The Missing Link(s): Women and Intergenerational Mobility*

Lukas Althoff† Harriet Brookes Gray‡ Hugo Reichardt§

[Most recent version here]

First version: November 19, 2022.
This version: March 17, 2024.

Abstract

Research on intergenerational mobility in US history has focused on father-son income correlations. We build a new linked census panel to include daughters (1850-1940). To also incorporate the role of mothers, we propose a mobility measure that considers parental human capital alongside income (R^2) and a semi-parametric latent variable method to estimate this measure from historical data. Our approach reveals increasing mobility, overturning conclusions based on income alone. Mothers' human capital was more predictive than fathers' and accounted for the increase in mobility. Aligning with their historical role in homeschooling, mothers were especially important when school access was limited.

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

†Stanford Institute for Economic Policy Research, Stanford University. lalthoff@stanford.edu
‡Department of Economics, Yale University. harriet.brookesgray@yale.edu
§Department of Economics, London School of Economics. h.a.reichardt@lse.ac.uk

1. Introduction

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

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

By constructing one of the first linked census panels to include women, we trace the parental backgrounds of sons and daughters. We overcome the challenge of linking women's census records despite name changes by leveraging historical administrative data from Social Security Number applications. These applications provide both married and maiden names for applicants' mothers and married female applicants. Using these data, we link the census records of 21 million women along with a similar number of men, resulting in a highly representative panel. We will make this dataset publicly available.

We also develop a novel methodology to account for multiple dimensions of parental background in the intergenerational analysis. To assess the joint importance of mothers and fathers, we propose measuring intergenerational mobility as the share of variation in child outcomes explained by parental background: R^2. Unlike traditional mobility measures, such as the parent-child coefficient, this measure accommodates multiple parental inputs. We show that the R^2 has many desirable properties and—in the special case of using only one parental input—has a one-to-one relationship with the rank-rank coefficient. Another advantage of R^2 is that it can be separated into each parent's predictive power using a statistical decomposition method (Shapley, 1953; Owen, 1977).

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

example, literacy can serve as a proxy for human capital. We extensively validate this method and discuss the assumptions it imposes on the joint distribution of parent and child outcomes.

Our first main finding is that intergenerational mobility increased from the 19th to the early 20th century, challenging previous evidence. Specifically, we find that parents' backgrounds, incorporating human capital alongside income, became less predictive of their children's income over time. The separate importance of parental human capital and income is a central aspect of intergenerational mobility theory (Becker et al., 2018), but prior empirical studies focus on income-to-income transmission alone.

Our second main finding is that maternal human capital is the main driver of increasing intergenerational mobility over time. The predictive power of mothers' human capital initially exceeded fathers', but it gradually declined to make both parents' contributions comparable. Decomposing our R^2 measure, we show that mobility would have decreased had it not been for the diminishing predictive power of mothers' human capital. This finding highlights mothers' key role in intergenerational mobility and shows that previous evidence of declining mobility is due to a focus on paternal factors.[1]

As a potential mechanism for the historically large and declining role of maternal human capital, we explore the shift from home-education to formal schooling. Until around 1900, public schooling was limited in many places and home education was common. Historians have highlighted the pivotal role of parental human capital in child development during this period (Kaestle and Vinovskis, 1978). Mothers, who primarily engaged in home production in this era, were key educators of their children (Dreilinger, 2021). "[T]he middle class mother was advised that she and she alone had the weighty mission of transforming her children into the model citizens of the day" (Margolis, 1984, p. 13). The spread of school access could therefore be a reason why parents' human capital—especially mothers'—became less important and intergenerational mobility increased over time.

We find that, indeed, intergenerational mobility increased with school access and that maternal human capital accounted for this trend. Specifically, mothers' (but not fathers') human capital was more predictive for children whose school access was low due to their race, sex, or place. For example, we find that Black children who lacked equal access to schools during the Jim Crow era relied more on their mother's human capital than white children. Similarly, we find that as school access expanded over time, mothers' predictive power declined. These findings offer an explanation for the importance of maternal human capital in early US history: as the main educators of their time, mothers were key contributors to their children's human capital and, as a consequence, to their broader economic status.

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

This paper deepens our insights into how mothers shaped Americans' life chances throughout history. Earlier studies focused either on father-child correlations (Olivetti and Paserman, 2015; Abramitzky et al., 2021a; Ward, 2023; Craig et al., 2019; Jácome et al., 2021; Buckles et al., 2023b) or the correlation between parents' average status and child outcomes (Chetty et al., 2014b; Card et al., 2022). None of these prior studies assesses mothers' importance in the intergenerational transmission of economic outcomes. Our paper emphasizes mothers' separate role in shaping child outcomes, uncovering that maternal human capital is a stronger predictor than father-based proxies. Espín-Sánchez et al. (2023) develop parametric assumptions under which the role of women in intergenerational mobility can be inferred from the outcomes of male family members. Instead, our methodology overcomes critical measurement issues to estimate women's role in intergenerational mobility directly, allowing us to highlight the mechanisms underlying their impact.

Including mothers in the study of mobility in US history is especially pressing given that evidence from other contexts suggests mothers are key determinants of child outcomes. For Norway, Black et al. (2005) find a child's education is positively impacted by their mother's but not their father's education. García and Heckman (2023) show that programs to increase mothers' parenting skills increase intergenerational mobility. Leibowitz (1974) shows that mothers' education is a strong predictor of child human capital whereas fathers' education is not, which they argue is a result of mothers spending more time with their children than fathers.

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

Lastly, a key contribution of this paper is to construct one of the most extensive and representative panels on intergenerational mobility that includes women, building on the foundations of previous work. Craig et al. (2019) and Bailey et al. (2022) initiated the effort to link women's records by expanding automated record linkage developed for men by Abramitzky et al. (2021b). However, the information they use to do so—historical birth, marriage, and death certificates—are available only for selected states and periods. Buckles et al. (2023b) innovatively use crowd-sourced family trees, leading to vastly

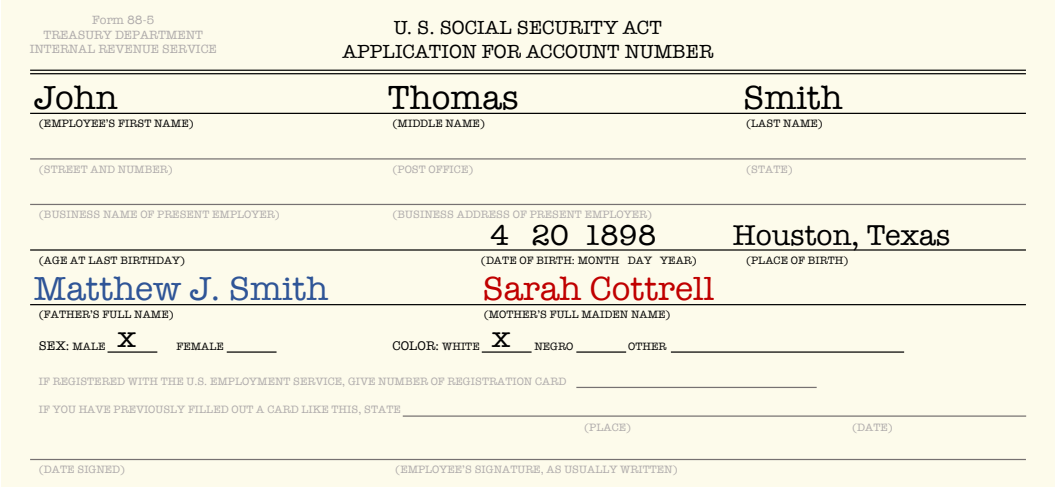
larger sample sizes. In contrast to prior work, we leverage historical *administrative* data, allowing for both scale and representativeness.[2]

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

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

2.1 Historical Administrative Data (Social Security Administration)

FIGURE 1: Social Security Application Form


Form 88-5
TREASURY DEPARTMENT
INTERNAL REVENUE SERVICE

U. S. SOCIAL SECURITY ACT
APPLICATION FOR ACCOUNT NUMBER

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

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

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

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

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

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

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

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

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

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

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

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

Representativeness. Initially, SSN applicants were not representative of the US population, as the SSN system was launched in 1935 to register employed individuals, excluding self-employed and certain other occupations (Puckett, 2009). However, its scope rapidly expanded; for example, Executive Order 9397 in 1943 and the IRS's adoption of SSNs for tax reporting in 1962 increased its coverage to almost 100 percent. Throughout, the share of female applicants has been close to 50 percent (see Appendix Figure D.1). The representativeness of our sample is further improved by parents who enter our sample irrespective of whether they applied for an SSN.

