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: December 1, 2023.

Abstract

Throughout US history, mothers played an important role in educating their children. However, this role is obscured because limitations in data and methods 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 key determinant of her child's outcomes, often surpassing the influence of fathers, especially for daughters and Black children. More generally, maternal human capital is more important for children with limited schools access. 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, Pablo Valenzuela, and numerous seminar participants 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 limitations in data and methods. Existing evidence on how parental background 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 background, 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 census 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 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 key predictor of child outcomes, surpassing the importance of paternal human capital. The significance of mothers is particularly pronounced in groups and places with low school access. These results are consistent with mothers' central role as educators during the 19^{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 women. Using these data, we link the census records of 21 million women along with a similar number of men, resulting in a highly representative panel. We make this dataset publicly available to further the reassessment of women's contributions to US history.

Second, to assess the joint importance of 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 using only one 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 a statistical decomposition method (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 underlying outcomes of interest are observed. In the historical data, such binary proxies are common—for example, literacy as a proxy for human capital. Using this method, we recover rank-rank relationships between the latent outcomes of interest while imposing only mild assumptions on their distributions.

Our first key finding is that parental human capital plays a key role in the intergenerational transmission of income. Most prior empirical studies focus on income-to-income transmission alone. Yet, the separate importance of parental human capital and income is a central aspect of intergenerational mobility theory (Becker et al., 2018). Our findings show that after incorporating the role of parental human capital, intergenerational mobility increased over the 19$^{\text{th}}$ century, 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, who 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).

In line with the qualitative historical evidence, our second main finding shows that mothers' human capital often more strongly predicts their children's outcomes than fathers'. We find that mothers on average account for 13 percent more of the variation in

child human capital than fathers—for daughters and Black children one-third to two-thirds more. These findings suggest that mothers play a critical role in the intergenerational transmission of human capital, especially for female and Black children.

As a potential mechanism for the importance of maternal human capital, we explore variation in access to schools across race, sex, place, and time. We find that when school access is low, mothers are particularly important predictors of child human capital. For children born before widespread public education, school access explains half of the variation in mothers' predictive power relative to fathers'. Similarly, we find that as school access 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 results in two main ways. First, we relate children's outcomes (ages 13–16) 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. Second, we validate our semi-parametric latent variable method and show that, in contrast to ordinary least squares, it correctly identifies levels and trends of rank-rank mobility based on binary proxies (e.g., literacy) for unobserved underlying outcomes of interest (e.g., human capital).

The first key contribution of this paper is to construct one of the most extensive and representative panels on intergenerational mobility that includes 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. (2023b) innovatively use crowd-sourced family trees, leading to vastly larger sample sizes. In contrast to prior work, we leverage historical *administrative* data, allowing for both scale and representativeness.[1]

[1] Espín-Sánchez et al. (2023) employ a small subset of the same administrative data.

This paper also deepens our insights into how mothers shaped Americans' life chances throughout history. Earlier studies focused either on father-child correlations (Abramitzky et al., 2021a; Ward, 2023; Craig et al., 2019; Jácome et al., 2021; Buckles et al., 2023b) or the correlation between parents' average status and child outcomes (Chetty et al., 2014b; Card et al., 2022). None of these prior studies assess mothers' importance in the intergenerational transmission of economic outcomes. Our paper emphasizes mothers' separate role in shaping child outcomes, uncovering that maternal human capital is a stronger predictor than father-based proxies, especially for female and Black children. Espín-Sánchez et al. (2023) develop parametric assumptions under which the role of women in intergenerational mobility can be inferred from the outcomes of male family members. Instead, our latent variable method overcomes critical measurement issues to estimate women's role in intergenerational mobility directly, allowing us to highlight the mechanisms underlying their impact.

Including mothers in the study of mobility in US history is especially pressing given that evidence from other contexts suggests mothers are key determinants of child outcomes. For Norway, Black et al. (2005) find a child's education is positively impacted by their mother's but not their father's education. 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 human capital whereas fathers' education is not, which they argue is a result of mothers spending more time with their children than fathers.

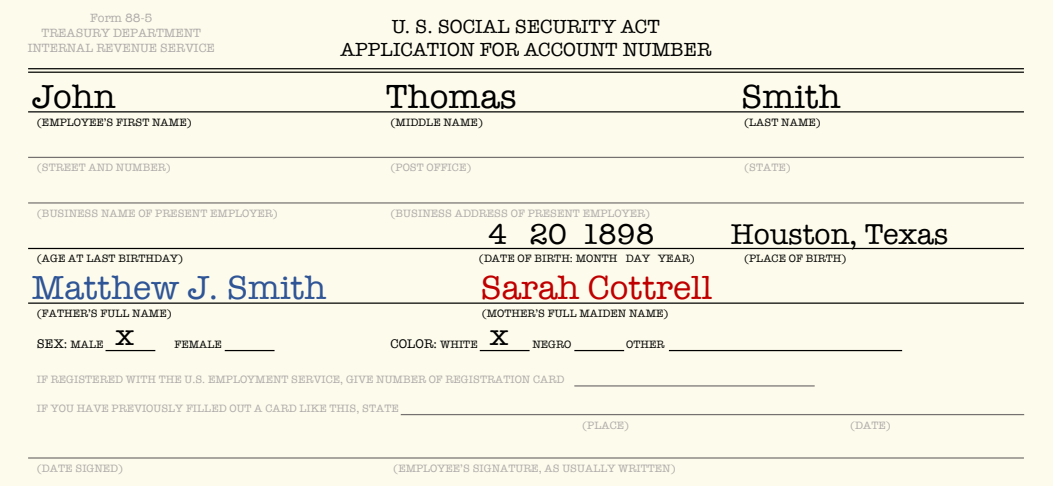
Lastly, this paper is part of an ongoing reassessment of intergenerational mobility in US history. Ward (2023) has illuminated the impact of measurement errors on mobility estimates. Jácome et al. (2021) demonstrate that excluding certain groups, notably Black daughters, skews perceptions of mobility trends. Eshaghnia et al. (2022) show that mobility measured in lifetime outcomes is magnitudes lower than 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)

A 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 stands out in its coverage and representativeness, particularly because it includes women.

2.1 Historical Administrative Data (Social Security Administration)

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.

The 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 each 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 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 to almost 100 percent. Throughout, the share of female applicants has been close to 50 percent (see Appendix Figure C.1). 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. The inclusion of parents in the SSN application files extend this coverage further back.

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 name, 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 census records over time.

First stage: Applicant SSN ↔ 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 observation.

Second stage: Parent SSN ↔ census. After linking SSN applicants to their census records, we focus on linking their parents to the census. Since specific birth details for applicants' parents are not available in the SSN applications, we cannot directly link them as we do for applicants. However, if a child's SSN application is successfully matched to a census record, and that census record shows the child residing with their parents, we can link the parents from an SSN application to that specific census household. For parents who are not SSN applicants themselves, we create a synthetic identifier similar to an SSN.

