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: August 20, 2024.

Abstract

Research on intergenerational mobility in US history has focused on father-son income correlations. To incorporate the role of mothers, we propose a mobility measure that considers parental human capital alongside income (R^2) and a semiparametric latent variable method to accurately estimate this measure from historical data. We also construct a new linked census panel that includes women (1850-1940). Our approach reveals increasing mobility, challenging previous studies that do not explicitly consider mothers' role. Parental human capital, especially mothers', became less predictive over time, accounting for the rise in mobility. We provide evidence that increased school access reduced the importance of family background by substituting for maternal human capital transmission.

*We thank Leah Boustan, Ellora Derenoncourt, Rebecca Diamond, Alice Evans, Claudia Goldin, John Grigsby, Ilyana Kuziemko, Robert Margo, Petra Moser, Daniele Paserman, Pablo Valenzuela, Daniel Wilhelm, 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

Studies on the evolution of intergenerational mobility in US history have focused on men, studying the link between fathers' and sons' economic status. This male-centric focus has two main reasons: a lack of intergenerational datasets that include women and the emphasis on income as the primary measure of parental background, which fails to capture mothers' contributions in an era of limited female labor force participation. Other literatures, in contrast, highlight mothers' key role in child development, for example by serving as primary educators before the widespread establishment of schools.

In this paper, we study how both mothers and fathers shaped children's life chances in the US between 1850 and 1940. We find that intergenerational mobility increased from the 19th to the early 20th century when considering a measure of parental background that incorporates human capital alongside income. This finding challenges previous evidence of declining mobility based on fathers' and sons' incomes alone. The rise in mobility is driven by the substantial role of mothers' human capital in the early period, which diminished as formal schooling gradually substituted for maternal home-education.

By constructing one of the first linked census panels to include women, we trace the parental backgrounds of sons and daughters. We 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 applicants' mothers and married female applicants. 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 will make this dataset publicly available.

We also develop a novel methodology to account for multiple dimensions of parental background in the intergenerational analysis. 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, this measure accommodates multiple parental inputs. We show that the R^2 has many desirable properties 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).

Finally, we use a cutting-edge statistical technique to accurately estimate intergenerational mobility despite limitations in the historical data. Specifically, we build on a recently developed semiparametric latent variable method to study rank-rank relationships between parents and children when only binary proxies of the underlying outcomes are observed (Fan et al., 2017). In the historical data, such binary proxies are common; in

our case, literacy serves as a proxy for human capital. We discuss the assumptions it imposes on the joint distribution of parent and child outcomes and extensively validate the method, including the use of modern datasets (PSID and NLSY) where we directly observe continuous measures of human capital through cognitive test scores.

Our first main finding is that intergenerational mobility increased from the 19th to the early 20th century. Specifically, we find that parental background, incorporating human capital alongside income, became less predictive of children's income over time. While the separate importance of parental human capital and income is a central aspect of intergenerational mobility theory (Becker et al., 2018), prior empirical studies focus on income-to-income transmission alone. We find that accounting for parental human capital and including daughters in the analysis reverses the trend in mobility over time.

Our second main finding is that the changing role of maternal human capital accounts for the increase in intergenerational mobility over time. The predictive power of mothers' human capital initially exceeded fathers', but it gradually declined to make both parents' contributions comparable. Statistically decomposing our R^2-measure, we show that mobility would have decreased without the diminishing predictive power of mothers' human capital. This finding is consistent with mothers' key role in the transmission of human capital and shows that previous evidence of declining mobility is due to a focus on paternal factors.[1]

As a potential mechanism for the historically large and declining role of maternal human capital, we explore the role of formal schooling. Until the late 19th century, public schooling was limited in many places and home education was common. From 1880 to 1900, the share of children (ages 6–13) in school rose from 60 to 90 percent. Historians have highlighted the pivotal role of parental human capital in child development before this transition (Kaestle and Vinovskis, 1978). 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). The expansion of school access could therefore be a reason why parents' human capital—especially mothers'—became less predictive and mobility increased over time.

We find that, indeed, intergenerational mobility increased with school access and that maternal human capital accounted for this trend. Specifically, mothers' (but not fathers') human capital was more predictive for children whose school access was low. For example, we find that Black children who lacked equal access to schools during the Jim Crow era relied more on their mother's human capital than white children. We also leverage variation from state compulsory schooling laws late in our sample period, supporting the conclusion that the rise in formal schooling caused a decline in mothers' predictive

[1]We validate our panel-based findings on human capital mobility using the cross-section of children aged 13–16 in their parents' household, bypassing the need for record linkage.

power over time. These findings offer an explanation for the importance of maternal human capital in early US history: as the main educators of their time, mothers were key contributors to their children's human capital and, as a consequence, to their broader economic status.

This paper deepens our insights into how mothers shaped Americans' life chances throughout history. Earlier studies focused on father-child correlations (e.g., Abramitzky et al., 2021a; Ward, 2023 focus on sons while Olivetti and Paserman, 2015; Craig et al., 2019; Jácome et al., 2021; Buckles et al., 2023b also include daughters) or the correlation between parents' average status and child outcomes (Chetty et al., 2014b; Card et al., 2022). None of these prior studies assesses 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. 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 methodology overcomes critical measurement issues to estimate women's role in intergenerational mobility directly, allowing us to highlight the underlying mechanisms.

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. Mothers spend more time with their children than other adults almost anywhere worldwide (Evans and Jakiela, 2024). Perhaps as a result, Black et al. (2005), Holmlund et al. (2011), Lundborg et al. (2014), and Abrahamsson et al. (2024) find that Scandinavian education and health interventions had positive intergenerational spillovers to the children of treated mothers but not treated fathers.[2] Lundborg et al. (2024) use data from randomly assigned donor children and find that only maternal human capital affects child outcomes, suggesting mothers' importance stems from childhood environment rather than genetics (see also Leibowitz, 1974).

This paper also expands our understanding of women's contribution to the economy throughout US history. Goldin (1977, 1990, 2006) pioneered the effort to study women's contributions when their labor force participation rose mid-20th century (see also Fernández et al., 2004; Olivetti, 2006; Fogli and Veldkamp, 2011; Fernández, 2013; Modalsli et al., 2024). For the era before the rise of female labor force participation, evidence on women's contribution is largely limited to documenting their hours worked in home production (Greenwood et al., 2005; Ramey, 2009; Ngai et al., 2024). While the output of home production is typically hard to measure, we uncover the product of one key aspect: the home-education of children. We find that through their unique role in child development, women made a critical contribution to human capital accumulation

[2] García and Heckman (2023) also show that programs to increase mothers' parenting skills increase intergenerational mobility.

in the US economy, even before the rise of female labor force participation.

Lastly, a key contribution of this paper is to construct one of the most extensive and representative panels that include women, building on the foundations of previous work. Craig et al. (2019) and Bailey et al. (2022) initiated the effort to link women's records by expanding automated record linkage developed for men by Abramitzky et al. (2021b). However, the historical birth, marriage, and death certificates they use to do so are available only for selected states and periods. Buckles et al. (2023b) innovatively use crowdsourced family trees, leading to vastly larger sample sizes. However, representativeness remains relatively low owing to the selectivity of users who contribute to these genealogies (Abramitzky et al., 2024). In contrast to prior work, we leverage historical *administrative* data, allowing for both scale and representativeness.[3]

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.

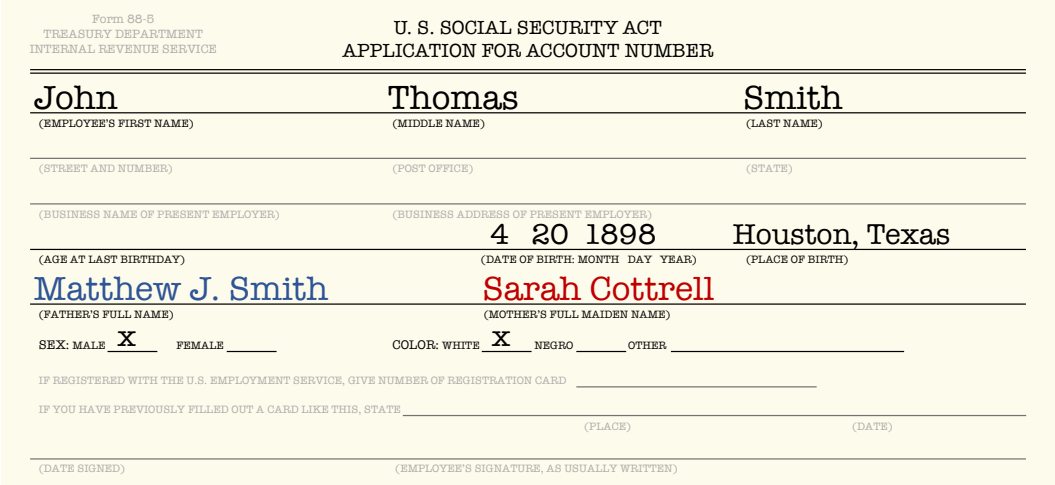
2.1 Historical Administrative Data (Social Security Administration)

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 and a 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

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

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.

share of female applicants has been close to 50 percent (see Appendix Figure D.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. Appendix D.1 describes our linking procedure in greater detail.

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. A link is established if a unique match is found; if dual matches occur, we discard the observation. For married female applicants, we conduct searches under both maiden and married names; however, if links to a census can be established with both names, we establish no link due to the non-uniqueness of the matches.

Leveraging the combination of both applicants' and their parents' names helps us establish *unique* matches for SSN applicants recorded in the same census household as their parents. Historically, this approach is not only effective for children but also adults in the many existing multi-generational households. During our sample period, 80 to 90 percent of Americans lived in multi-generational households. By the end of our sample period in 1940, 60 percent of 21-year-olds and 20 percent of 30-year-olds lived with at least one parent. Note that while using parental names increases the uniqueness of potential matches of those residing with their parents, we also link adults not observed with their parents.

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. A person only enters the linked census panel if their SSN application record is linked to at least two different census decades.

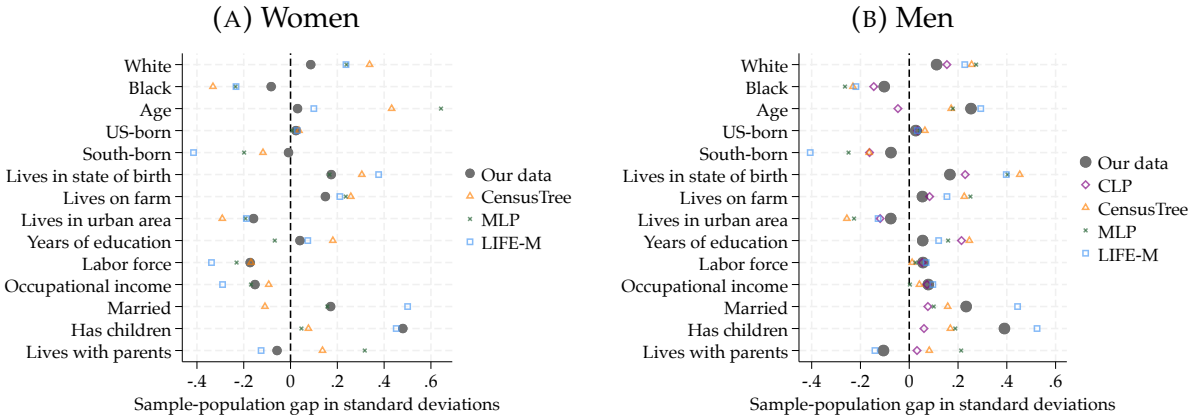
In principle, it would be possible to establish additional links across census records by using standard or machine learning methods. These methods would be particularly useful for men and never-married women, where the issue of name changes does not apply. However, we choose not to use these methods 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 married women would be linked based on a different set of criteria than other groups.

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 more detailed information, notably parents' and spouses' 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: Balance of Linked Sample (1910 to 1940)



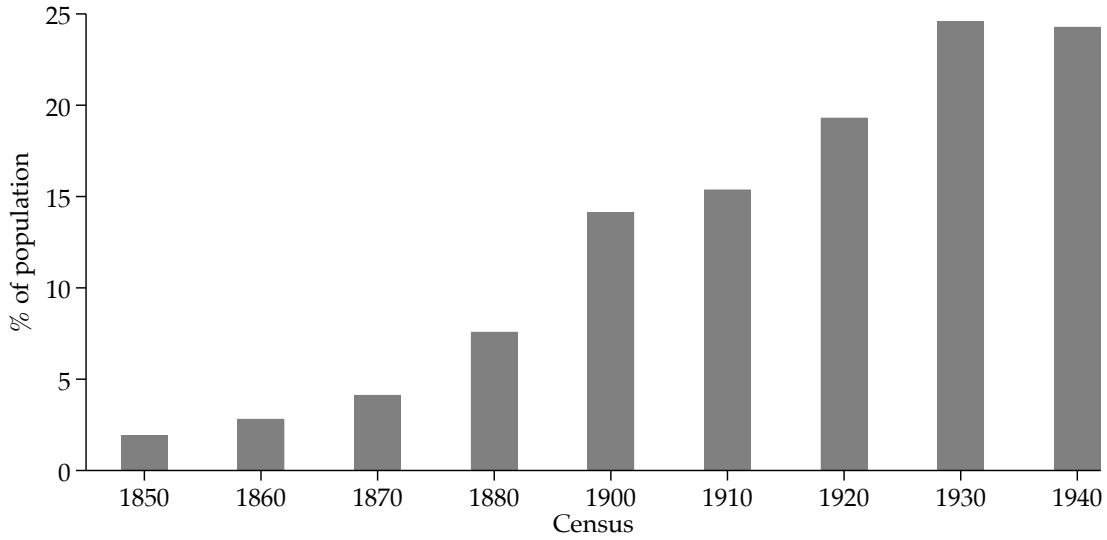
Notes: This figure shows the representativeness of characteristics among individuals in the 1940 census who we successfully link to the 1910 census compared to the 1940 census population of individuals aged 30+ (and therefore alive in 1910); following Abramitzky et al. (2024). We regress each outcome on a dummy for whether we link this individual back to 1910 (outcomes are standardized to have a mean of 0 and a standard deviation of 1). The sample is exceptionally representative compared to existing panels with an average absolute deviation of 0.12 standard deviations (compared to 0.19 to 0.22 among existing data); for this exercise we pool men and women and include "female" and a characteristic. Appendix Figure D.2 repeats this exercise for individuals linked from 1850 to 1880 and those linked form 1880 to 1910. CLP only includes men (Abramitzky et al., 2020). CensusTree uses genealogical data from the user-generated FamilyTree (Buckles et al., 2023a). MLP (Helgertz et al., 2023) contains decade-to-decade links that we append iteratively. LIFE-M (Bailey et al., 2022) covers only Ohio and North Carolina.

Our panel's representativeness of the overall US population exceeds that of other existing linked datasets (see Figure 2). Across a wide range of characteristics, our linked sample (1910–1940) differs on average by 0.12 standard deviations (in absolute terms) from the 1940 population of individuals aged 30 or above. In other linked datasets, this average ranges between 0.19 and 0.22. Representativeness remains exceptional in earlier samples linked from 1850 to 1880 and 1880 to 1910 (see Appendix Figure D.2). Our sample particularly stands out in the key dimensions of sex, race, birthplace, farm and

urban status, and education. Our sample over-represents married individuals and those with children, 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.[4]

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.

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.

We weight our sample to more closely resemble the US population's characteristics in our empirical analysis. Specifically, we use a flexible non-parametric method to construct inverse propensity weights separately for each birth cohort (see Appendix D.2).

Moreover, our panel offers broad coverage. It captures 16–24 percent of the US population from 1910–1940 and 2–14 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 high agreement rates with existing data for overlapping individuals (see Appendix Figure D.3). 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).

[4] Appendix Figure D.5 compares the characteristics of individuals in the census whom we successfully assign an SSN with those of the overall population.