Coverage. The data has extensive coverage of men and women born in the 1880s or after. The majority of Americans born in or after 1915 were assigned an SSN and therefore enter our data as applicants—a fact we establish by comparing each cohort's number of births and SSNs (CDC, 2023; SSA, 2023). The share of Americans with an SSN rises from 64 percent for those born in 1915 to 80 percent for those born in 1920, 90 percent for 1935, and close to 100 percent starting with those born in 1950. The inclusion of parents in the SSN application files extend this coverage further back.

2.2 Census Data

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

2.3 Linking Method

We use a multi-stage linking process to maximize the utility of SSN application data, building on existing methods of automated record linkage (Abramitzky et al., 2021b). This procedure consists of three stages: linking SSN applicants to census records, linking applicants' parents to census records, and tracking census records over time. Appendix D.1 describes our linking procedure in greater detail.

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

establish no link due to the non-uniqueness of the matches.

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

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

Third stage: Census $\leftrightarrow$ census. Having assigned unique identifiers to millions of individuals in the census records, we can link these records over time irrespective of name changes. We cover all possible pairs of census decades from 1850 to 1940. A person only enters the linked census panel if their SSN application record is linked to at least two different census decades.

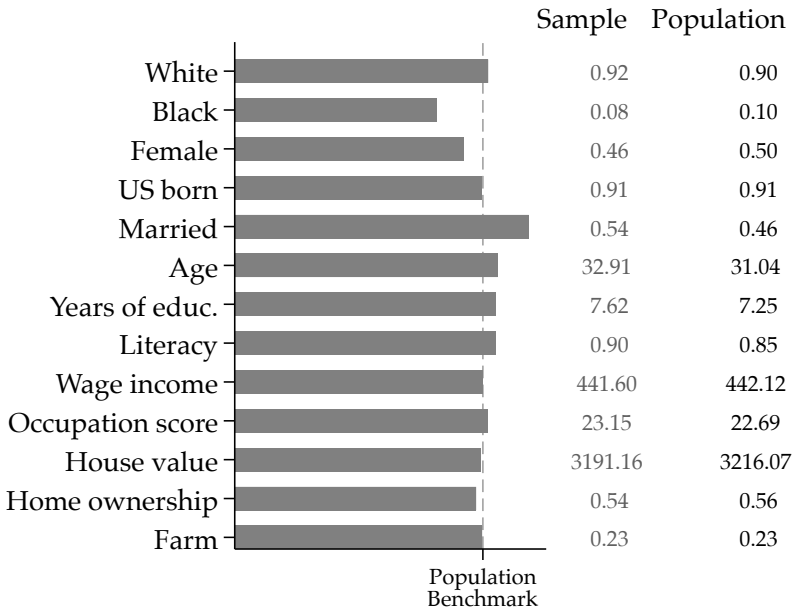
In principle, it would be possible to establish additional links across census records by using standard or machine learning methods. These methods would be particularly useful for men and never-married women, where the issue of name changes does not apply. However, we choose not to use these methods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

2.4 Our New Panel

In the first two stages, our process assigns SSNs to 36 million census records—16 million applicants and 20 million parents. Our linking rate is 40 percent for applicants, surpassing the more typical 25 percent of prior studies thanks to our use of more detailed information, notably parent names. In the third stage, we link 112 million census records

over time, tracking each of the 36 million individuals through more than three census decade pairs on average.

FIGURE 2: Sample Balance Prior to Weighting (1940)



Notes: This figure shows the representativeness of characteristics among individuals in the 1940 census who we successfully assign an SSN compared to the full population in the 1940 census. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Because of the large sample sizes, even economically small differences are statistically significant. In the 1940 census, instead of literacy, we observe the highest year of school or degree completed. We classify individuals who have completed at least two grades of school as literate; others we classify as illiterate.

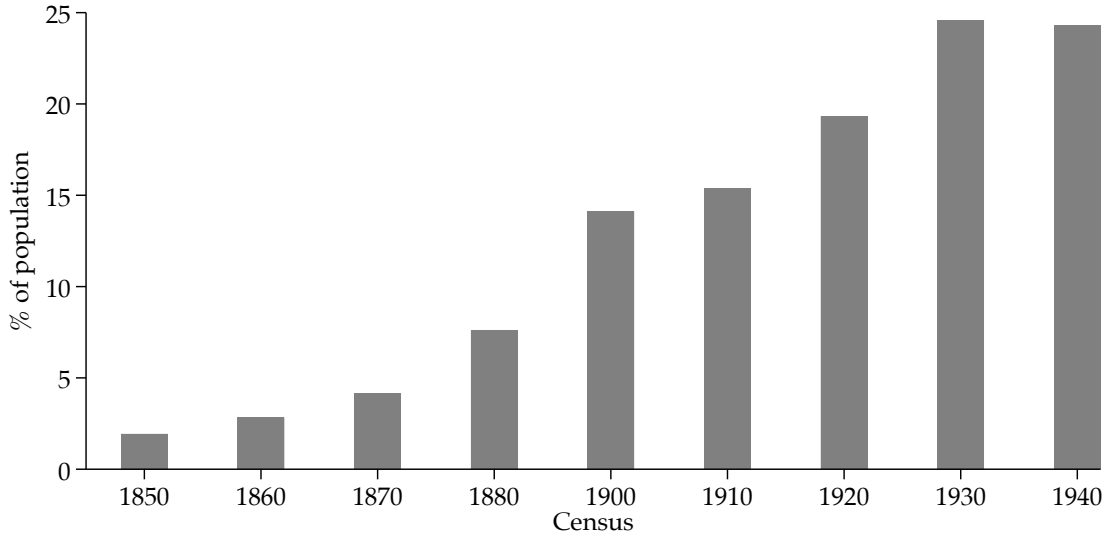
A standout feature of the panel is the inclusion of 12 million women for whom we observe pre- and post-marriage data. The sample sizes are largest for people born between the 1890s and the 1920s, with each birth decade containing 1.5 to 3 million women. These data allows us to overcome critical data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

Our panel is highly representative of the overall US population across several metrics, including gender and race (see Figure 2). Women comprise 46 percent of our linked sample in 1940. The sample mirrors the US-born and foreign-born shares of the population. While Black Americans are slightly underrepresented, our panel exceeds the representativeness of other samples in this dimension as well. Socioeconomic factors like income, home ownership, years of education, and literacy also align well with the broader population. Our sample over-represents married individuals, possibly because we use the names of a person's children or spouse in the linking procedure if they are known to us, improving linking rates for those who have children, a spouse, or both.

We reweight our sample to more closely resemble the US population's characteristics in our empirical analysis.[3] Our reweighted sample is close to perfectly representa-

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

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



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

tive of the full population, even in characteristics not directly targeted by the reweighting method. The panel maintains its representative quality even in the earliest census decades (see Appendix Figure D.2).

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

Compared to existing linked census data, our new panel covers a substantial number of individuals whose records have not previously been linked, while maintaining high agreement rates with existing data for overlapping individuals (see Appendix Figure D.3). Our panel shares the most data with the novel Census Tree—an innovative, extensive panel that includes women through genealogical data (Buckles et al., 2023a). Agreement rates vary from 80 to nearly 100 percent and are highest with LIFE-M—a panel that leverages vital records in the linking process (Bailey et al., 2022).

2.5 Economic Outcomes

To understand the role of mothers and fathers in shaping child outcomes, we require separate measures of each parent's outcomes. We therefore focus on human capital measures, such as literacy or years of education, reflecting the status of both men and women.

