors $\boldsymbol{x}_j$ is obtained by normalizing each marginal contribution so that they sum to the total R-squared:

$$\phi_j = \frac{\Delta_j}{k!}, \tag{7}$$

where k is the number of vectors of regressors in V (i.e., $k = |V|$). Each ϕ_j then corresponds to the goodness-of-fit of a given vector of regressor, and they sum up to equal the model's total R^2. Using this method, perfect statistical substitutes will receive the same Shapley value.

B.2.1 Example with two inputs

Table B.1 shows an example for the Shapley-Owen decomposition of the R^2 for the case of two parental inputs, omitting their interaction. We add variables at every column, leading up to the full two-parent model containing the outcomes of both fathers and mothers. Note that the individual parental contributions (i.e., Shapley values) sum up to the total R^2 of 0.25 in the two-parent model. In this case, mothers account for 64% of the variation in child outcomes explained by parental background.

TABLE B.1: Example of Shapley-Owen Decomposition

Empty Model		One-Parent Model		Two-Parent Model		Marginal Contribution (Δ_j)	
Regressors	R^2	Regressors	R^2	Regressors	R^2	Father	Mother
$\emptyset$	0.0	Father	0.08	Father, Mother	0.25	$0.08 - 0 = 0.08$	$0.25 - 0.08 = 0.17$
$\emptyset$	0.0	Mother	0.15	Father, Mother	0.25	$0.25 - 0.15 = 0.10$	$0.15 – 0 = 0.15$
				Shapley Value (ϕ_j)		$\frac{0.08+0.1}{2!} = 0.09$	$\frac{0.17+0.15}{2!} = 0.16$

B.2.2 Unpacking the Shapley-value with two inputs

To better understand what the Shapley-value for each parental input comprises, we express it as a function of regression coefficients, variances, and covariances in the two-input case. Let ϕ_1 be one parent's Shapley value—i.e., the contribution that the parent's input makes to the overall R^2 when regressing child outcomes on both parents' inputs. Applying equation (7), we have

$$\phi_1 = \frac{1}{2}\left(R^2(\{x_1, x_2\}) - R^2(\{x_2\}) + R^2(\{x_1\}) - R^2(\{\emptyset\})\right).$$

Further, using equation (6), we have

$$\phi_1 = \frac{1}{2}\left(\left[\widehat{\beta}_1^2 + \widehat{\beta}_{1,univ}^2\right]\frac{Var(x_1)}{Var(y)} + \left[\widehat{\beta}_2^2 + \widehat{\beta}_{2,univ}^2\right]\frac{Var(x_2)}{Var(y)} + 2\widehat{\beta}_1\widehat{\beta}_2\frac{Cov(x_1, x_2)}{Var(y)}\right),$$

where $\widehat{\beta}_{1,univ}^2$ is the coefficient on the mother's input in a univariate regression and $\widehat{\beta}_1^2$ the coefficient on the mother's input in the multivariate regression including the father's input. Using the omitted variable bias formula, $\widehat{\beta}_{1,univ}^2 = \widehat{\beta}_1 + \widehat{\beta}_2\frac{Cov(x_1,x_2)}{Var(x_1)}$, we have

$$\phi_1 = \frac{1}{2Var(y)}\left(2\widehat{\beta}_1^2 Var(x_1) + \{Cov(x_1, x_2)\}^2\left[\frac{\widehat{\beta}_2^2}{Var(x_1)} - \frac{\widehat{\beta}_1^2}{Var(x_2)}\right] + 2\widehat{\beta}_1\widehat{\beta}_2 Cov(x_1, x_2)\right).$$