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 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 build on a state-of-the-art semiparametric latent variable method to estimate the R^2 from a rank-rank regression when only binary proxies of underlying outcomes are 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 outcome is a linear function of parental inputs:

$$\text{rank}\left(Y_i\right) = \alpha + \boldsymbol{\beta}' \mathbf{rank}\left(\mathbf{Y}_i^{\text{parental}}\right) + \varepsilon_i, \tag{1}$$

where rank (Y_i) is the percentile rank of outcome of i and $\mathbf{rank}\left(\boldsymbol{Y_i^{\text{parental}}}\right)$ is a $k \times 1$ vector of i's ranked parental outcomes. Parental outcomes can include information on mothers, fathers, or both parents.

There are several advantages to the rank-rank approach, which considers mobility in relative positions in the distribution (Chetty et al., 2014a). 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.

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. While in this paper we focus on human capital and income, the model can be extended to accommodate many different inputs including parents' wealth, grandparents' or other relatives' backgrounds, or neighborhood characteristics.

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

We propose using the R^2 of equation (1) as an intuitive mobility measure that can account for multiple inputs. It summarizes the joint importance of mothers and fathers:

$$R^2 = \frac{\sum_{i=1}^{N}\left[\widehat{\text{rank}}\left(Y_i\right) - 50\right]^2}{\sum_{i=1}^{N}\left[\text{rank}\left(Y_i\right) - 50\right]^2} = \frac{\text{Variance in child outcomes explained by parents}}{\text{Variance in child outcomes}},$$

where $\widehat{\text{rank}}\left(Y_i\right)$ is the predicted rank of i from equation (1) and 50 is the average rank by construction.[5]

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 ($\widehat{\beta}$).[6] In Appendix C.1, we show that in such univariate rank-rank regressions, there is a one-to-one mapping between the parent-child coefficient and our mobility measure: $R^2 = \widehat{\beta}^2$.

The advantage of R^2 is that it can provide an intuitive and easily interpretable measure of mobility even when considering multiple parental inputs. We use this advantage

[5] Note, because the distribution of ranked outcomes is fixed, the variance in child outcomes is constant.

[6] The parent-child coefficient $\widehat{\beta}$ is the OLS estimate of β: $\text{rank}\left(Y_i\right) = \alpha + \beta \cdot \text{rank}\left(Y_i^{\text{parental}}\right) + \varepsilon_i$.

to include both mothers' and fathers' outcomes, and 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 decompose 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 (e.g., rank (Y_i)) on a set of variables $T \subseteq V$ (e.g., $V = \left\{\text{rank}\left(Y_i^{\text{mother}}\right), \text{rank}\left(Y_i^{\text{father}}\right)\right\}$), 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 C.2, we describe the decomposition method in more detail and, for the special case of two parental inputs, provide a closed-form expression for ϕ_j in (1) in terms of the estimated coefficients and the correlation between the inputs.

The Shapley-Owen decomposition offers several unique advantages, being the only that satisfies three formal conditions defined by Young (1985) and 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 economics (recent exceptions are Biasi and Ma, 2023; Fourrey, 2023; Redding and Weinstein, 2023).

3.4 Measuring Mobility with Latent Inputs

Our goal is to estimate intergenerational mobility (R^2) using ranked variables like child and parental human capital. However, historical data often provides only sparse information for key variables, such as binary indicators (e.g., literacy status). This section outlines our methodology for estimating mobility under these data constraints.

Consider the following rank regression with a single input:

$$\text{rank}(Y_i) = \alpha + \beta \cdot \text{rank}(X_i) + \epsilon_i \tag{2}$$

where Y_i represents the child's human capital and X_i represents the parental human capital.[7] This is the simplest version of equation (1); Appendix C.3 generalizes this framework and the discussion below to multiple inputs $X_{i1}, \ldots, X_{ik}$ and provides further formal detail.

3.4.1 Identification challenge

In our data, we do not observe the continuous human capital measures Y_i and X_i. Instead, we only observe binary indicators (literacy status) Y_i^* and X_i^*:

$$Y_i^* = \mathbb{1}[Y_i > \delta_y] \tag{3}$$
$$X_i^* = \mathbb{1}[X_i > \delta_x] \tag{4}$$

where δ_y and δ_x are unknown thresholds that may differ between child and parent. The rank correlation we aim to estimate is a function of the copula (a function that describes the dependence structure between random variables) of the latent variables Y_i and X_i.

However, there exists a large class of copulas that are compatible with the observed empirical distributions of the binary indicators. Without further assumptions, the rank correlation is not identified from the binary data alone.

3.4.2 Gaussian copula assumption

We obtain identification by assuming that the joint distribution of the latent variables follows a Gaussian copula. That is, we assume that there exists unknown monotonic functions $f_Y(\cdot)$ and $f_X(\cdot)$ such that $f_Y(Y_i), f_X(X_i) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\text{diag}(\boldsymbol{\Sigma}) = \mathbb{1}$.[8] Note that this does not impose that the latent variables of interest (e.g., human capital) are normally distributed. Rather, it requires that there exists some monotonic transformations of the latent variables that are jointly normally distributed.

The Gaussian copula distribution is commonly used in the statistics literature due to its flexibility in capturing a wide range of dependence structures, including those in socioeconomic variables (e.g. Liu et al., 2009, 2012; Zue and Zou, 2012). It sufficiently restricts the class of possible copulas to resolve the identification problem.

[7] Both variables are expressed in percentile ranks that range from 0 to 100.

[8] Because we allow for any monotonic transformation, the assumption that the marginal distributions have zero mean and variance equal to 1 is without loss of generality.

3.4.3 Identification of rank correlations from binary indicators

Under the Gaussian copula assumption, we can identify ρ—the correlation between the jointly normal random variables $f_Y(Y_i)$ and $f_X(X_i)$—using the Kendall's tau correlation coefficient of the observed binary variables:

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} (X_i^* - X_{i'}^*)(Y_i^* - Y_{i'}^*).$$

Denote $\Delta_X \equiv f_X(\delta_x)$ and $\Delta_Y \equiv f_Y(\delta_y)$. Then,

$$\begin{aligned} \mathbb{E}[\hat{\tau}] &= 2\left[\mathbb{E}(X_i^* Y_i^*) - \mathbb{E}(X_i^*)\mathbb{E}(Y_i^*)\right] \\ &= 2\left[\mathbb{P}\{X_i > \delta_x, Y_i > \delta_y\} - \mathbb{P}\{X_i > \delta_x\}\mathbb{P}\{Y_i > \delta_y\}\right] \\ &= 2\left[\Phi_2(\Delta_X, \Delta_Y, \rho) - \Phi\left(\Delta_X\right)\Phi\left(\Delta_Y\right)\right]. \end{aligned} \tag{5}$$

where $\Phi_2(\Delta_X, \Delta_Y, \rho)$ is the cumulative distribution function of the bivariate standard normal distribution with correlation ρ (evaluated at Δ_X, Δ_Y), and $\Phi(\cdot)$ is the standard normal CDF. The last equation in (5) follows from the Gaussian copula assumption that $f_X(X_i)$ and $f_Y(Y_i)$ are jointly standard normal.

We can estimate Δ_X, Δ_Y, and τ from the observed binary data and $\Phi_2(\Delta_X, \Delta_Y, \rho)$ is strictly increasing in ρ for any Δ_X, Δ_Y.[9] Therefore, the estimator $\hat{\rho}$ is the unique solution to

$$2\left[\Phi_2(\hat{\Delta}_X, \hat{\Delta}_Y, \hat{\rho}) - \Phi\left(\hat{\Delta}_X\right)\Phi\left(\hat{\Delta}_Y\right)\right] = \hat{\tau}.$$

The rank correlation of two jointly normal random variables with correlation ρ is identified as $\rho_r = \frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)$. Finally, since ranks are preserved under monotone transformations, the rank correlation between the non-transformed latent variables Y_i and X_i are identical. Thus, $R^2 = \rho_r^2$ of equation (2) is identified. Note that while we identify rank correlations, the individual ranks themselves are not identified.

In Appendix C.3, we discuss how R^2 is identified under multiple inputs and mixtures of binary and continuous inputs.[10] 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.

3.4.4 Validation

We extensively validate the semiparametric latent variable method using (1) simulated data satisfying the identifying distributional assumption, (2) modern survey data (NLSY79)

[9] See Fan et al. (2017) for the proof. We can estimate Δ_X (and Δ_Y) from the binary data as $\widehat{\Delta}_X = \Phi^{-1}(1 - \overline{X}^*)$ where $\overline{X}^* = \sum_{i=1}^n X_i^*/n$.

[10] The method can be further extended to allow for non-binary ordinal and truncated variables (Dey and Zipunnikov, 2022).

where human capital ranks are observed directly via test scores (AFQT and other cognitive tests), and (3) historical census data (1940) where years of education are observed as an approximation of human capital.

First, we show that the method accurately estimates rank correlations from binary proxies of simulated data drawn from a Gaussian copula distribution, even when the cut-off thresholds vary over time. In our context, an important concern stems from literacy increasing over time, changing the information that it contains about a person's human capital rank. To address this concern, we simulate jointly normally distributed data, transform them into ranks, and dichotimize these ranks according to historical literacy rates for each decade from 1870 to 1940. We show that, in contrast to a naive ordinary least squares approach, our semiparametric latent variable method yields correct estimates of mobility (R^2) over time, despite changing cut-offs (see Appendix Figure A.1).

Second, we use the 1979 National Longitudinal Survey of Youth (NLSY79), which provides continuous measures of human capital for both parents and children. We show that our method accurately estimates the key rank mobility parameters presented in sections 4 and 5, even when only binary proxies are available. The exercise strengthens confidence in the methodology's performance in practice. We provide more detail on the data and the specification in the relevant sections.

Third, we show that the latent variable method accurately captures variation in educational mobility across US states using data on years of education from the 1940 census. We first compute rank mobility using parents' and children's years of education. We then create binary proxies of those ranks, using different cutoffs for children, mothers, and fathers (e.g., 11 years for children, 9 for mothers, 7 for fathers). Our method's mobility estimates by state align well with those derived from the original, undichotomized data (see Appendix Figure A.2). This demonstrates the method's performance in relevant historical data.

4. Income Mobility & Parental Human Capital

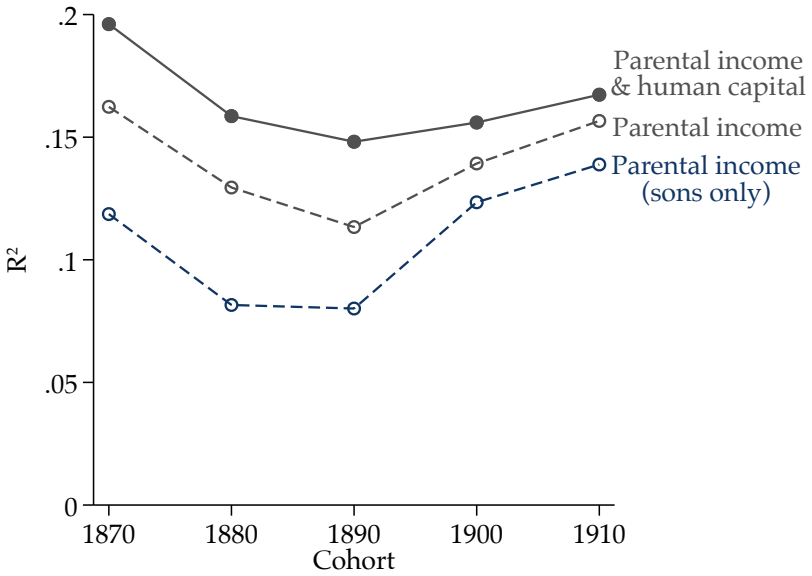
We measure intergenerational mobility as the share of variation in child outcomes that is attributable to parental background. We leverage our new panel that allows us to relate both men's and women's outcomes in adulthood with their parental background measured during childhood. We find that accounting for parental human capital alongside income reveals a trend of rising mobility across US history, challenging earlier findings that considered only income. This shift is largely accounted for by the evolving role of maternal human capital—a finding corroborated by historical literature.

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 incomes (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 parental income and do not take 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.

FIGURE 4: Share of Variation in Income 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) the share accounted for by parents' household income ranks. 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 (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied (see Appendix D.2). Appendix Figure A.3 replicates this figure using "occscore" instead of LIDO.

We account for both parental income and human capital by measuring intergenerational mobility as the R^2 in the following version of equation (1):

$$\text{rank}\left(inc_i\right) = \alpha + \beta_p \text{rank}\left(inc_i^{\text{parents}}\right) + \beta_m \text{rank}\left(h_i^{\text{mother}}\right) + \beta_f \text{rank}\left(h_i^{\text{father}}\right) + \varepsilon_i, \quad (6)$$

where inc is household income and h is (latent) human capital. We measure household income as the household head's LIDO occupational income score. Literacy serves as a binary proxy for latent human capital. Literacy offers a more direct measure of human

capital than, say, years of schooling, which only captures formal educational investment. This distinction is particularly important in the historical context, where limited access to formal schooling meant children often acquired human capital through alternative means, such as maternal home-education. We estimate this model using the semiparametric latent variable method described in section 3.4 and our new representative panel dataset described in section 2.4.[11]

Our estimates suggest that mobility increased throughout US history (as shown by a declining R^2). The magnitude of this increase was around 15 percent from the 1870 to the 1910 cohort. Those estimates differ significantly from previous evidence of declining mobility based solely on fathers' and sons' incomes or occupations (Ferrie, 2005; Long and Ferrie, 2013; Feigenbaum, 2018; Song et al., 2020). While we replicate the finding of decreasing income mobility among sons, we show that including daughters in the analysis flattens the trend in mobility. Further, when we incorporate parental human capital, we observe an increase in mobility over time. We find similar patterns when using occupational income scores that are not specific to sex, race, age, or region ("occscore"; see Appendix Figure A.3).

To understand the drivers of increasing intergenerational mobility, we decompose our mobility measure into multiple components and analyze their individual contributions. Specifically, we decompose R^2 in equation (6) into

$$R^2 = \widehat{\beta}_p^2 + \widehat{\beta}_m^2 + \widehat{\beta}_f^2 + 2\left(\widehat{\beta}_p\widehat{\beta}_m\hat{\rho}_{p,m} + \widehat{\beta}_p\widehat{\beta}_f\hat{\rho}_{p,f} + \widehat{\beta}_m\widehat{\beta}_f\hat{\rho}_{m,f}\right) \tag{7}$$

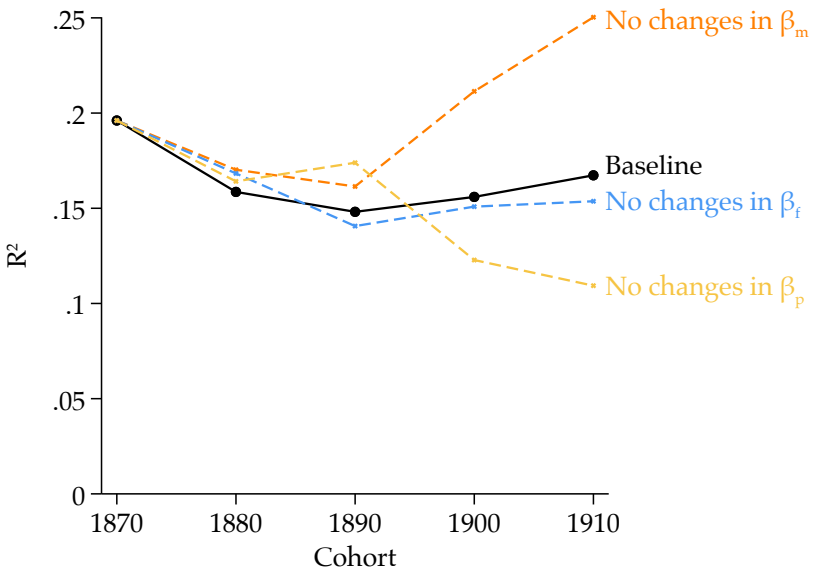
where $\hat{\rho}_{p,m}$, $\hat{\rho}_{p,f}$, and $\hat{\rho}_{m,f}$ are the correlations between parental income and mother's human capital, between parental income and father's human capital, and between mother's and father's human capital.[12] The latter correlation, $\hat{\rho}_{m,f}$, is a measure of assortative mating based on human capital. Using this decomposition, we compute the counterfactual R^2 holding a given parameter constant over time.