To measure parental background, we additionally consider household-level measures such as income. We incorporate household-level alongside individual-level information only when considering the overall importance of parental background, not when we aim

to distinguish mothers' and fathers' separate contributions.

For children, we consider outcomes during both child- and adulthood. During childhood (ages 13–16), we measure literacy (as a proxy for human capital), school attendance, and total years of schooling completed. During adulthood (ages 20–54), we measure literacy, years of education, and occupational income scores.

3. Measuring Intergenerational Mobility with Multiple Inputs

In this section, we propose a statistical model of intergenerational mobility that accounts for the contributions of both fathers' and mothers' human capital to their children's economic outcomes. First, we propose using the R^2 of a regression of child outcomes on multiple parental inputs as a mobility measure that integrates the roles of both parents. Second, we use a simple decomposition method that allows to separate the contributions of mothers and fathers to the overall R^2. Third, we build on a state-of-the-art semiparametric latent variable method to estimate the R^2 from a rank-rank regression when only binary proxies of underlying outcomes are observed (e.g., literacy as a proxy for human capital).

3.1 A Simple Model of Intergenerational Mobility

We build on standard statistical models of intergenerational mobility where a child's economic outcome is a linear function of parental inputs:

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

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

There are several advantages to the rank-rank approach, which considers mobility in relative positions in the distribution (Chetty et al., 2014a). First, correlations in ranks are not affected by changes in the marginal distribution of outcomes which, given the long time horizon of our study, enhances the interpretability of the coefficients. Second, using ranked outcomes ensures that the marginal distributions of mother's and father's outcomes are identical, so that their relative contributions can be effectively compared.

This statistical model differs from most previous research by allowing for multiple parental inputs—most importantly to explicitly incorporate mothers alongside fathers as contributors to a child's outcomes. While in this paper we focus on human capital and

income, the model can be extended to accommodate many different inputs including parents' wealth, grandparents' or other relatives' backgrounds, or neighborhood characteristics.

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

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

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

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

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

The R^2 has a direct relationship with traditional mobility measures—parent-child coefficients or, most commonly, father-son coefficients ($\widehat{\beta}$).[4] In Appendix C.1, we show that in such univariate rank-rank regressions, there is a one-to-one mapping between the parent-child coefficient and our mobility measure: $R^2 = \widehat{\beta}^2$.

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

3.3 Measuring Individual Inputs' Contribution to R^2

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

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

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

where $R^2(T)$ represents the R^2 of regressing the dependent variable (e.g., $\text{rank}\left(y_i\right)$) on a

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

set of variables $T \subseteq V$ (e.g., $V = \left\{\text{rank}\left(y_i^{\text{mother}}\right), \text{rank}\left(y_i^{\text{father}}\right)\right\}$), and k is the number of variables in V (i.e., $k = |V|$). Intuitively, ϕ_j represents the weighted sum of marginal contributions that a parent makes to the variation in child outcomes explained by different combinations of parental inputs. In Appendix C.2, we describe the decomposition method in more detail and, for the special case of two parental inputs, provide a closed-form expression for ϕ_j in (1) in terms of the estimated coefficients and the correlation between the inputs.

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

1. *Additivity.* Individual contributions to the R^2 add up to the total R^2.
2. *Equal treatment.* Regressors that are equally predictive receive equal values.
3. *Monotonicity.* More predictive regressors receive larger values.

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

3.4 Measuring Mobility with Latent Inputs

To estimate rank-rank mobility (R^2) when we only observe binary proxies of the rank variables in equation (1), we propose a method based on Fan et al. (2017). Appendix C.3 discusses the method in detail.

Many binary variables can be interpreted as a function of a continuous underlying latent variable that is equal to one if that variable exceeds an unknown threshold and zero otherwise. In our application, we interpret literacy—the only information on human capital in pre-1940 censuses—as such a proxy for human capital.

Under distributional assumptions, we can use the observed binary proxies to identify the parameters and R^2 in equation (1). Specifically, we assume that parental and child outcomes in equation (1) are drawn from a joint Gaussian copula distribution. That is, we assume that there exists a set of unknown monotonic transformations $f_c, f_{p_1}, \ldots, f_{p_k}$ such that $\left(f_c(y_i), f_{p_1}\left(y_{i,1}^{\text{parental}}\right), \ldots, f_{p_k}\left(y_{i,k}^{\text{parental}}\right)\right)' \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$.[5] We do not require information on the monotonic transformation themselves. Note that because ranks are themselves monotonic transformations, this assumption implies that not only the outcomes but also their ranks follow the Gaussian copula distribution.

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

The Gaussian copula distribution is commonly used in the statistics literature due to its flexibility and good performance in practice (e.g. Liu et al., 2009, 2012; Zue and Zou, 2012). It is a family of probability distributions that includes but is not limited to the normal distribution. For instance, since it includes any monotonic transformation of normally distributed random variables, it allows for skewed and multi-modal distributions. Importantly, the Gaussian copula assumption does not impose that the latent variables of interest (e.g., human capital) are themselves normally distributed.

We show that this semi-parametric latent variable method allows us to estimate the rank-rank regression in equation (1) even if only binary proxies of the rank variables are observed. Specifically, Fan et al. (2017) show how to estimate $\mathbf{\Sigma}$—the correlations between each underlying variable—under such data limitations.[6] $\mathbf{\Sigma}$ in turn identifies the pairwise correlations between the ranked variables. We show that any rank-rank regression is identified by the pairwise correlations, and that therefore $\mathbf{\Sigma}$ is sufficient to identify equation (1) including its R^2. In Appendix C.3, we present an explicit formula for $\widehat{\boldsymbol{\beta}}$ and R^2 as a function of $\widehat{\mathbf{\Sigma}}$.

We extensively validate this method and show that it correctly recovers rank-rank mobility by simulation.

First, when observing rank variables to estimate rank-rank mobility directly, we show that our method correctly identifies mobility even after the rank variables are dichotomized arbitrarily. Specifically, we use ranks in educational attainment from the 1940 census and dichotomize this data. We use different cutoffs for children, mothers, and fathers (e.g., 11 years for children, 9 for mothers, 7 for fathers). Our method's mobility estimates by state align well with those derived from the original, undichotomized data (see Panel A, Appendix Figure A.1). This shows the method's performance in relevant historical data.

Second, we show that the method is robust to cut-offs changing over time, even shifting towards tail ends of the distribution. In our context, an important concern stems from literacy increasing to close to 100 percent over time, changing the information that it contains about a person's human capital rank. To address this concern, we simulate jointly normally distributed data, transform them in ranks, and dichotimize these ranks according to historical literacy rates for each decade from 1870 to 1940. We show that, in contrast to Ordinary Least Squares, our semi-parametric latent variable method yields correct estimates of mobility (R^2) over time, despite changing cut-offs (see Panel B, Appendix Figure A.1).

We apply the semi-parametric latent variable method not only to measuring rank-rank mobility in human capital (through literacy), but also to measuring educational rank-rank mobility (through school attendance at a given age). Because we anticipate

[6]The method in Fan et al. (2017) allows for a combination of binary and continuous variables. It can be extended to non-binary ordinal and truncated variables (Dey and Zipunnikov, 2022). Furthermore, they derive statistical properties of the estimator of $\mathbf{\Sigma}$, notably $\sqrt{n}$-consistency.

this method to be useful for future research facing similar data limitations, we developed a Stata command for easy implementation by others.