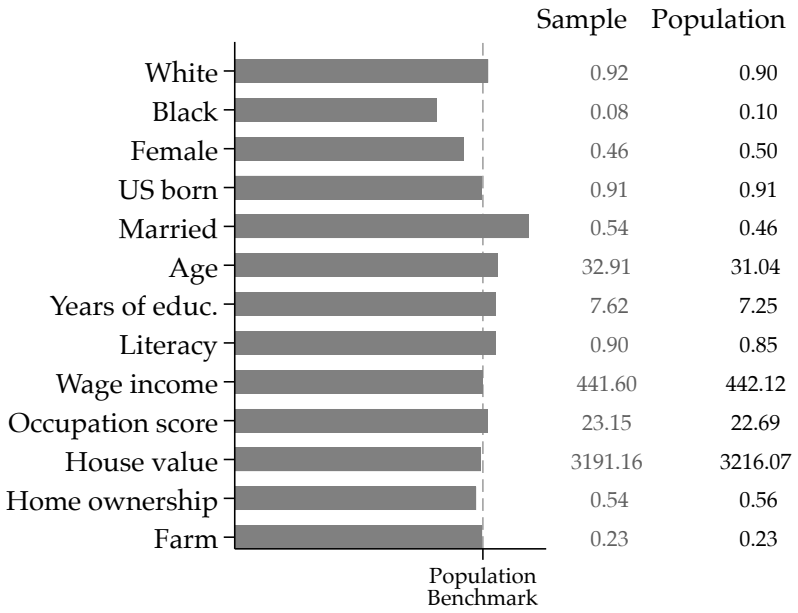
Third stage: Census ↔ census. Having assigned unique identifiers to millions of individuals in the census records, we can link these records over time irrespective of name changes. We cover all possible pairs of census decades from 1850 to 1940.

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 despite name changes—a feature not replicable by standard linking or machine learning methods. Second, using different methods for different subgroups would compromise the 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. Our 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.

FIGURE 2: Sample Balance Prior to Weighting (1940)

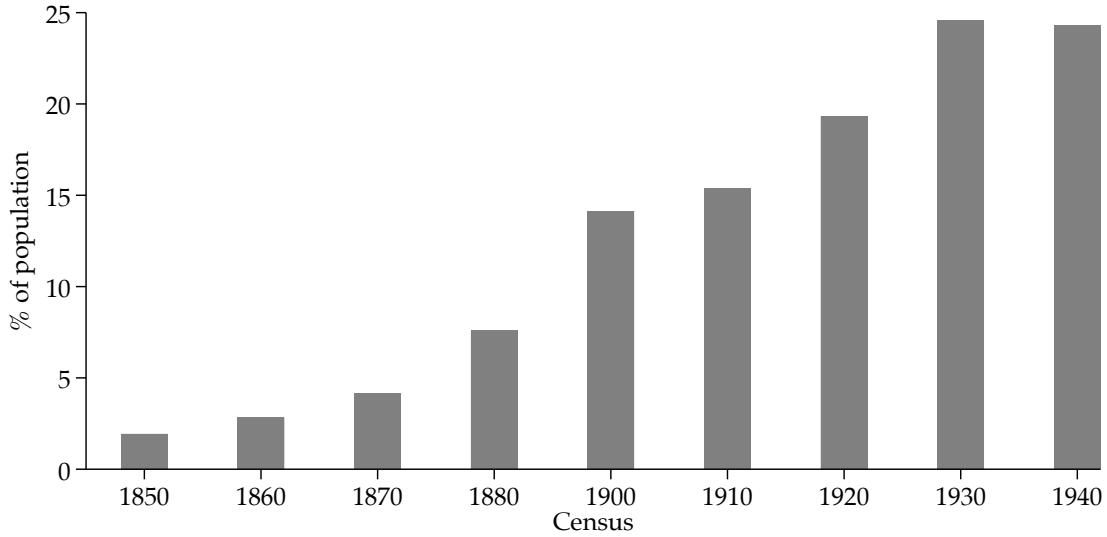


Notes: This figure shows the representativeness of characteristics among individuals in the 1940 census who we successfully assign an SSN compared to the full population in the 1940 census. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race.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

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 people 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.

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.

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 we use the names of a person's children or spouse in the linking procedure if they are known to us, improving linking rates for those who have children, a spouse, or both.

We reweight our sample to more closely resemble the US population's characteristics in our empirical analysis.[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.2).

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.

Compared to existing linked census data, our new panel covers a substantial number of individuals whose records have not previously been linked, while maintaining

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

high agreement rates with existing data for overlapping individuals (see Appendix Figure C.4). Our panel shares the most data with the novel Census Tree—an innovative, extensive panel that includes women through genealogical data (Buckles et al., 2023a). Agreement rates vary from 80 to nearly 100 percent and are highest with LIFE-M—a panel that leverages vital records in the linking process (Bailey et al., 2022).

2.5 Validation via Census Cross-Section

Results based on the census cross-section provide a valuable benchmark for results derived from our panel. The census cross-section is useful because studying intergenerational mobility in childhood outcomes does not require census linking. Specifically, we use such 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 of 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.

2.6 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 measures, such as literacy or years of education, reflecting the status of both men and women.

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.

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 economic outcomes. First, we propose using the R^2 of a regression of child outcomes on multiple parental inputs as a mobility measure that integrates the roles of both parents. Second, we use a simple decomposition method that allows to separate the contributions of mothers and fathers to the overall R^2. Third, we use a state-of-the-art semiparametric latent variable method to identify rank-rank mobility 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 economic outcomes are a linear function of parental inputs:

$$y_i^{\text{child}} = \alpha + \beta_1 y_i^{\text{mother}} + \beta_2 y_i^{\text{father}} + \varepsilon_i, \tag{1}$$

where y_i^{child}, y_i^{mother}, and y_i^{father} are the ranked outcomes of child i, their mother, and their father, respectively. The focus on ranked outcomes means that we only consider mobility based on relative positions in the distribution (Chetty et al., 2014a).

There are several advantages to the rank-rank approach in our setting. First, correlations in ranks 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 mobility measure that can account for multiple inputs. It summarizes the joint importance of 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 $\widehat{y}_i^{\text{child}}$ is the predicted rank of child i from equation (1) and $\bar{y}^{child}$ is the average child rank.

We argue that predictability as captured by the R^2 is an intuitive measure of intergenerational mobility. In a perfectly mobile society, child outcomes cannot be predicted by parental background ($R^2 = 0$). In contrast, if child outcomes can be perfectly predicted by parental background ($R^2 = 1$), society is perfect immobile.

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

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

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 across generations.

The advantage of R^2 is that it can provide an intuitive and easily interpretable measure

[3]The parent-child coefficient $\widehat{\beta}$ is the OLS estimate of β in the equation $y_i^{\text{child}} = \alpha + \beta y_i^{\text{parent}} + \varepsilon_i$.

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 dimensions of parental background. Another advantage is that the R^2 can be decomposed into the contributions of individual inputs, as described in the next section.

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 based on Shapley (1953); Owen (1977).

This decomposition method defines the contribution ϕ_j of each set of inputs $\boldsymbol{x}_j \subseteq V$ to the overall R^2:

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

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 in 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 (Lundberg and Lee, 2017; Redell, 2019), it has not been widely used in eco-

nomics (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 can serve as proxies for these unobserved continuous outcomes. In this section, we propose a method to recover rank-rank relationships in the unobserved continuous outcomes under such data limitations.

Most importantly, while human capital is not directly observable, we observe literacy as a binary proxy thereof. Specifically, we assume that a person is literate if their human capital is above a given threshold.

We further assume that parental and child outcomes are drawn from a joint Gaussian copula distribution, i.e., that there exists a set of unknown monotonically increasing transformations f_c, f_m, f_f so that $\left\{f_c\left(y_i^{\text{child}}\right), f_m\left(y_i^{\text{mother}}\right), f_f\left(y_i^{\text{father}}\right)\right\}^T \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$. 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, 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 $\mathbf{\Sigma}$ of the latent variables when only binary proxies are observed. 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).

Intuitively, the method leverages the information that parent-child correlations in literacy contain about the correlations in underlying human capital. More specifically, Fan

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

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.[5] Because the marginal distributions 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 $\widehat{\mathbf{\Sigma}}$, we then estimate the parameters in equation (1) by simulating from $\mathcal{N}(0, \widehat{\mathbf{\Sigma}})$, 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. 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 binarized 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 (Figure A.1, Panel A). 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, its trends are confounded by changing literacy rates across years (Figure A.1, Panel B).

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 developed a Stata command for easy implementation by others.

