America's Rise in Human Capital Mobility*

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

[Most recent version here]

February 19, 2025.

Abstract

How did the US become a land of opportunity? We show that the country's pioneering role in mass education was key. Unlike previous research, which has focused on father-son income correlations, we incorporate both parents in a new measure of intergenerational mobility that considers multiple inputs, including mothers' and fathers' human capital. To estimate mobility despite limitations in historical data, we introduce a latent variable method and construct a representative linked panel that includes women. Our findings reveal that human capital mobility rose sharply from 1850 to 1950, driven by a declining reliance on maternal human capital, which had been most predictive of child outcomes before widespread schooling. Broadening schooling weakened this reliance on mothers, raising mobility in both human capital and income over time.

*We thank Leah Boustan, Raj Chetty, Ellora Derenoncourt, Rebecca Diamond, Christian Dustmann, Alice Evans, Claudia Goldin, Ilyana Kuziemko, Robert Margo, Bhash Mazumder, Pablo Mitnik, Petra Moser, Daniele Paserman, Matthew Staiger, Pablo Valenzuela, Nico Voigtländer, Fabian Waldinger, Monia Tomasella, Francesca Truffa, and numerous seminar participants for insightful comments. Pedro Carvalho and Alex Shaffer provided excellent research assistance. This paper previously circulated under the title "The Missing Link(s): Women and Intergenerational Mobility" (first version: Nov 2022).

[†]Department of Economics, Stanford University. lalthoff@stanford.edu

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

[§]Centre de Recerca en Economica Internacional (CREI). hreichardt@crei.cat

1. INTRODUCTION

Has the US been a land of opportunity, and what factors have shaped Americans' intergenerational mobility? Evidence on these questions remains limited, particularly regarding mothers' role in mobility. Yet their importance is abundantly clear from history: figures like Abraham Lincoln, Thomas Edison, and Katherine Johnson all thrived thanks to their mothers' teaching, especially crucial since their fathers lacked formal education.[1] Past studies have nonetheless focused on father-son comparisons due to two constraints: a lack of panel data that include women and an emphasis on income as the measure of parental background—missing mothers' contributions when few worked outside the home. This focus contrasts with both theory, which highlights parental human capital beyond income (Becker et al., 2018), and empirical work from other contexts showing mothers' key role in child development.

This paper studies the evolution of intergenerational mobility in the US from 1850 to 1950, highlighting the role of mothers' human capital and the expansion of mass schooling. We introduce a new measure of mobility that incorporates both parents' human capital, develop a latent variable method to estimate mobility despite historical data limitations, and construct a representative linked panel that includes women. Our findings reveal that human capital mobility rose sharply over this period, driven by a strong reliance on maternal human capital that declined as schooling expanded. Incorporating mothers alters key conclusions about long-term trends in and geographic patterns of intergenerational mobility. For example, while father-child comparisons suggest the South was relatively mobile during this period, including mothers reveals it as the least mobile region, underscoring the role of maternal human capital in areas where schools were scarce.

We first introduce a simple new methodology to account for multiple dimensions of parental background in the intergenerational analysis. Specifically, we propose measuring intergenerational mobility as the share of variation in child outcomes left unexplained by parental background: $1 - R^2$. Unlike traditional mobility measures, such as the parent-child coefficient, this measure accommodates multiple parental inputs. We show that the R^2-based measure 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 is that it can be separated into each parent's pre-

[1] Abraham Lincoln (1809-1865) was encouraged in his education by his mother and stepmother, while his illiterate father, a farmer, showed little interest. Similarly, Thomas Edison (1847-1931) was home-educated by his mother, a trained teacher, while his father had no formal schooling. Katherine Johnson (1918-2020), a NASA mathematician, had a mother who was a teacher and a father who worked as a janitor. To access quality education, her family moved across the Jim Crow South. Her daughter later became a NASA mathematician as well.

dictive power using a statistical decomposition method (Shapley, 1953; Owen, 1977).

Second, to accurately estimate mobility despite limitations in the historical data, we use a recently developed latent variable method from the statistics literature (Fan et al., 2017). This method allows us to study rank-rank relationships between parents and children when only binary proxies of the underlying outcomes are observed. In the historical data, such binary proxies are common; in our case, literacy provides information about a person's human capital. We discuss the assumptions this semi-parametric method imposes on the joint distribution of parent and child outcomes. We then extensively validate the method using (1) modern datasets (PSID and NLSY) where we directly observe continuous measures of human capital through cognitive test scores and (2) the 1940 census where we observe years of education as a more continuous outcome to benchmark mobility by state and demographic group.

Our first main finding is that human capital mobility sharply increased between 1850 and 1950. Over time, parents' human capital became less predictive of children's, with mobility ($1 - R^2$) increasing from 0.3 to 0.65. While both Black and white Americans experienced rising mobility, the timing differed. Black mobility surged from 1850 to 1880, coinciding with the end of slavery, which had previously excluded most from formal education. However, Black mobility then declined during a time when Jim Crow policies restricted educational access. In contrast, white mobility only began rising after 1890, aligning with the expansion of near-universal school attendance from 1890 to 1910. All results rely on full-count census data of 13- to 16-year-olds living with their parents, allowing us to observe parent and child outcomes jointly without requiring census linkage.

Our second main finding is that mothers play a central role in the intergenerational transmission of human capital. Mothers' human capital is more predictive of children's outcomes than fathers', particularly for female and Black children. Although maternal human capital remained disproportionately important over time, its influence began to decline around 1890. A statistical decomposition shows that the declining predictive power of maternal human capital fully accounts for the increase in mobility over time. While our main analysis focuses on two-parent households, we also show that maternal human capital is especially predictive in single- and widowed-mother households, among mothers who do not work outside the home, and in families with fewer children.

Observing that the rise in human capital mobility and the declining predictive power of maternal human capital coincided with a rapid expansion of schooling in the US, we explore universal education as a key mechanism. Between 1880 and 1900, school attendance among children ages 6-13 rose rapidly from below 60 to over 90 percent, cementing the country's lead in mass education (Goldin, 2001, 2016). Historians emphasize that before this transition, parental human capital—especially mothers'—played a central role in child development (Kaestle and Vinovskis, 1978; Dreilinger, 2021). As one scholar notes, "[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 schooling likely reduced reliance on parental human capital.

Indeed, we identify schooling as a key driver of intergenerational mobility. First, we document that school attendance strongly correlates with higher human capital mobility over time and across states. Even within the same period and location, groups with limited schooling were less mobile. For example, as Jim Crow segregation intensified educational barriers for Black children, their mobility declined while white mobility surged. Examining parental roles separately, we find that mothers' (but not fathers') human capital was especially predictive when schooling was low. These findings help explain why maternal human capital was so influential in early US history: as the primary educators of their time, mothers played a crucial role in shaping their children's human capital. Second, to further corroborate this mechanism, we exploit quasi-random variation in cohorts' exposure to state compulsory schooling laws. Instrumental variable estimates confirm that expanding formal education was a causal driver of rapidly rising human capital mobility.

To assess how income mobility evolved over this period of rising human capital mobility, we construct one of the first linked census panels to trace both men and women from child- to adulthood. A key challenge in linking historical records is name changes after marriage, which we overcome by leveraging administrative data from Social Security Number applications. These data provide both the birth and married names of applicants' mothers (plus those of married female applicants), enabling us to link individuals across generations regardless of name changes. Using this information, we link 186 million census records from 1850 to 1950, covering 42 million Americans. The resulting panel sets a new benchmark for representativeness, particularly for sex and race.

Using this new panel, we document that income mobility rose in tandem with human capital mobility from 1850 to 1950. This evidence leverages our new method to measure mobility based on multiple parental inputs, income and both parents' human capital, and holds across different occupational income proxies used in the literature. We also find that daughters tended to be more mobile than sons across the century. Black Americans were the most mobile group in the decades following slavery, but as Jim Crow policies intensified after 1890 and human capital mobility declined, they lost this advantage. By 1920, Black sons had become the least mobile group.

Statistical decompositions suggest that rising mobility in income and human capital was driven by the declining influence of maternal human capital. In contrast, changes in the importance of other factors, most importantly assortative mating, fathers' human capital, and income, if anything worked against this trend.[2] While intergenerational mobility theory highlights the separate importance of parental human capital and income (Becker et al., 2018), prior empirical studies have focused almost exclusively on income-to-income transmission. We show that incorporating parental human capital—especially mothers'—can alter conclusions about mobility trends and geographic patterns.

This paper deepens our understanding of how the US became a land of opportunity by documenting rising mobility from 1850 to 1950 and identifying a key driver: the spread of mass schooling. Previous studies have documented father-child correlations (e.g., Abramitzky et al., 2021a; Ward, 2023; Olivetti and Paserman, 2015; Craig et al., 2019; Jácome et al., 2021; Buckles et al., 2023b) or used parents' average status (Chetty et al., 2014b; Card et al., 2022) but have not assessed mothers' distinct role in economic transmission. While Espín-Sánchez et al. (2023) infer women's influence through male relatives using parametric assumptions, our methodology directly estimates women's role, overcoming measurement challenges and identifying the mechanisms behind mobility. We find that maternal human capital was a stronger predictor of child outcomes than father-based proxies, especially where schooling was limited. As schooling expanded, it replaced home education, reducing reliance on maternal human capital and driving mobility gains. Institutional change has reshaped mobility elsewhere (Chen et al., 2015), yet evidence for the US had been limited.

Incorporating mothers into mobility studies is particularly important given evidence from other contexts that mothers are key determinants of child outcomes. Mothers spend more time with their children than other adults almost anywhere worldwide (Evans and Jakiela, 2024), and evidence from Scandinavia shows that interventions that improve maternal health or education have disproportionately large intergenerational effects (e.g., Black et al., 2005; Holmlund et al., 2011; Lundborg et al., 2014; Abrahamsson et al., 2024; Björkegren et al., 2024).[3] Using data from randomly assigned donor children, Lundborg et al. (2024) show that only mothers' human capital (not fathers') affects child outcomes, suggesting that 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;

[2] Changes in parental assortative mating have a negligible impact on mobility trends.

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

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). However, the output of home production is typically difficult to measure. We uncover one critical product: the home education of children. Our findings reveal that women made a major contribution to human capital accumulation in the US economy, even before their large-scale entry into the workforce.

Finally, we construct and publicly provide an extensive and representative linked panels including women, building on the foundations of previous work. Craig et al. (2019) and Bailey et al. (2022) expand automated record linkage (Abramitzky et al., 2021b) for women using historical birth, marriage, and death certificates, but these are limited to selected states and periods. Buckles et al. (2023b) leverage crowd-sourced family trees, significantly increasing sample sizes, though with selectivity concerns (Abramitzky et al., 2024). In contrast, we use historical *administrative* data, achieving both scale and representativeness.

2. INTERGENERATIONAL DATASETS

This section describes the data we use to study how parental background affects child outcomes. First, using census cross-sections of children aged 13-16 who live in their parents' household allows us to analyze the transmission of human capital during childhood. Second, constructing a representative panel that includes women allows us to study how parental background affects both human capital and income in adulthood.

2.1 Childhood Outcomes: Census Cross-Sections (1850–1950)

We use full-count census data from 1850 to 1950 (Ruggles et al., 2024a,b), which provide each person's socioeconomic characteristics and identify family interrelationships within households. This allows us to link parent and child outcomes for all children who live with their parents.[4]

For children aged 13-16, we measure literacy and total years of schooling. For parents aged 20-54, we measure literacy and years of education. Literacy data are available in every census between 1850 and 1930. Years of education are only reported in 1940 and 1950. A remaining limitation of the historical data is that they do not capture aspects of human capital independent of literacy.

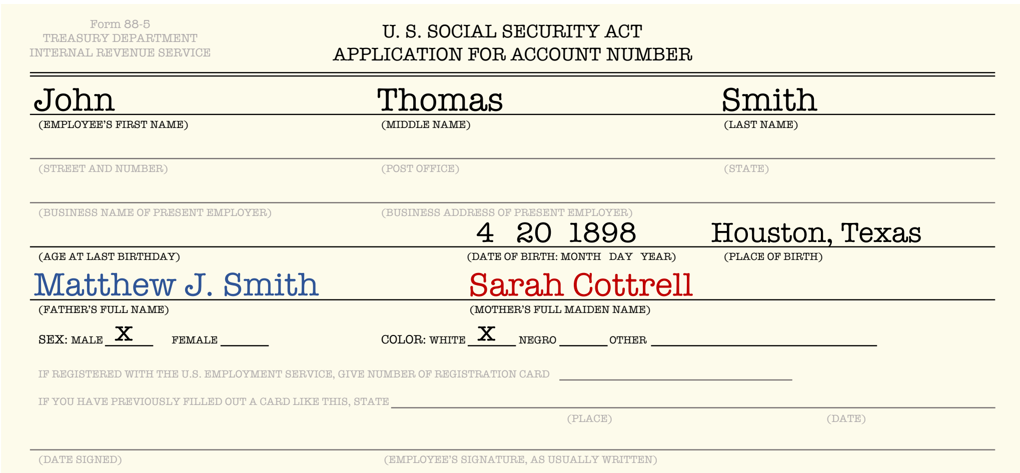
[4] 77 percent of children aged 13-16 are recorded with both of their parents between 1850 and 1950, 91 percent with at least one parent.

2.2 Adult Outcomes: A New Panel that Includes Women (1850–1950)

A main empirical challenge in including adult women to study the long-run evolution of intergenerational mobility is the lack of suitable panel data. We overcome this hurdle by combining census records with historical administrative data that contain the birth and married names of millions of women. Using these data, we link adult men and women in historical censuses (1850-1950) to their childhood census records. The result is one of the first historical panel datasets to include women and one that stands out in its representativeness.