4. Income Mobility & Parental Human Capital

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

4.1 Income Mobility Accounting for Parental Human Capital

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

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

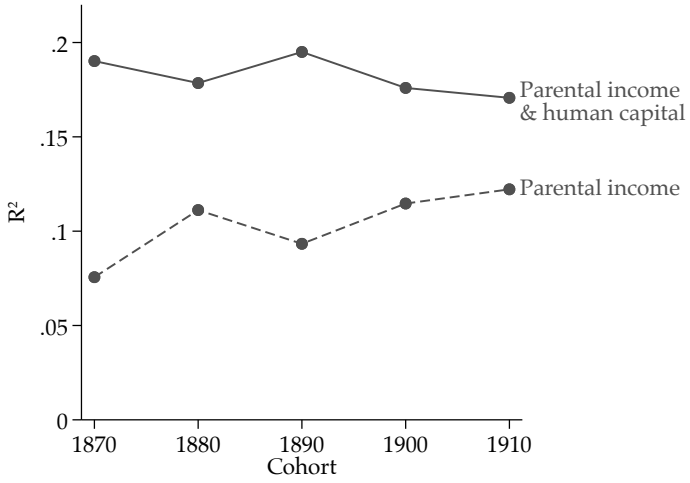
We account for both parental income and human capital by measuring intergenerational mobility as the R^2 in the following version of equation (1):

$$\text{rank}\left(inc_i\right) = \alpha + \beta_p \text{rank}\left(inc_i^{\text{parents}}\right) + \beta_m \text{rank}\left(h_i^{\text{mother}}\right) + \beta_f \text{rank}\left(h_i^{\text{father}}\right) + \varepsilon_i, \quad (2)$$

where inc is household income and h is (latent) human capital. We measure household income as the household head's LIDO occupational income score. Literacy serves as a binary proxy for latent human capital ranks. We estimate this model using the semi-parametric latent variable method described in section 3.4 and our new representative panel dataset described in section 2.4.

We find that parental human capital accounts for a large share of variation in children's incomes, even conditional on parents' incomes (see Figure 4). In some periods, the predictive power of parental background doubles after incorporating human capital. Most importantly, the broader measure of parental background that includes both

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



Notes: This figure shows the share of the variance in a child's household income rank explained by (1) parents' household income ranks and their (latent) human capital ranks (R^2) and (2) parents' household income ranks alone. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied.

income and human capital suggests that intergenerational mobility in the United States increased over time—challenging the conclusion of declining mobility derived from measures based on income alone (Ferrie, 2005; Long and Ferrie, 2013; Feigenbaum, 2018; Song et al., 2020). We document a similar pattern when using more occupational income scores that are not specific to sex, race, age, or region ("occscore"; see Appendix Figure A.2).

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

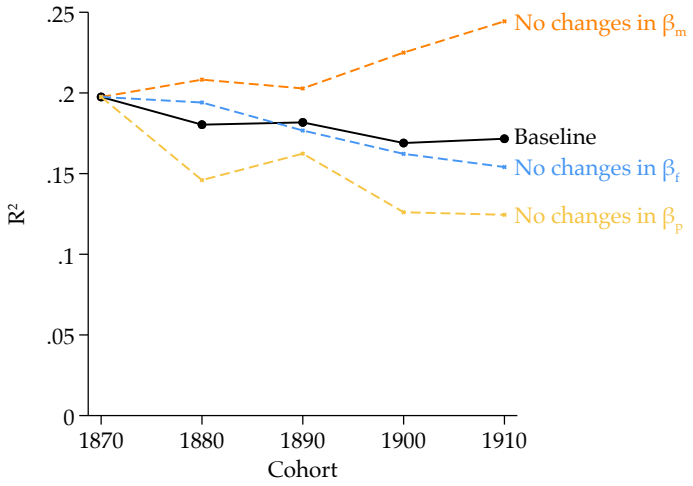
$$R^2 = \widehat{\beta}_p^2 + \widehat{\beta}_m^2 + \widehat{\beta}_f^2 + 2\left(\widehat{\beta}_p\widehat{\beta}_m\hat{\rho}_{p,m} + \widehat{\beta}_p\widehat{\beta}_f\hat{\rho}_{p,f} + \widehat{\beta}_m\widehat{\beta}_f\hat{\rho}_{m,f}\right) \tag{3}$$

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.[7] The latter correlation, $\hat{\rho}_{m,f}$, is a measure of assortative mating based on human capital. Using this decomposition, we compute the counterfactual R^2 holding a given parameter constant over time.

Our decomposition shows that the evolving role of maternal human capital ($\widehat{\beta}_m$) is the main reason why intergenerational mobility increased over time (see Figure 5). Specifically, R^2 would have increased without the changing coefficient of maternal human capital. The importance of father's human capital ($\widehat{\beta}_f$) did not affect mobility significantly. Without changes in the importance of parental income ($\widehat{\beta}_p$) mobility would have

[7] For a similar decomposition of R^2 in a rank-rank regression with an arbitrary number of independent variables, see equation (8) in Appendix C.1.2.

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



Notes: This figure shows the role of each parameter on the R^2 in equation (2). The baseline represents the observed R^2 shown in Figure 4. The other three lines represent the counterfactual R^2, had the respective parameter not changed over time, computed using the decomposition in equation (3). For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied.

increased even further. The rise in $\widehat{\beta}_p$ aligns with decreasing income mobility in previous research. However, we find that the focus of that research on income alone masked important changes in the role of parental background in shaping the outcomes of children (see also Ward, 2023, who documents that accounting for measurement error also reverses the trend).

In contrast to the slope coefficients ($\widehat{\boldsymbol{\beta}}$), none of the correlations between parental inputs ($\widehat{\boldsymbol{\rho}}$)—including assortative mating—had a significant impact on R^2 (see Appendix Figure A.3). For instance, while patters in assortative mating decreased before 1880 and remained constant after (see Appendix Figure A.4), these changes played a negligible role for intergenerational mobility.

Mobility by group. We show that the predictive power of parental background varies considerably across children of different sex and race (see Appendix Figure A.5). Sons generally exhibit lower intergenerational mobility compared to daughters, with R^2 around twice as high for sons as for daughters (around 0.3 versus 0.15). White sons are least mobile, with 13 to 19 percent of variation in household incomes linked to parental background. Black sons are more mobile than white sons, followed by White daughters and Black daughters. Black daughters are not only the most mobile group, they are also the only group whose mobility increased over time. It is important to recognize that (1) high within-group mobility does not imply high mobility within the general population and that (2) high mobility does not necessarily equate to high *upward* mobility.

4.2 The Historical Role of Parental Human Capital

Our finding that parental human capital was important—and especially so in the late 19th century—is consistent with the historical role of parents. Prior to public school access becoming universal in the late 19th and early 20th centuries, parental home education was central for children's human capital development. Even children who were enrolled in school in the late 19th century attended school less than four months a year on average (Dreilinger, 2021).

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

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

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

Motivated by our finding of the importance of maternal human capital for intergenerational mobility and the historical literature, the subsequent analysis studies the specific role of mothers' human capital in shaping their children's outcomes.

5. Mothers & Human Capital Transmission

Motivated by our results in the previous section, we now zero in on the intergenerational transmission of human capital. We find that, mirroring our results on income mobility,

human capital mobility increased significantly from the 1850s to 1910s birth cohorts. We decompose the overall predictive power of paternal human capital into the contributions of mothers and fathers. Our findings show that mothers' human capital more strongly predicts child human capital than fathers'. This difference is particularly pronounced for female and Black children.

5.1 Parental Human Capital and Child Outcomes

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

$$\text{rank}\left(h_i\right) = \delta + \gamma_m \text{rank}\left(h_i^{\text{mother}}\right) + \gamma_f \text{rank}\left(h_i^{\text{father}}\right) + \eta_i, \tag{4}$$

where h is (latent) human capital. We estimate this model using the semi-parametric latent variable method described in section 3.4 and use the census cross-section of children in their parents' households. We then use the Shapley-Owen decomposition described in section 3.3 to separate mothers' and fathers' contributions to predicting children's human capital (see Appendix Figure A.6 for an illustration of the method).