4. Income Mobility & Parental Human Capital

We measure intergenerational mobility as the share of variation in child outcomes that is attributable to parental background. We demonstrate the theoretical, historical, and

[5] We refer to Fan et al. (2017) for a more detailed and formal description of the estimator and its properties, notably $\sqrt{n}$-consistency.

empirical motivations for incorporating parental human capital in the study of intergenerational mobility. Our results 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 that corroborates the vital importance of parental human capital for child outcomes.

4.1 Income Mobility Accounting for Parental Human Capital

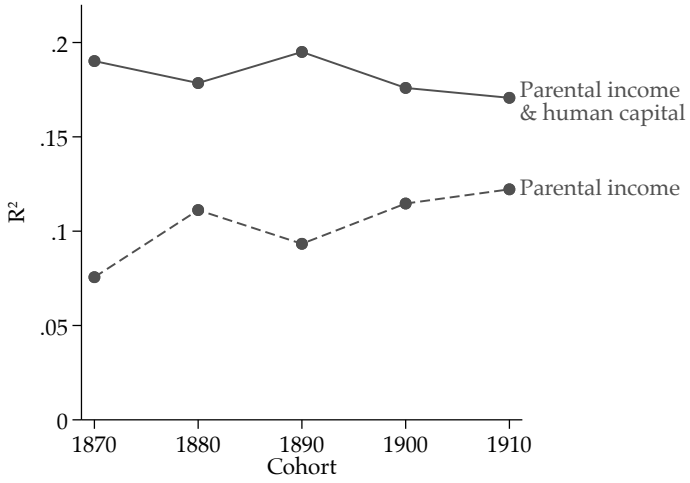
Theories of intergenerational mobility indicate that parental human capital, in addition to income, is a critical determinant of children's outcomes (Becker et al., 2018). Human capital may not only increase parents' capacity for monetary investments in their children but may also shape their children's human capital directly. However, existing empirical studies focus on income and do not take parental human capital into account.

In addition to the theoretical rationale for including parental human capital, there are significant empirical reasons. The lack of detailed data on economic outcomes in historical US data has forced researchers to rely on occupational income proxies. Factoring in human capital can therefore substantially enhance the measurement of parental background in historical data.

We find that parental human capital accounts for a large share of variation in children's incomes, even conditional on parents' incomes (see Figure 4). Most importantly, the broader measure of parental background that includes both income and human capital suggests that intergenerational mobility increased over time—opposite to the conclusion derived from measures that ignore parental human capital (see also Ward, 2023).

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. White sons are least mobile, 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 Black daughters. Black daughters are not only the most mobile group, they are also the only group whose mobility increased over time. It is important to recognize that (1) high within-group mobility

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.

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

Parental human capital was 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).

Women bore most of the responsibility to educate children in the home during the 19th century—a time marked by women's specialization in home production and a 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 for 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 home-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).

Motivated by this historical literature, the subsequent analysis studies the specific role of mothers' human capital in shaping their children's outcomes.

5. Maternal Human Capital & Child Outcomes

We decompose the predictive power of parental human capital into contributions from mothers and fathers. Our findings show that mothers' human capital more strongly predicts child outcomes than fathers'. 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.

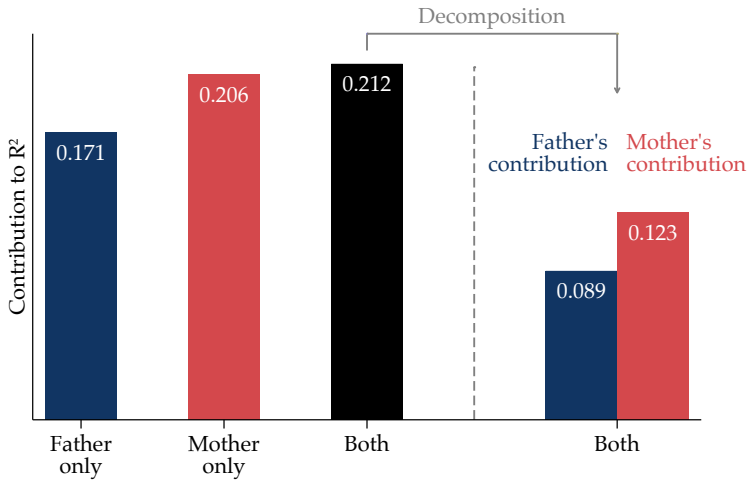
5.1 Parental Human Capital and Child Outcomes

Integrating our baseline model (equation 1) and our method for latent inputs (section 3.4), we evaluate the impact of mothers' and fathers' human capital ranks on their children's ranks in human capital and formal schooling. Figure 5 demonstrates the Shapley-Owen

method to decompose parents' overall predictive power into mothers' and fathers' separate contributions. For children born in the 1880s, mothers' human capital alone accounts for 20.6 percent of the variation in child human capital. Fathers and mothers together predict 21.2 percent of the variation. Notably, using the Shapley-Owen decomposition, we find that mothers account for more than half (58 percent) of the joint predictive power.

FIGURE 5: Illustrating our Decomposition Method

Intergenerational Transmission of Human Capital

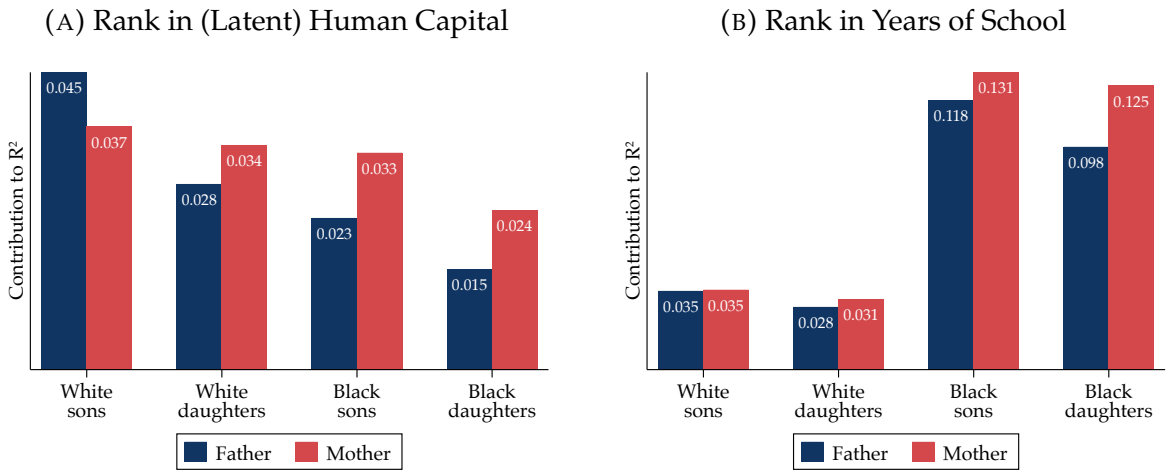


Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. We decompose the 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.

Next, we examine how the transmission varies by child gender and race. 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. Panel B shows this relationship for ranks in years of education. The distinction between human capital and formal schooling is particularly crucial historically, when parental education was the primary source of learning.

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

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



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.

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 percent of variation in school education can be explained by parental education. In contrast, among Black children, parents account for 22 to 25 percent of variation in school education. In line with this finding, school access among white children had massively expanded in the early 1900s (see Appendix Figure A.6). In contrast, most Black children—especially those whose ancestors were enslaved and largely denied literacy until the Civil War—lived in the Jim Crow South with restricted school access, shorter school years, and poor school quality (Card and Krueger, 1992; Althoff and Reichardt, 2023).

Our results also show that racial differences in educational mobility are larger than those in human capital mobility. This finding highlights that formal school education can partly be substituted via parental education. Thus, while the lack of formal schooling was highly persistent across generations among Black families, their mobility in human capital was similar to that of white families.