For rank-rank regressions, we have

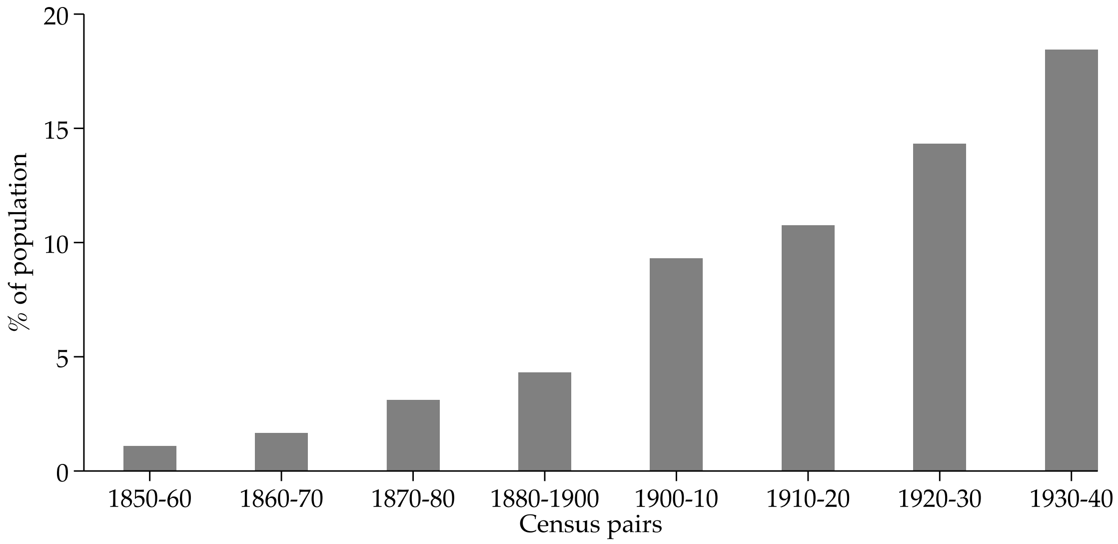
$$\begin{aligned}\phi_1 &= \widehat{\beta}_1^2 + \frac{1}{2}\left(\widehat{\beta}_2^2 - \widehat{\beta}_1^2\right)\left(\frac{Cov(x_1, x_2)}{Var(y)}\right)^2 + \widehat{\beta}_1\widehat{\beta}_2\frac{Cov(x_1, x_2)}{Var(y)} \\ &= \widehat{\beta}_1^2 + \frac{\widehat{\rho}_{1,2}^2}{2}\left(\widehat{\beta}_2^2 - \widehat{\beta}_1^2\right) + \widehat{\beta}_1\widehat{\beta}_2\widehat{\rho}_{1,2}.\end{aligned}$$

C. DATA APPENDIX

FIGURE C.7: Sample Balance Prior to Weighting (1850–1920)

*Notes:*This figure shows the representativeness of characteristics among individuals who we successfully assign an SSN compared to the full population in each census before 1940. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Because of the large sample sizes, even economically small differences are statistically significant.

FIGURE C.8: Fraction of US Population Linked in Our New Panel



Notes: This figure shows the fraction of the full population of men and women that we successfully link from one census decade to the next. Our empirical analysis also leverages links across non-adjacent census pairs, further increasing coverage.

C.1 Linking Procedure

We develop a multi-stage linking process built on the procedural record linkage method developed by Abramitzky et al. (2021b). Our process consists of three stages. 1) linking SSN applications to census records. 2) Identifying the applicant's parents in the census. 3) Tracking these parents' census records over time. With our linking method, we are able to maximize the number of SSN-census links and subsequently build a multigenerational family tree for each linked SSN applicant.

First stage: Applicant SSN ↔ census.

- *Preparing SSN data*: We use a digitized version of the Social Security Number application data from the National Archives and Records Administration (NARA) known as the Numerical Identification Files (NUMIDENT). We harmonize the application, death and claims files to capture all the available information of each SSN record. These data include each applicant's name, age, race, place of birth, and the maiden names of their parents. We recode certain variables to align with census data, for example, we ensure codes for countries of birth, race and sex are consistent across the SSN and Census. Additionally, we apply the ABE name cleaning method to names of applicants and their parents resulting in an "exact" and a NYSIIS cleaned version of all names (Abramitzky et al., 2021a)[6].