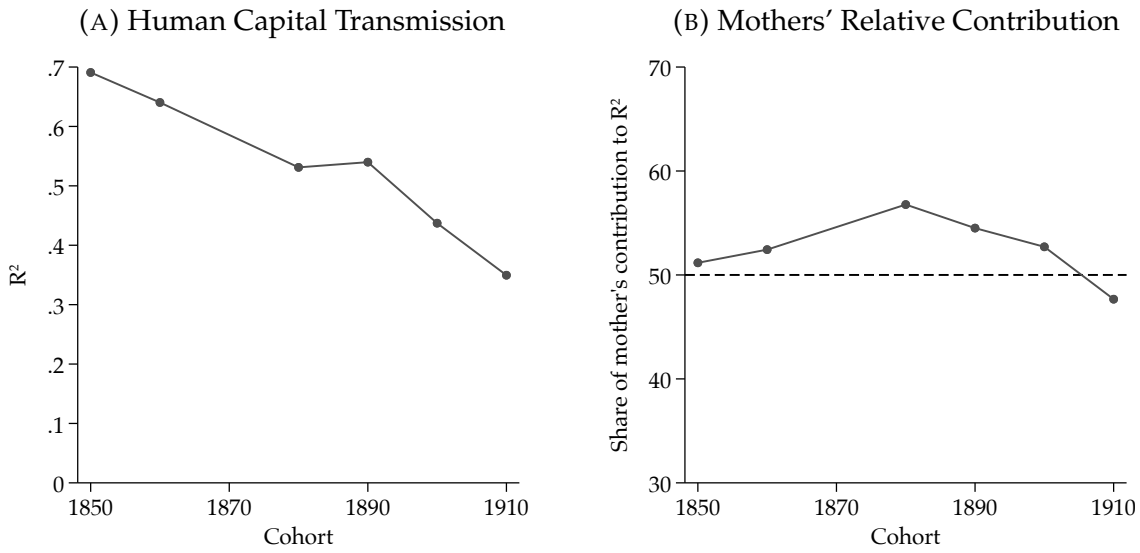
Census cross-sections of children who reside with their parents allow us to study intergenerational mobility in certain outcomes without census linking. Specifically, we use such cross-sections to relate parental background to their children's early life outcomes of literacy and school attendance at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

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

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

Our findings highlight the role of human capital transmission, especially from mothers, in enhancing income mobility over time. Our analysis in section 4 revealed that the declining predictive power of maternal human capital for their child's income led

FIGURE 6: Transmission of (Latent) Human Capital Ranks Across Cohorts



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

to increased mobility. We show in this section that the diminished predictive power of maternal human capital for income is accounted for by its reduced predictive power for the child's human capital.

We successfully replicate the cross-sectional patterns of human capital mobility using our new panel (see Appendix Figure A.7). We find that the relative changes in human capital mobility (R^2) match perfectly across both datasets. Similarly, the proportion of human capital transmission attributed to mothers decreases by a similar amount in both datasets. Our panel, while confirming the patterns of *relative changes* over time observed in the cross-section, interestingly shows higher *levels* of human capital mobility. This difference can be explained by two main factors. First, the similarity between parental and child human capital is likely more pronounced in childhood than in adulthood, due to human capital accumulation or depreciation in adult life (intra-generational mobility). Unlike the cross-sectional analysis, our panel includes adult children and accounts for such intra-generational shifts, potentially leading to lower estimates of intergenerational mobility. Second, inaccuracies in automated record linkage might understate the degree of intergenerational persistence through measurement error in parental background.

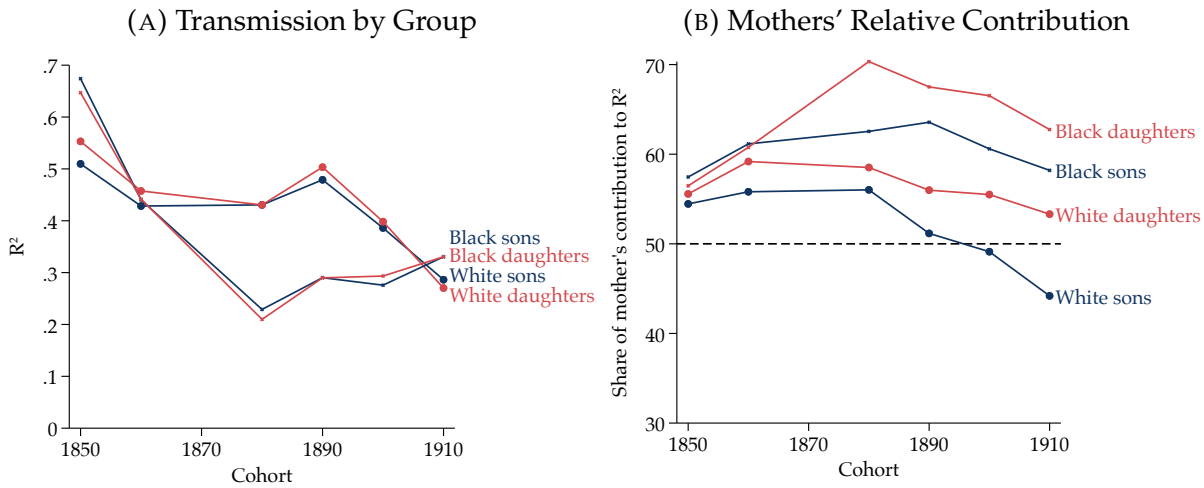
5.2 Human Capital Mobility by Group

We estimate equation (4) separately by race and sex and find that human capital mobility varied significantly for Black and white Americans. The human capital rank of Black

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

In line with this finding, school access among white children became almost universal in the early 1900s (see Appendix Figure A.8). In contrast, most Black children—especially those whose ancestors were enslaved and largely denied literacy until 1865—lived in the Jim Crow South with restricted school access, shorter school years, and poor school quality (Card and Krueger, 1992; Althoff and Reichardt, 2023). The denial of equal access to high-quality schooling under Jim Crow may explain why human capital mobility among Black Americans decreased starting around 1880.

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



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

The finding that mothers' contributions to their children's human capital are generally larger than fathers' is particularly pronounced among female and Black children (see Panel B of Figure 7).[8] Mother's large influence on daughters and Black children aligns with the historical lack of access to educational resources for these groups (Kober and Rentner, 2020). For daughters, it could also suggest the presence of gender-specific role model effects (e.g., Bettinger and Long, 2005; Olivetti et al., 2020).

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

We also estimate a version of equation (4) where (latent) human capital ranks are replaced with ranks in formal school attendance completed from the 1940 census. We find that racial differences in educational mobility are larger than those in human capital mobility (see Appendix Figure A.9). This result underscores the fact that the lack of access to formal schooling was even more persistent across generations among Black families than the racial differences in human capital. In contrast, white Americans, who had nearly universal access to schools, were able to substitute parental homeschooling with formal schooling, thereby generating even higher mobility than that observed in human capital.

6. The Role of Mothers as Educators

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

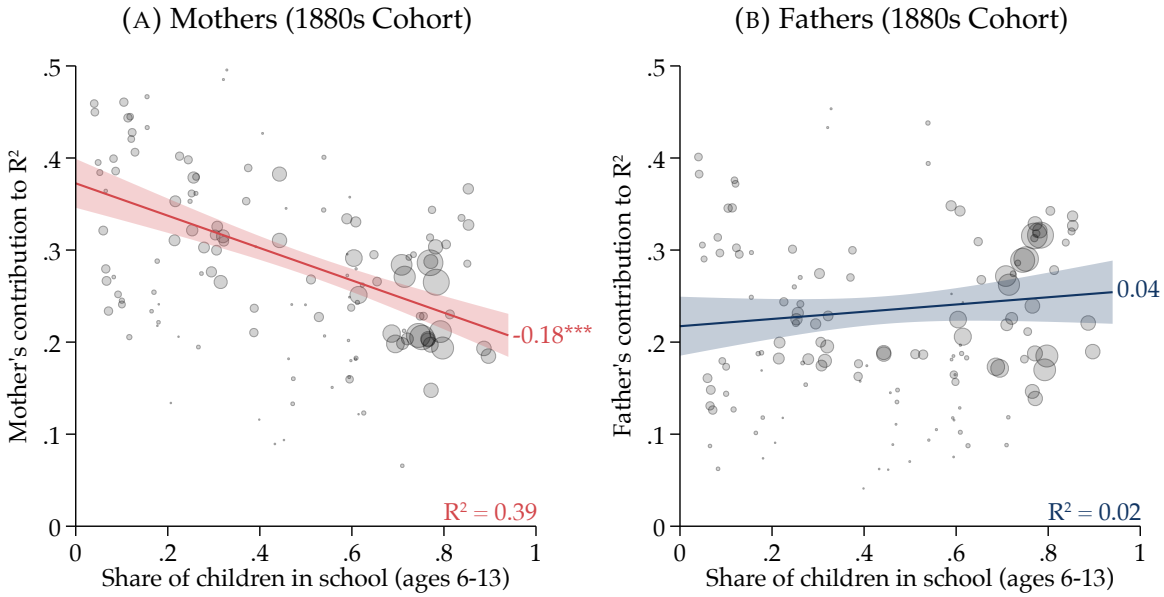
6.1 Schools and the Rise of Human Capital Mobility

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

Consistent with mothers' importance in home schooling, mothers are more predictive of child outcomes in areas with limited school access (see Figure 8). Maternal human capital explains almost 40 percent of variation in child human capital when school access is minimal, and around 20 percent when school access is universal. Conversely, fathers' contribution was lower and showed no correlation with school access. In fact, the contributions of mothers and fathers were comparable only when school access was universal.