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. Measuring human capital during childhood in the cross-section suggests stronger intergenerational persistence than our panel-based findings. The disparity partly reflects the intra-generational mobility that is captured in panel data. For example, gaining literacy during adulthood was not uncommon during this period.

6. The Role of Mothers as Educators

So far, we have established that mothers' human capital is more predictive of their child's human capital than fathers'. This section examines mothers' role in home education before the advent of widespread schooling as an explanation for mothers' disproportionate importance, 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

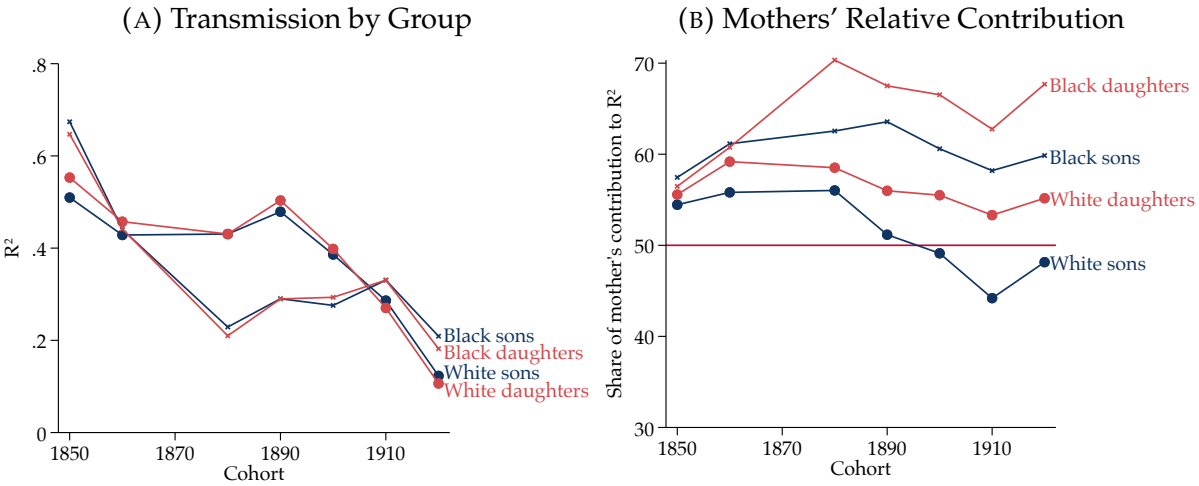
Our comparison of intergenerational human capital transmission over time reveals increasing educational mobility for American children. In the 1850s, parental background accounted for 50 to 70 percent of variation in child human capital, while for those born in the 1920s, this figure dropped to 10 to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

The trend, however, varied significantly by race. After slavery ended, Black children

saw a rapid increase in mobility, which then plateaued over the next half-century. White children's human capital mobility remained low and stable until the late 1800s. However, around the turn of the century, their mobility sharply increased, marking the first time since the Civil War that white children surpassed Black children in educational mobility.

Mothers, consistently more than fathers, have 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.

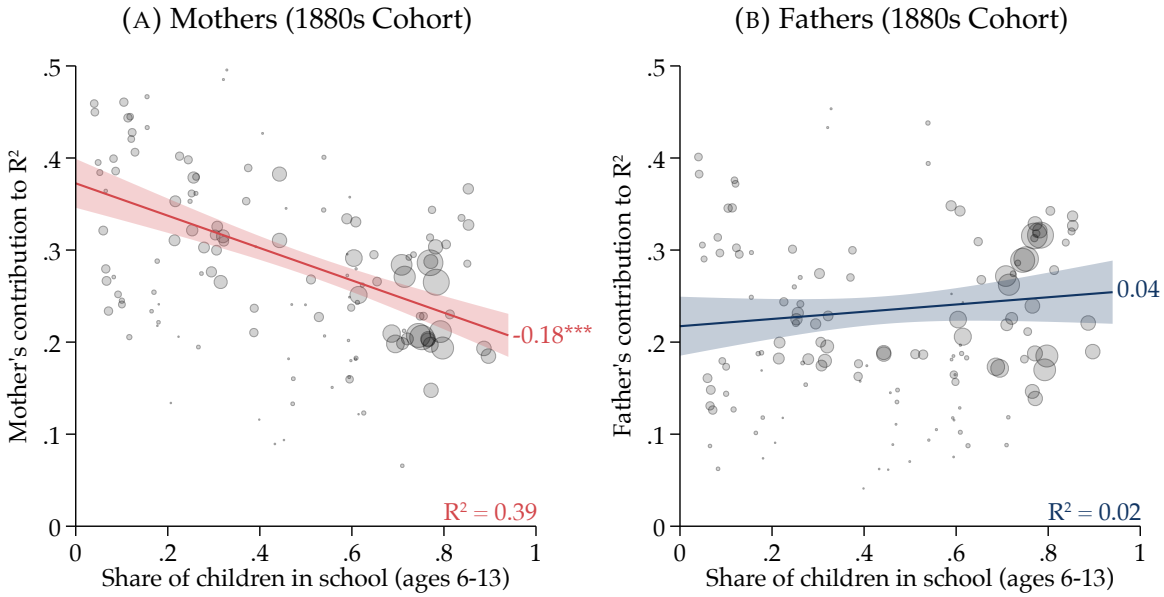
FIGURE 7: Transmission of (Latent) Human Capital Ranks Across Cohorts



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 have highlighted 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).

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 respectively show mothers' and fathers' contributions to the overall R^2 using the Shapley-Owen method. Each dot represents a group of children born in 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.

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' contributions relative to fathers', school access is even more negatively correlated, explaining 51 percent of the variation across groups (see Panel A of Appendix Figure A.5). As school access spread over time, this correlation remained similar but school access accounted for smaller shares of the variation in mothers' importance (see Panel B of Figure A.5). This reduced impact of parental human capital as public education access improved is consistent with Biasi (2023), who shows that equalizing school resources can reduce disparities in intergenerational mobility.

Our results are robust to using alternative measures of school access (see Appendix Table A.1). We newly digitize data on state-specific school ages, enrollment, attendance, and term lengths from the 1880s Census Statistical Abstracts. From these data, we compute the average likelihood of attending school on any given day in the year (ages 6–16), specific to each state. This alternative measure reveals even stronger results: Nearly 60 percent of the variation in mothers' relative contributions to child outcomes can by explained by disparities in school access. Again, we find no correlation between fathers' contributions and school access.

7. Conclusion

This paper studies the influence of maternal and paternal background on child outcomes in the US from 1850 to 1940, emphasizing the role of maternal human capital. We construct a representative panel that includes women in early US history, introduce the R^2 mobility measure to accommodate multiple parental inputs, leverage advanced statistical techniques to analyze intergenerational transmission under data constraints, and separate the impact of maternal and paternal inputs. Our findings highlight the significant influence of maternal human capital on children's outcomes, particularly for daughters and Black children. We propose that gaps in school access may explain why the importance of mothers' human capital for child outcomes varies across race, location, and time.

There are several promising avenues for future research. We expanded the parental status measurement to separately encompass maternal and paternal roles. 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 work (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., A. HAWS, J. PRICE, AND H. WILBERT (2023a): "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project," Working Paper.

BUCKLES, K., J. PRICE, Z. WARD, AND H. WILBERT (2023b): "Family Trees and Falling Apples: Historical Intergenerational Mobility Estimates for Women and Men," 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, AND E. SAEZ (2014a): "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 129, 1553–1623.

CHETTY, R., N. HENDREN, P. KLINE, E. SAEZ, AND N. TURNER (2014b): "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.