- *Preparing Census data*: Within each census decade from 1850 and 1940, we apply the same name cleaning algorithm used to clean the SSN data. Where available, we extract parent and spouse names from each individual's census record to create crosswalks that are later used in the linking process. Each cleaned census decade is subsequently divided into individual birthplace files for easing the computational intensity of the linking procedure.

- *Linking SSN to Census records*: Our goal is to achieve a high linkage rate of SSN applications to the census, while ensuring the accuracy of each link. Our linking algorithm has the following steps:

  1. We first create a pool of potential matches by finding all possible links between an SSN application and census record using first and last name (NYSIIS), place of birth, marital status and birth year within a 5-year age band. In the census, we identify marital status from the census variable "marst" or whether her position in the household is described as spouse. In the SSN data, we identify marital status if the applicants last name is different from that of her father.

  2. Once we have established our pool of potential matches, we essentially rerun our linking process. However, we use additional matching variables in order to pin down the most likely correct link among the potential matches. In our first round of this process, we aim to pin down the correct link by matching using the following set of matching characteristics: exact first, middle and last names of both the applicant and their parents, exact birth month (when available), state or country of birth, race, and sex. An SSN application is either uniquely matched to a census record or not.

  3. We attempt a second round of the matching described in point 2. for all SSN applicants who were *not* uniquely matched to a census record. In this round,

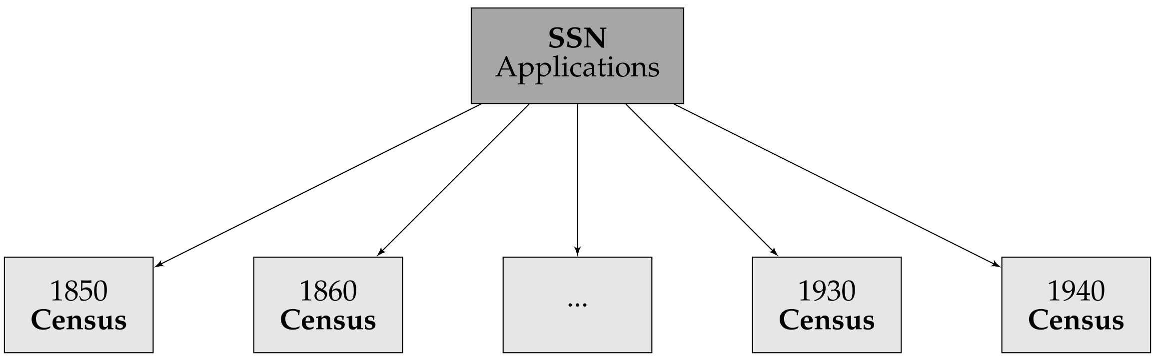
[6]The use of the NYSIIS phonetic algorithm helps in matching names with minor spelling differences, as mentioned in Abramitzky et al. (2021a)

we keep all matching variables the same, however, we use the phonetically standardized version of the middle name to account for spelling discrepancies. Once again, we separate those SSN applications that were uniquely matched to the census and those that were not.

4. We repeat this matching process where we remove successfully matched individuals and attempt to rematch unmatched applications from our pool of potential matches. As we progress through the rounds of linking, the additional matching criteria become less stringent. We allow for misspellings or remove one or more variables in each subsequent iteration until we arrive at the literature standard, which involves only first and last name with spelling variations allowed, state of birth, and year of birth within a 5-year band.

We attempt to match each SSN record to all the census decades available as an individual may appear in the 1900 and 1910 census, for example. For married women applicants, we search for potential census matches using both their maiden and married names. As a result, if we are able to find both records, married women appear in our data twice. We assign these links a slightly altered SSN to differentiate between the married and unmarried SSN-Census link. We do not link married women in the census who are below the age of 16.