Our decomposition shows that the evolving role of maternal human capital ($\widehat{\beta}_m$) is the main reason why intergenerational mobility increased over time (see Figure 5). Specifically, R^2 would have increased without the changing coefficient of maternal human capital. The importance of father's human capital ($\widehat{\beta}_f$) did not affect mobility significantly. Without changes in the predictive power of parental income ($\widehat{\beta}_p$) mobility would have increased even further. The rise in $\widehat{\beta}_p$ aligns with decreasing income mobility in previous research. However, we find that the focus of that research on income alone masked important changes in the role of parental background in shaping the outcomes of children (see also Ward, 2023, who documents that accounting for measurement error also reverses the trend).

[11] Note that this method identifies the parameters in equation (6), but not individual human capital ranks.

[12] For a similar decomposition of R^2 in a rank-rank regression with an arbitrary number of independent variables, see equation (12) in Appendix C.1.2.

FIGURE 5: The Changing Role of Parental Inputs in Intergenerational Mobility



Notes: This figure shows the role of each parameter on the R^2 in equation (6). The baseline represents the observed R^2 shown in Figure 4. The other three lines represent the counterfactual R^2, had the respective parameter not changed over time, computed using the decomposition in equation (7). 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 (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied (see Appendix D.2). Appendix Figure A.3 replicates this figure using "occscore" instead of LIDO.

In contrast to the slope coefficients ($\widehat{\boldsymbol{\beta}}$), none of the correlations between parental inputs ($\widehat{\boldsymbol{\rho}}$)—including assortative mating—had a significant impact on R^2 (see Appendix Figure A.5). For instance, while patters in assortative mating decreased before 1880 and remained constant after (see Appendix Figure A.6), these changes played a negligible role for intergenerational mobility.

Mobility by group. The predictive power of parental background varies across children of different sex and race (see Appendix Figure A.7). Sons generally exhibit higher intergenerational persistence than daughters, with R^2 around twice as high for sons as for daughters. White sons are least mobile, with 10 to 16 percent of variation in household incomes linked to parental background. Black sons are more mobile than white sons; Black and white daughters are the most mobile groups. It is important to recognize that (1) high within-group mobility does not imply high mobility within the general population and that (2) high mobility does not necessarily equate to high *upward* mobility.

Validation. We validate the semiparametric latent variable method to identify equation (6) using data from the NLSY79 Child and Young Adult cohort. The key value of the NLSY79 for our purpose is that it contains the AFQT score of the mother. where a continuous measure of the mother's human capital are observed. Using this data, we compare the estimated R^2 after dichotomizing the AFQT score using a range of rank cutoffs with the R^2 as estimated by OLS on the continuous variables. Appendix Figure A.4 shows the

results: regardless of the position in the distribution where the AFQT score is binarized, the semiparametric latent variable method accurately estimates the "true" R^2.

4.2 The Historical Role of Parental Human Capital

Our finding that parental human capital was important—and especially so in the late 19th century—is consistent with the historical role of parents. Prior to public school access becoming universal 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).

The specific importance of the mothers' human capital to her children's outcomes also aligns with historical evidence. 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, especially among men, leading home production to be increasingly done by women. 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 our finding of the importance of maternal human capital for intergenerational mobility and the historical literature, the subsequent analysis studies the specific role of mothers' human capital in shaping their children's outcomes.

5. Mothers & Human Capital Transmission

Motivated by our results in the previous section, we now zero in on the intergenerational transmission of human capital. We find that, mirroring our results on income mobility, human capital mobility increased significantly from the 1850s to 1910s birth cohorts. We decompose the overall predictive power of parental human capital into the contributions of mothers and fathers. Our findings show that mothers' human capital more strongly predicts child human capital than fathers'. This difference is particularly pronounced for female and Black children.

5.1 Parental Human Capital and Child Outcomes

We estimate human capital mobility (R^2) in the following version of equation (1):

$$\text{rank}\left(h_i\right) = \delta + \gamma_m \text{rank}\left(h_i^{\text{mother}}\right) + \gamma_f \text{rank}\left(h_i^{\text{father}}\right) + \eta_i, \tag{8}$$

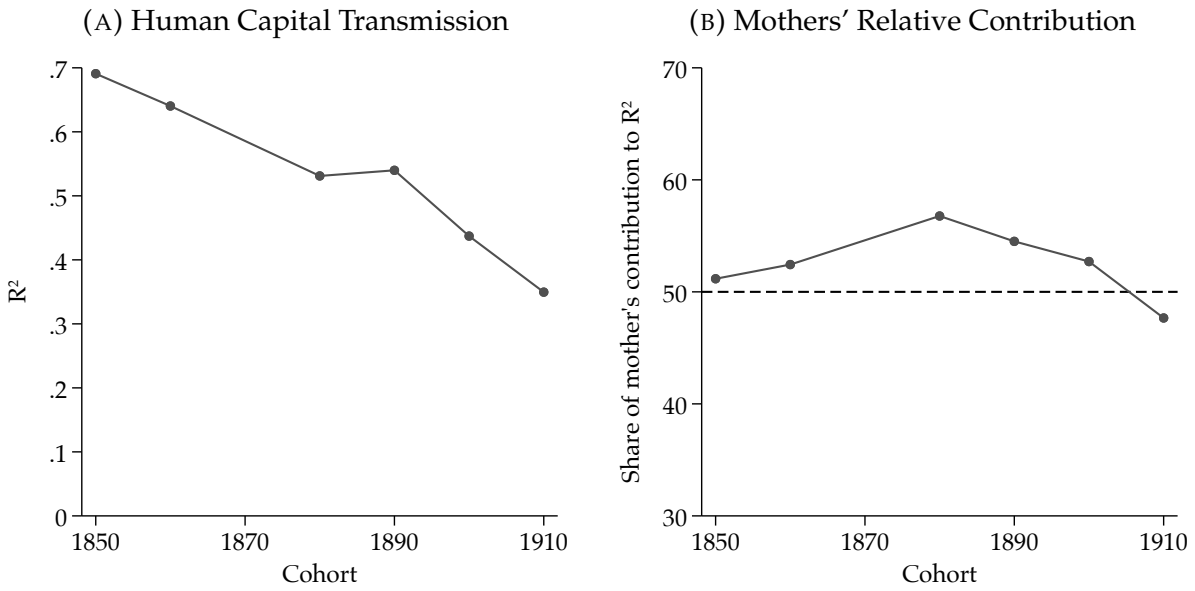
where h is (latent) human capital. We estimate this model using the semiparametric latent variable method described in section 3.4 and use the census cross-section of children in their parents' households. We then use the Shapley-Owen decomposition described in section 3.3 to separate mothers' and fathers' contributions to predicting children's human capital (see Appendix Figure A.8 for an illustration of the method).

Census cross-sections of children who reside with their parents allow us to study intergenerational mobility in certain outcomes without census linking. Specifically, we use such cross-sections to relate parental background to their children's early life outcomes of literacy and school attendance at ages 13–16. Within this age range, the likelihood of a child living apart from their parents is small, minimizing selection into the sample. Our results based on such census cross-sections provide a valuable benchmark for results derived from our new linked census panel. We also replicate those child-based results for adults using our new panel dataset described in section 2.4.

First, our estimates reveal increasing human capital mobility for American children born from the 1850s to the 1910s (see Panel A of Figure 6). While parental background accounted for 70 percent of variation in human capital in the earliest cohort, this figure halved to 35 percent for those born in the latest cohort. The largest increases in human capital mobility took place around the end of slavery (1850–1880) and in the era of rapidly rising school attendance (around 1900).

Second, mothers' human capital was more predictive of child human capital than the fathers' (see Panel B of Figure 6). For cohorts born before 1910, mothers' human capital contributed the majority of the predictive power of child outcomes. Over time, mothers' relative influence on children has diminished and fell below 50 percent for the first time

FIGURE 6: 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. 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.

among children born in the 1910s.

Our findings highlight the role of human capital transmission, especially from mothers, in enhancing income mobility over time. Our analysis in section 4 revealed that the declining predictive power of maternal human capital for their child's income led to increased mobility. We show in this section that the diminished predictive power of maternal human capital for income is accounted for by its reduced predictive power for the child's human capital.

We successfully replicate the cross-sectional patterns of human capital mobility using our new panel (see Appendix Figure A.10). We find that the relative changes in human capital mobility (R^2) match perfectly across both datasets. Similarly, the proportion of human capital transmission attributed to mothers decreases by a similar amount in both datasets. Our panel, while confirming the patterns of *relative changes* over time observed in the cross-section, interestingly shows higher *levels* of human capital mobility. This difference can be explained by two main factors. First, the similarity between parental and child human capital is likely more pronounced in childhood than in adulthood, due to human capital accumulation or depreciation in adult life (intra-generational mobility). Unlike the cross-sectional analysis, our panel includes adult children and accounts for such intra-generational shifts, potentially leading to lower estimates of intergenerational mobility. Second, inaccuracies in automated record linkage might understate the degree of intergenerational persistence through measurement error in parental background.

We further validate that the semiparameric latent variable method accurately estimates human capital rank mobility using the NLSY79 Child and Young Adult Cohort. We estimate the R^2 of a rank regression between the Armed Forces Qualification Test score of the mother and the average percentile score of the child across five cognitive tests they took. We then binarized the scores of the parent and child using various rank cutoffs and estimated the R^2 using our latent variable method. Appendix Figure A.11 shows the "true" R^2—the horizontal dashed line—and the estimates using binary data. Reassuringly, the method captures the rank correlation of human capital well even when only binary indicators are observed. We also run this validation exercise separately for each component of children's cognitive test scores, including reading & verbal, math, and memory. Each of those exercises validates our approach (results available upon request).

5.2 Human Capital Mobility by Group

We estimate equation (8) separately by race and sex and find that human capital mobility varied significantly for Black and white Americans. The human capital rank of Black children born in the earliest cohort (1850s) was highly predictable by their parents' ($R^2 = 0.7$). However, Black children saw a rapid increase in mobility after slavery ended in 1865 ($R^2 = 0.2$ by 1880). After 1880, Black human capital mobility began to decline again. In contrast, white children's human capital mobility remained low and stable until around 1890 ($R^2 = 0.55$) before it sharply increased around 1900—four decades after the increase in Black mobility had started. The 1910s cohort marked the first time since the Civil War that white children's human capital mobility surpassed Black children's ($R^2 = 0.3$).

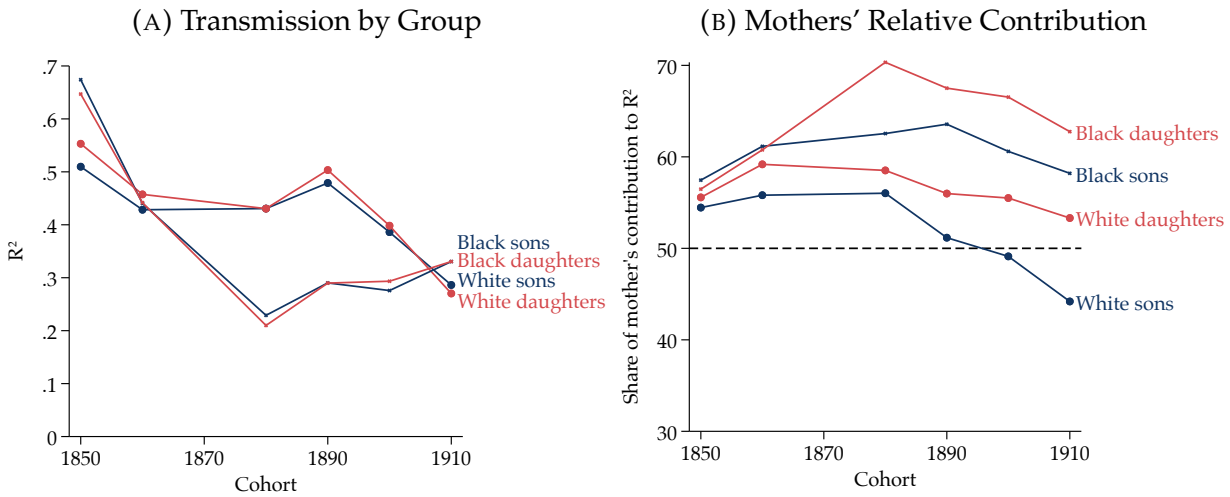
In line with this finding, school access among white children became almost universal in the early 1900s (see Appendix Figure A.12). In contrast, most Black children—especially those whose ancestors were enslaved and largely denied literacy until 1865—lived in the Jim Crow South with restricted school access, shorter school years, and poor school quality (Card and Krueger, 1992; Althoff and Reichardt, 2024). The denial of equal access to high-quality schooling under Jim Crow may explain why human capital mobility among Black Americans decreased starting around 1880.

The finding that mothers' contributions to their children's human capital are generally larger than fathers' is particularly pronounced among female and Black children (see Panel B of Figure 7).[13] Mother's large influence on daughters and Black children aligns with the historical lack of access to educational resources for these groups (Kober and Rentner, 2020). For daughters, it could also suggest the presence of gender-specific role model effects (e.g., Bettinger and Long, 2005; Olivetti et al., 2020).

We also estimate a version of equation (8) where (latent) human capital ranks are re-

[13] Olivetti et al. (2018) find similar gender-specific transmission from paternal and maternal grandparents to their grandsons and granddaughters.

FIGURE 7: Transmission of (Latent) Human Capital Ranks By Group



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.

placed with ranks in formal school attendance completed from the 1940 census. We find that racial differences in educational mobility are larger than those in human capital mobility (see Appendix Figure A.13). This result underscores the fact that the lack of access to formal schooling was even more persistent across generations among Black families than the racial differences in human capital. In contrast, white Americans, who had nearly universal access to schools, were able to substitute parental homeschooling with formal schooling, thereby generating even higher mobility than that observed in human capital.

Lastly, while our analysis so far has focused on two-parent families, we also assess human capital mobility across family types (see Appendix Figure A.14). Single parents have greater predictive power than those in two-parent families, likely due to undivided parental responsibilities. Single fathers' predictive power remains below that of mothers in two-parent families. Working mothers are less predictive of child outcomes than non-working mothers, possibly reflecting differences in time spent with children. Lastly, a larger number of siblings is associated with lower predictive power of mothers, possibly due to weaker human capital transmission when resources are shared across multiple children.

5.3 Mothers' Impact on State-Level Mobility Patterns

To evaluate the significance of including mothers in the analysis of geographic variation in intergenerational mobility, we examine human capital mobility across states. Using our latent variable method, we compare the rank-rank transmission of human capital

from fathers to children against that from both parents to children in census cross-sections from 1870 to 1930, focusing on children aged 13-16 living with their parents.

The incorporation of mothers' influence alters the landscape of intergenerational mobility across the United States, as illustrated in Figure A.9. The most pronounced shifts occur in the South, while the Northeast experiences minimal changes. This pattern aligns with the historical context of limited school access in the South, which amplifies the importance of maternal human capital. Indeed, we observe a strong negative correlation ($\rho = -0.85$) between school access and the increased predictive power gained by including mothers in our analysis.