2.2.1 Historical Administrative Data from the Social Security Administration

FIGURE 1: Social Security Number 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 Number application form. Besides the applicants' information, the form includes fathers' names and mothers' birth names ("maiden names").

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 birth names of their parents (see Figure 1). Based on these data, we can derive the birth and married 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).

The Social Security Number (SSN) system was introduced in 1935 to register employed individuals, initially excluding the self-employed and certain other occupations (Puckett, 2009). However, its coverage expanded rapidly—Executive Order 9397 (1943) and the IRS's adoption of SSNs for tax reporting (1962) extended it to nearly all individuals.[5] Most linked individuals with name changes are mothers of SSN applicants. This improves our sample's representativeness, as parents are included regardless of whether they applied for an SSN themselves, as long as they had at least one child who did.

The SSN data on applicants alone (excluding parents) covers large shares of Americans born after the 1880s, exceeding 50 percent for those born in or after the 1910s. Coverage further rises from 64 percent for the 1915 birth cohort to 80 percent for 1920, 90 percent for 1935, and nearly 100 percent from 1950 onward (comparing each cohort's births to SSN records; CDC, 2023; SSA, 2023). Including parents extends coverage further back.

2.2.2 Linking Method

We use an automated multi-stage linking process to maximize the use of SSN application data, building on existing methods (Abramitzky et al., 2021b). This process involves three stages: linking SSN applicants to census records, linking applicants' parents to census records, and tracking census records over time. Appendix E.1 provides further details.

First, we link each applicant to their census record using a rich set of information: applicants' and parents' names, year and state of birth, race, and sex. For those not matched, we progressively relax criteria to the literature standard, matching applicants' first and last name (with spelling variations), state of birth, and birth year within a 5-year band. For married female applicants, we search under both birth and married names. Non-unique matches are discarded. Using both applicants' and parents' names improves match uniqueness. This approach is effective not only for children but also for adults in multi-generational households, which accounted for 80-90 percent of Americans during our sample period. While parental names improve match uniqueness for those living with their parents, we also link adults who are not co-resident.

Second, we link applicants' parents to census records. Since SSN applications lack detailed birth information for parents, direct matching is not possible. However, when a child's application is successfully linked to a census record where they reside with their parents, we can link those parents to that census household. For parents without a known SSN, we assign a synthetic identifier.

Last, after assigning unique identifiers despite possible name changes, we track individuals across censuses from 1850 to 1950. While standard or machine learning

[5] Throughout this period, female applicants consistently made up close to 50 percent (see Appendix Figure E.1).

methods could expand linkages—particularly for men and never-married women—we deliberately avoid them. A key strength of our dataset is its ability to trace women from childhood to adulthood despite name changes. The set of known name changes is determined by the data and cannot be enhanced with these methods. Moreover, using different linking techniques for different subgroups would undermine sample representativeness by applying different criteria to married women than to others.

2.2.3 Our New Panel

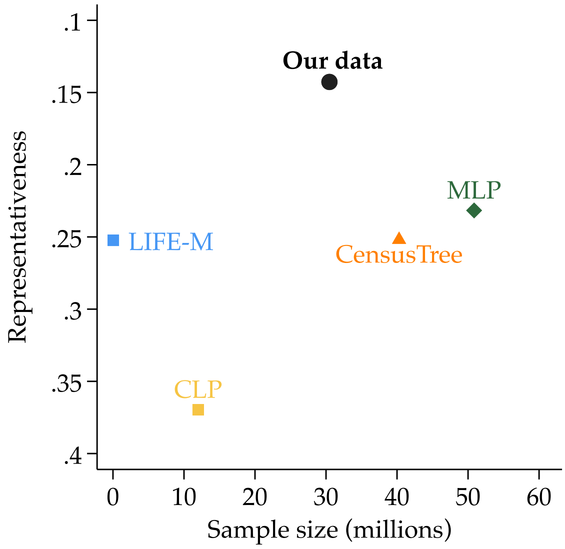
Our panel covers 42 million individuals who are linked across 186 million census records between 1850 and 1950. This implies that we trace each person across more than four census decade pairs on average. In our first linking stage, we assign SSNs to 54 million individuals recorded across censuses,[6] of whom 42 million enter our panel by being observed in at least two census decades. Our linking rate of 40 percent among applicants surpasses the more typical 25 percent of prior studies thanks to the more detailed information available in the SSN application data, notably parents' and spouses' names.

Representativeness. The value of intergenerational datasets depends on how well individuals linked over long time horizons (child- to adulthood) represent the population. We therefore compare the socioeconomic and demographic characteristics of individuals linked across any 30-year period covered by the census between 1850 and 1950 to those of the full population aged 30 or above in the later census. Our panel is more representative of the US population than existing datasets (see Figure 2). Across a large variety of socioeconomic characteristics, our data on average differs by 0.13 (absolute) standard deviations from the population; existing datasets range between 0.21 and 0.36.

To provide further details on the representativeness of our panel, we assess differences separately across various demographic and socioeconomic characteristics over different periods (see Appendix Figure E.2). Our sample stands out from other datasets especially in the key dimensions of sex, race, education, and geographic representation. While we achieve better representativeness across most characteristics, our sample over-represents married individuals and those with children. These deviations stem from our linking procedure's use of spouse and children's names when available, which improves match rates for individuals with families. We provide sample weights to correct for those factors. Specifically, we use a flexible non-parametric method to construct inverse propensity weights separately for each birth cohort (see Appendix E.2).

[6] 32 million of the 54 million people identified are SSN applicants; 22 million are applicants' parents.

FIGURE 2: Coverage & Representativeness of Long-Run Panels (1850-1950)



Notes: This figure compares coverage and representativeness of census panels that cover any 30-year period between 1850 and 1950 (1850-80, 1870-1900, 1880-1910, 1900-30, 1910-40, and 1920-50). The horizontal axis shows the total number of links each panel provides across these periods. The vertical axis shows the representativeness of each panel, measured as the average absolute standard deviation from a comprehensive set of population characteristics (listed in Appendix Figure E.2). Americans aged 30+ in the later census serve as the benchmark population. Our sample achieves a smaller average absolute deviation (0.13) than existing panels (0.21-0.36). Comparison panels include: CLP (men only; Abramitzky et al., 2020), CensusTree (FamilyTree data generated by users of online genealogy; Buckles et al., 2023a), MLP (iterative decade links; Helgertz et al., 2023), and LIFE-M (OH/NC only; Bailey et al., 2022).

Coverage. In addition to its high representativeness, our samples are large, ranging in size between those of the widely used Census Linking Project (CLP) and datasets that combine machine learning and vast online genealogy data (CensusTree). A standout feature of our panel is the inclusion of 9 million women for whom we observe pre- and post-marriage data. These data allows us to overcome critical data limitations to study the role of women in mobility throughout US history. More generally, our panel covers large shares of the US population: 15–25 percent from 1910–1950 and 2–14 percent from 1850–1900 (see Appendix Figure E.3). Our panel has high agreement rates with existing data for overlapping individuals but also covers a high number of individuals whose records had not previously been linked (see Appendix Figure E.4). Agreement rates vary from 80 to nearly 100 percent and are highest with LIFE-M—a smaller panel that leverages vital records in the linking process (Bailey et al., 2022).

3. Measuring Intergenerational Mobility with Multiple Inputs

In this section, we first propose measuring mobility as the share of variance in child outcomes that cannot be explained by (potentially multiple) measures of parental background: $1 - R^2$. Second, we build on a state-of-the-art semiparametric latent variable method to estimate $1 - R^2$ from a rank-rank regression when only binary proxies of underlying outcomes are observed (e.g., literacy as a proxy for human capital). Third, we lay out a decomposition method to separate the statistical contribution of various inputs to R^2 (e.g., the relative contribution of mother's and father's 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:

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

where $\text{rank}\left(Y_i\right)$ is the percentile rank of outcome of i and $\mathbf{rank}\left(\boldsymbol{Y}_{\boldsymbol{i}}^{\textbf{parental}}\right)$ 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, other relatives' backgrounds, or neighborhood characteristics.

3.2 $1 - R^2$ as a Measure of Mobility with Multiple Inputs

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

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

where $\widehat{\text{rank}}\,(Y_i)$ is the predicted rank of i from equation (1) and 50 is the average rank by construction.[7]

We argue that lack of predictability as captured by $1 - R^2$ 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.

Our R^2-based measure has a direct relationship with traditional mobility measures—parent-child coefficients or, most commonly, father-son coefficients ($\widehat{\beta}$).[8] 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 the R^2-based measure is that it can provide an intuitive and easily interpretable measure of mobility even when considering multiple parental inputs. We use this advantage to include both mothers' and fathers' outcomes, and to include multiple dimensions of parental background. Another advantage is that it can be decomposed into the contributions of individual inputs.

3.3 Measuring Mobility with Latent Inputs

Our goal is to estimate intergenerational mobility ($1 - R^2$) 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:

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

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

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

where Y_i represents the child's human capital and X_i represents the parental human capital.[9] This is the simplest version of equation (1); Appendix C.3 generalizes this framework and the discussion below to multiple inputs and provides further formal detail.

3.3.1 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^*:

$$Y_i^* = \mathbb{1}[Y_i > \delta_y] \quad (3)$$

$$X_i^* = \mathbb{1}[X_i > \delta_x] \quad (4)$$

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.

3.3.2 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 $f_Y(Y_i), f_X(X_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$.[10] 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. Note that this does not impose that the latent variables of interest (e.g., human capital) are normally distributed.

3.3.3 Identification of rank correlations from binary indicators

Under the Gaussian copula assumption, ρ—the correlation between the jointly normal random variables $f_Y(Y_i)$ and $f_X(X_i)$—is identified and can be estimated using the Ken-

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

[10] 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.

dall's tau correlation coefficient of the observed binary variables:

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

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.[11] 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.[12] 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.

3.3.4 Illustration and validation

To illustrate the method, we begin by applying it to simulated data that satisfy the identifying distributional assumption by construction. Specifically, we simulate jointly normally distributed parent-child data. Based on these data, we estimate benchmark rank-rank mobility. Then, we dichotomize the continuous data and attempt to recover rank mobility using our latent variable method. We show that, unlike a naive OLS

[11] 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 $\overline{X}^* = \sum_{i=1}^n X_i^*/n$.

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

approach, our latent variable method accurately recovers rank mobility from binary proxies (see Appendix Figure A.1; see also Appendix Figure C.1 for a conceptual illustration of this method). This is true even when cut-offs vary over time, which is particularly relevant in our context, where rising literacy rates may change literacy's informativeness as a proxy for human capital.

Turning to validation using real data (which may not follow distributional assumptions), we first show that our method produces mobility patterns across groups closely matching those from standard rank-rank regressions when continuous data is available. Specifically, we compare our latent variable method estimates based on 1930 literacy with 1940 educational mobility estimates based on standard rank-rank regressions (see Appendix Figure A.2). While literacy and years of education capture different concepts, one may expect mobility patterns based on the two to be broadly consistent. Indeed, we find that the two are highly correlated across states, sex, and race ($\rho = 0.85$), supporting the reliability of our approach. Notably, literacy rates in 1930 exceed 95 percent, demonstrating that our method performs well in practice even when binary proxies have extreme cutoffs.

We conduct a similar validation by arbitrarily binarizing ranks of educational attainment in the 1940 census. After computing rank-rank mobility from the continuous education rank data, we dichotomize these ranks with different cut-offs for children, mothers, and fathers. Our method's state-level mobility estimates align closely with those from the original undichotomized data (see Appendix Figure A.3), further demonstrating the method's performance in relevant historical data.

Lastly, we assess whether our method can recover mobility estimates when continuous human capital measures are arbitrarily dichotomized at a range of cutoffs. We first use modern data that provide continuous human capital measures for both parents and children: the 1979 National Longitudinal Survey of Youth (NLSY79) and the Panel Study of Income Dynamics (PSID). We estimate rank-rank mobility using the raw continuous data as a benchmark; we then apply various dichotomization thresholds, including extreme cutoffs, and use our latent variable method to re-estimate mobility. Across all cases, our method accurately recovers correct estimates of mobility (see Appendix Figure A.4). We also conduct 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).

3.4 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., $\text{rank}\,(Y_i)$) 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).

4. The Rise in Human Capital Mobility

This section applies our new methodology to measure mobility with multiple parental inputs and uncovers a steep rise in human capital mobility across our sample period (covering birth cohorts from the 1860s to the 1910s). Parental human capital—especially mothers' but not fathers'—became less predictive over time, most rapidly for Black children early in our sample period and for white children much later.