FIGURE C.9: First & Second Linking Stages

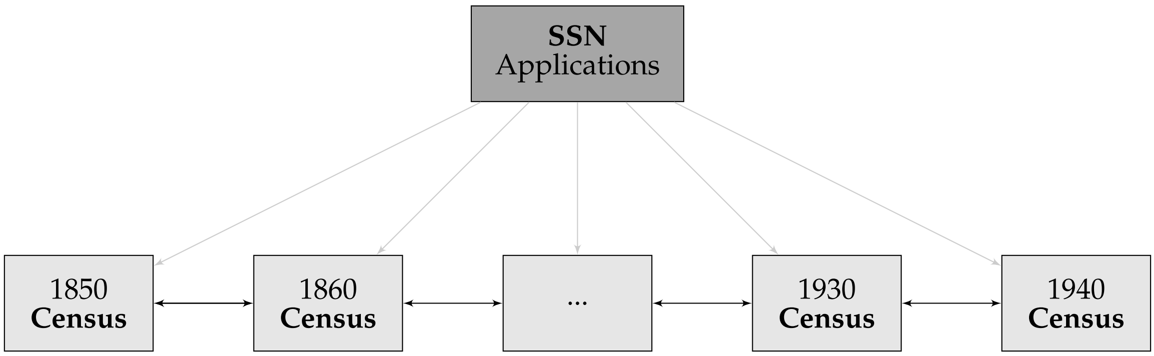


Notes: This figure shows the first and second step of our linking procedure—linking individuals' Social Security Numbers to their census records.

Second stage: SSN applicant parents $\leftrightarrow$ census. Specific birth details for mothers and fathers are not available in the SSN applications meaning we cannot directly link them like we do for the applicants. However, if we can successfully link an SSN applicant to their childhood census record, it is possible to identify and link their parents to other census decades. This process also allows us to identify grandparents. Importantly, we have mother's maiden in the SSN application data, allowing us to link a married mother to her unmarried census record. For parents that we are able to identify in the census from a successful SSN-census link, we apply the same matching procedure described above. However, an important difference is that we do not use parent names (as we no longer have that information), but we are able to use spouse name and information on their parents' birthplace (i.e., the SSN applicant's grandparents birthplace) which is available from the census records. For parents who are not SSN applicants themselves, we create a synthetic identifier similar to an SSN.

Third stage: Census ↔ census. Having assigned unique SSNs or synthetic identifiers to millions of individuals in the census records, we can link these records over time. We cover all possible pairs of census decades from 1850 to 1940.

FIGURE C.10: Final Linking Stage



Notes: This figure shows the final step of our linking procedure—linking individuals' census records over time. Once we have linked SSN applications to the census as well as linked their parents where possible (stage one and two), we link individuals across censuses despite potential name changes upon marriage.

C.2 Sample Weight Construction

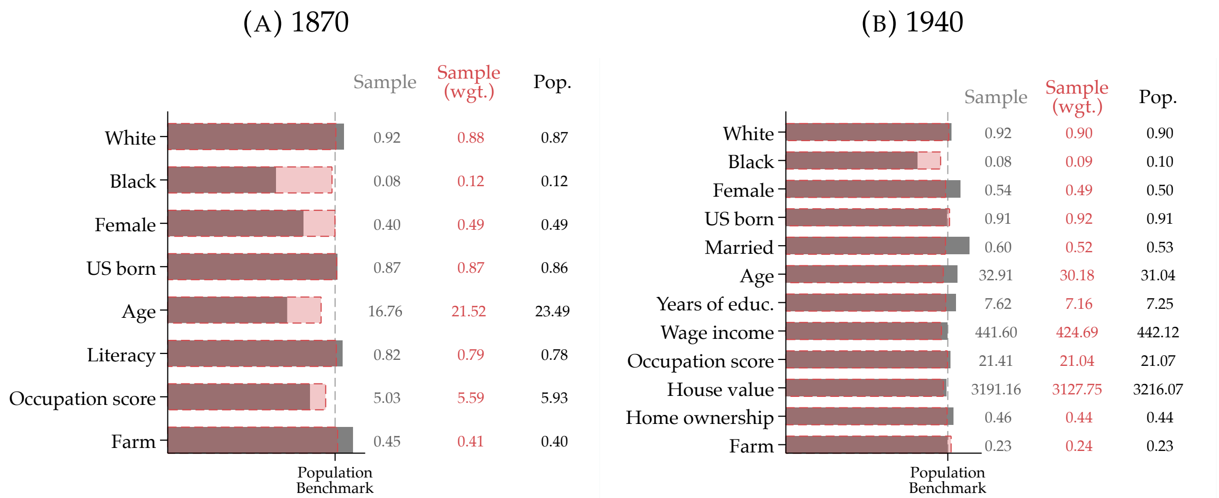
We use inverse propensity score weights so that our sample is representative of the overall population across key observable characteristics.