Our findings highlight the potential for misinterpretation when relying solely on a father-centered approach to assess intergenerational mobility. For instance, considering only fathers' human capital suggests higher mobility in Mississippi compared to Maine. However, when maternal influence is factored in, Mississippi's mobility rate falls below that of almost every state outside the South. This stark contrast emphasizes the crucial role mothers play in shaping children's opportunities to build human capital.

6. The Role of Mothers as Educators

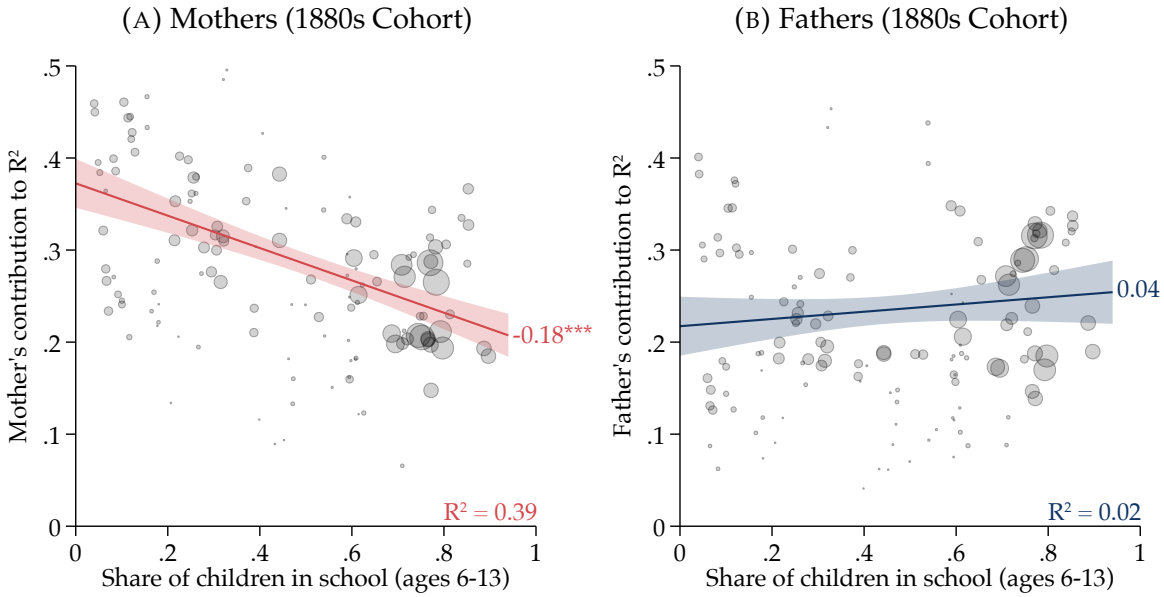
The previous section showed that mothers' human capital is more predictive of their child's human capital than fathers'. This section examines whether mothers' disproportionate importance can be explained by their historical role in home education. We correlate the predictive power of mother's human capital with local school access. Consistent with the role of mothers as home educators, we find that the predictive power of maternal (but not paternal) human capital was substantially greater for groups with limited access to schools.

6.1 Schools and the Rise of Human Capital Mobility

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 in the late 19th century was rapid, it was highly unequal. Specifically, Black children and girls were slower to gain access than white 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, 2024).

Consistent with mothers' importance in home schooling, mothers are more predictive of child outcomes in areas with limited school access (see Figure 8). Maternal human

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.

capital explains almost 40 percent of variation in child human capital when school access is minimal, and around 20 percent when school access is universal. Conversely, fathers' contribution was lower and showed no correlation with school access. In fact, the contributions of mothers and fathers were comparable only when school access was universal.

As school access expanded, it diminished the disparities in human capital mobility previously observed among groups with varying levels of school access (see Panel B of Figure A.15). The reduced influence of parental human capital with improved public school access aligns with Biasi (2023), who shows that equalizing school resources can reduce disparities in intergenerational mobility.

Our analysis reveals an even stronger correlation between school access and human capital mobility when refining our measure of school access to reflect children's daily attendance. By digitizing data on state-specific school ages, enrollment, attendance, and term lengths from the 1880s Census Statistical Abstracts, we calculate the percentage of children aged 6 to 16 attending school on any given day within each state. This refined measure shows that disparities in school access explain nearly 60 percent of the variation in mothers' contributions to human capital transmission (see Appendix Table B.1). Conversely, we observe no correlation between fathers' contributions and school access.

To provide further evidence on schools' role in shaping human capital mobility, we

leverage the staggered implementation of compulsory schooling laws across US states post-1913 (Acemoglu and Angrist, 2000; Goldin and Katz, 2008; Stephens and Yang, 2014). We instrument a state's share of children in school (by sex and race) with the number of years a child was exposed to compulsory schooling (see Appendix Table B.2). A strong first stage ($F = 43.9$) confirms that compulsory schooling laws significantly increased school attendance. Our IV estimates reveal a substantial rise in human capital mobility following the introduction of these laws. We interpret this as evidence of a fundamental shift in the primary source of human capital from parents to formal schooling, which was instrumental in boosting human capital mobility.

In sum, our results suggest that broadening school access in the late 19th and early 20th century contributed to increasing intergenerational mobility. The increase in mobility was driven by a declining role of maternal human capital as schools substituted for home-education. The critical role of schools in increasing intergenerational mobility is consistent with Card et al. (2022) who show that state-level school quality are correlated with higher educational upward mobility in the 1940 census, and with more modern work on the role of education in intergenerational mobility (Chetty et al., 2020; Barrios Fernández et al., 2021; Zheng and Graham, 2022; Black et al., 2023).

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. Our results also illuminate why maternal human capital played such a crucial role in early US history: functioning as primary educators, mothers significantly shaped their children's human capital development, which in turn influenced their overall economic outcomes.

While school access increases intergenerational mobility, parents—especially mothers—likely remained crucial in shaping child human capital even as schooling expanded. Heckman (2000, 2006); Cunha and Heckman (2007); Becker (2009); Cunha et al. (2010) emphasize that early childhood environments significantly influence returns to later education. This dynamic complementarity between early parental inputs and later formal education underscores the lasting importance of family background in determining children's long-term outcomes.

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 or consider the role of other relatives including grandparents. Given the importance of the location in which a person grows up—as documented in previous work (e.g., Chetty et al., 2016; Chetty and Hendren, 2018)—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; Olivetti, 2014) and single-motherhood (Althoff, 2023).

Lastly, 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 women.

REFERENCES

ABRAHAMSSON, S., A. BÜTIKOFER, K. V. LØKEN, AND M. PAGE (2024): "Sources of Generational Persistence in Education and Income," Working paper.

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, H. BROOKES GRAY, K. ERIKSSON, S. PÉREZ, AND M. RASHID (2024): "Finding John Smith: Using Extra Information for Historical Record Linkage," Working paper.

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.

ACEMOGLU, D. AND J. ANGRIST (2000): "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," *NBER Macroeconomics Annual*, 15, 9–59.

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

ALTHOFF, L. AND H. REICHARDT (2024): "Jim Crow and Black Economic Progress After Slavery," *Quarterly Journal of Economics*.

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`.

BARRIOS FERNÁNDEZ, A., C. NEILSON, AND S. D. ZIMMERMAN (2021): "Elite universities and the intergenerational transmission of human and social capital," *Available at SSRN 4071712*.

BECKER, G. S. (2009): *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, University of Chicago Press.

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.

BIASI, B. AND S. MA (2023): "The Education-Innovation Gap," Working Paper 29853, National Bureau of Economic Research.

BLACK, S. E., J. T. DENNING, AND J. ROTHSTEIN (2023): "Winners and Losers? The Effect of Gaining and Losing Access to Selective Colleges on Education and Labor Market Outcomes," *American Economic Journal: Applied Economics*, 15, 26–67.

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," *The 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., J. N. FRIEDMAN, E. SAEZ, N. TURNER, AND D. YAGAN (2020): "Income segregation and intergenerational mobility across colleges in the United States," *The Quarterly Journal of Economics*, 135, 1567–1633.

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, AND L. F. KATZ (2016): "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment," *American Economic Review*, 106, 855–902.

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.

CUNHA, F. AND J. HECKMAN (2007): "The technology of skill formation," *American economic review*, 97, 31–47.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883–931.

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.

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.

EVANS, D. K. AND P. JAKIELA (2024): "The Role of Fathers in Promoting Early Childhood Development in Low- and Middle-Income Countries: A Review of the Evidence," *The World Bank Research Observer*.

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.

FEIGENBAUM, J. J. (2018): "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940," *The Economic Journal*, 128, F446–F481.

FERNÁNDEZ, R. (2013): "Cultural change as learning: The evolution of female labor force participation over a century," *American Economic Review*, 103, 472–500.

FERNÁNDEZ, R., A. FOGLI, AND C. OLIVETTI (2004): "Mothers and Sons: Preference Formation and Female Labor Force Dynamics*," *The Quarterly Journal of Economics*, 119, 1249–1299.

FERRIE, J. P. (2005): "History lessons: The end of American exceptionalism? Mobility in the United States since 1850," *Journal of Economic Perspectives*, 19, 199–215.

FOGLI, A. AND L. VELDKAMP (2011): "Nature or nurture? Learning and the geography of female labor force participation," *Econometrica*, 79, 1103–1138.

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.

GOLDIN, C. AND L. F. KATZ (2008): "Mass Secondary Schooling and the State: The Role of State Compulsion in the High School Movement," in *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy*, ed. by D. L. Costa and N. R. Lamoreaux, University of Chicago Press, 275–310.

GREENWOOD, J., A. SESHADRI, AND M. YORUKOGLU (2005): "Engines of Liberation," *The Review of Economic Studies*, 72, 109–133.

HECKMAN, J. J. (2000): "Policies to Foster Human Capital," *Research in Economics*, 54, 3–56.

——— (2006): "Skill Formation and the Economics of Investing in Disadvantaged Children," *Science*, 312, 1900–1902.

HELGERTZ, J., S. RUGGLES, J. R. WARREN, C. A. FITCH, J. D. HACKER, M. A. NELSON, J. P. PRICE, E. ROBERTS, AND M. SOBEK (2023): "IPUMS Multigenerational Longitudinal Panel: Version 1.1 [dataset]," .

HOLMLUND, H., M. LINDAHL, AND E. PLUG (2011): "The causal effect of parents' schooling on children's schooling: A comparison of estimation methods," *Journal of Economic Literature*, 49, 615–651.

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.

LONG, J. AND J. FERRIE (2013): "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, 103, 1109–1137.

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

LUNDBORG, P., A. NILSSON, AND D.-O. ROOTH (2014): "Parental education and offspring outcomes: evidence from the Swedish compulsory School Reform," *American Economic Journal: Applied Economics*, 6, 253–278.

LUNDBORG, P., E. PLUG, AND A. W. RASMUSSEN (2024): "On the Family Origins of Human Capital Formation: Evidence from Donor Children," *Review of Economic Studies*, forthcoming.

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

MODALSLI, J., C. OLIVETTI, M. D. PASERMAN, AND L. SALISBURY (2024): "Female Labor Force Participation and Intergenerational Mobility," Working paper.

NGAI, L. R., C. OLIVETTI, AND B. PETRONGOLO (2024): "Gendered Change: 150 Years of Transformation in US Hours," Working Paper 32475, National Bureau of Economic Research.

OLIVETTI, C. (2006): "Changes in Women's Hours of Market Work: The Role of Returns to Experience," *Review of Economic Dynamics*, 9, 557–587.

——— (2014): *The Female Labor Force and Long-Run Development: The American Experience in Comparative Perspective*, University of Chicago Press, 161–197.

OLIVETTI, C. AND M. D. PASERMAN (2015): "In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850–1940," *American Economic Review*, 105, 2695–2724.

OLIVETTI, C., M. D. PASERMAN, AND L. SALISBURY (2018): "Three-generation mobility in the United States, 1850–1940: The role of maternal and paternal grandparents," *Explorations in Economic History*, 70, 73–90.

OLIVETTI, C., E. PATACCHINI, AND Y. ZENOU (2020): "Mothers, Peers, and Gender-Role Identity," *Journal of the European Economic Association*, 18, 266–301.

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.

RAMEY, V. A. (2009): "Time Spent in Home Production in the Twentieth-Century United States: New Estimates from Old Data," *The Journal of Economic History*, 69, 1–47.

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`.

SAAVEDRA, M. AND T. TWINAM (2020): "A machine learning approach to improving occupational income scores," *Explorations in Economic History*, 75, 101304.

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.

SONG, X., C. G. MASSEY, K. A. ROLF, J. P. FERRIE, J. L. ROTHBAUM, AND Y. XIE (2020): "Long-term decline in intergenerational mobility in the United States since the 1850s," *PNAS*, 117, 251–258.

STEPHENS, M. J. AND D.-Y. YANG (2014): "Compulsory Education and the Benefits of Schooling," *The American Economic Review*, 104, 1777–1792.

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.

ZHENG, A. AND J. GRAHAM (2022): "Public education inequality and intergenerational mobility," *American Economic Journal: Macroeconomics*, 14, 250–282.

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 **35**

B Appendix Tables **44**

C Methods Appendix **45**

C.1 Relation Between R^2 and Coefficients 45

C.2 Shapley-Owen Decomposition of the R^2 46

C.3 Semiparametric latent variable method 47

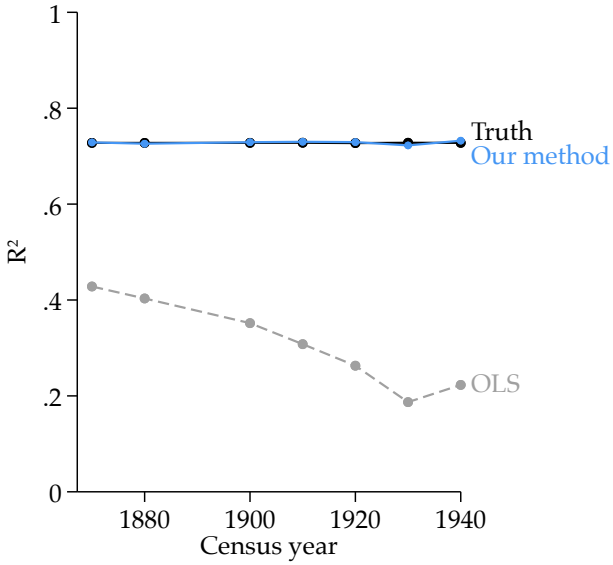
D Data Appendix **50**

D.1 Linking Procedure 52

D.2 Sample Weight Construction 55

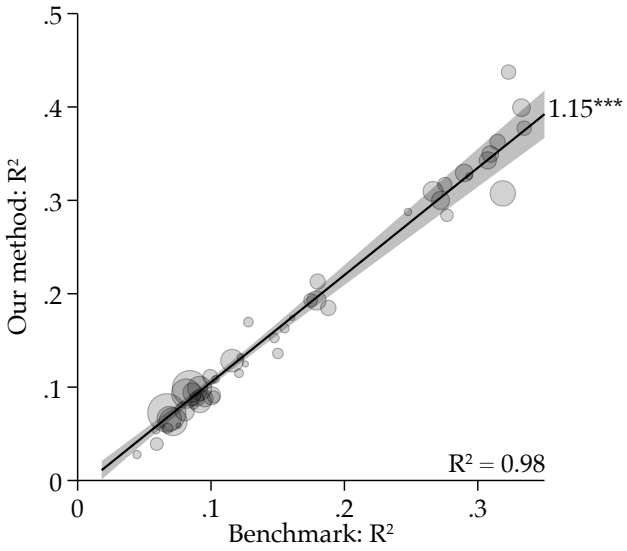
A. Appendix Figures

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



Notes: This figure demonstrates the effectiveness of our semiparametric latent variable method in identifying rank-rank relationships from binary proxies of continuous variables. We simulate jointly normal random variables, convert them into dummies such that the binary distribution reflects historical literacy rates, and compare the R^2 on the continuous rank rank regression with the estimated R^2 based on the dichotimized data. The "Truth" line represents the R^2 from a continuous 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. In the 1940 census, instead of literacy, we observe the highest year of school or degree completed. We classify individuals who have completed at least two grades of school as literate; others we classify as illiterate.