4.1 Parental Human Capital and Child Outcomes

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

$$\text{rank}\,(h_i) = \delta + \gamma_m \text{rank}\left(h_i^{\text{mother}}\right) + \gamma_f \text{rank}\left(h_i^{\text{father}}\right) + \eta_i, \tag{6}$$

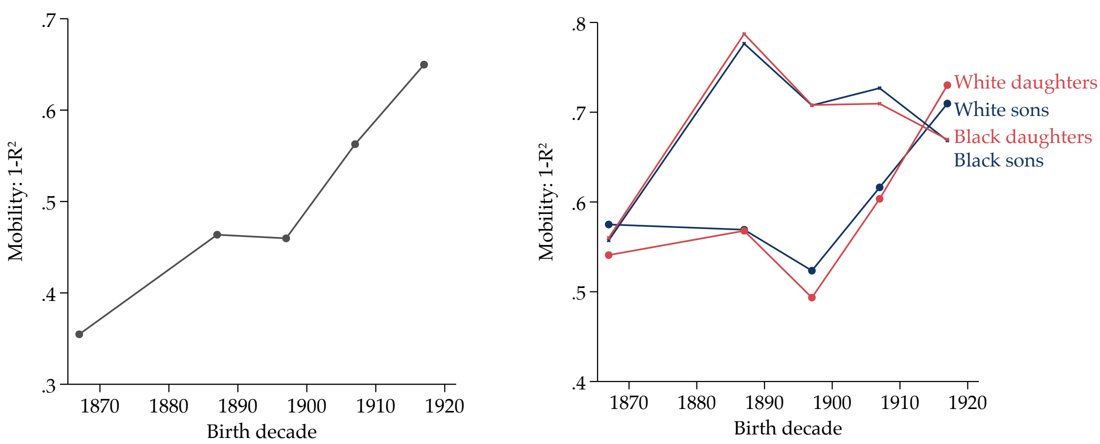
where h is (latent) human capital. We estimate this model using the semiparametric latent variable method described in Section 3.3, which allows us to infer rank-rank mobility from binary proxies of human capital. Section 3.3.4 lays out underlying assumptions and provides validation exercises that demonstrate the method's ability to recover mobility patterns in practice.

Census cross-sections of children who reside with their parents allow us to study human capital mobility without record linkage. Specifically, we use such cross-sections to relate parental background to 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. We successfully replicate cross-sectional estimates using our new panel (see Appendix Figure D.1).[13]

FIGURE 3: Human Capital Mobility

(A) Mobility Across Groups

(B) Racial Differences in Mobility Patterns



Notes: This figure shows trends in intergenerational human capital mobility. Panel (A) plots the share of variance in children's (latent) human capital rank unexplained by both parents' (latent) human capital ranks $(1 - R^2)$. Panel (B) shows those estimates separately by race and sex. We recover rank-rank mobility using information on literacy and the latent variable method introduced in section 3.3. Results are based on census cross-sections of children ages 13–16 in their parents' household.

Our first set of results reveals rapidly increasing human capital mobility for chil-

[13] While the trends match, our panel consistently shows higher levels of human capital mobility. This difference likely reflects two factors. First, intergenerational persistence may appear stronger in childhood-based cross-sectional data than in our panel of adults, whose human capital may evolve to become less tied to their parents' human capital over time. Second, automated record linkage may introduce measurement error in parental background, overestimating levels of mobility. However, since trends are consistent across datasets, any overestimation would appear to be stable over time.

dren over our sample period (see Panel A of Figure 3). While parental human capital accounted for 65 percent of variation in child human capital in the earliest cohort (born in the late 1860s), it only accounted for 35 percent among children born in the latest cohort (born in the late 1910s). This rise in mobility is large in magnitude, exceeding that of the income mobility gap between the US versus Denmark or Sweden.[14]

Second, we document stark racial differences in the evolution of human capital mobility (see Panel B of Figure 4). Estimating equation (6) separately by race and sex, we find that Black children saw a sharp rise in mobility around the time that slavery ended, increasing from the 1860s cohort ($1 - R^2 = 0.55$) to the 1880s cohort ($1 - R^2 = 0.78$). This rise was followed by a decline in mobility. In contrast, white children's mobility remained low and stable until around 1890 ($1 - R^2 = 0.55$) and then surged after 1900, with magnitudes mirroring the earlier gains of Black children in the 1870s and 1880s. By the 1910s cohort, white children's mobility had surpassed that of Black children for the first time since the Civil War.

Mobility patterns align with broader historical shifts. Under slavery, most Black Americans were denied formal education and remained illiterate. After emancipation in 1865, Black Americans' literacy surged for the first time in generations, consistent with the rise in human capital mobility we document. However, beginning in 1877, Southern states imposed new restrictions on Black education. In the Jim Crow South, school access declined, school years shortened, and quality deteriorated (Card and Krueger, 1992; Althoff and Reichardt, 2024). This persistent lack of high-quality education may explain why Black human capital mobility declined after the 1870s, diverging from the rising mobility of white Americans. White children saw rapid expansions in schooling, especially in the South. Around 1900, the share of children not in school fell from one-third to 10 percent within a decade (see Appendix Figure A.5). This surge in formal schooling likely contributed to the sharp rise in white Americans' mobility.

Lastly, while our analysis so far has focused on two-parent families, we also assess human capital mobility across family types (see Appendix Figure D.2). Single parents have greater predictive power than those in two-parent families, likely due to undivided parental responsibilities. However, 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. A larger number of siblings is also associated with lower predictive power of mothers, possibly due to weaker human capital transmission when resources are shared across multiple children.

[14] The difference in mobility ($1 - R^2$) of $0.65 - 0.35$ corresponds (in magnitudes) to a difference in parent-child coefficients of $\sqrt{0.65} - \sqrt{0.35} \approx 0.21$. Denmark's and Sweden's rank-rank income mobility coefficients are around 0.19, whereas the US's is 0.36, a difference of 0.17 (Britto et al., 2024).

4.2 Drivers of Changing Human Capital Mobility

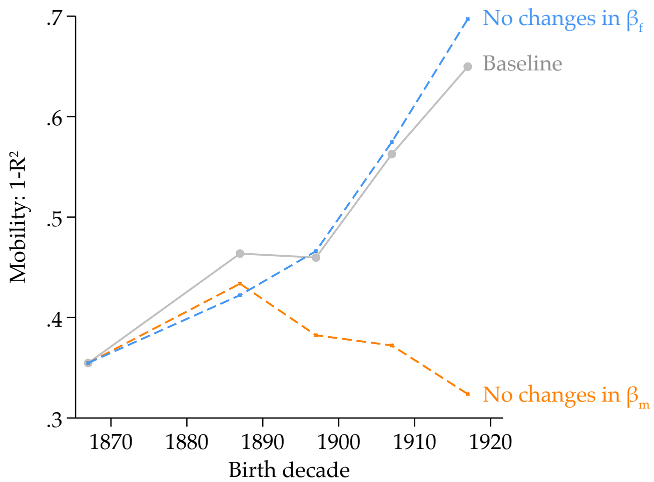
We next assess the drivers of changing mobility in more detail, focusing on the role of maternal versus paternal inputs. First, we separate mothers' and fathers' contributions to predicting children's human capital using the Shapley-Owen decomposition described in section 3.4 (see Appendix Figure D.3 for an illustration of the method).

Second, to understand how the changing role of mothers' and fathers' contributions affected the evolution of human capital mobility over time, we separate our mobility measure into multiple components and analyze their individual contributions. Specifically, we decompose R^2 in equation (6) into

$$R^2 = \widehat{\beta}_m^2 + \widehat{\beta}_f^2 + 2\widehat{\beta}_m\widehat{\beta}_f\hat{\rho}_{m,f} \tag{7}$$

where $\hat{\rho}_{m,f}$ is the correlation between mother's and father's human capital—a measure of assortative mating. We provide a more general decomposition of R^2 in rank-rank regressions with an arbitrary number of independent variables in Appendix C.1.2. Using this decomposition, we compute the (statistical) counterfactual R^2 holding a given parameter constant over time.

FIGURE 4: Changing Role of Mothers as Driver of Rising Mobility



Notes: This figure shows the results of decomposing changes in human capital mobility over time via equation (7). We decompose mobility $(1 - R^2)$ into its components and compute (statistical) counterfactual mobility by holding a given parameter constant over time. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.3. Results are based on the census cross-section of children ages 13–16 in their parents' household.

Our first result is that mothers' human capital tends to be a stronger predictor than fathers' human capital (see Appendix Figure D.4). This is particularly true for female and black children, with maternal human capital accounting for up to 70 percent of parental human capital's overall predictive power. This finding is consistent with narrative evidence of mothers' key role as educators, especially where formal schooling was not universal, and previous evidence by Olivetti et al. (2018), who find similar gender-specific transmission from paternal and maternal grandparents to their grandsons and granddaughters. Mothers' disproportionate 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).

Our second result is that the rise in human capital mobility over time is fully accounted for by the large but declining role of maternal human capital, $\widehat{\beta}_m$ (see Figure 4). Specifically, mobility would have decreased, had it not been for the weakening link between child and maternal human capital. In contrast, changes in the role of father's human capital ($\widehat{\beta}_f$) barely affected mobility trends. Changes in assortative mating also did not have a significant impact (see Appendix Figure D.5).

In addition to driving trends in human capital mobility over time, including mothers significantly alters conclusions about the geography of mobility (see Appendix Figure A.6). Using our latent variable method, we compare rank-rank mobility estimates based on father-child transmission to those incorporating both parents. The largest shifts occur in the South, where maternal human capital likely played a greater role in shaping child outcomes due to the region's scarcity of schools, while the Northeast remains largely unchanged. Indeed, we document a strong negative correlation ($\rho = -0.85$) between schooling and the additional predictive power gained by incorporating mothers. These findings highlight the limitations of a father-centered approach to studying intergenerational mobility.

5. Mothers, Schools, and Human Capital Mobility

The previous section shows that mothers' human capital was more predictive of child outcomes than fathers' and that the changing role of maternal human capital accounts for trends in human capital mobility. This section focuses on the mechanism of rising mobility and mothers' role as educators of their children. We first show that maternal human capital was especially predictive where schooling was limited. As schooling expanded, mobility increased. Using variation in state compulsory schooling laws, we corroborate the expansion of formal schooling as a causal driver of reduced reliance on maternal human capital, driving gains in human capital mobility.

5.1 The Historical Role of Maternal Human Capital

Historians have highlighted mothers' important role in educating their children in the 19th century (Kaestle and Vinovskis, 1978; Margolis, 1984; Dreilinger, 2021). Prior to public schooling becoming universal outside the Northern states in the late 19th and early 20th centuries, parental home education was central for children's human capital development. Even children who were enrolled in school in the late 19th century attended school less than four months a year on average (Dreilinger, 2021).

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

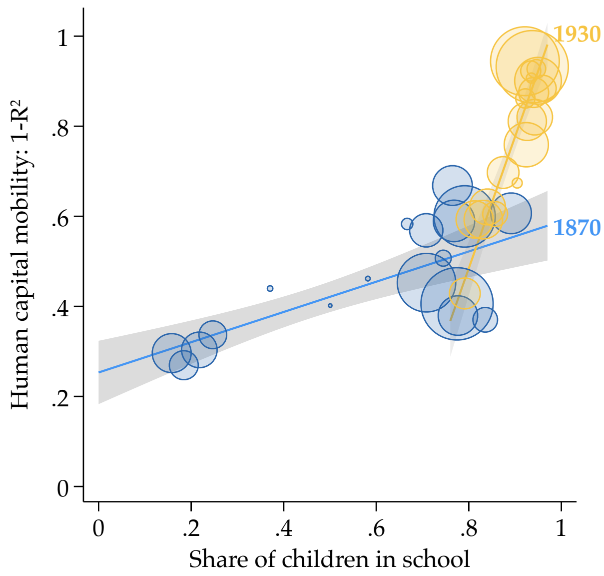
Mothers' pivotal role gained recognition from contemporary intellectuals, who advocated for the professionalization of women's role as home-educators. "The mother forms the character of the future man," Catharine Beecher, a famous American educator, wrote (Beecher, 1842). "The mother may, in the unconscious 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).

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

5.2 Schools and the Rise of Human Capital Mobility

While the share of children attending school rose rapidly in the late 19th century, the spread of schooling was also 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).

FIGURE 5: Schools and the Rise of Human Capital Mobility



Notes: This figure shows the relationship between local school attendance and human capital mobility by state over time (1870 and 1930). Schooling is measured as a state's share of children in school at ages 6–13. Mobility is measured as the share of the variance in a child's (latent) human capital rank left unexplained by parents' (latent) human capital ranks ($1 - R^2$) across cohorts and states. We recover rank-rank mobility using information on literacy and the latent variable method introduced in section 3.3. The size of each state's bubble is proportional to its population.

We first document a strong correlation between human capital mobility and school attendance across both time and place (see Figure 5). In 1870, school attendance varied widely by state, race, and sex. Children with the lowest school access were also the least mobile, with parental human capital explaining 70 percent of the variation in child human capital. In contrast, among children with the highest school attendance, parental human capital accounted for only 40 percent of the variation. By 1930, schooling had expanded significantly, yet mobility remained closely linked to remaining disparities in education.[15] As schooling approached universal levels, human capital mobility also neared its peak. The reduced influence of parental human capital with improved public schooling aligns with Biasi (2023), who shows that equalizing school resources can reduce disparities in intergenerational mobility.