For each census between 1850 to 1940, we create indicator variables for whether (1) we have identified an individual's Social Security Number, (2–4) whether we have been able to measure the economic status of the individual's (2) mother, (3) father, or (4) both parents. Measuring parental economic status may itself involve census linking and does not rely on observing parents in the same census wave.

In a second step, we then divide the population into groups based on their observable characteristics and (non-parametrically) compute the propensity of each group to be included in our sample via indicators (1–4). Those groups are comprised of individuals with equal (i) sex, (ii) race, (iii) age in decades, (iv) region, (v) farm-status, (vi) literacy, (vii) rural-urban status, (viii) state of birth, (ix) homeownership, (x) marital status, (xi) school attendance, (xii) occupational group, and (xiii) industry group.

As the final sample weight, we assign an individual the inverse propensity of being observed in our linked panel given the characteristic-based group to which they belong. We use different sample weights depending on whether we require only the individual to be linked across time (1), observing the person's and their mother's economic status (2), observing the person's and their father's economic status (3), or observing the person's and both of their parents' economic status (4).

FIGURE C.11: Sample Balance After Inverse Propensity Weighting (1870 & 1940)



*Notes:*This figure shows the representativeness of characteristics among individuals who we successfully assign an SSN compared to the full population in each census before 1940. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Our inverse propensity weights produce an almost perfectly representative sample. Panel A shows the 1870—typically the first year we include in our results—and Panel B shows 1940—the last year of our panel.

Figure C.11 shows average sample characteristics after applying our new inverse propensity weights. The reweighted sample is almost perfectly representative of the full population in all dimensions, even those not targeted by our reweighting method. For example, wage income and occupational income scores match close to perfectly despite only having included coarse occupation and industry categories in our reweighting procedure. Similarly, housing wealth is not targeted but our reweighted sample closely mirrors the overall population.

References

ABRAMITZKY, R., L. BOUSTAN, E. JACOME, AND S. PEREZ (2021a): "Intergenerational Mobility of Immigrants in the United States over Two Centuries," *American Economic Review*, 111, 580–608.

ABRAMITZKY, R., L. P. BOUSTAN, K. ERIKSSON, J. J. FEIGENBAUM, AND S. PÉREZ (2021b): "Automated Linking of Historical Data," *Journal of Economic Literature*, 59, 865–918.

AZEVEDO, J. P., V. SANFELICE, AND M. C. NGUYEN (2012): "Shapley Decomposition by Components of a Welfare Aggregate," .

FOURREY, K. (2023): "A Regression-Based Shapley Decomposition for Inequality Measures," *Annals of Economics and Statistics*, 39–62.

LUNDBERG, S. M. AND S.-I. LEE (2017): "A Unified Approach to Interpreting Model Predictions," Working Paper.

OWEN, G. (1977): "Values of games with a priori unions," in *Essays in Mathematical Economics and Game Theory*, ed. by R. Heim and O. Moeschlin, New York: Springer.

REDELL, N. (2019): "Shapley Decomposition of R-Squared in Machine Learning Models," Working Paper.

SHAPLEY, L. (1953): "A value for n-person games," in *Contributions to the Theory of Games*, ed. by H. Kuhn and A. Tucker, Princeton University Press, vol. 2.