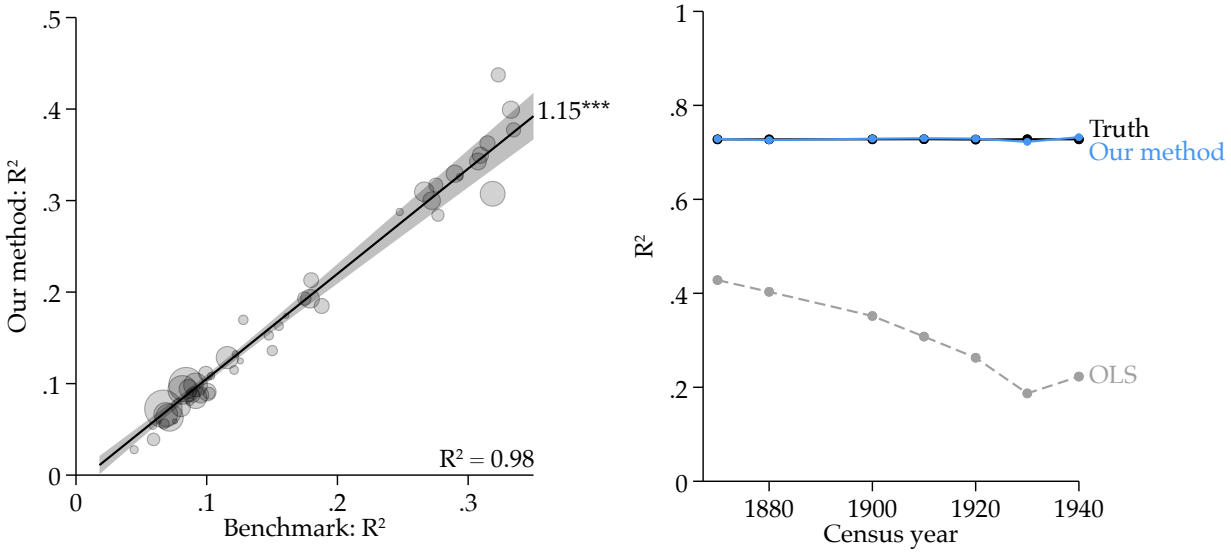
ZUE, L. AND H. ZOU (2012): "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571.

APPENDIX

A. Appendix Figures

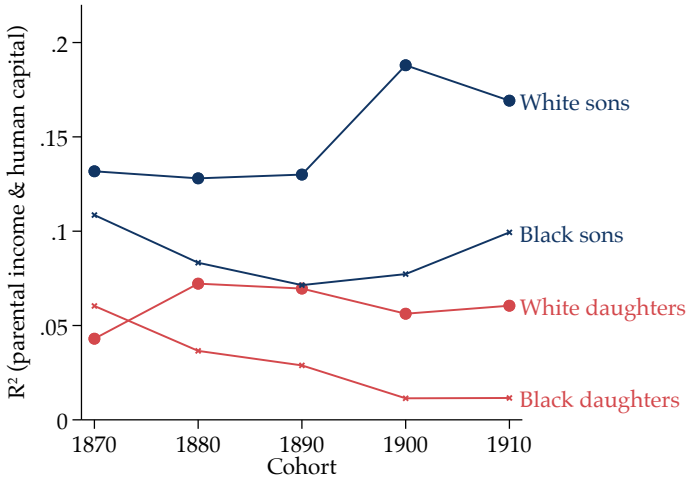
FIGURE A.1: Validation of the Semiparametric Latent Variable Method

(A) Education ranks vs. dummies (1940 census) (B) Literacy dummies over time (simulation)



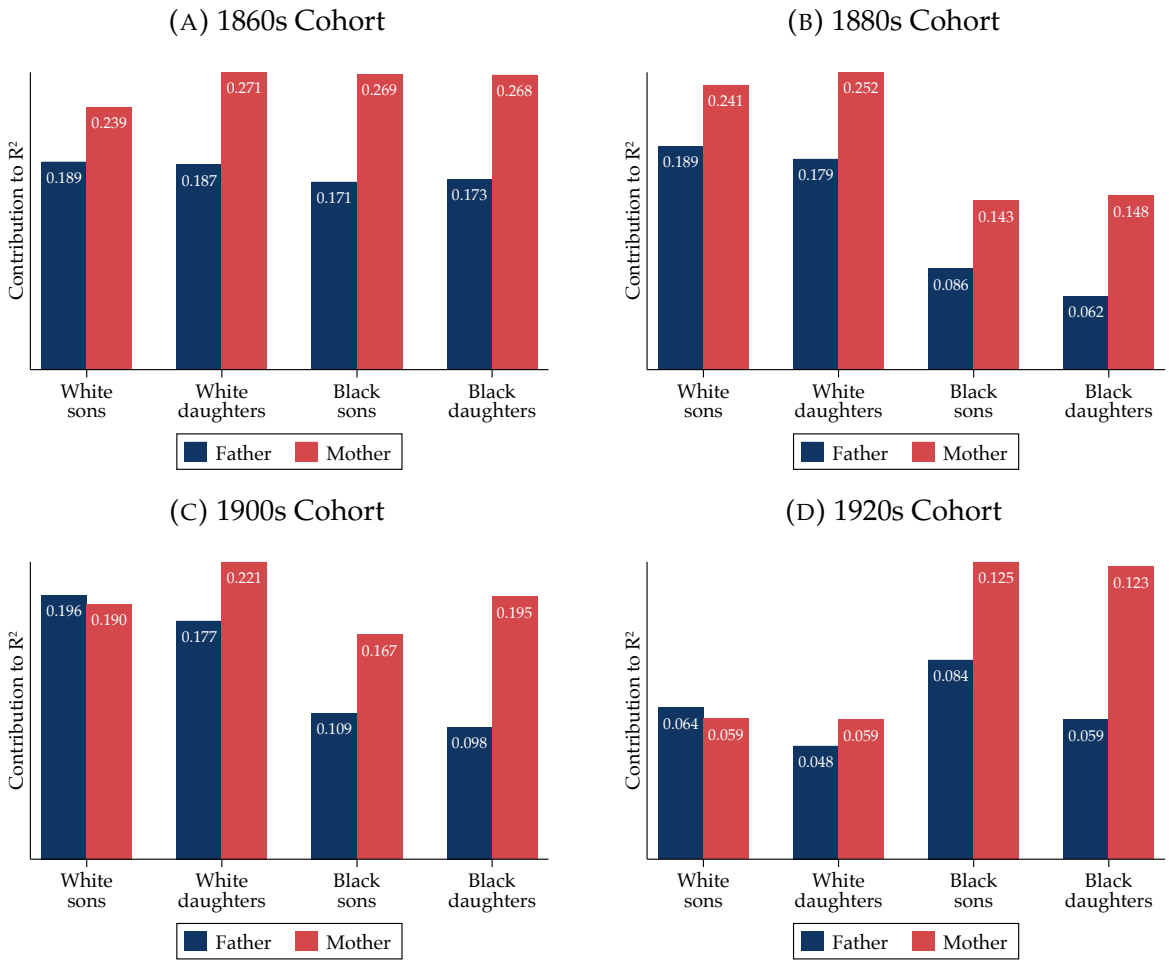
Notes: This figure demonstrates the effectiveness of our semiparametric latent variable method in identifying rank-rank relationships from binary proxies. Panel A contrasts the R^2 values from rank-rank regressions using actual and binarized educational data from the 1940 census. We binarize the data 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 a human capital rank-rank regression, "Our method" from our latent variable method using literacy dummies, and "OLS" from a standard OLS regression with the same literacy dummies.