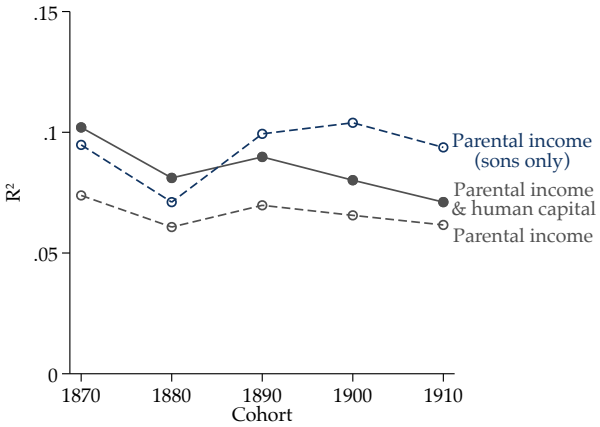
FIGURE A.2: Validation of the Semiparametric Latent Variable Method on Census Data



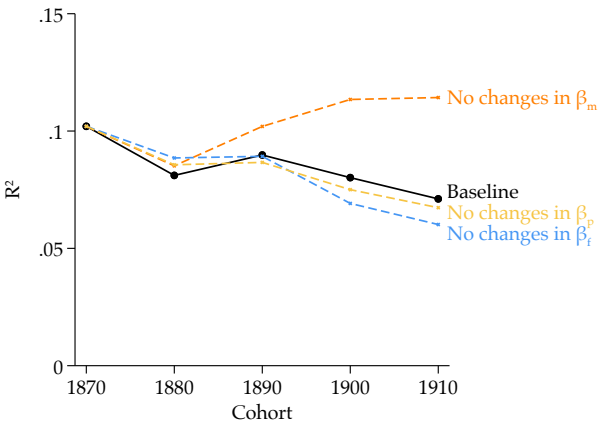
Notes: This figure contrasts the R^2 values from rank-rank regressions using actual and binarized educational data from the 1940 census. We binarize the data by arbitrarily categorizing individuals based on their educational attainment: more than 11 years for children, 9 for mothers, and 7 for fathers. Each dot represents a US state, weighted by sample size and focusing on children aged 13—21 living with parents.

FIGURE A.3: Estimates Based on "occscores"

(A) Share of Variation in Income Explained

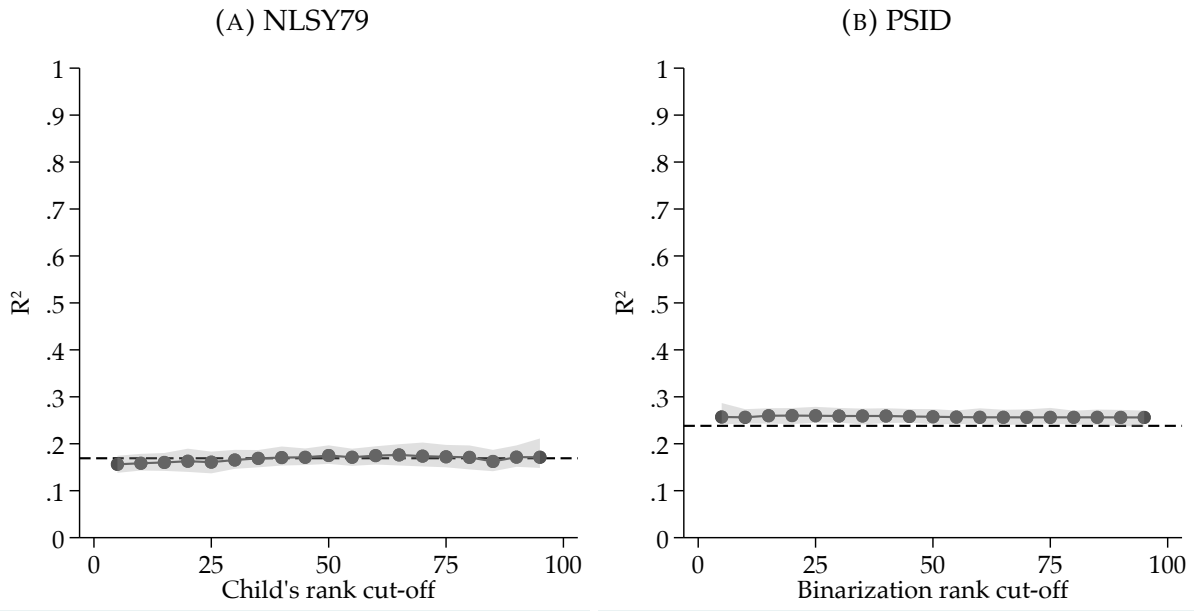


(B) Changing role of Parental Inputs



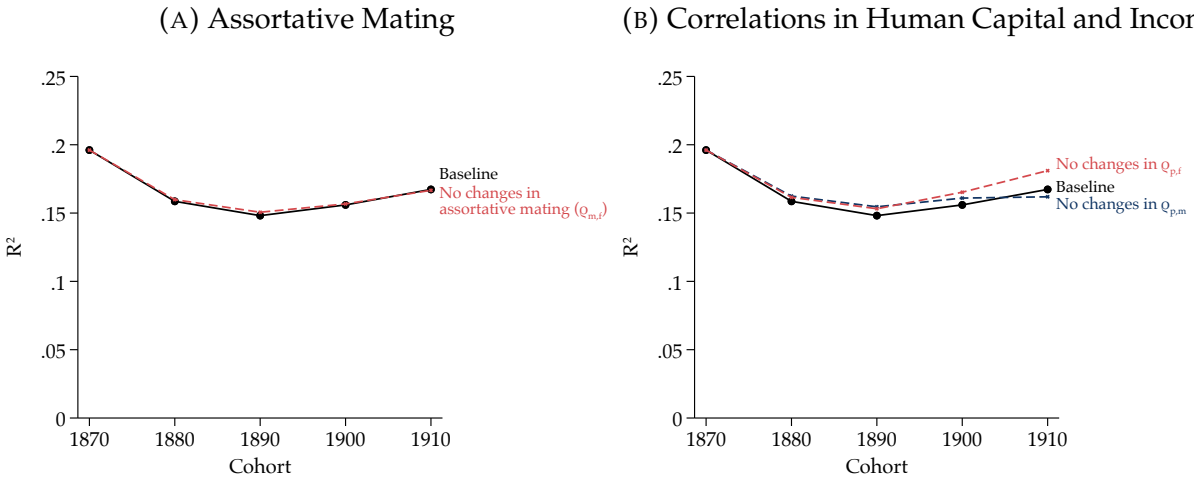
Notes: This figure replicates Figures 4 and 5 using "occscore" instead of LIDO. It shows the share of the variance in a child's household income rank explained by (1) parents' household income ranks and their (latent) human capital ranks (R^2) and (2) the share accounted for by parents' household income ranks. 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 "occscore" occupational income score. Results are based on our new panel and sample weights are applied (see Appendix D.2).

FIGURE A.4: Validation of Income Mobility Estimation using Cognitive Test Scores



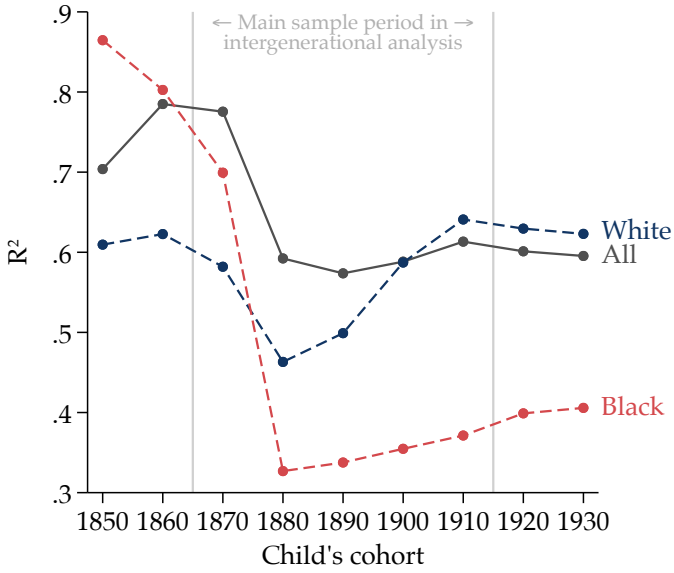
Notes: This figure shows the accuracy of the semiparametric latent variable method in estimating equation (6) in the NLSY79 (Panel A) and the PSID (Panel B). The dashed lines represents the estimated R^2 of a regression of income of the child on family income of the parents and a cognitive test score of the mother. The solid lines represent the estimated R^2 after binarization of the mother's cognitive test score, using varying cutoffs. Panel A uses the NLSY Child and Young Adult Cohort. The cognitive test score of the mother is the Armed Forces Qualification Test (AFQT). Panel B uses the PSID Child Development Supplement 1997. The cognitive test score of the mother is the passsage comprehension test. Shaded area are 95% bootstrapped confidence intervals.

FIGURE A.5: Mobility and the Impact of Evolving Parental Input Correlations



Notes: This figure shows the role of each parameter on the R^2 in equation (6). The baseline represents the observed R^2 shown in Figure 4. The other three lines represent the counterfactual R^2, had the respective parameter not changed over time, computed using the decomposition in equation (7). 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 (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied (see Appendix D.2).

FIGURE A.6: Assortative Mating Estimates by Group

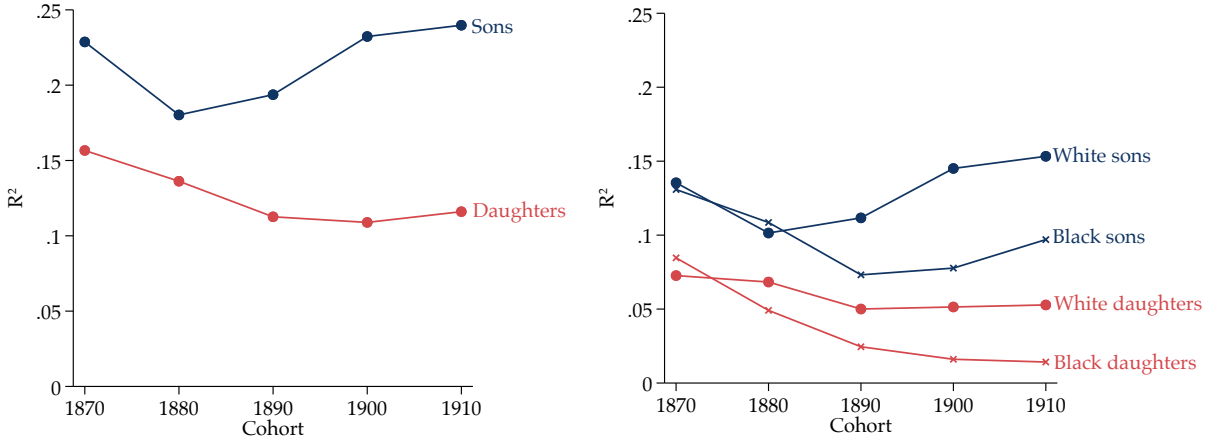


Notes: This Figure shows the share of the variance in a person's (latent) human capital rank explained by their spouse's (latent) human capital rank (R^2) across their child's cohort. For human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. Results are based on the full census cross-section of two-parent households with children aged 1 to 16. Note that as we show in Appendix C.1, in this univariate rank-rank model, $R^2 = \beta^2 = \rho^2_{x,y}$, allowing researchers to directly compare our estimates of assortative mating to (the square of) conventional rank-rank correlations.

FIGURE A.7: Within-Group Mobility Estimates

(A) Mobility by Sex

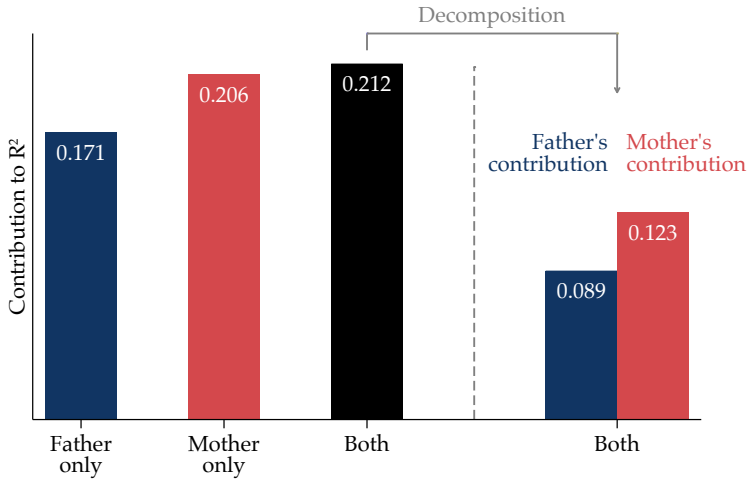
(B) Mobility by Sex & Race



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 (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied (see Appendix D.2).

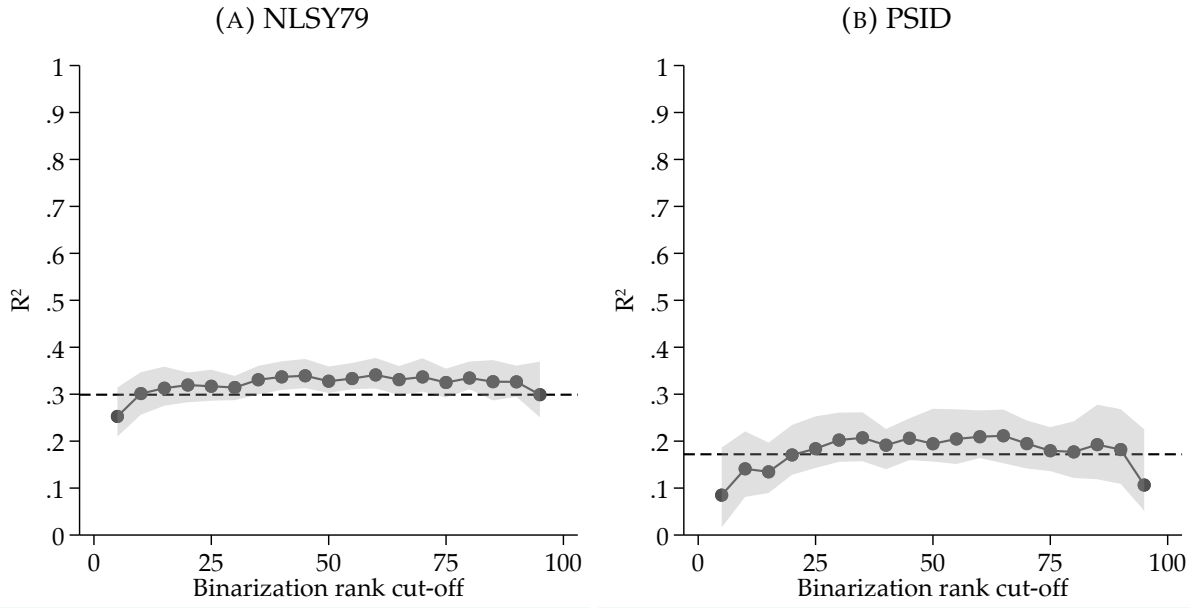
FIGURE A.8: Illustrating our Decomposition Method