As school access expanded, it diminished the disparities in human capital mobility previously observed among groups with varying levels of school access (see Panel B of Figure A.10). The reduced influence of parental human capital with improved public

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



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

school access aligns with Biasi (2023), who shows that equalizing school resources can reduce disparities in intergenerational mobility.

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

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

7. Conclusion

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

There are several promising avenues for future research. We expanded the parental status measurement to separately encompass maternal and paternal roles. Future research could integrate broader parental background measures like wealth or social norms or consider the role of other relatives including grandparents. Given the importance of the location in which a person grows up—as documented in previous work (e.g., Chetty et al., 2016; Chetty and Hendren, 2018)—future research could also use the R^2 mobility metric to factor in neighborhood quality alongside parental background. Another promising avenue for future work would be to assess changes in maternal transmission of economic outcomes over the 20th century, especially amid rising female labor participation (Goldin, 1977, 1990, 2006; Olivetti, 2014) and single-motherhood (Althoff, 2023).

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

BUCKLES, K., J. PRICE, Z. WARD, AND H. WILBERT (2023b): "Family Trees and Falling Apples: Historical Intergenerational Mobility Estimates for Women and Men," Working Paper.

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

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

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

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

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

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

CHETTY, R., N. HENDREN, P. KLINE, AND E. SAEZ (2014a): "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 129, 1553–1623.

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

NGAI, R., C. OLIVETTI, AND B. PETRONGOLO (2024): "Structural Transformation over 150 years of Women's and Men's Work," *Unpublished Working Paper*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 **30**

B Appendix Table **35**

C Methods Appendix **36**

C.1 Relation Between R^2 and Coefficients 36

C.2 Shapley-Owen Decomposition of the R^2 37

C.3 Semi-parametric latent variable method 38

D Data Appendix **41**

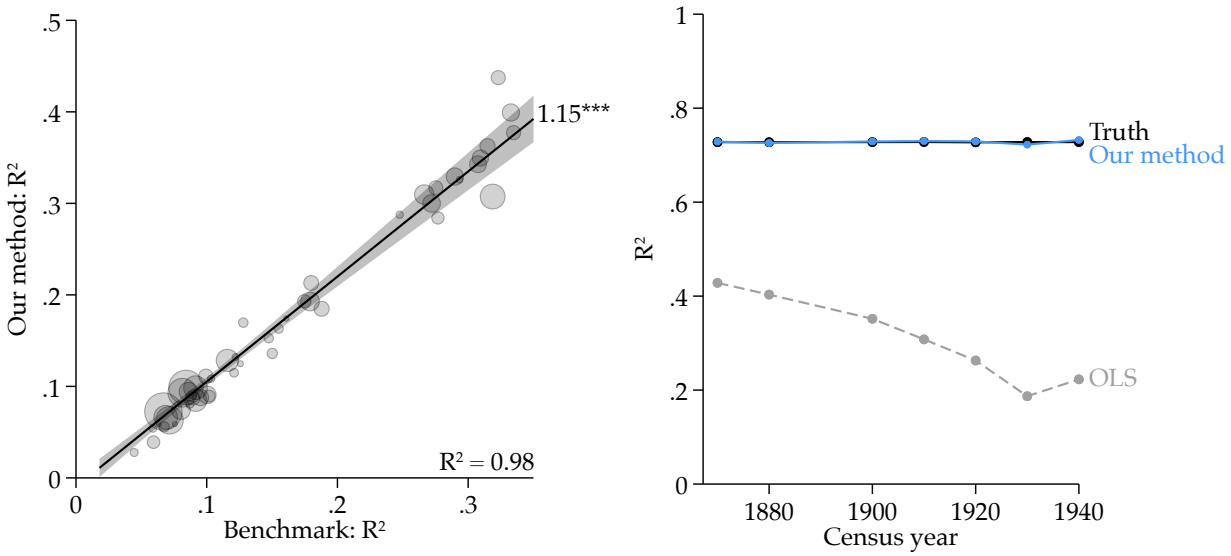
D.1 Linking Procedure 44

D.2 Sample Weight Construction 46

A. Appendix Figures

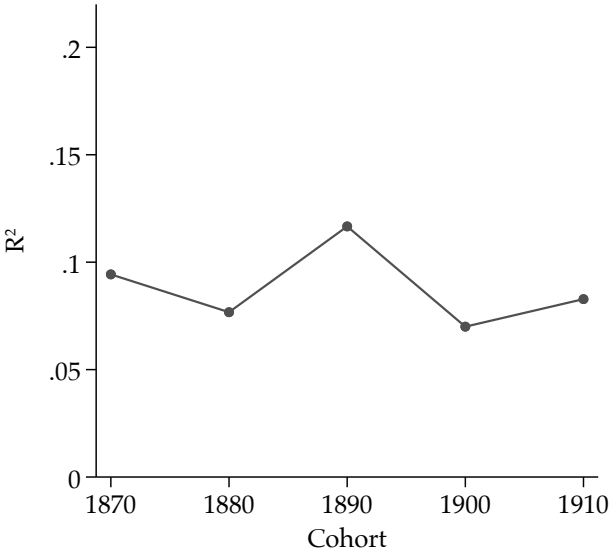
FIGURE A.1: Validation of the Semi-parametric Latent Variable Method

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



Notes: This figure demonstrates the effectiveness of our semi-parametric latent variable method in identifying rank-rank relationships from binary proxies. Panel A contrasts the R^2 values from rank-rank regressions using actual and binarized educational data from the 1940 census. We binarize the data by arbitrarily categorizing individuals based on their educational attainment: more than 11 years for children, 9 for mothers, and 7 for fathers. Each dot represents a US state, weighted by sample size and focusing on children aged 13—21 living with parents. Panel B illustrates a simulation where literacy serves as a binary proxy for human capital. We simulate human capital ranks, convert them into literacy dummies based on historical literacy rates, and compare the R^2 values from regressions using these dummies. The "Truth" line represents the R^2 from a human capital rank-rank regression, "Our method" from our latent variable method using literacy dummies, and "OLS" from a standard OLS regression with the same literacy dummies. In the 1940 census, instead of literacy, we observe the highest year of school or degree completed. We classify individuals who have completed at least two grades of school as literate; others we classify as illiterate.