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



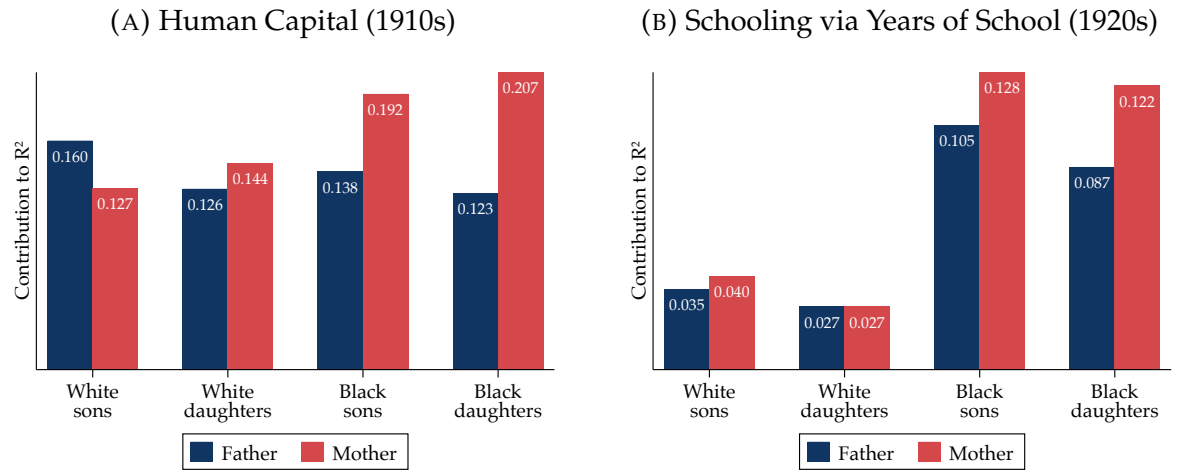
Notes: This Figure shows the share of the variance in a child's household income rank explained by parents' household income ranks and their (latent) human capital ranks (R^2) across cohorts and groups. For parental human capital ranks, we use information on parental literacy and the latent variable 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



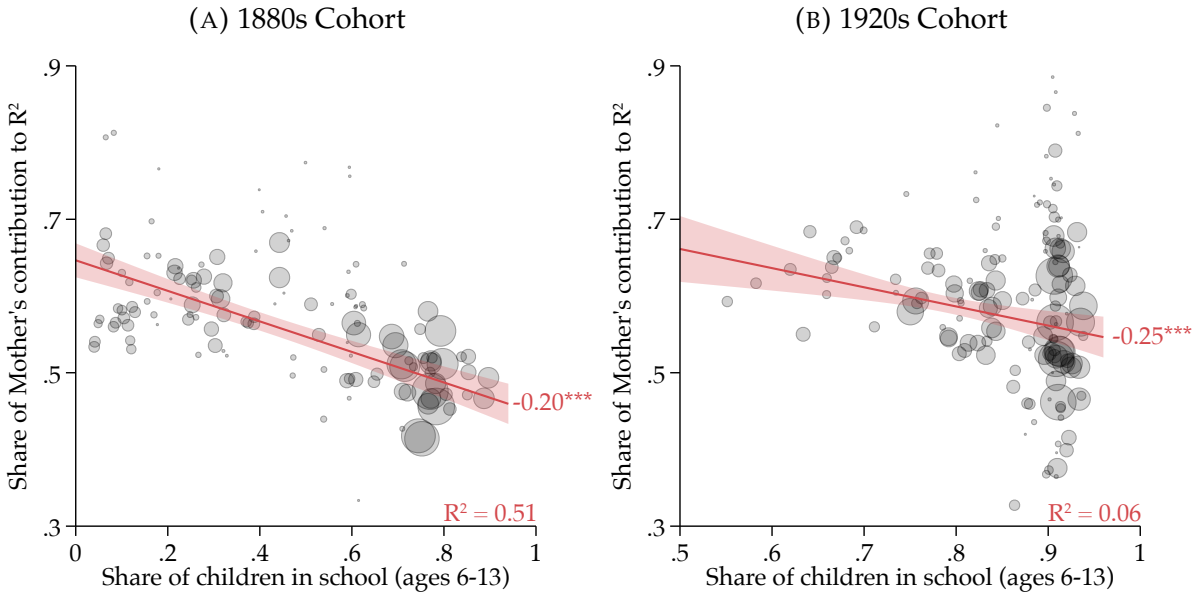
Notes: This figure shows the share of variation in child human capital explained by paternal and maternal human capital across cohorts. We use literacy to measure the rank-based transmission of human capital based on the method we introduce in section 3.4. Results are based on the census cross-section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

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



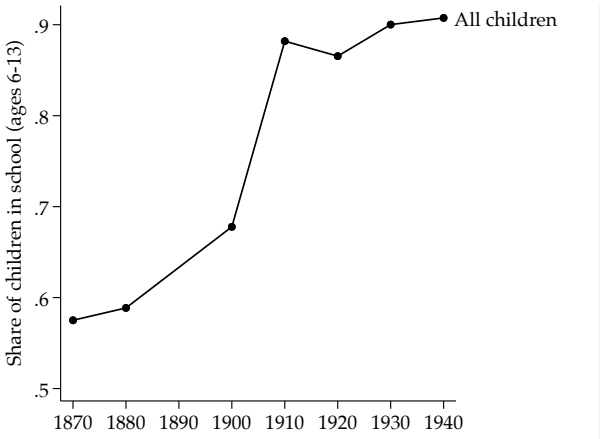
Notes: This figure shows the share of variation in child human capital explained by paternal and maternal human capital for children at ages 13-16 in the 1930 census (1910s cohort); and variation in child schooling explained by paternal and maternal schooling for children at ages 13-16 in the 1940 census (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 group of children born in the 1880s or 1920s, 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.

TABLE A.1: Mothers & Schools—Robustness to Measures of School Access

	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$
School access (baseline)	-0.18***	0.04	-0.20***			
	(0.03)	(0.05)	(0.03)			
Adjusted school access				-0.47***	0.15	-0.58***
				(0.08)	(0.11)	(0.10)
R^2	0.39	0.02	0.51	0.37	0.04	0.57
Observations	133	133	133	128	128	128

Notes: This table shows the relationship between local school access and parents' contributions to child human capital. Columns 1–3 (baseline) contain the results from Figure 8 and Panel A of Appendix Figure A.5. For this baseline, school access is determined by the race- and sex-specific share of children aged 6–13 in school according to the 1880 census. Columns 4–6 (robustness) show that these results are even stronger when we use an alternative measure of school access. For this measure, we newly digitized data on state-specific school ages, enrollment, attendance, and term lengths from the Census Statistical Abstracts. From these data, we compute the average likelihood of attending school on any given day in the year between ages 6–16, specific to each state. These data are incomplete for Arkansas and Wyoming, leading to slightly lower sample sizes.

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:

$$\beta = \text{cor}(x, y) \cdot \sqrt{\frac{\text{Var}(y)}{\text{Var}(x)}} \tag{3}$$

$$R^2 = \text{cor}(x, y)^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}(x)}{\text{Var}(y)}, \tag{4}$$

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:[6]

$$R^2 = \widehat{\beta}_1^2 \frac{Var(x_1)}{Var(y)} + \widehat{\beta}_2^2 \frac{Var(x_2)}{Var(y)} + 2\widehat{\beta}_1 \widehat{\beta}_2 \frac{Cov(x_1, x_2)}{Var(y)}. \tag{5}$$

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:

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

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

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

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

marginal contribution so that they sum to the total R-squared:

$$\phi_j = \frac{\Delta_j}{k!}, \tag{6}$$

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.2 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 percent of the variation in child outcomes explained by parental background.