Second, we find that mothers, but not fathers, were more predictive of child outcomes in areas with limited schooling (see Figure 6). Maternal human capital alone explains almost 40 percent of variation in child human capital when schooling is minimal, consistent with mothers' importance in home schooling, and around 20 percent

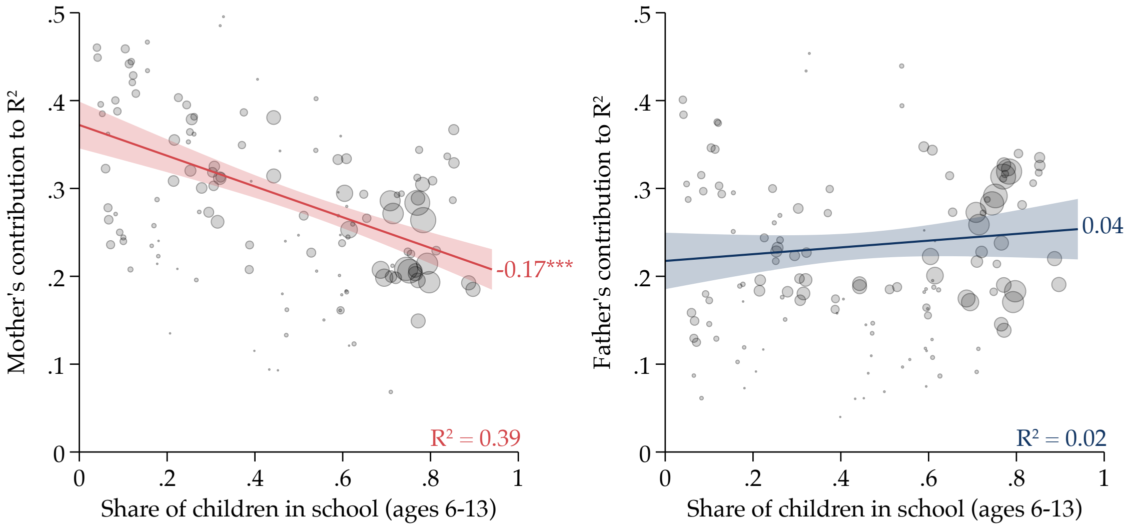
[15] This also confirms that our latent variable method based on literacy captures meaningful patterns in human capital mobility, even when literacy rates are high, as in 1930.

when it is universal. Conversely, fathers' contribution was lower and showed no correlation with schooling. In fact, the contributions of mothers and fathers were comparable only when schooling was near-universal.

Our analysis reveals an even stronger correlation between schooling and human capital mobility when refining our measure of schooling 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 schooling 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 schooling.

FIGURE 6: Dependence on Maternal Human Capital Before Universal Schooling

(A) Mothers' Contribution (1870) (B) Fathers' Contribution (1870)



Notes: This figure relates local school attendance to parental contributions to child human capital. Panels A and B respectively show mothers' and fathers' contributions to the overall R^2 using the Shapley-Owen method. 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 rank-rank mobility using information on literacy and the latent variable method introduced in section 3.3. Each dot represents a group of children born in the 1870s, categorized by race, sex, and state. Sample size weights are applied. Schooling is determined by the race- and sex-specific share of children aged 6-13 in school.

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 caused by the introduction of these laws. We interpret this as evidence that increased access to schooling led to 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 schooling in the late 19$^{\text{th}}$ and early 20$^{\text{th}}$ century contributed to increasing intergenerational mobility in human capital. 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).

6. Income Mobility & Parental Human Capital

Having established that human capital mobility increased over time, we now focus on income mobility. We find that income mobility increased: parent's income and human capital became jointly less predictive of their children's income. Similar to the findings on human capital mobility, the decreasing predictive power of maternal human capital accounts for the rise in income mobility.

6.1 Income Mobility Accounting for Parental Human Capital

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

$$\text{rank}\left(inc_i\right) = \alpha + \beta_p \text{rank}\left(inc_i^{\text{parents}}\right) + \beta_m \text{rank}\left(h_i^{\text{mother}}\right) + \beta_f \text{rank}\left(h_i^{\text{father}}\right) + \varepsilon_i, \quad (8)$$

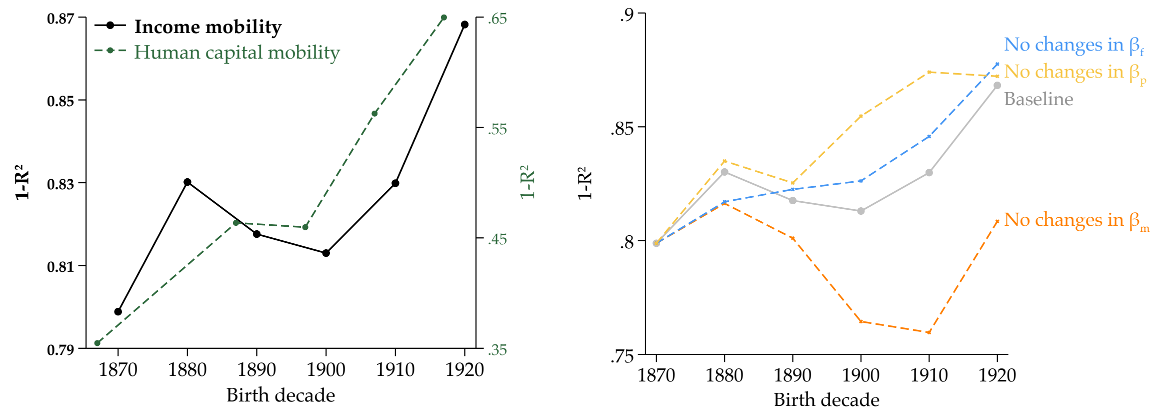
where inc is household income and h is (latent) human capital. We measure household income as the household head's LIDO occupational income score (Saavedra and Twinam, 2020). Literacy serves as a binary proxy for latent human capital. We estimate this model using the semiparametric latent variable method described in section 3.3 and our new representative panel dataset described in section 2.2.3.[16]

Including parental human capital alongside income in measures of mobility has both theoretical and empirical motivations. First, theories of intergenerational mobil-

[16] Note that this method identifies the parameters in equation (8), but not individual human capital ranks.

FIGURE 7: The Rise in Income Mobility

(A) Income Mobility (B) Driver: Changing Role of Mothers



Notes: This figure shows trends in income mobility. Panel A plots the share of variance in children's household income rank unexplained by their parents' income and (latent) human capital ($1 - R^2$). The dashed line repeats the estimates of human capital mobility from Figure 3. Panel B decomposes changes in mobility into contributions from mothers' human capital (β_m), fathers' human capital (β_f), and parental income (β_p), with counterfactual mobility estimated assuming each factor remained constant; see equation (9). We recover rank-rank mobility using information on literacy and the latent variable method introduced in section 3.3. Income is measured by the LIDO occupational score (Saavedra and Twinam, 2020), residualized by age. Results are based on our new panel with sample weights (Appendix E.2).

ity emphasize that parental human capital not only affects parents' ability to invest in their children monetarily but also directly shapes their children's human capital, thereby independently influencing children's outcomes (Becker et al., 2018). Empirically, incorporating parents' human capital can help measuring parental background more accurately. Increased accuracy is particularly important in the historical context where US data lack detailed economic measures, which has forced researchers to almost exclusively focus on occupational income scores.

Our estimates suggest that income mobility rose in tandem with human capital mobility through our sample period (see Figure 7). The share of variation in household income scores explained by parental income scores and human capital dropped from 20 percent to 13 percent from the 1870 to the 1920 cohort. We show that both including daughters in the analysis and incorporating parental human capital are important to accurately document this rise in mobility, contrasting with traditional father-son comparisons (see Panel A of Appendix Figure A.7). Consistent with the rise in human capital mobility and the role of schooling documented in the previous section, we find that parental human capital was especially predictive of the child's income before the widespread access of schooling. These results are also consistent with a literature that links expanding school access to increasing income mobility (Mitnik, 2020; van de Werfhorst, 2024).

Our mobility estimates are robust to alternative measures of occupational status (see Panel B of Appendix Figure A.7). A key challenge in historical mobility analysis is the large share of farmers, for whom occupational income scores mask economic heterogeneity. Following Song et al. (2020), we construct human capital-based occupational rankings by computing average literacy rates and education levels for each occupation, birth cohort, race, and region using census data from 1850 to 2010. Assigning individuals cohort-specific percentile ranks based on these averages yields mobility estimates nearly identical to our baseline results. We also find similar trends using traditional occupational income scores that do not account for sex, race, age, or region ("occscore"). Finally, we confirm that the observed rise in mobility is not driven by differences in age at measurement; restricting the sample to children observed at ages 20–29 yields even stronger mobility trends (see Appendix Figure D.6).

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 (8) into

$$R^2 = \widehat{\beta}_p^2 + \widehat{\beta}_m^2 + \widehat{\beta}_f^2 + 2\left(\widehat{\beta}_p\widehat{\beta}_m\hat{\rho}_{p,m} + \widehat{\beta}_p\widehat{\beta}_f\hat{\rho}_{p,f} + \widehat{\beta}_m\widehat{\beta}_f\hat{\rho}_{m,f}\right) \tag{9}$$

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 (see Appendix C.1.2). The latter correlation, $\hat{\rho}_{m,f}$, is a measure of assortative mating based on human capital. Using this decomposition, we compute the counterfactual mobility $(1 - R^2)$ holding a given parameter constant over time.

Our decomposition shows that the evolving role of maternal human capital ($\widehat{\beta}_m$) accounts for the rise in intergenerational income mobility over time (see Panel B of Figure 7). Specifically, mobility would have decreased 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 (Ferrie, 2005; Long and Ferrie, 2013; Feigenbaum, 2018; Song et al., 2020). 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).

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 D.7). For instance, while patters in assortative mating decreased before

1880 and remained constant after (see Appendix Figure D.8), these changes played a negligible role for intergenerational mobility.

We also document that the predictive power of parental background varies across children of different sex and race (see Appendix Figure A.8). Daughters generally exhibit higher mobility than daughters, with the share of variation in household incomes explained by parental background being 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 (Jácome et al., 2021; Buckles et al., 2023b).

Last, having extensively validated our semiparametric latent variable method for human capital mobility in previous sections, we separately validate it for income mobility. Specifically, we identify equation (8) using data from the NLSY79 as well as the PSID, where continuous measures of mother's and children's human capital are observed. Using this data, we compare the estimated $1 - R^2$ after binarizing originally continuous AFQT scores using a range of rank cutoffs with the $1 - R^2$ as estimated by OLS on the continuous variables. We find that regardless of the position in the distribution where the AFQT score is dichotomized, the semiparametric latent variable method accurately estimates mobility (see Appendix Figure D.9).

7. Conclusion

This paper examines the evolution of intergenerational mobility in the US from 1850 to 1950, highlighting the role of maternal human capital and the expansion of mass schooling. We introduce a new measure of mobility that incorporates both parents' human capital, develop a latent variable method to estimate mobility despite historical data limitations, and construct a representative linked panel that includes women despite name changes. Our findings reveal that human capital mobility rose sharply over this period, driven by a declining reliance on maternal human capital, which had been most predictive of child outcomes before widespread schooling. As schooling expanded, it replaced home education, weakening the intergenerational persistence of parental human capital and fueling gains in both human capital and income mobility.

Our findings highlight that high intergenerational mobility in the US is not guaranteed—it depends on the public provision of schooling. Historically, mass education played a key role in weakening the reliance on parental human capital, allowing mobility to rise. Access to quality schooling disproportionately benefits children whose parents have lower levels of human capital, which may not be easily substituted with income or wealth (Løken, 2010; Borra et al., Forthcoming). These findings highlight the continued importance of public investment in education, particularly for children from disadvantaged backgrounds (Goldman et al., 2023).

While schooling played a central role in reducing reliance on parental background, parents likely remained critical in shaping children's early human capital, even as formal education expanded. Theories of skill formation emphasize the complementarity between early parental investments and later schooling (Heckman, 2000, 2006; Cunha and Heckman, 2007; Becker, 2009; Cunha et al., 2010). This dynamic complementarity between early parental inputs and later education points to a lasting importance of family background in determining children's long-term outcomes.

There are several promising avenues for future research. First, our measure of parental background aligns closely with theoretical models emphasizing the distinct roles of parental human capital and income in shaping mobility. However, future work could extend this framework by incorporating additional parental factors such as wealth, occupation, or social networks. Second, given the well-documented importance of neighborhood environments in shaping mobility (e.g., Chetty et al., 2016; Chetty and Hendren, 2018), future research could use our R^2-based measure to incorporate geographic and institutional factors alongside family background. Third, 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, we make our new panel dataset publicly available to facilitate 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.

BJÖRKEGREN, E., M. LINDAHL, M. PALME, AND E. SIMEONOVA (2024): "Decomposing the Parental Education Gradient in Health: Lessons from a Large Sample of Adoptees," *Journal of Labor Economics*, Forthcoming.

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.

BORRA, C., A. COSTA-RAMÓN, L. GONZALEZ, AND A. SEVILLA (Forthcoming): "The Causal Effect of an Income Shock on Children's Human Capital," *Journal of Labor Economics*.

BRITTO, D. G. C., A. FONSECA, P. PINOTTI, B. SAMPAIO, AND L. WARWAR (2024): "Intergenerational Mobility in the Land of Inequality," Working Paper.

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`.

CHEN, Y., S. NAIDU, T. YU, AND N. YUCHTMAN (2015): "Intergenerational Mobility and Institutional Change in 20th Century China," *Explorations in Economic History*, 58, 44–73.

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.

——— (2001): "The Human-Capital Century and American Leadership: Virtues of the Past," *The Journal of Economic History*, 61, 263–292.

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

——— (2016): *Human Capital*, Heidelberg, Germany: Springer Verlag.

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.

GOLDMAN, B., J. GRACIE, AND S. R. PORTER (2023): "Can Individualized Student Supports Improve Economic Outcomes for Children in High Poverty Schools?" Working paper.

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.

LØKEN, K. V. (2010): "Family income and children's education: Using the Norwegian oil boom as a natural experiment," *Labour Economics*, 17, 118–129.

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.

MITNIK, P. A. (2020): "Intergenerational Income Elasticities, Instrumental Variable Estimation, and Bracketing Strategies," *Sociological Methodology*, 50, 1–46.