FIGURE A.2: Mobility Estimates Based on "occscores"

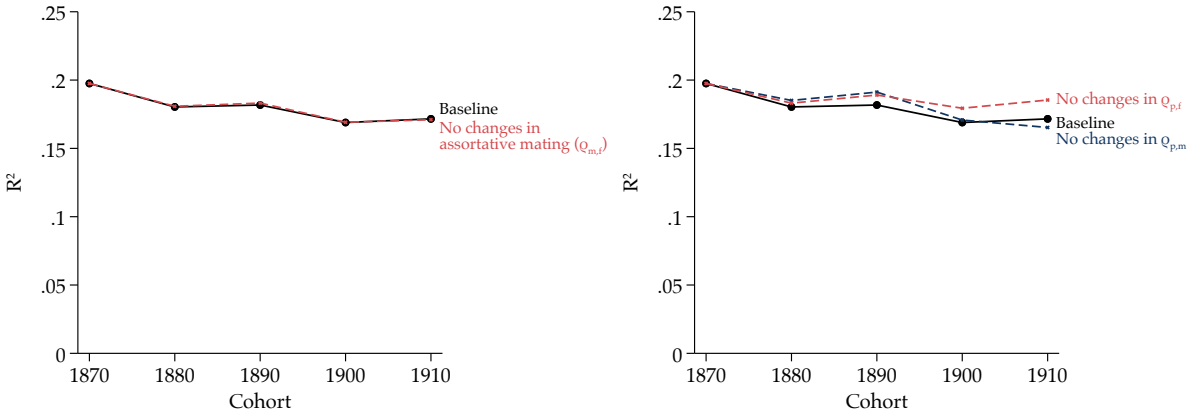


Notes: This figure shows the share of the variance in a child's household income rank explained by (1) parents' household income ranks and their (latent) human capital ranks (R^2) and (2) parents' household income ranks alone. For parental human capital ranks, we use information on parental literacy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

FIGURE A.3: Mobility and the Impact of Evolving Parental Input Correlations

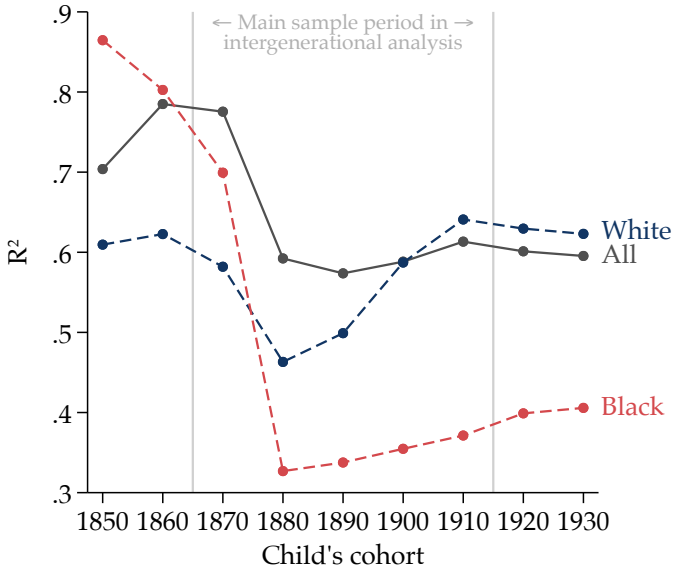
(A) Assortative mating

(B) Correlations in human capital and income



Notes: This figure shows the role of each parameter on the R^2 in equation (2). The baseline represents the observed R^2 shown in Figure 4. The other three lines represent the counterfactual R^2, had the respective parameter not changed over time, computed using the decomposition in equation (3). For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied.

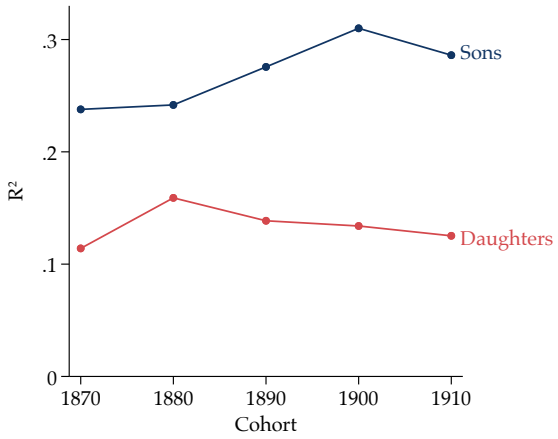
FIGURE A.4: Assortative Mating Estimates by Group



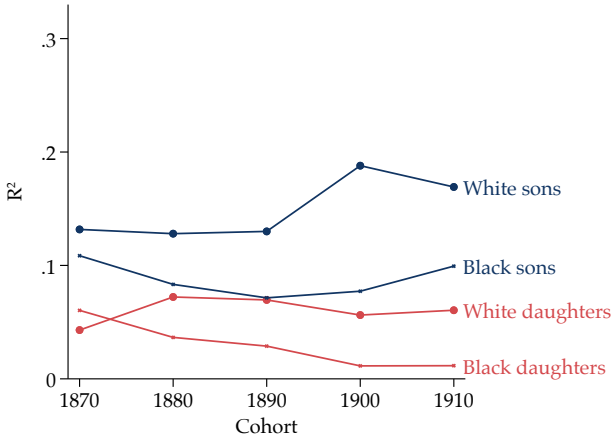
Notes: This Figure shows the share of the variance in a person's (latent) human capital rank explained by their spouse's (latent) human capital rank (R^2) across their child's cohort. For human capital ranks, we use information on parental literacy and the latent variable method introduced in section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

FIGURE A.5: Within-Group Mobility Estimates

(A) Mobility by sex



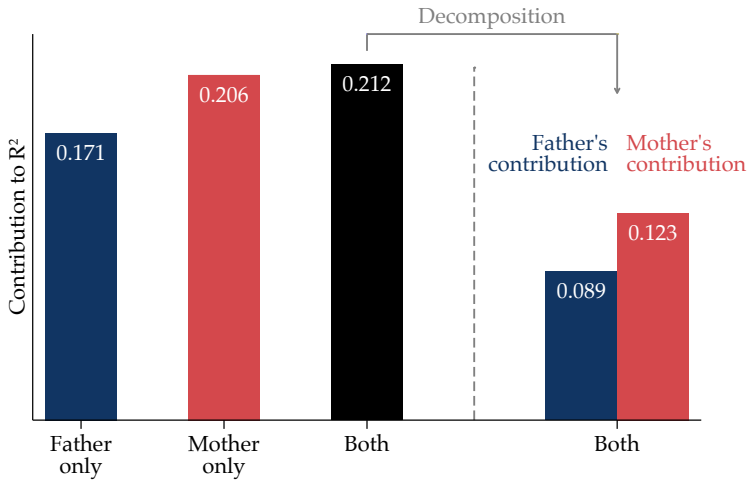
(B) Mobility by sex and race



Notes: This Figure shows the share of the variance in a child's household income rank explained by parents' household income ranks and their (latent) human capital ranks (R^2) across cohorts and groups. For parental human capital ranks, we use information on parental literacy and the latent variable method introduced in section 3.4. We use the household head's LIDO occupational income score (Saavedra and Twinam, 2020). Results are based on our new panel and sample weights are applied.

FIGURE A.6: Illustrating our Decomposition Method

Intergenerational Transmission of Human Capital

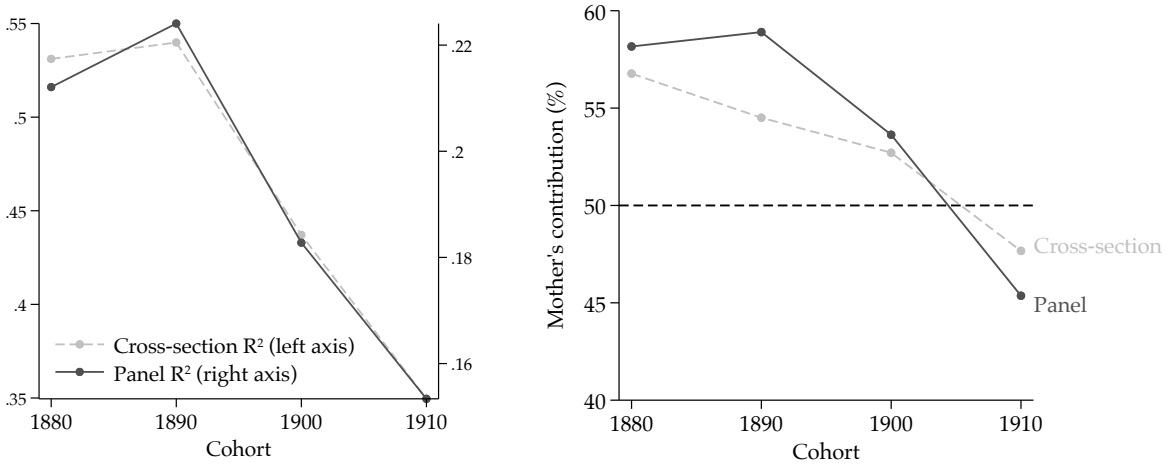


Notes: This figure shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2). We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. We decompose the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