TABLE B.2: 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 (6), 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 (5), 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

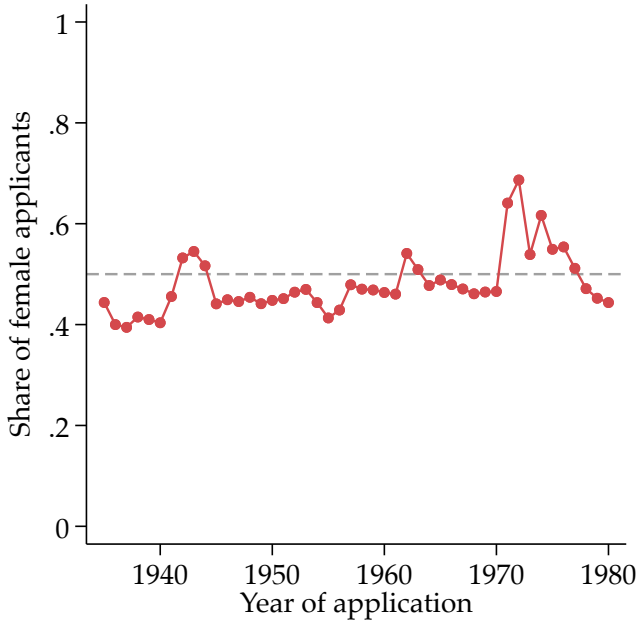
$$\phi_1 = \frac{1}{2Var(y)}\left(2\widehat{\beta}_1^2 Var(x_1) + \{Cov(x_1, x_2)\}^2\left[\frac{\widehat{\beta}_2^2}{Var(x_1)} - \frac{\widehat{\beta}_1^2}{Var(x_2)}\right] + 2\widehat{\beta}_1\widehat{\beta}_2 Cov(x_1, x_2)\right).$$

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

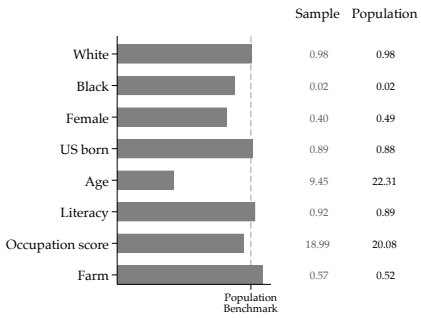
FIGURE C.1: Share of Female Applicants



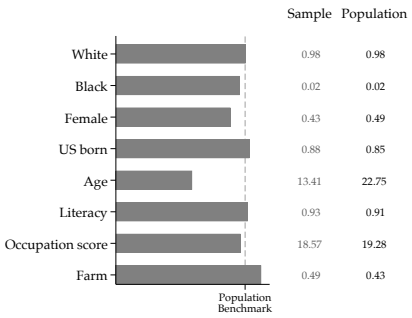
Notes: This figure shows the share of SSN applicants who are female by year of application.

FIGURE C.2: Sample Balance Prior to Weighting (1850–1920)