Intergenerational Transmission of Human Capital



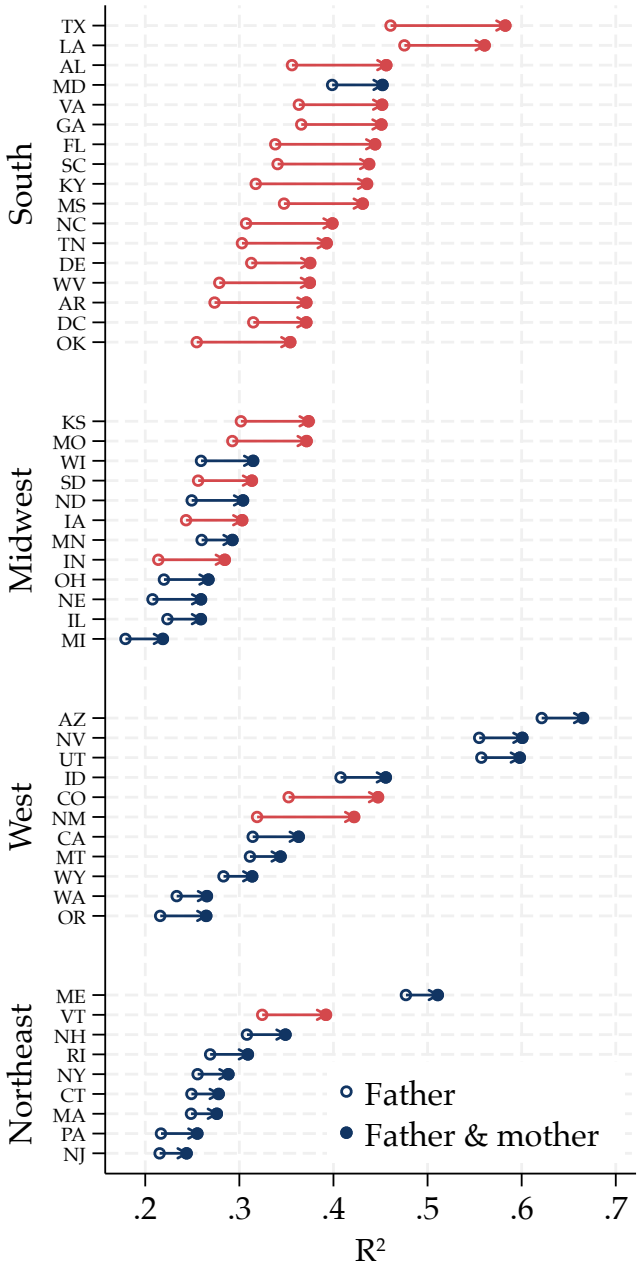
Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. We decompose the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

FIGURE A.11: Validation of Human Capital Mobility using Cognitive Test Scores



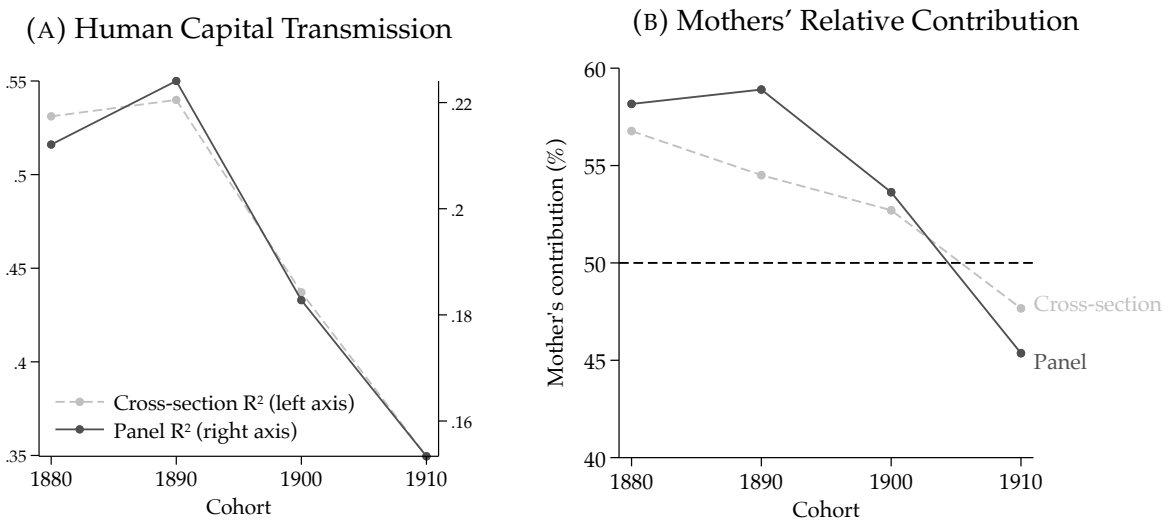
Notes: This figure shows the accuracy of the semiparametric latent variable method in estimating equation (8) in the NLSY79 (Panel A) and the PSID (Panel B). The dashed lines represent the estimated R^2 on the observed continuous cognitive test measures. The solid lines represent the estimated R^2 after binarization of the mother's and child's score, using varying cutoffs for the child and the median for the mother's cutoff. Shaded area are 95% bootstrapped confidence intervals. Panel A uses the NLSY Child and Young Adult Cohort. The cognitive test score of the mother is the Armed Forces Qualification Test (AFQT). For the children, we use the average percentile score across the five cognitive tests: reading recognition, reading comprehension, math, vocabulary, and memory. Panel B uses the PSID Child Development Supplement 1997. The cognitive test score of the mother is the passsage comprehension test. For the children, we use the average across three cognitive tests: letter word identification, applied problems, and broad math.

FIGURE A.9: Human Capital Mobility by State Before vs. After Incorporating Mothers



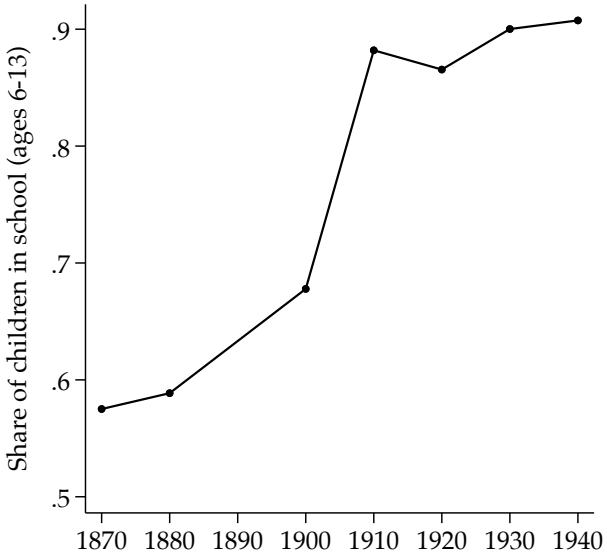
Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by their (1) father's or (2) father's and mother's (latent) human capital rank (R^2) across states. States with above-median changes in R^2 are displayed in red; states with below-median changes are displayed in blue. Each estimate is the average R^2 across the census cross-sections from 1870 to 1930 of children aged 13–16 in their parents' household. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4.

FIGURE A.10: Panel-Based Estimates of Human Capital Mobility Across Cohorts



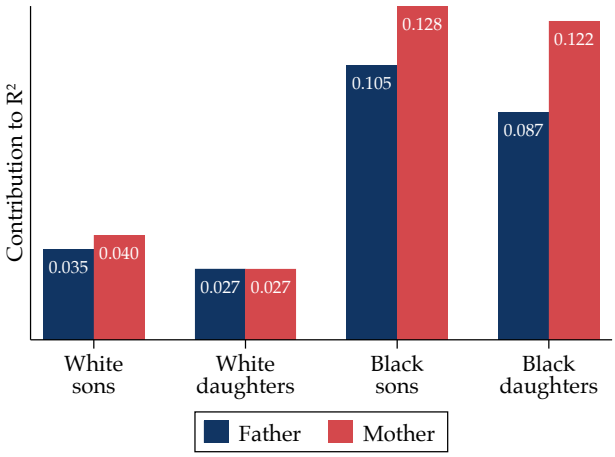
Notes: This figure compares our baseline results of human capital transmission from the cross-section of children who live with their parents to estimates based on our new panel. Panel A shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2) across cohorts. 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. Cross-sectional results are based on the census cross-section of children ages 13–16 in their parents' household; panel results are based on individuals of any age.

FIGURE A.12: Increasing Access to Schools



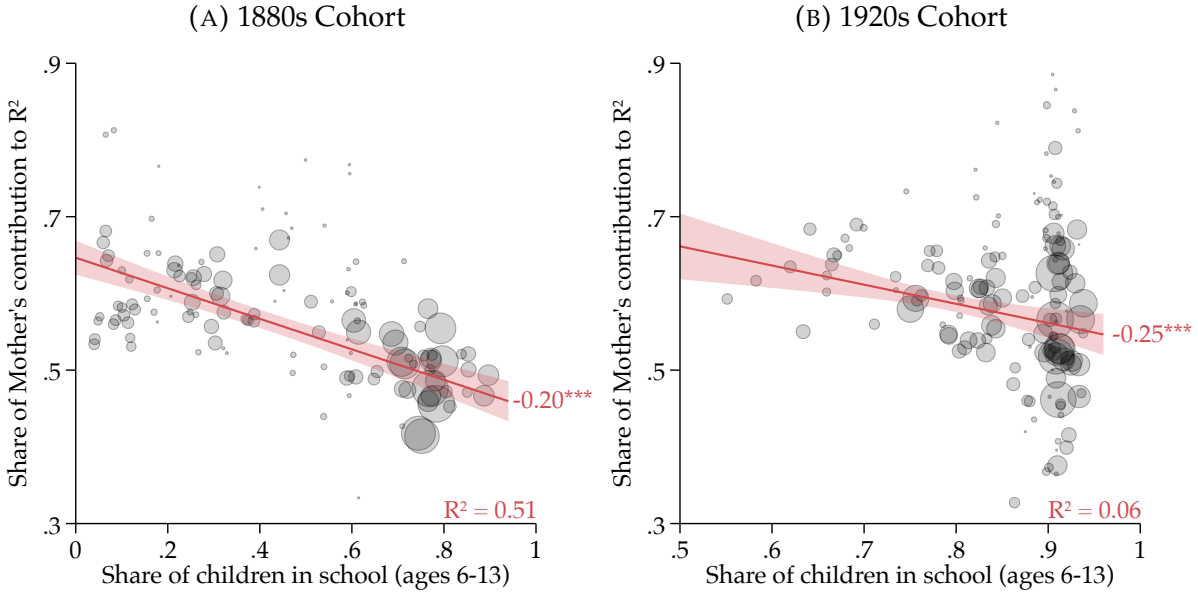
Notes: This figure shows the share of children aged 6–13 who attend school across time.

FIGURE A.13: Intergenerational Transmission of Formal Schooling (1920s Cohort)



Notes: This figure shows the share of the variance in a child's years of education rank explained by parents' years of education ranks (R^2). The figure focuses on the 1920s cohort (children aged 13–16 in the 1940 census—the only historical census that records years of education). We decompose the overall R^2 using the Shapley-Owen method to quantify each parent's contribution. Results are based on the census cross-section of children in their parents' household.

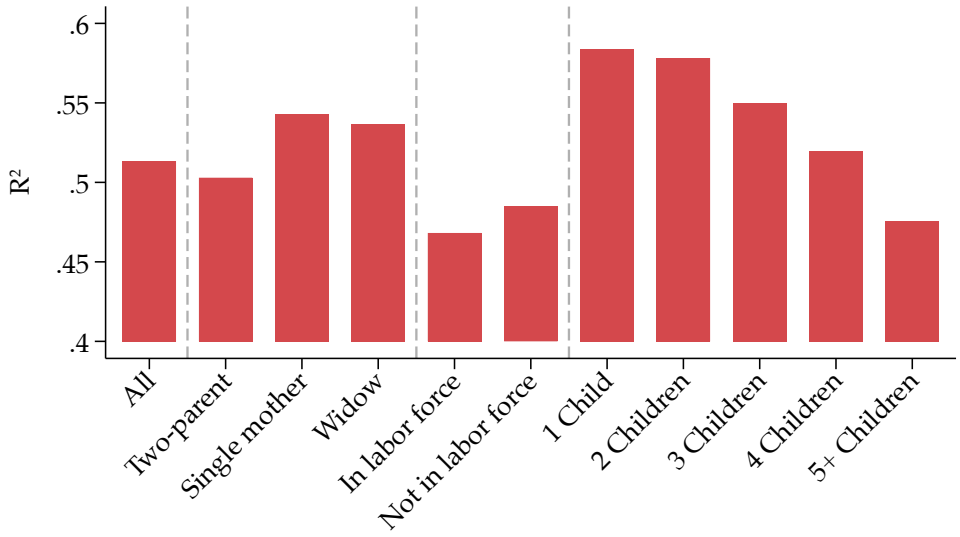
FIGURE A.15: Mothers' Human Capital as Substitute for Local Schools



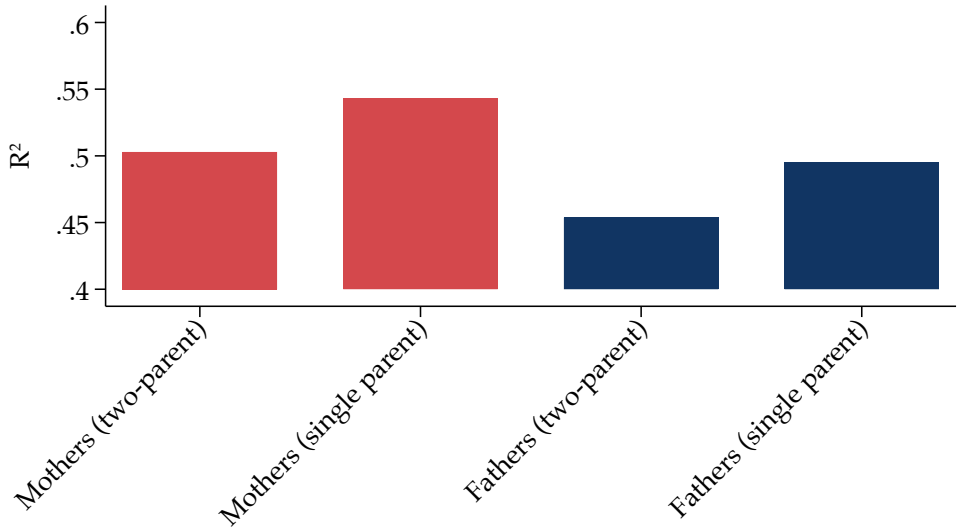
Notes: This figure shows the relationship between local school access and mothers' *relative* contributions to child human capital (as a share of total variation explained). Literacy is used as the measure for rank-based transmission of human capital (section 3.4). Each dot represents a group of children 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.14: Human Capital Mobility by Family Type

(A) Predictive Power of Maternal Human Capital



(B) Maternal vs. Paternal Human Capital



Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by mothers' or fathers' (latent) human capital rank (R^2) across family types. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. Panel A shows mothers' predictive power across family types; Panel B repeats two of those estimates and compares them to the equivalent for fathers. Results are based on the census cross-section of children ages 13–16 in their parents' household.

B. Appendix Tables

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

	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$
Baseline measure of school access	-0.18***	0.04	-0.20***			
	(0.03)	(0.05)	(0.03)			
Refined measure of school access				-0.47***	0.15	-0.58***
(accounts for attendance, term lengths, etc.)				(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.15. 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 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE B.2: Mothers & Schools—Impact of Mandatory Schooling Laws

	Outcome: ϕ_{Mother}			
	OLS	IV	OLS	IV
IV: Schooling via compulsory schooling laws	-0.23***	-0.92***	-0.73***	-0.92***
	(0.04)	(0.22)	(0.18)	(0.22)
Cohort Fixed Effects	Y	Y	Y	Y
Sample restricted to 1920–1940	N	N	Y	Y
F-statistic	–	35.52	–	35.39
R^2	0.47	–	0.38	–
Observations	1,049	1,049	465	465

Notes: This table presents OLS and instrumental variable (IV) estimates of the relationship between school attendance and mother's contribution to child human capital. The outcome variable is mother's contribution to R^2. In columns 2 and 4, school access is instrumented by years of exposure to compulsory schooling laws. Columns 3 and 4 present estimates for a restricted sample (1920-1940) to ensure results are not driven by zeros for the instrument before the first laws are recorded in the 1910s. Standard errors are in parentheses and are clustered at the state-cohort level. All specifications include cohort fixed effects. The F-statistic reported for the 2SLS estimations is the Kleibergen-Paap Wald F-statistic. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C. Methods Appendix

C.1 Relation Between R^2 and Coefficients

C.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:

$$\widehat{\beta} = \text{cor}(X, Y) \cdot \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}} \tag{9}$$

$$R^2 = \text{cor}(X, Y)^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}(X)}{\text{Var}(Y)}, \tag{10}$$

where $\text{cor}(X, Y)$ is the correlation between Y and X and $\text{Var}(Y)$ is the variance of Y_i.