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, M. SOBEK, D. BACKMAN, A. CHEN, G. COOPER, S. RICHARDS, R. RODGERS, AND M. SCHOUWEILER (2024a): "IPUMS USA: Version 15.0," dataset: `https://doi.org/10.18128/D010.V15.0`.

RUGGLES, S., M. A. NELSON, M. SOBEK, C. A. FITCH, R. GOEKEN, J. D. HACKER, E. ROBERTS, AND J. R. WARREN (2024b): "IPUMS Ancestry Full Count Data: Version 4.0," dataset: `https://doi.org/10.18128/D014.V4.0`.

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.

van de Werfhorst, H. G. (2024): "Is Meritocracy Not So Bad After All? Educational Expansion and Intergenerational Mobility in 40 Countries," *American Sociological Review*, 89, 1181–1213.

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.

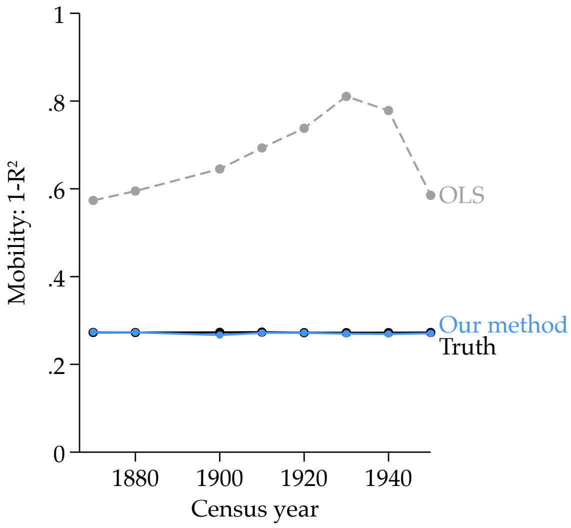
Zheng, A. and J. Graham (2022): "Public education inequality and intergenerational mobility," *American Economic Journal: Macroeconomics*, 14, 250–282.

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

APPENDIX

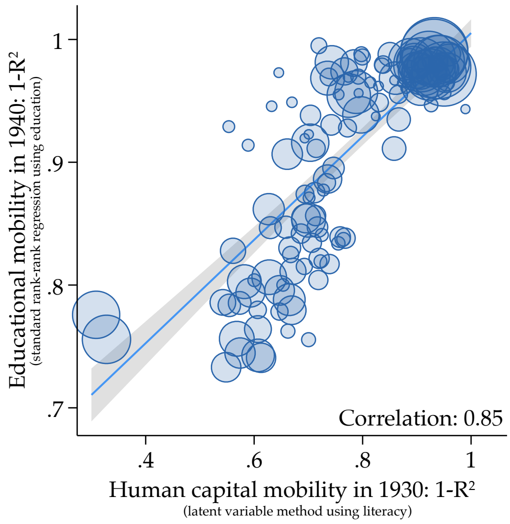
A. APPENDIX FIGURES

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



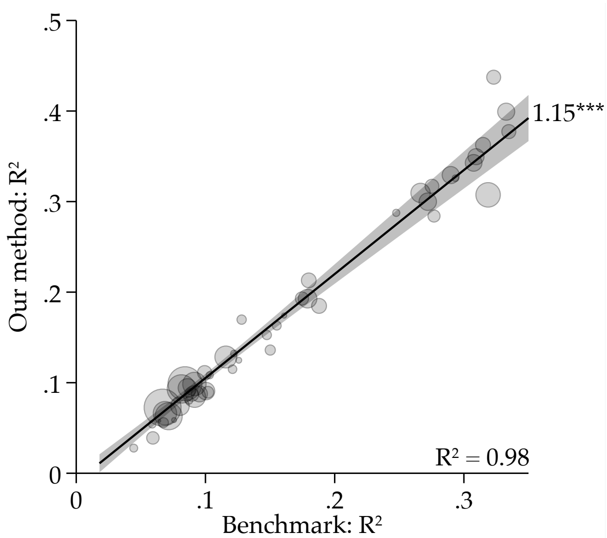
Notes: This figure demonstrates the effectiveness of our semiparametric latent variable method in identifying rank mobility from binary proxies of continuous variables. We simulate jointly normal random variables, dichotomize them such that their distributions reflect historical literacy rates, and use our latent variable method to estimate mobility. The "truth" line represents the continuous rank-rank regression, "our method" uses literacy dummies via our latent variable method introduced in section 3.3, and "OLS" uses the same literacy dummies via standard OLS. In the 1940 and 1950 censuses, we classify individuals who have completed at least two grades of school as literate; others we classify as illiterate.

FIGURE A.2: Validation of Latent Variable Method via Rank-Rank Mobility in Educational Attainment by State, Sex, and Race



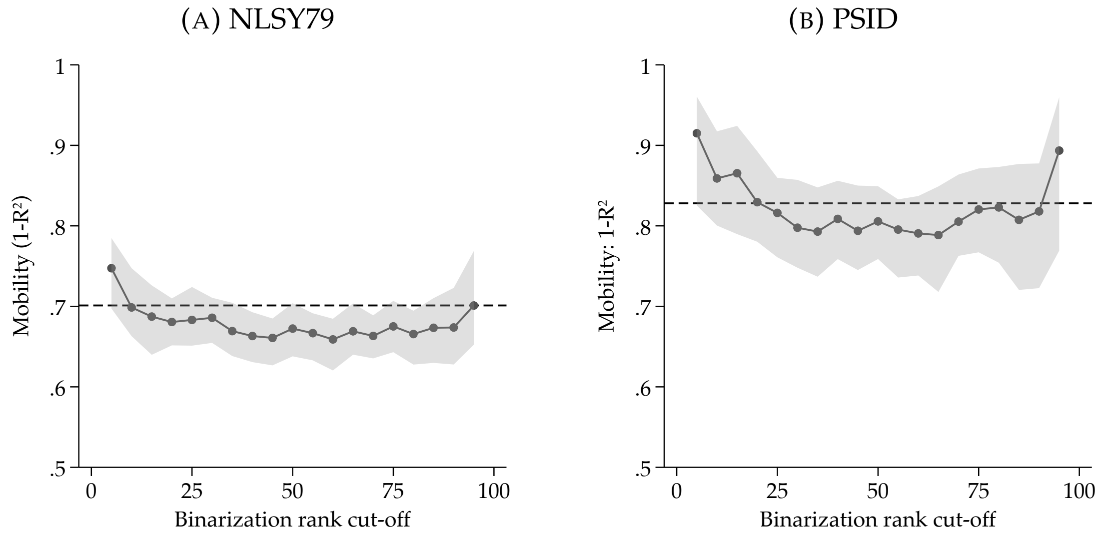
Notes: This figure compares estimates of human capital mobility based on 1930 literacy data and our semiparametric latent variable method introduced in section 3.3 with estimates of educational mobility based on 1940 educational attainment and standard rank-rank regressions. Estimates are separate by state, sex, and race, each bubble's size corresponding to the groups sample size. All results are based on the census cross-section of children ages 13–16 in their parents' household.

FIGURE A.3: Validation of Latent Variable Method via Dichotomization of Education



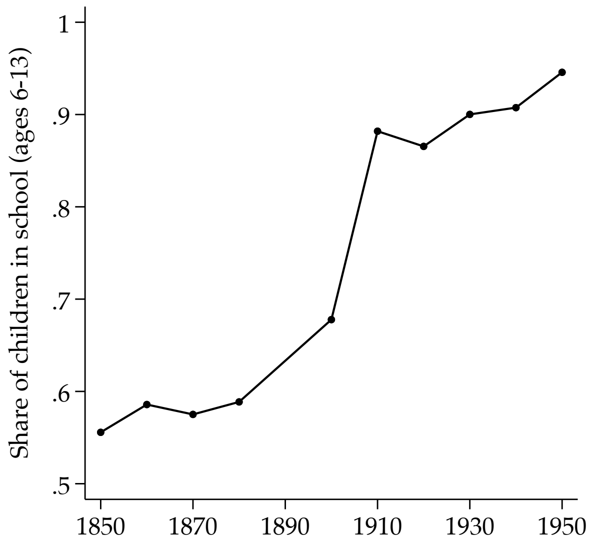
Notes: This figure contrasts the R^2 values from rank-rank regressions using actual and dichotomized educational data from the 1940 census. We dichotomize the data by arbitrarily categorizing individuals based on their educational attainment: more than 11 years for children, 9 for mothers, and 7 for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

FIGURE A.4: Validation of Latent Variable Method via Arbitrary Dichotomization of Continuous Test Scores



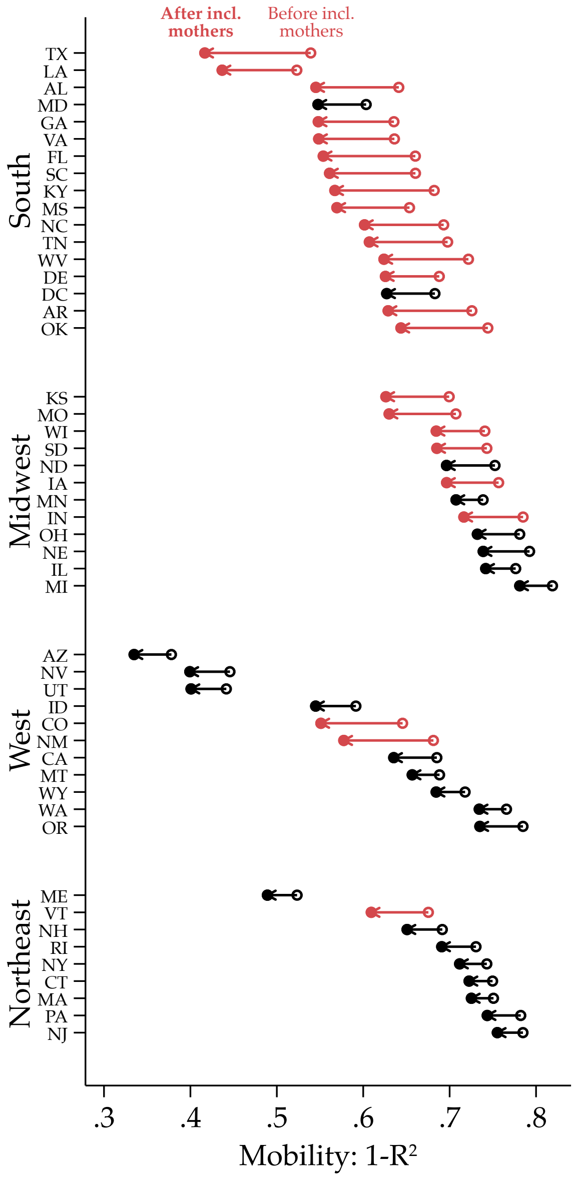
Notes: This figure shows the accuracy of the latent variable method in estimating equation (6) in the NLSY79 and the PSID. Dashed lines represent the estimated $1 - R^2$ on the observed continuous cognitive test measures. The solid lines represent estimates after dichotomization of the mother's and child's score, using varying cutoffs for the child and the median for the mother's cutoff. Shaded area are 95% bootstrapped confidence intervals. In the NLSY, mothers' cognitive test scores are the Armed Forces Qualification Test (AFQT); for children we use the average across scores for reading recognition, reading comprehension, math, vocabulary, and memory. In the PSID, mothers' cognitive test scores are the passage comprehension test; for children, we use the average across scores for letter word identification, applied problems, and broad math.

FIGURE A.5: The Rise of Mass Schooling



Notes: This figure shows the share of children aged 6–13 who attend school across time. We account for the fact that in 1850 and 1860, enslaved children (not recorded in the census) could not attend school.

FIGURE A.6: Human Capital Mobility Before vs. After Incorporating Mothers

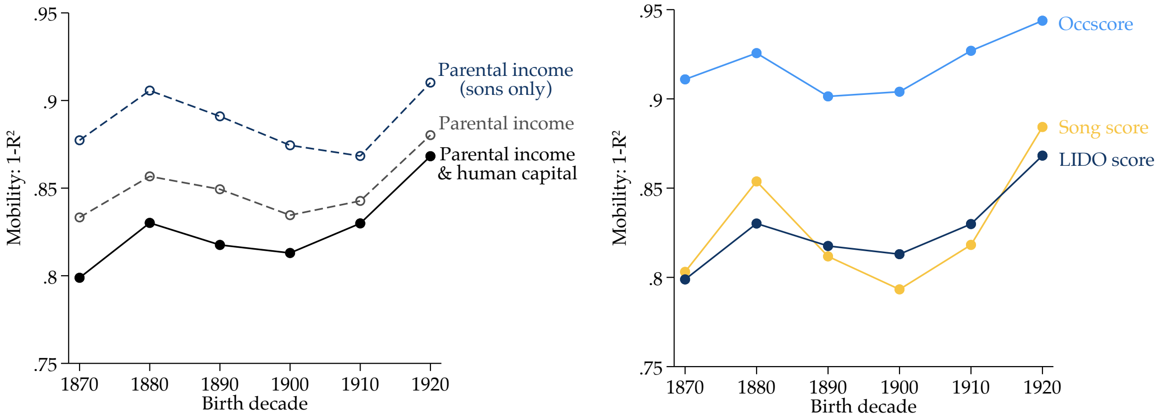


Notes: This figure shows the share of the variance in a child's (latent) human capital rank left unexplained by their (1) father's or (2) father's and mother's (latent) human capital rank ($1 - R^2$) across states. States with above-median changes are displayed in red. Each estimate is the average mobility ($1 - R^2$) 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 3.3.