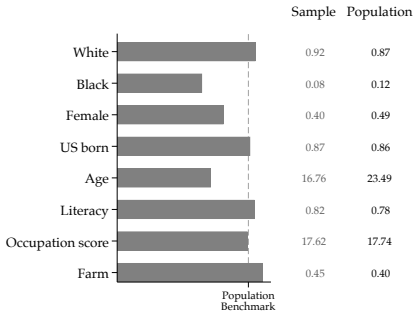
(A) 1850



(B) 1860

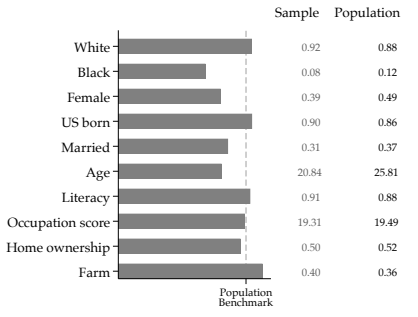


(C) 1870

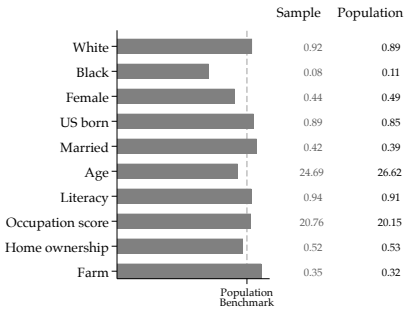


(D) 1880

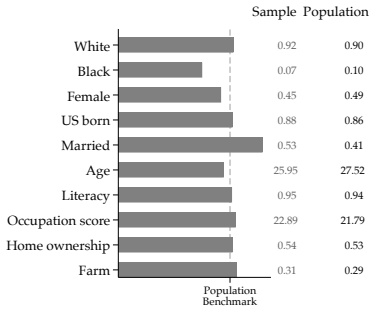
(E) 1900



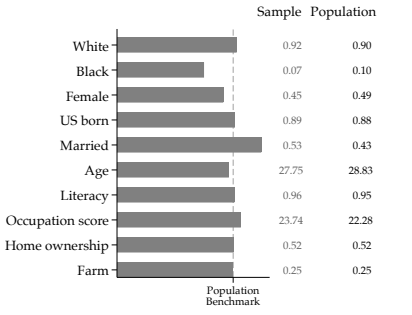
(F) 1910



(G) 1920

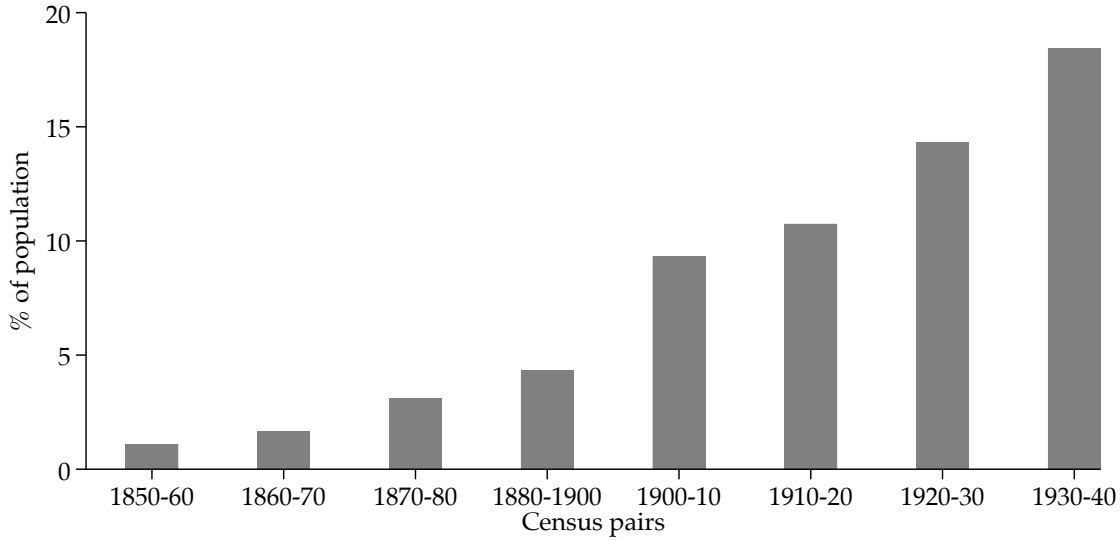


(H) 1930



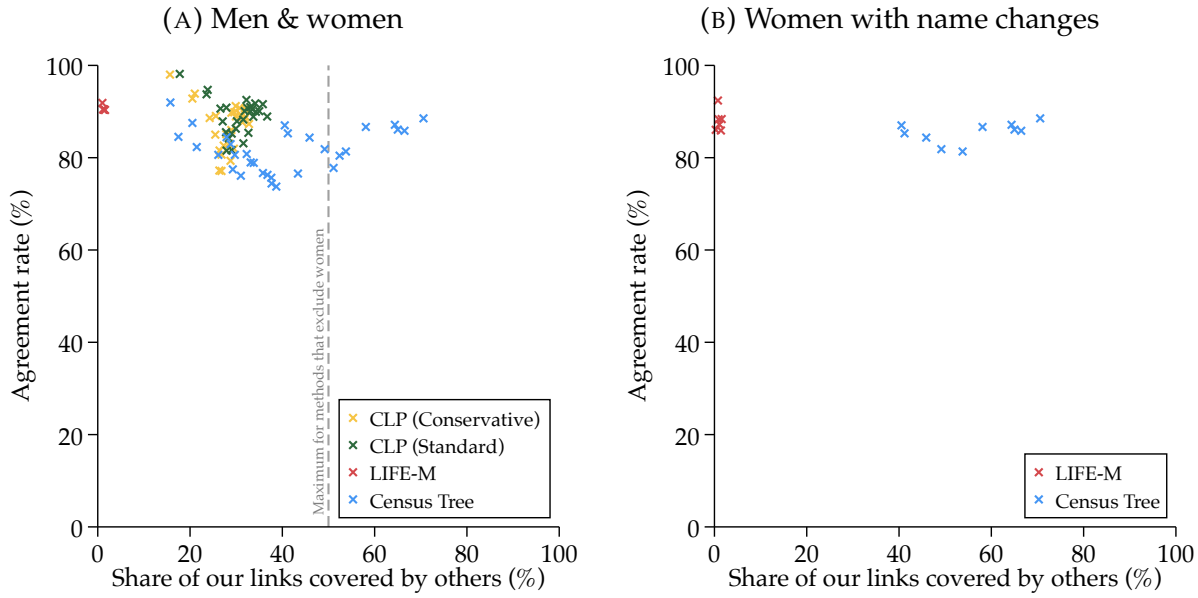
*Notes:*This figure shows the representativeness of characteristics among individuals who we successfully assign an SSN compared to the full population in each census before 1940. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Because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

FIGURE C.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 link from one census decade to the next. Our empirical analysis also leverages links across non-adjacent census pairs, further increasing coverage.

FIGURE C.4: Our New Panel Compared to Existing Data



Notes: This figure compares our linked panel (1850–1940) to those of the Census Linking Project (CLP, Abramitzky et al., 2020), LIFE-M (Bailey et al., 2022), and the Census Tree (Buckles et al., 2023). Each point represents a link from one census decade to another (potentially non-adjacent). The x-axis shows the share of individuals in our panel who were not yet captured by previously existing datasets. The y-axis shows the share of agreement with previously existing datasets on which precise records are linked, conditional on having established any link.

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 $\leftrightarrow$ 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)[7].

- *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

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

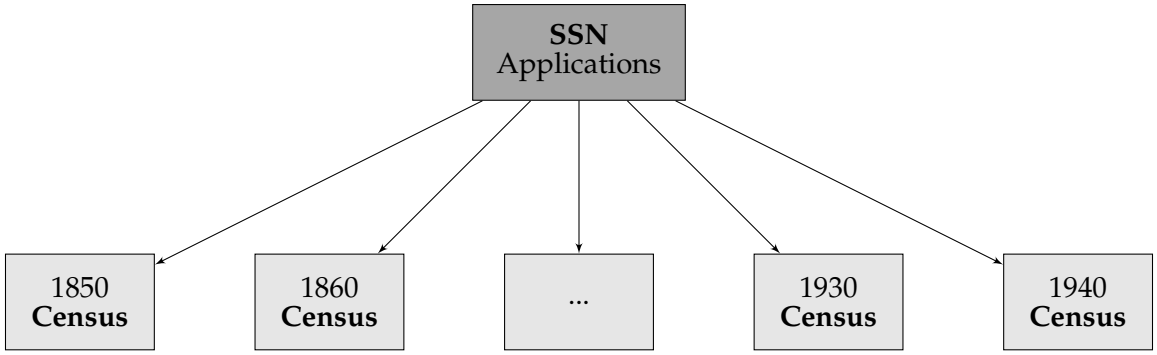
algorithm has the following steps:

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, 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 indi-

vidual 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.5: First & Second Linking Stages



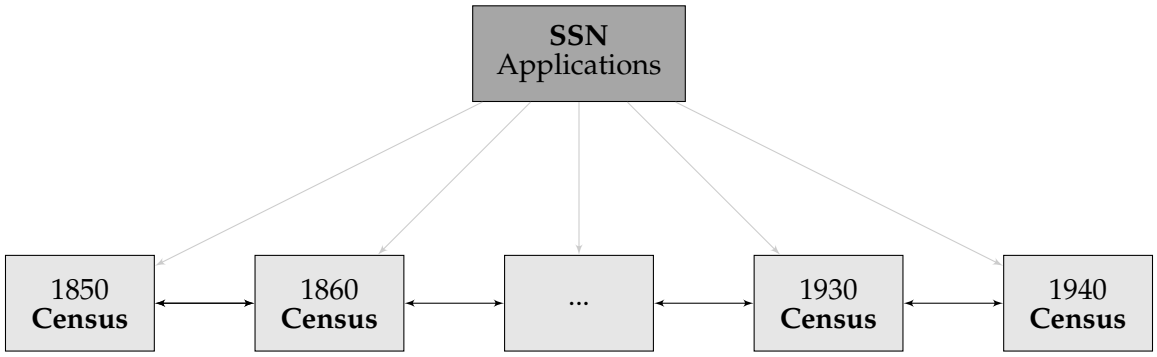
Notes: This figure shows the first and second step of our linking procedure—linking individuals' Social Security Numbers to their census records.

Second stage: SSN applicant parents ↔ 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.6: 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.

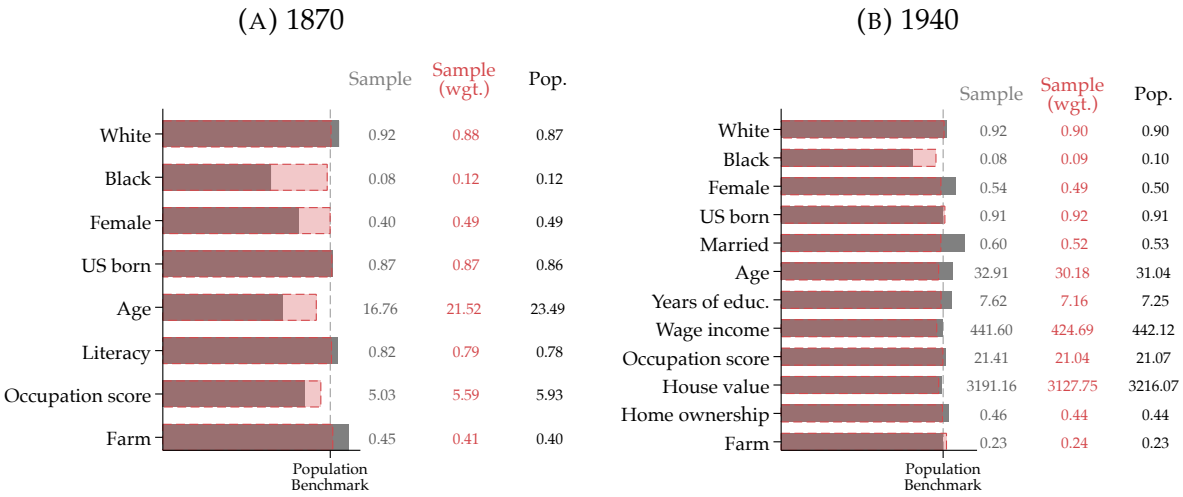
For each census between 1850 to 1940, we create indicator variables for whether (1) we have identified an individual's Social Security Number, (2–4) whether we have been able to measure the economic status of the individual's (2) mother, (3) father, or (4) both parents. Measuring parental economic status 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.7: 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.7 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. BOUSTAN, AND M. RASHID (2020): "Census Linking Project: Version 1.0," dataset: https://censuslinkingproject.org.

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," .

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.

BUCKLES, K., A. HAWS, J. PRICE, AND H. WILBERT (2023): "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project," Working Paper.

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.