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 $b > 0$.

C.1.2 Multiple inputs

In a multivariate linear regression, $Y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \varepsilon_i$, the R^2 depends on the parameters $\beta_1, \ldots, \beta_k$ and the variance-covariance matrix of the explanatory variables. That is,

$$R^2 = \frac{Var\left(\sum_{j=1}^{k} \widehat{\beta}_j X_{i,j}\right)}{Var(Y)} = \frac{\sum_{j=1}^{k} \widehat{\beta}_j^2 Var(X_j) + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} \widehat{\beta}_j \widehat{\beta}_l Cov\left(X_j, X_l\right)}{Var(Y)}. \tag{11}$$

Rank-rank coefficients. Again, using that quantile-ranked outcomes share the same distribution by construction—i.e., $\text{Var}(Y) = \text{Var}(X_j) \;\; \forall j = 1, \ldots, k$—we obtain

$$R^2 = \sum_{j=1}^{k} \widehat{\beta}_j^2 + 2\sum_{j=1}^{k-1}\sum_{j=i+1}^{k} \widehat{\beta}_j \widehat{\beta}_l \widehat{\rho}_{j,l} \tag{12}$$

where $\widehat{\rho}_{j,l}$ is the correlation between X_j and X_l.

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

$$\phi_j = \frac{\Delta_j}{k!}, \tag{13}$$

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.

C.2.1 Example with two inputs

Table C.3 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 C.3: 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)		$\frac{0.08+0.1}{2!} = 0.09$	$\frac{0.17+0.15}{2!} = 0.16$

C.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 (13), 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 (11), 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

$$\phi_1 = \frac{1}{2Var(Y)}\left(2\widehat{\beta}_1^2 Var(X_1) + \{Cov(X_1, X_2)\}^2\left[\frac{\widehat{\beta}_2^2}{Var(X_1)} - \frac{\widehat{\beta}_1^2}{Var(X_2)}\right] + 2\widehat{\beta}_1\widehat{\beta}_2 Cov(X_1, X_2)\right).$$

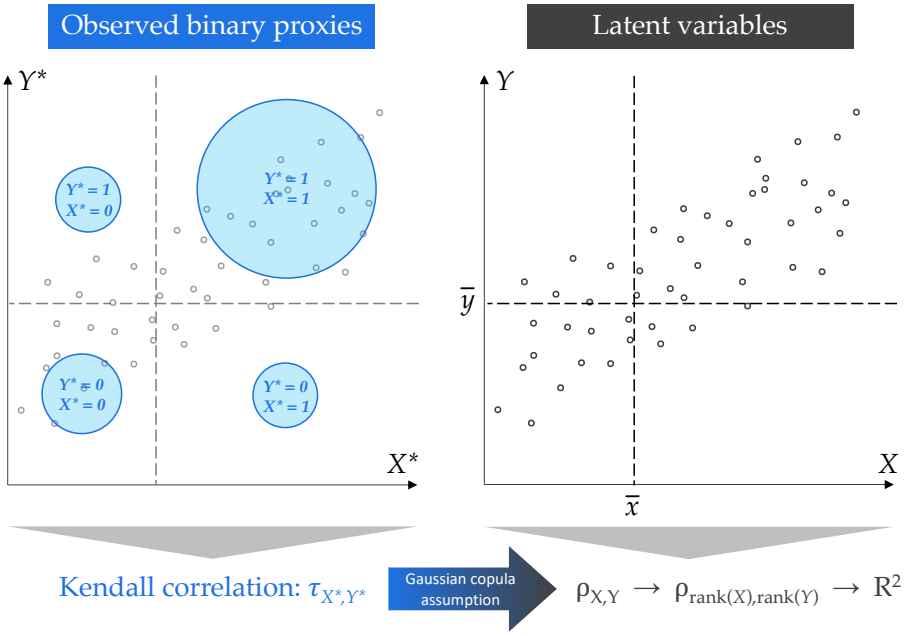
For rank-rank regressions, we have

$$\begin{aligned}\phi_1 &= \widehat{\beta}_1^2 + \frac{1}{2}\left(\widehat{\beta}_2^2 - \widehat{\beta}_1^2\right)\left(\frac{Cov(X_1, X_2)}{Var(Y)}\right)^2 + \widehat{\beta}_1\widehat{\beta}_2\frac{Cov(X_1, X_2)}{Var(Y)} \\ &= \widehat{\beta}_1^2 + \frac{\widehat{\rho}_{1,2}^2}{2}\left(\widehat{\beta}_2^2 - \widehat{\beta}_1^2\right) + \widehat{\beta}_1\widehat{\beta}_2\widehat{\rho}_{1,2}.\end{aligned}$$

C.3 Semiparametric latent variable method

We use the semiparametric latent variable method introduced by Fan et al. (2017) to estimate rank-rank mobility (R^2) when only binary proxies of the underlying rank variable are observed. The rank-rank regression of interest is that in equation (1).

FIGURE C.1: Illustrating the semiparametric Latent Variable Method



Notes: This figure illustrates the semiparametric latent variable method, recovering rank-rank mobility (R^2) in latent variables from observed binary proxies. Assuming that the underlying latent variables are drawn from a joint Gaussian copula distribution, pairwise rank-rank correlations can be identified from Kendall's correlation between the observed binary proxies using the bridging function in (16). Rank-rank regressions can be identified from the pairwise correlation matrix using equations and (17) and (18).

We assume that the dependent and independent variables are drawn from a joint Gaussian copula distribution. That is, we assume that there exists a set of unknown monotonic transformations f_y, $f_1, \cdot\cdot, f_k$ such that $f_Y(Y_i), f_1(X_{1i}), f_k(X_{ki}) \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$.

Fan et al. (2017) show how to estimate all elements of $\mathbf{\Sigma}$ even if only binary proxies of the rank variables of interest are available. For example, let us consider Σ_{12}, the correlation between $f_Y(Y_i)$ and $f_1(X_{1i})$. We summarize the more formal arguments by Fan et al. (2017). Three cases are considered. First, that both Y_i and X_{1i} are observed. Second, that Y_i is observed, but only a binary proxy of X_{1i} is observed. That is, we observe only X^*_{1i} which is one if X_{1i} is above an arbitrary cut-off and zero otherwise. Third, that only observe binary proxies of each variable are observed.

Case 1: Both rank variables observed. Fan et al. (2017) show that Σ_{12} is an increasing function of the Kendall's rank correlation coefficient τ_{12}. Therefore, observing the ranked variables is sufficient to identify Σ_{12}. Specifically, the "bridging function" between Kendall's rank correlation coefficient and Σ_{12} is

$$\Sigma_{12} = \sin\left(\frac{\pi}{2}\tau_{12}\right). \tag{14}$$

Therefore, our estimate $\hat{\Sigma}_{12}$ is the sample equivalent of equation (14).

Case 2: One rank variable and one binary proxy observed. In this case, we observe $\text{rank}(Y_i)$ but we only observe the binary proxy X^*_{1i}. In such cases, Fan et al. (2017) show that

$$\tau_{12} = 4\Phi_2\left(\Delta_2, 0, \frac{\Sigma_{12}}{\sqrt{2}}\right) - 2\Phi\left(\Delta_2\right) \tag{15}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, $\Phi_2(u, v, t)$ is the CDF of a bivariate normal distribution with correlation coefficient t, evaluated at u and v. Δ_2 is the cut-off value above which the binary proxy is 1 and can be estimated as $\hat{\Delta}_2 = \Phi^{-1}\left(1 - \overline{X}^*_1\right)$ where $\overline{X}^*_1 \equiv \frac{1}{n}\sum_{i=1}^n X^*_{1i}$. Because equation (15) is strictly increasing in Σ_{12} (see (Fan et al., 2017) for the proof), Σ_{12} is identified as the unique root of equation (15) where τ_{12} and Δ_2 are replaced with their finite sample analogues.

Case 3: Only binary proxies observed. For two binary proxies, the bridging function is

$$\tau_{12} = 2\Phi_2\left(\Delta_1, \Delta_2, \Sigma_{12}\right) - 2\Phi\left(\Delta_1\right)\Phi\left(\Delta_2\right). \tag{16}$$

The right hand side of this equation is increasing in Σ_{12}. Since Δ_1, Δ_2, and τ_{12} can be estimated, Σ_{12} is identified as the unique root of equation (16) where τ_{12}, Δ_1, and Δ_2 are replaced with their finite sample analogues.

The last step of the method is to estimate the parameters and R^2 of equation (1) from the pairwise correlations between the underlying random variables that are jointly normal. First, given two jointly normal random variables with correlation ρ, the correlation of their ranks (Spearman's rank correlation ρ_s) is equal to $\rho_s = \frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)$. Let $\widehat{\boldsymbol{R}}$ be the rank-rank correlation matrix, i.e. $\widehat{\boldsymbol{R}}_{jl} = \frac{6}{\pi}\sin^{-1}\left(\frac{\hat{\Sigma}_{jl}}{2}\right)$ for each $l, j = 1, \ldots, k+1$. We use that the coefficients and R^2 in rank-rank regressions are identified from the rank-rank correlation matrix (again using that the marginal distributions of all ranked variables are equal). Specifically,

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{R}}_x\right)^{-1}\widehat{\boldsymbol{R}}_{xy} \tag{17}$$

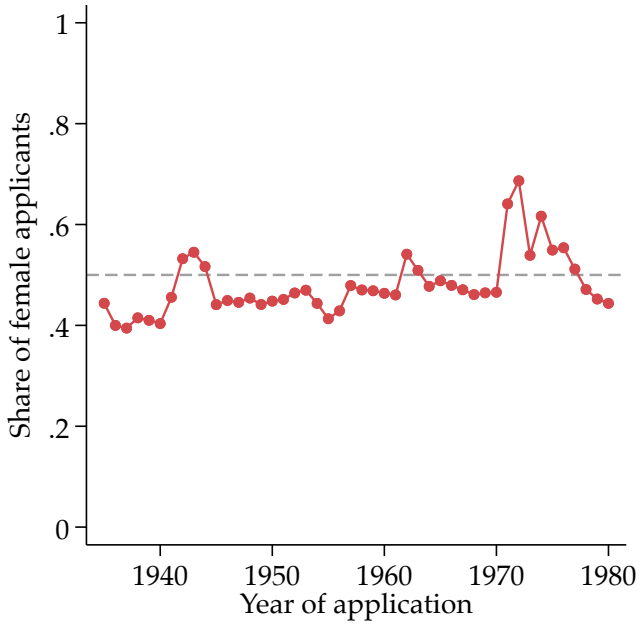
where $\widehat{\boldsymbol{R}}_x$ is a $k \times k$ rank-rank correlation matrix of the independent variables and $\widehat{\boldsymbol{R}}_{xy}$ is a $k \times 1$ vector of rank-correlations between the independent variable and dependent variable. $\hat{\alpha}$ is then computed as $\overline{Y} - \widehat{\boldsymbol{\beta}}'\overline{\boldsymbol{X}}$. Similarly, R^2 is estimated as

$$R^2 = \widehat{\boldsymbol{R}}'_{xy}\left(\widehat{\boldsymbol{R}}_x\right)^{-1}\widehat{\boldsymbol{R}}_{xy}. \tag{18}$$

Equations (17) and (18) are numerically equivalent to the rank-rank coefficient vector and R^2 in the case without latent variables (for a proof, see O'Neill (2021) and impose that the marginal distributions of the variables are identical). From equations (17) and (18), we also see the relation between the slope coefficient and R^2 in the univariate case discussed in Appendix C.1.1: $\hat{\beta} = \sqrt{R^2}$.

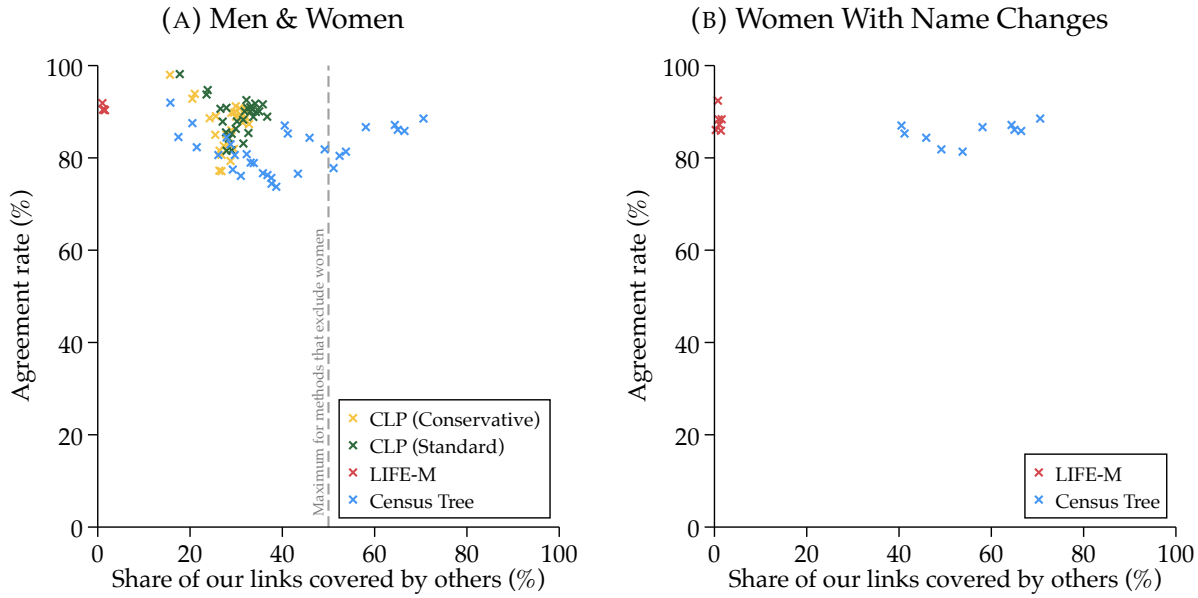
D. Data Appendix

FIGURE D.1: Share of Female Applicants



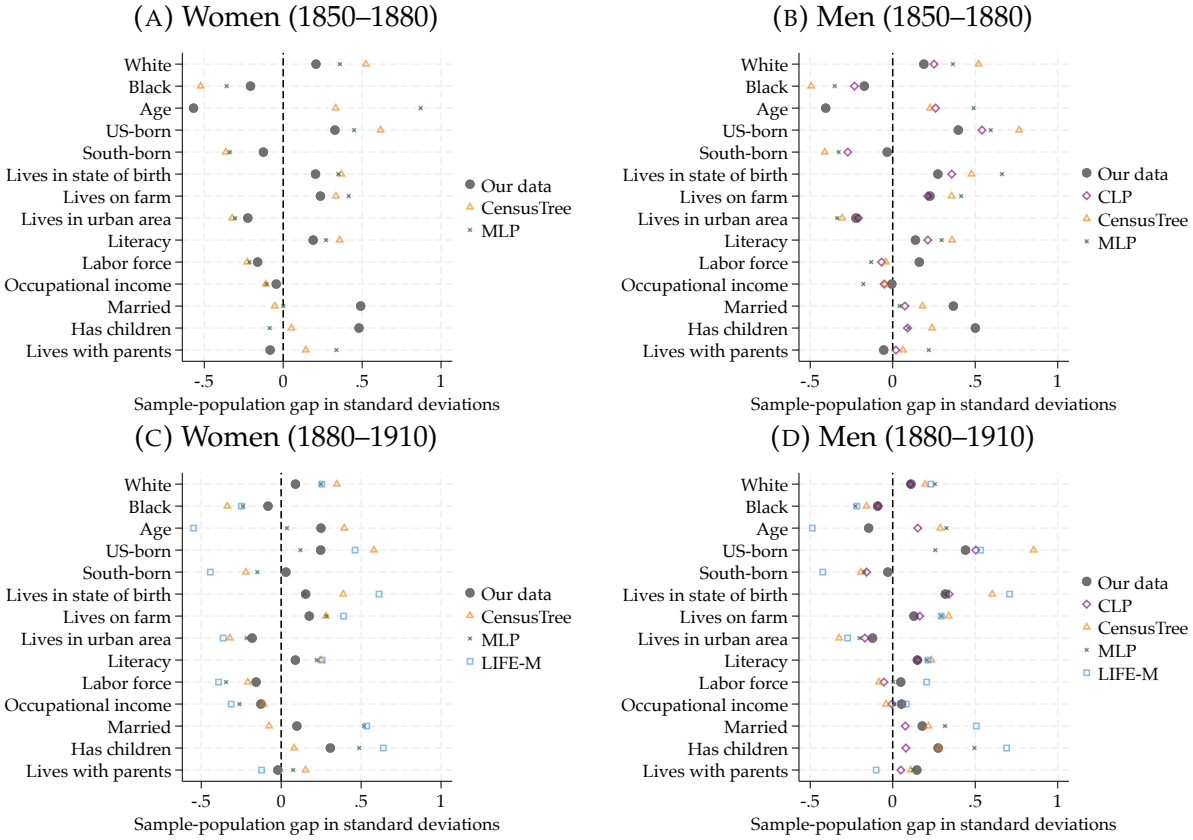
Notes: This figure shows the share of SSN applicants who are female by year of application.