FIGURE A.7: Share of Variation in Income Explained by Parental Background

(A) Mobility Across Parental Inputs

(B) Mobility Across Income Scores

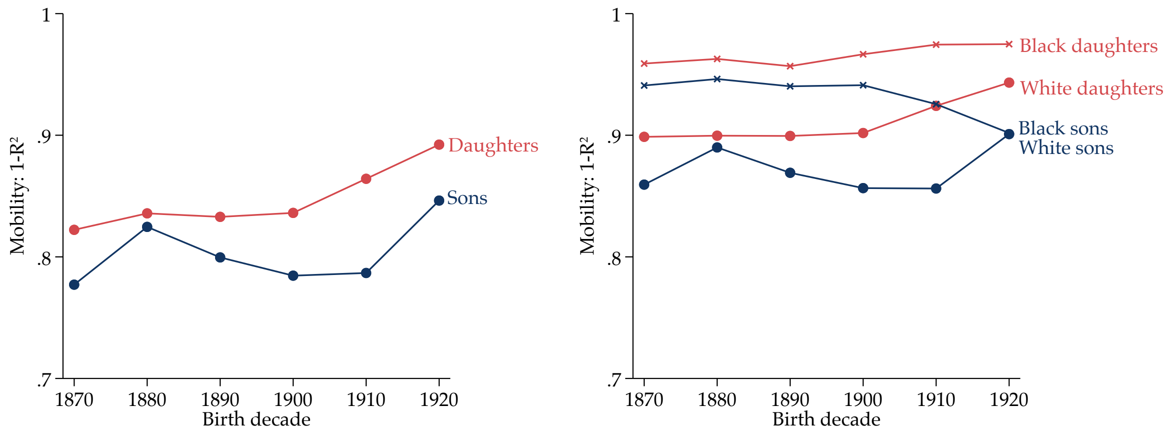


Notes: Panel (A) of this figure shows the share of the variance in a child's household income rank left unexplained by (1) parental income and human capital (2) parental income alone, and (3) parental income alone for sons only. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.3. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020), residualized by a quadratic function of age. Panel (B) shows that the trends in mobility are similar for two popular alternative occupational income measures: occupational income scores ("occscore") and Song scores (Song et al., 2020). Results are based on our new panel and sample weights are applied (see Appendix E.2).

FIGURE A.8: 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 left unexplained by parents' household income ranks and their (latent) human capital ranks ($1 - R^2$) across cohorts and groups. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.3. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020), residualized by a quadratic function of age. Results are based on our new panel and sample weights are applied (see Appendix E.2).

B. Appendix Tables

Table B.1: Mothers & Schools—Robustness to Measures of Schooling

	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$
Baseline measure of schooling	-0.18***	0.04	-0.20***			
	(0.03)	(0.05)	(0.03)			
Refined measure of schooling				-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 schooling and parents' contributions to child human capital. Columns 1–3 (baseline) contain the results from Figure 6. For this baseline, schooling reflects 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 schooling. 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	-0.23***	-0.92***	-0.73***	-0.92***
compulsory schooling laws	(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 schooling and mother's contribution to child human capital. The outcome variable is mother's contribution to R^2. In columns 2 and 4, schooling 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:

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

$$R^2 = \text{cor}(X,Y)^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}(X)}{\text{Var}(Y)}, \tag{11}$$

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 $\beta_1, \ldots, \beta_k$ and the variance-covariance matrix of the explanatory variables:

$$R^2 = \frac{Var\left(\sum_{j=1}^{k} \widehat{\beta}_j X_{i,j}\right)}{Var(Y)} = \frac{\sum_{j=1}^{k} \widehat{\beta}_j^2 Var(X_j) + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} \widehat{\beta}_j \widehat{\beta}_l Cov\left(X_j, X_l\right)}{Var(Y)}. \tag{12}$$

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

$$R^2 = \sum_{j=1}^{k} \widehat{\beta}_j^2 + 2\sum_{j=1}^{k-1}\sum_{j=i+1}^{k} \widehat{\beta}_j \widehat{\beta}_l \widehat{\rho}_{j,l} \tag{13}$$

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:

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

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

$$\phi_j = \frac{\Delta_j}{k!}, \tag{14}$$

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

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.

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

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
$\varnothing$	0.0	Father	0.08	Father, Mother	0.25	$0.08-0=0.08$	$0.25-0.08=0.17$
$\varnothing$	0.0	Mother	0.15	Father, Mother	0.25	$0.25-0.15=0.10$	$0.15-0=0.15$
				Shapley Value (ϕ_j)		$\frac{0.08+0.1}{2!}=0.09$	$\frac{0.17+0.15}{2!}=0.16$

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 (14), we have

$$\phi_1 = \frac{1}{2}\left(R^2(\{X_1,X_2\}) - R^2(\{X_2\}) + R^2(\{X_1\}) - R^2(\{\varnothing\})\right).$$

Further, using equation (12), 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 mothers' input in a univariate regression and $\widehat{\beta}_1^2$ the coefficient on mothers' input in the multivariate regression including the fathers' 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

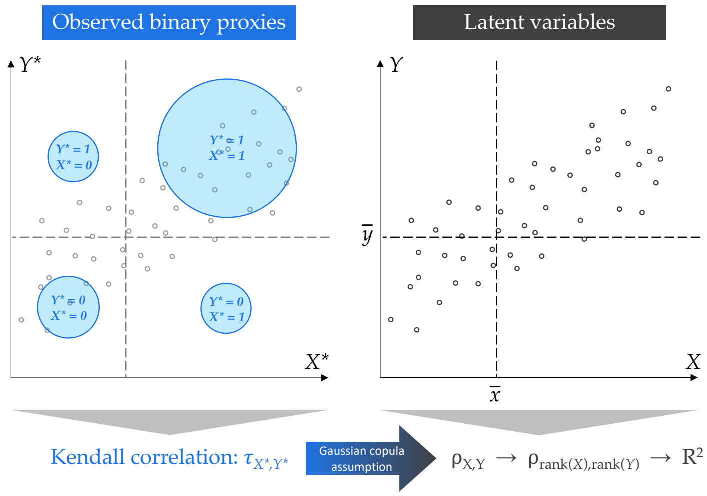
$$\begin{aligned}\phi_1 &= \widehat{\beta}_1^2 + \frac{1}{2}\left(\widehat{\beta}_2^2 - \widehat{\beta}_1^2\right)\left(\frac{Cov(X_1,X_2)}{Var(Y)}\right)^2 + \widehat{\beta}_1\widehat{\beta}_2\frac{Cov(X_1,X_2)}{Var(Y)} \\ &= \widehat{\beta}_1^2 + \frac{\widehat{\rho}_{1,2}^2}{2}\left(\widehat{\beta}_2^2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

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).

We assume that the dependent and independent variables are drawn from a joint Gaussian copula—i.e., 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 $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$.

FIGURE C.1: Illustrating the semiparametric Latent Variable Method



Notes: This figure illustrates the latent variable method, recovering rank-rank mobility ($1 - R^2$) in latent variables from observed binary proxies. Assuming that the underlying latent variables are drawn from a joint Gaussian copula distribution, pairwise rank correlations can be identified from Kendall's correlation between the observed binary proxies using the bridging function in (17). Rank-rank regressions can be identified from the pairwise correlation matrix using equations and (18) and (19).

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: Two rank variables. 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

$$\Sigma_{12} = \sin\left(\frac{\pi}{2}\tau_{12}\right). \tag{15}$$

Therefore, our estimate $\hat{\Sigma}_{12}$ is the sample equivalent of equation (15).

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

$$\tau_{12} = 4\Phi_2\left(\Delta_2, 0, \frac{\Sigma_{12}}{\sqrt{2}}\right) - 2\Phi\left(\Delta_2\right) \tag{16}$$

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 $\overline{X}_1^* \equiv \frac{1}{n}\sum_{i=1}^{n} X_{1i}^*$. As equation (16) is strictly increasing in Σ_{12} (see Fan et al., 2017), Σ_{12} is identified as the unique root of equation (16) where τ_{12} and Δ_2 are replaced with their finite sample analogues.

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

$$\tau_{12} = 2\Phi_2\left(\Delta_1, \Delta_2, \Sigma_{12}\right) - 2\Phi\left(\Delta_1\right)\Phi\left(\Delta_2\right). \tag{17}$$

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 (17) 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,

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

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

$$R^2 = \widehat{\boldsymbol{R}}_{xy}'\left(\widehat{\boldsymbol{R}}_x\right)^{-1}\widehat{\boldsymbol{R}}_{xy}. \tag{19}$$

Equations (18) and (19) 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 (18) and (19), we also see the relation between the slope coefficient and R^2 in the univariate case discussed in Appendix C.1.1: $\hat{\beta} = \sqrt{R^2}$.

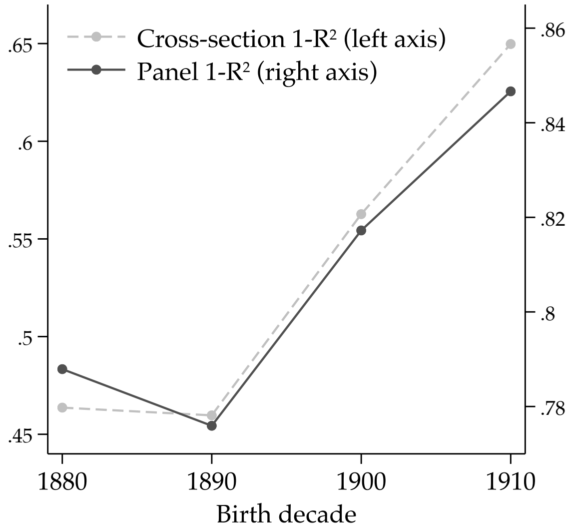
Online Appendix

America's Rise in Human Capital Mobility

Lukas Althoff
Harriet Brookes Gray
Hugo Reichardt

D. Online Figure Appendix

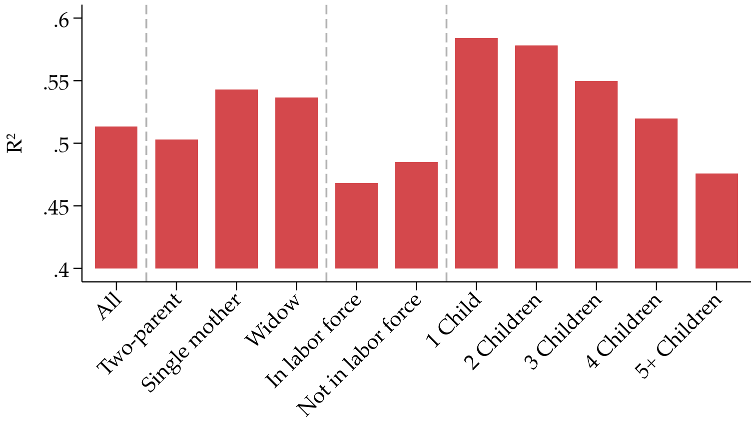
FIGURE D.1: Human Capital Mobility in Panel vs. Cross-Section



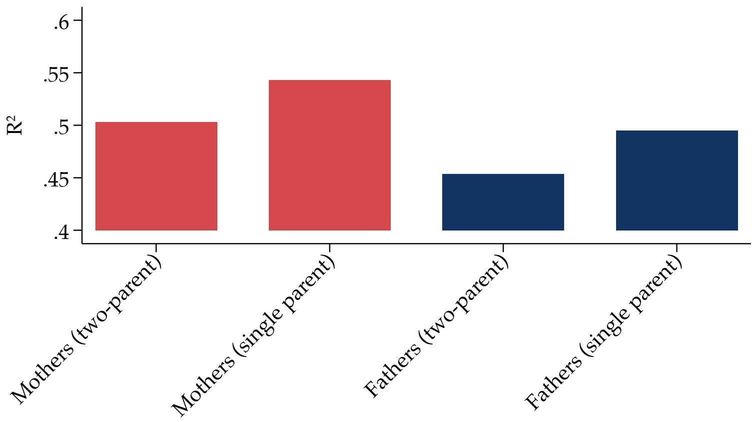
Notes: This figure compares our baseline results of human capital mobility from the cross-section of children who live with their parents to estimates based on our new panel. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.3. 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 and apply sample weights.

FIGURE D.2: Maternal Human Capital's Predictive Power

(A) Maternal Human Capital by Family Type

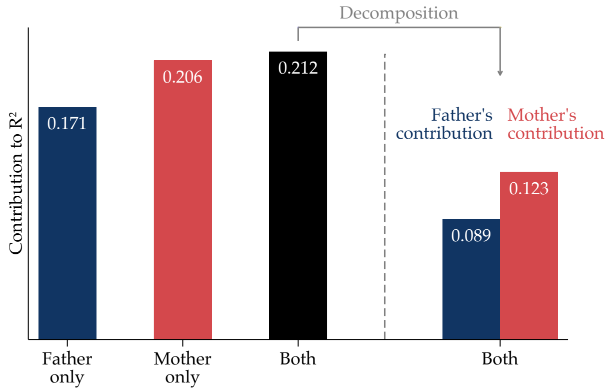


(B) Maternal vs. Paternal Human Capital



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 3.3. 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.