FIGURE D.3: Our New Panel Compared to Existing Data



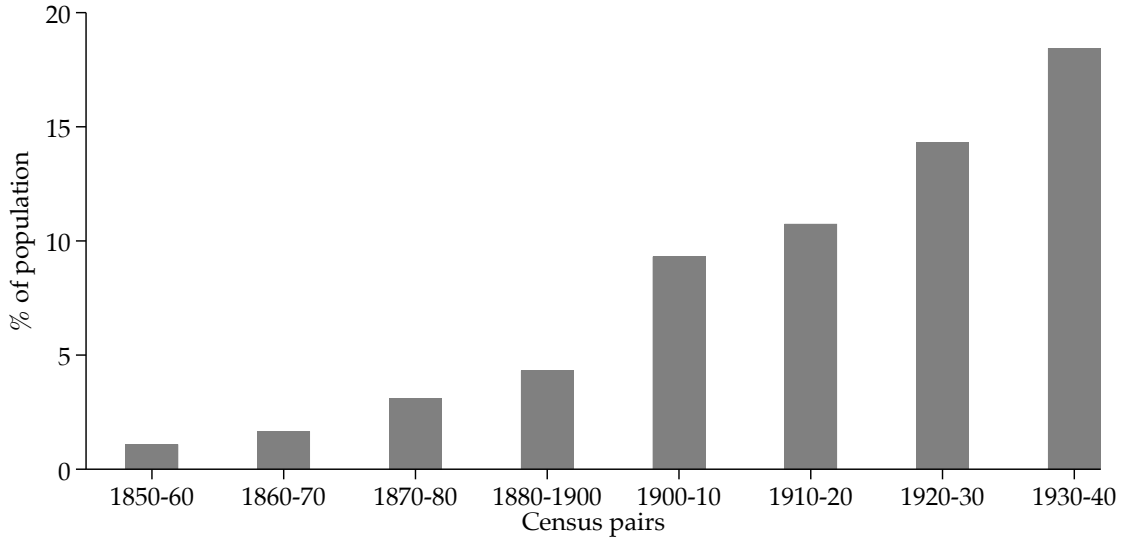
Notes: This figure compares our linked panel (1850–1940) to those of the Census Linking Project (CLP, Abramitzky et al., 2020), LIFE-M (Bailey et al., 2022), and the Census Tree (Buckles et al., 2023). Each point represents a link from one census decade to another (potentially non-adjacent). The x-axis 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.

FIGURE D.2: Balance of Linked Sample (1850–1880 & 1880–1910)



Notes: This figure shows the representativeness of characteristics among individuals in the 1880 or 1910 census who we successfully link to the 1850 or 1880 census compared to the full population in the 1880 or 1910 census of individuals aged 30 or above (and therefore alive in 1850 or 1880). We regress each outcome on a dummy for whether we link this individual back (outcomes are standardized to have a mean of 0 and a standard deviation of 1). The sample remains exceptionally representative in earlier years compared to existing panels with an average absolute deviation of 0.23 standard deviations from 1850 to 1880 (compared to 0.28 to 0.31 among existing data) and 0.15 standard deviations from 1880 to 1910 (compared to 0.21 to 0.34 among existing data); for this exercise we pool men and women and include "female" and a characteristic. CLP only includes men (Abramitzky et al., 2020). CensusTree uses genealogical data from the user-generated FamilyTree (Buckles et al., 2023). MLP (Helgertz et al., 2023) contains decade-to-decade links that we append iteratively. LIFE-M (Bailey et al., 2022) covers only Ohio and North Carolina and does not provide links for pre-1880 censuses.

FIGURE D.4: 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.

D.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)[14].

- *Preparing Census data*: Within each census decade from 1850 and 1940, we apply

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

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

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

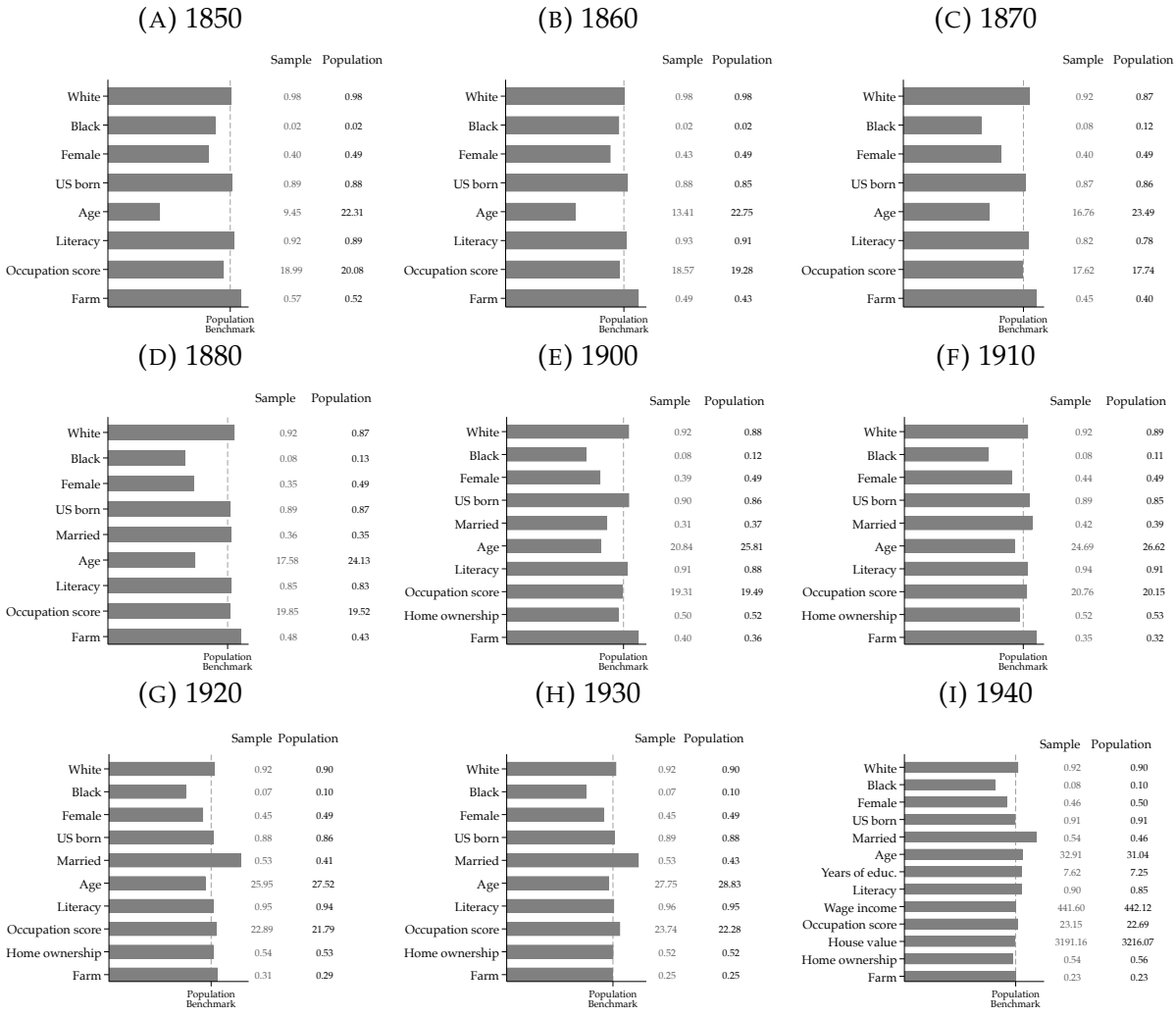
    1. We first create a pool of potential matches by finding all possible links between an SSN application and census record using first and last name (NYSIIS), place of birth, marital status and birth year within a 5-year age band. In the census, we identify marital status from the census variable "marst" or whether her position in the household is described as spouse. In the SSN data, we identify marital status if the applicants last name is different from that of her father.
    2. Once we have established our pool of potential matches, we essentially rerun our linking process. However, we use additional matching variables in order to pin down the most likely correct link among the potential matches. In our first round of this process, we aim to pin down the correct link by matching using the following set of matching characteristics: exact first, middle and last names of both the applicant and their parents, exact birth month (when available), state or country of birth, race, and sex. An SSN application is either uniquely matched to a census record or not.
    3. We attempt a second round of the matching described in point 2. for all SSN applicants who were *not* uniquely matched to a census record. In this round, we keep all matching variables the same, however, we use the phonetically standardized version of the middle name to account for spelling discrepancies. Once again, we separate those SSN applications that were uniquely matched to the census and those that were not.
    4. We repeat this matching process where we remove successfully matched individuals and attempt to rematch unmatched applications from our pool of potential matches. As we progress through the rounds of linking, the additional matching criteria become less stringent. We allow for misspellings or remove one or more variables in each subsequent iteration until we arrive at the literature standard, which involves only first and last name with spelling variations allowed, state of birth, and year of birth within a 5-year band.

We attempt to match each SSN record to all the census decades available as an individual may appear in the 1900 and 1910 census, for example. For married women

applicants, we search for potential census matches using both their maiden and married names. As a result, if we are able to find both records, married women appear in our data twice. We assign these links a slightly altered SSN to differentiate between the married and unmarried SSN-Census link. We do not link married women in the census who are below the age of 16.

The sample of individuals in the census who we successfully assign an SSN is highly representative of the overall population (see Appendix Figure D.5).

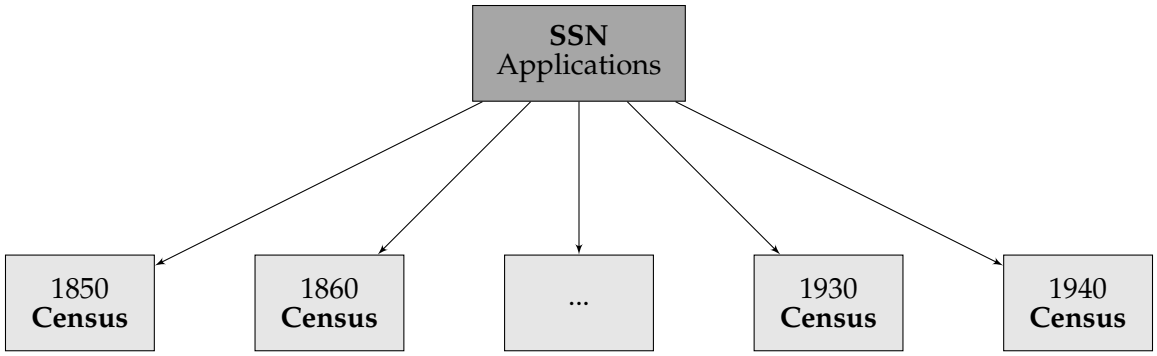
FIGURE D.5: Characteristics of Individuals Assigned an SSN (1850–1940)



Notes: This figure shows the representativeness of characteristics among individuals who we successfully assign an SSN compared to the full population in each census between 1850 and 1940.

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

FIGURE D.6: First & Second Linking Stages

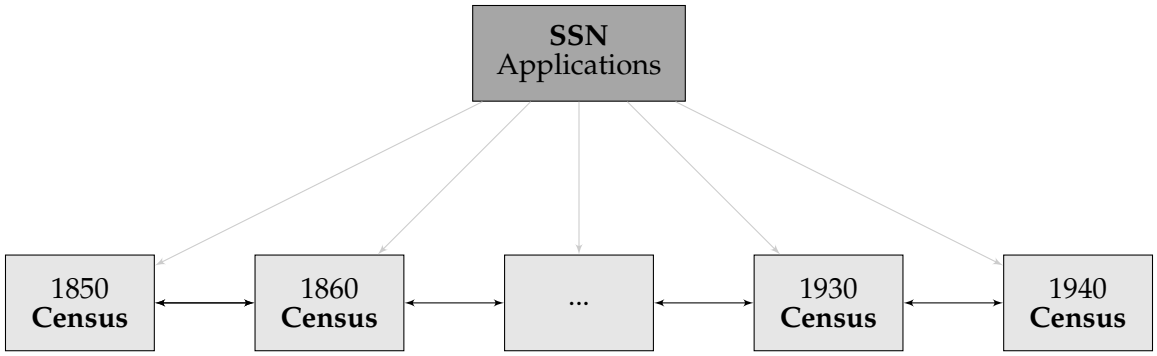


Notes: This figure shows the first and second step of our linking procedure—linking individuals' Social Security Numbers to their census records.

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

FIGURE D.7: 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.

D.2 Sample Weight Construction

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

Across all censuses between 1850 to 1940 and birth cohorts between 1870 and 1910, we create indicator variables for whether the individual enters our sample, i.e., whether we observe (1) their household's occupational income score in adulthood and (2) their parents' literacy and household occupational income score. We also create weights separately for individuals for whom we only observe one parent's outcomes, but our main analysis focuses on two-parent families. 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. Those groups are comprised of individuals with equal (i) sex, (ii) race, (iii) cohort in decades, (iv) state, (v) farm-status, (vi) rural-urban status, and (vii) occupational 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 observing the person's and their mother's economic status, observing the person's and their father's economic status, or observing the person's and both of their parents' economic status.

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.

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.

HELGERTZ, J., S. RUGGLES, J. R. WARREN, C. A. FITCH, J. D. HACKER, M. A. NELSON, J. P. PRICE, E. ROBERTS, AND M. SOBEK (2023): "IPUMS Multigenerational Longitudinal Panel: Version 1.1 [dataset]," .

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

O'NEILL, B. (2021): "Multiple Linear Regression and Correlation: A Geometric Analysis," *arXiv preprint arXiv:2109.08519*.

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.

SAAVEDRA, M. AND T. TWINAM (2020): "A machine learning approach to improving occupational income scores," *Explorations in Economic History*, 75, 101304.

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.