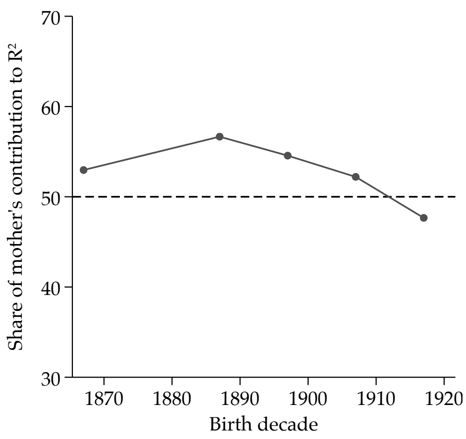
FIGURE D.3: Illustrating our Decomposition Method



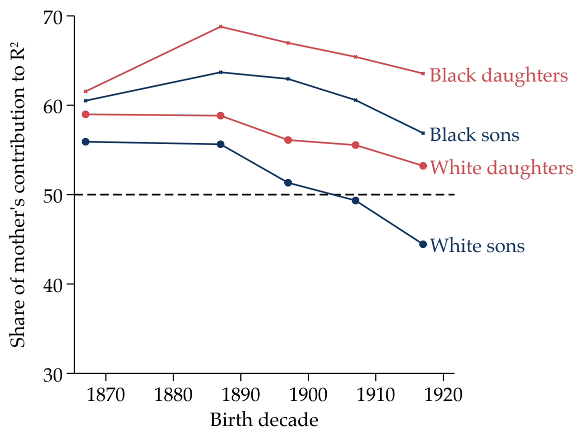
Notes: This figure shows the share of variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank transmission using information on literacy and the latent variable method introduced in section 3.3. We decompose the overall 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 E.2).

FIGURE D.4: Mothers' Relative Contribution to Human Capital Transmission

(A) Mothers Account for Majority of R^2

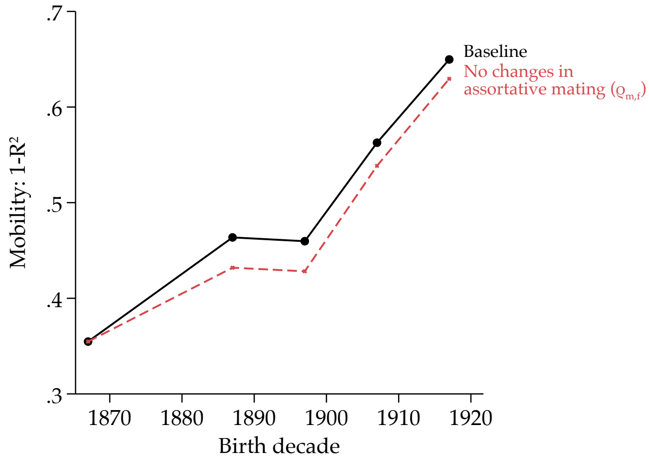


(B) Heterogeneity by Child's Race and Sex



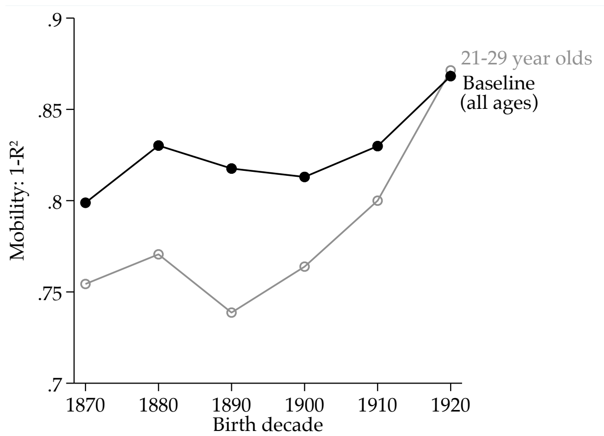
Notes: This figure shows mothers' relative contribution to the overall R^2 using the Shapley-Owen method introduced in Section 3.4. Panel A shows that mother is tend to account for the majority of parental human capital's predictive power, exceeding that our fathers'. Panel B shows that there is substantial heterogeneity in mothers' relative contribution across sex and race, with disproportionate contributions to female and Black children's human capital. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.3. Results are based on the census cross-section of children ages 13–16 in their parents' household.

FIGURE D.5: Changes in Assortative Mating and Human Capital Mobility



Notes: This figure shows that changes in parental assortative mating had a negligible impact on the evolution of human capital mobility. Mobility is measured as the share of the variance in a child's (latent) human capital rank left unexplained by parents' (latent) human capital ranks ($1 - R^2$) across cohorts. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.3. Results are based on the census cross-section of children ages 13–16 in their parents' household.

FIGURE D.6: Mobility Increase is Not Driven by Changing Age Composition

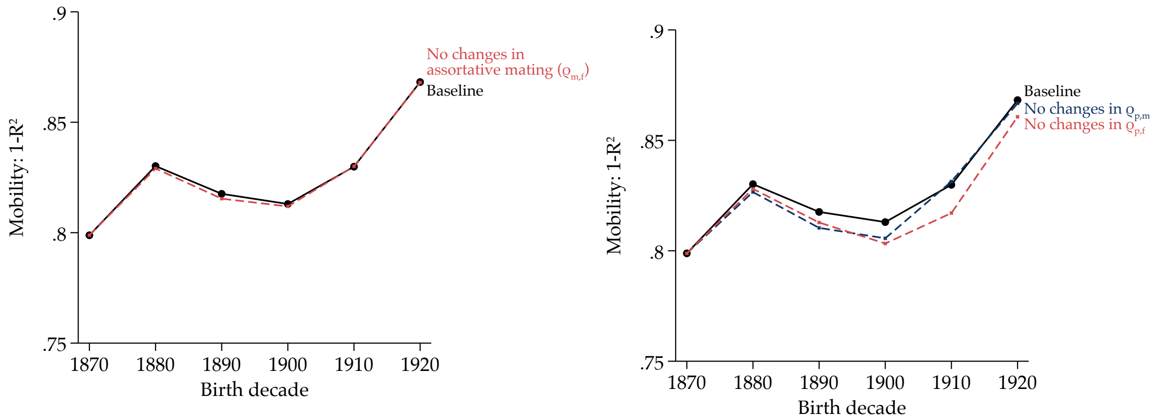


Notes: This figure shows the share of the variance in a 21-29 year old child's household income rank left unexplained by parents' household income ranks and their (latent) human capital ranks ($1 - R^2$). For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.3. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020), residualized by a quadratic function of age. Results are based on our new panel and sample weights are applied (see Appendix E.2).

FIGURE D.7: Mobility and the Impact of Evolving Parental Input Correlations

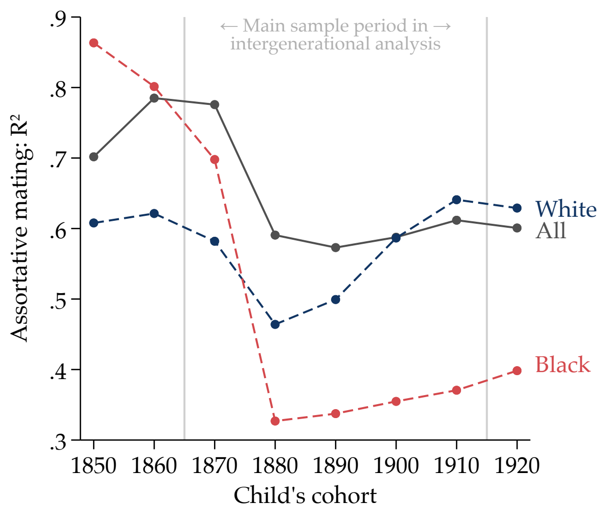
(A) Assortative Mating

(B) Correlations in Human Capital and Income



Notes: This figure shows the impact of each parameter on the mobility estimate $(1 - R^2)$ in equation (8). The baseline represents the observed mobility shown in Figure 7. The other three lines represent the counterfactual mobility, had the respective parameter not changed over time, computed using the decomposition in equation (9). $\rho_{m,f}$ is the correlation between mothers' and fathers' human capital ("assortative mating"); $\rho_{p,m}$ and $\rho_{p,f}$ are the correlations between parental income and mothers' and fathers' human capital, respectively. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.3. 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 E.2).

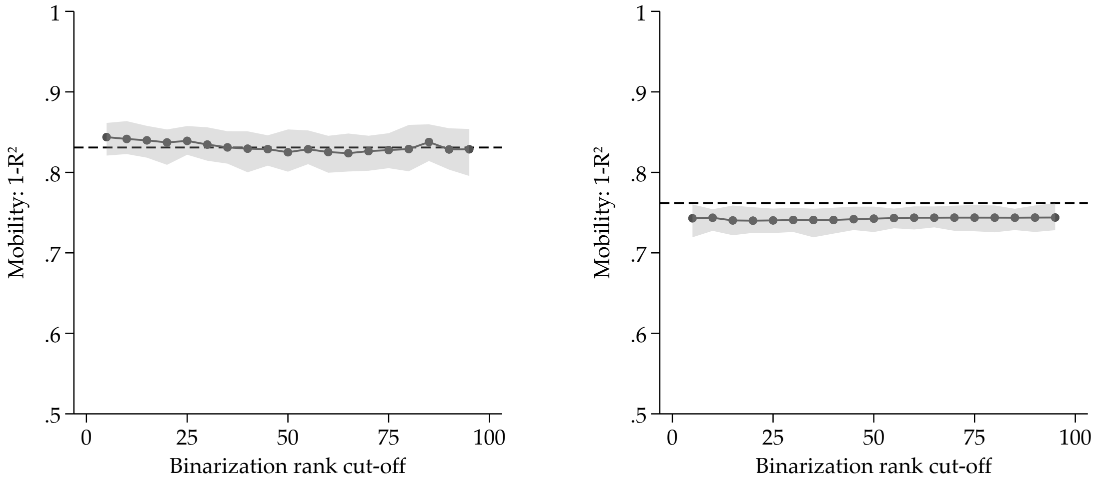
FIGURE D.8: Assortative Mating by Group



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 3.3. 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 D.9: Validation of Income Mobility via Latent Variable Method and Arbitrary Dichotomization of Continuous Test Scores

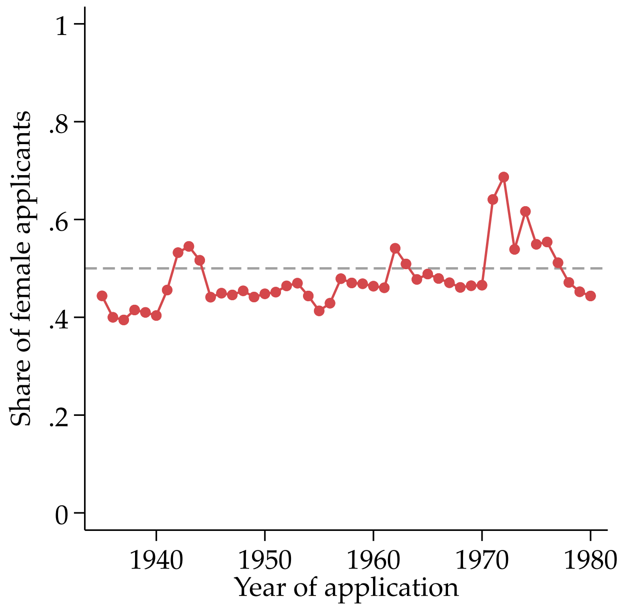
(A) NLSY79 (B) PSID



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 represents the estimated mobility ($1 - R^2$) 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 mobility ($1 - R^2$) after dichotomization 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.

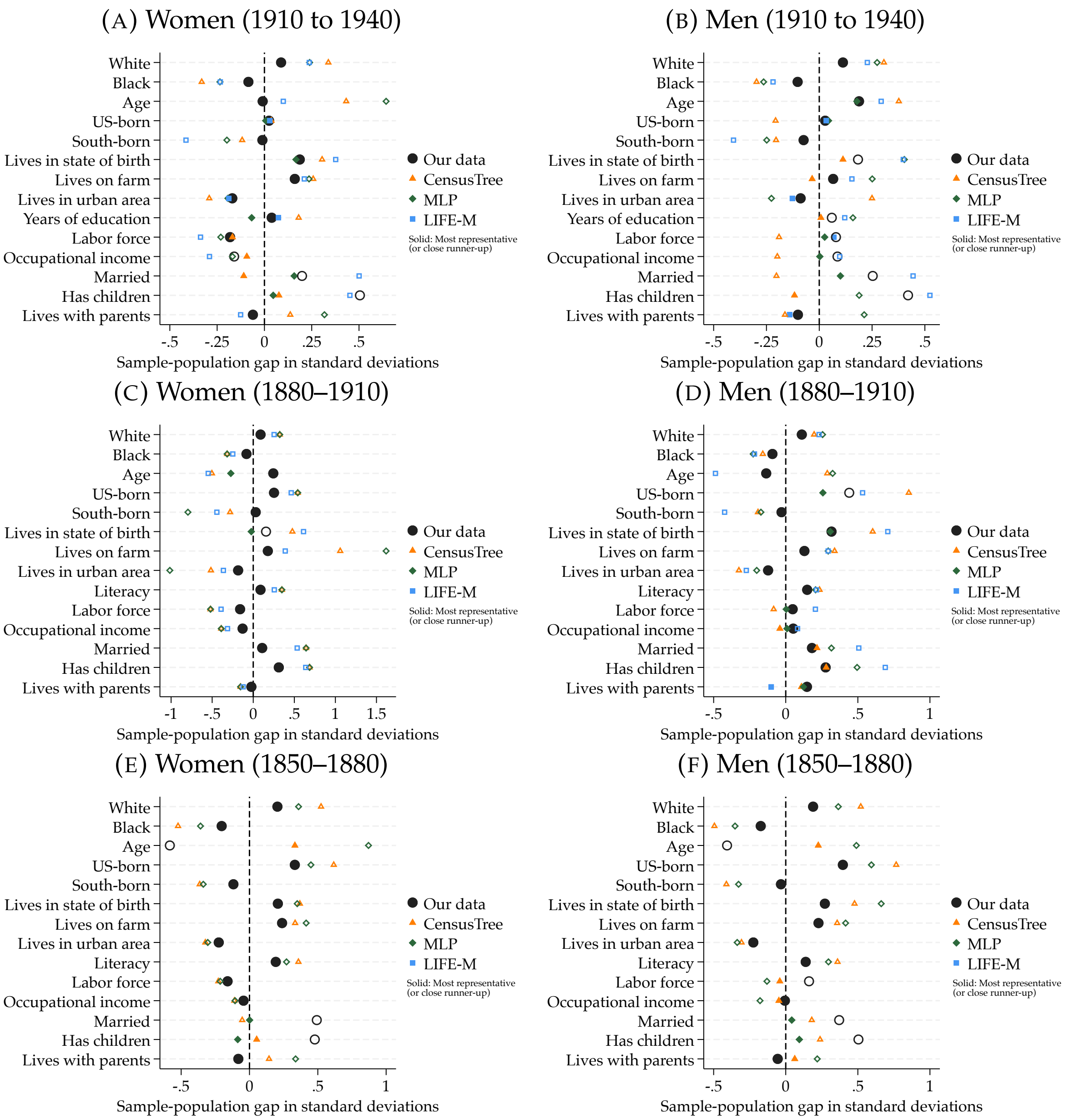
E. Online Data Appendix

Figure E.1: Share of Female Applicants



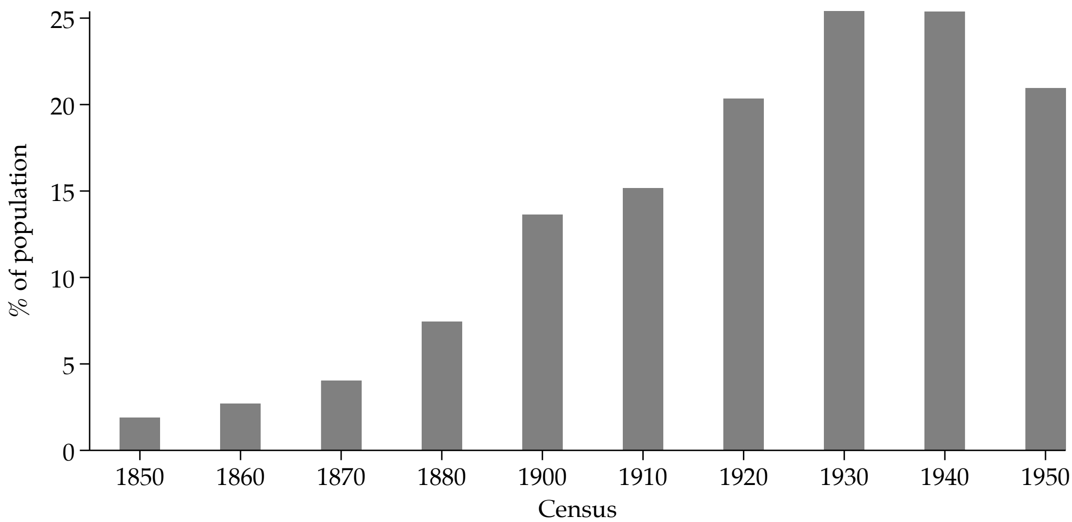
Notes: This figure shows the share of SSN applicants who are female by year of application.

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



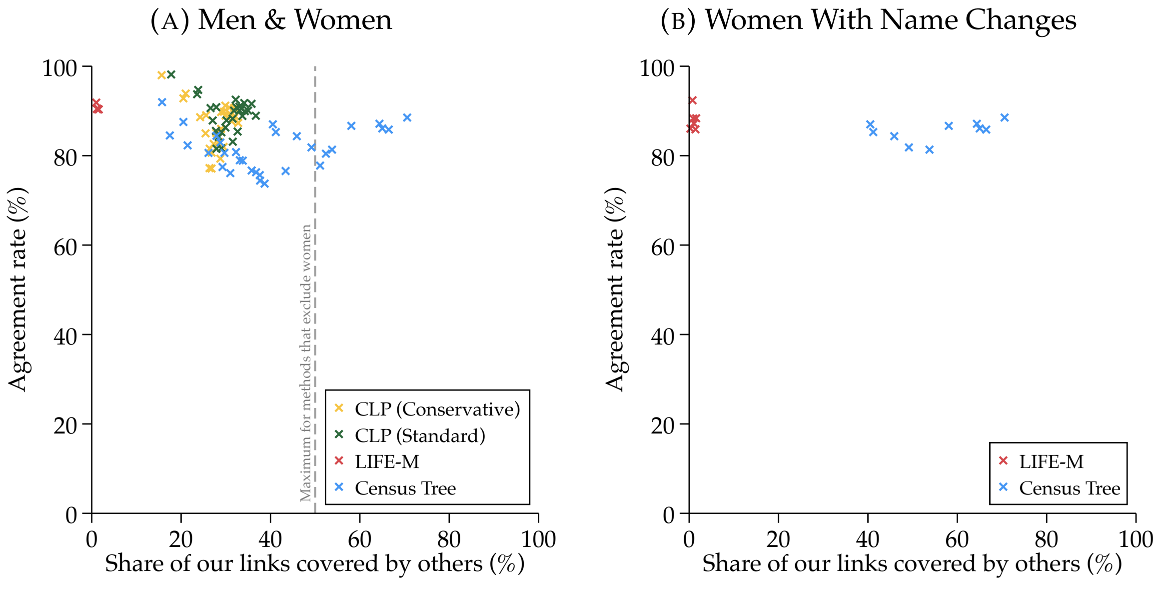
Notes: This figure shows demographic differences between individuals linked from the 1910 to 1940, 1880 to 1910, and 1850 to 1880 censuses and their respective full populations aged 30+ Abramitzky et al. (2024). Each point represents the standardized difference (mean = 0, SD = 1) between linked and full population for a given characteristic. Our sample achieves better average representativeness than existing panels, with average absolute deviations of 0.12 (vs. 0.19 − 0.22) for 1910 to 1940 links, 0.15 (vs. 0.21-0.34) for 1880-1910 links, and 0.24 (vs. 0.28-0.31) for 1850-1880 links, pooling men and women with gender as a characteristic. Comparison panels cover all available datasets that include women: CensusTree (FamilyTree data generated by users of online genealogy; Buckles et al., 2023), MLP (iterative decade links; Helgertz et al., 2023), and LIFE-M (OH/NC only, no pre-1880 links; Bailey et al., 2022).

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



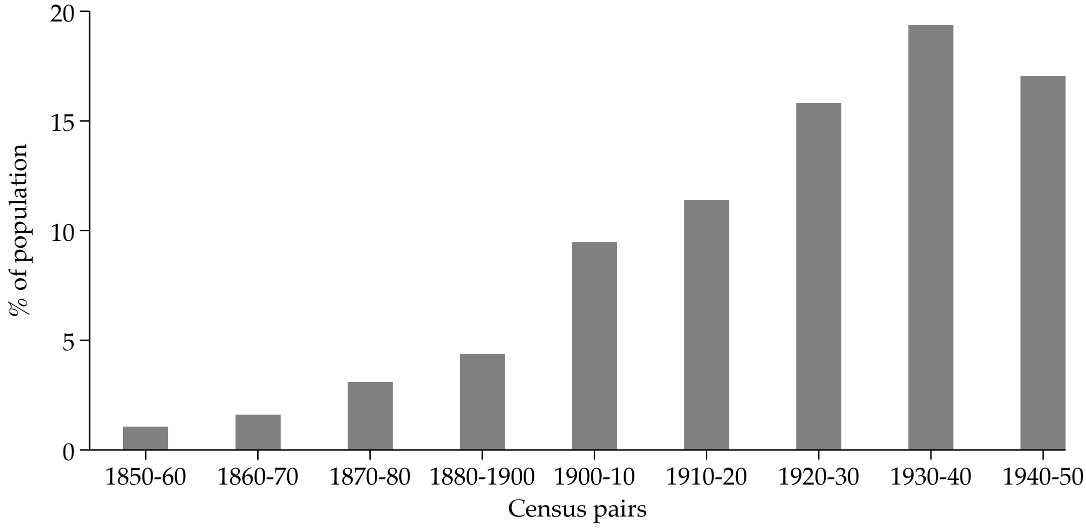
Notes: This figure shows the fraction of the full population of men and women that we successfully assign a Social Security Number (SSN). This includes parents of SSN applicants who did not apply for an SSN themselves and who we assign synthetic identifiers.

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



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

FIGURE E.5: 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.

E.1 Linking Procedure

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

First stage: Applicant SSN ↔ census.

- *Preparing SSN data*: We use a digitized version of the Social Security Number application data from the National Archives and Records Administration (NARA) known as the Numerical Identification Files (NUMIDENT). We harmonize the application, death and claims files to capture all the available information of each SSN record. These data include each applicant's name, age, race, place of birth, and the birth 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)[17].

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

- *Preparing Census data*: Within each census decade from 1850 and 1950, we apply the same name cleaning algorithm used to clean the SSN data. Where available, we extract parent and spouse names from each individual's census record to create crosswalks that are later used in the linking process. Each cleaned census decade is subsequently divided into individual birthplace files for easing the computational intensity of the linking procedure.

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

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

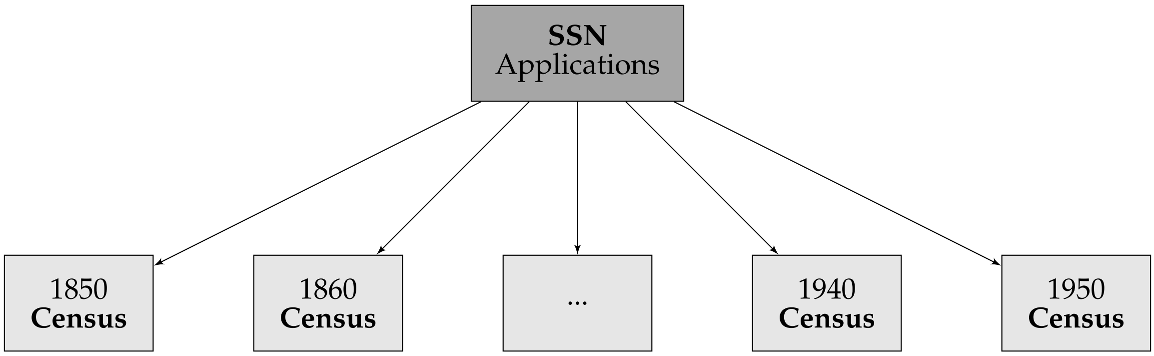
  2. Once we have established our pool of potential matches, we essentially rerun our linking process. However, we use additional matching variables in order to pin down the most likely correct link among the potential matches. In our first round of this process, we aim to pin down the correct link by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

  3. We attempt a second round of the matching described in point 2. for all SSN applicants who were *not* uniquely matched to a census record. In this round, 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 process, removing successfully matched individuals and attempting 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 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 birth and married names. As a result, if we are able to find both records, married women appear in our data twice. We assign these links a slightly altered SSN to differentiate between the married and unmarried SSN-Census link. We do not link married women in the census who are below the age of 16.

FIGURE E.6: First & Second Linking Stages

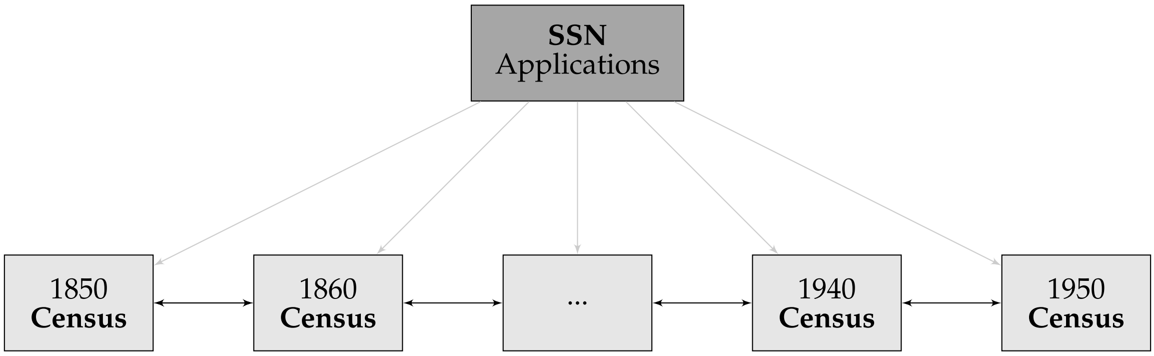


Notes: This figure shows the first and second step of our linking procedure—linking individuals' Social Security Numbers to their census records.

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 birth in the SSN application data, allowing us to link a married mother to her unmarried census record. For parents that we are able to identify in the census from a successful SSN-census link, we apply the same matching procedure described above. However, an important difference is that we do not use parent names (as we no longer have that information), but we are able to use spouse name and information on their parents' birthplace (i.e., the SSN applicant's grandparents birthplace) which is available from the census records. For parents who are not SSN applicants themselves, we create a synthetic identifier similar to an SSN.

Third stage: Census ↔ census. Having assigned unique SSNs or synthetic identifiers to millions of individuals in the census records, we can link these records over time. We cover all possible pairs of census decades from 1850 to 1950.

FIGURE E.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.

E.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 1950 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, 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.

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.

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.