FIGURE A.7: Panel-Based Estimates of Human Capital Mobility Across Cohorts

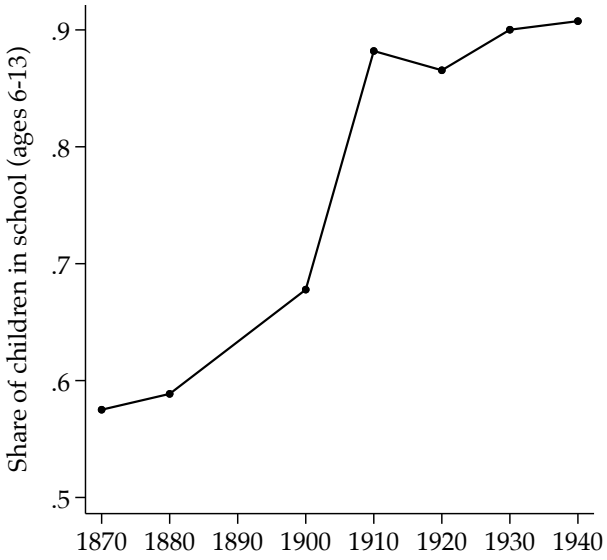
(A) Human Capital Transmission

(B) Mothers' Relative Contribution



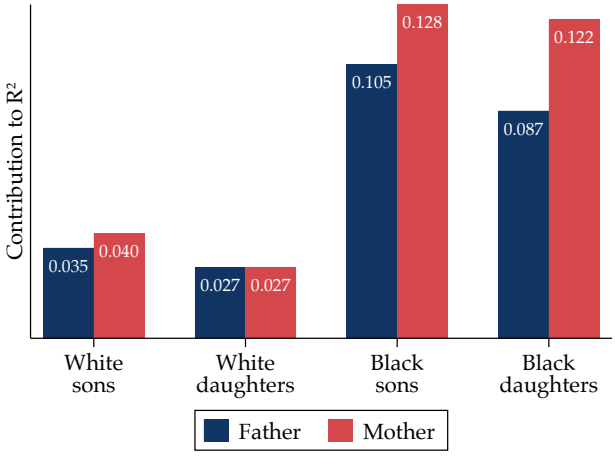
Notes: This figure compares our baseline results of human capital transmission from the cross-section of children who live with their parents to estimates based on our new panel. Panel A shows the share of the variance in a child's (latent) human capital rank explained by parents' (latent) human capital ranks (R^2) across cohorts. We recover human capital rank-rank transmission using information on literacy and the latent variable method introduced in section 3.4. Panel B shows mothers' relative contribution to the overall R^2 using the Shapley-Owen method. Cross-sectional results are based on the census cross-section of children ages 13–16 in their parents' household; panel results are based on individuals of any age.

FIGURE A.8: Increasing Access to Schools



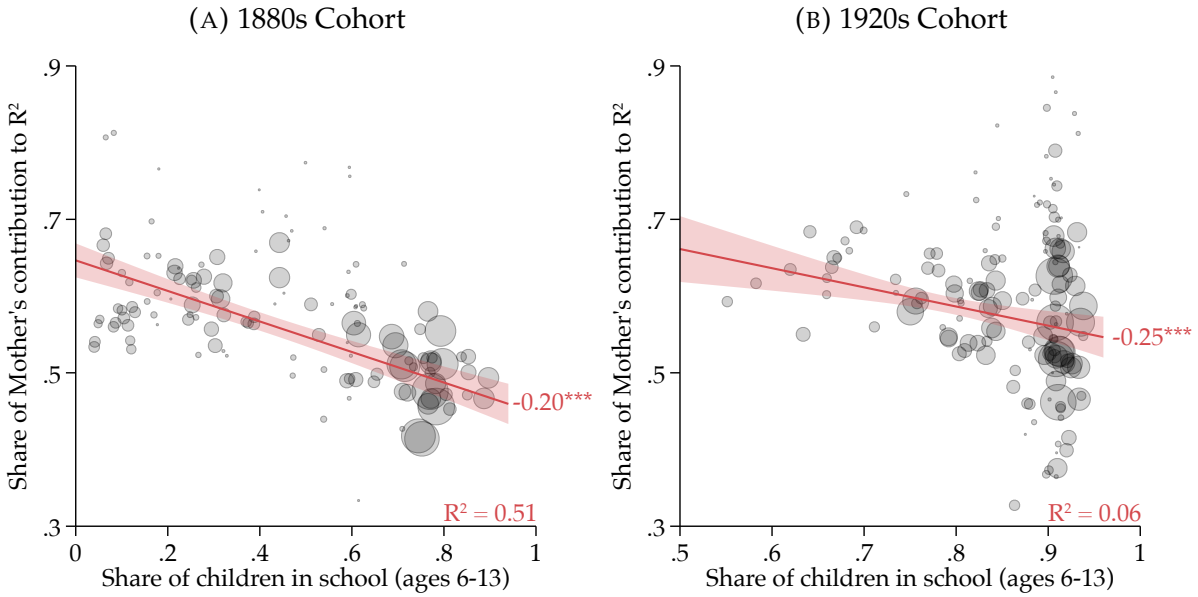
Notes: This figure shows the share of children aged 6–13 who attend school across time.

FIGURE A.9: Intergenerational Transmission of Formal Schooling (1920s cohort)



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

FIGURE A.10: Mothers' Human Capital as Substitute for Local Schools



Notes: This figure shows the relationship between local school access and mothers' *relative* contributions to child human capital (as a share of total variation explained). Literacy is used as the measure for rank-based transmission of human capital (section 3.4). Each dot represents a group of children born in the 1880s or 1920s, categorized by race, sex, and state. Sample size weights are applied. School access is determined by the race- and sex-specific share of children aged 6–13 in school. Results are based on the census cross-section of children ages 13–16 in their parents' household.

B. APPENDIX TABLE

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

	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$	ϕ_{Mother}	ϕ_{Father}	$\frac{\phi_{\text{Mother}}}{R^2}$
Baseline measure of school access	-0.18***	0.04	-0.20***			
	(0.03)	(0.05)	(0.03)			
Refined measure of school access				-0.47***	0.15	-0.58***
(accounts for attendance, term lengths, etc.)				(0.08)	(0.11)	(0.10)
R^2	0.39	0.02	0.51	0.37	0.04	0.57
Observations	133	133	133	128	128	128

Notes: This table shows the relationship between local school access and parents' contributions to child human capital. Columns 1–3 (baseline) contain the results from Figure 8 and Panel A of Appendix Figure A.10. For this baseline, school access is determined by the race- and sex-specific share of children aged 6–13 in school according to the 1880 census. Columns 4–6 show that these results are even stronger when we use an alternative measure of school access. For this measure, we newly digitized data on state-specific school ages, enrollment, attendance, and term lengths from the Census Statistical Abstracts. From these data, we compute the average likelihood of attending school on any given day in the year between ages 6–16, specific to each state. These data are incomplete for Arkansas and Wyoming, leading to slightly lower sample sizes.

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:

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

$$R^2 = \text{cor}(x, y)^2 = \widehat{\beta}^2 \cdot \frac{\text{Var}(x)}{\text{Var}(y)}, \tag{6}$$

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

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

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

C.1.2 Multiple inputs

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

$$R^2 = \frac{Var\left(\sum_{j=1}^{k} \widehat{\beta}_j x_{i,j}\right)}{Var(y)} = \frac{\sum_{j=1}^{k} \widehat{\beta}_j^2 Var(x_j) + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} \widehat{\beta}_j \widehat{\beta}_l Cov\left(x_j, x_l\right)}{Var(y)}. \tag{7}$$

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

$$R^2 = \sum_{j=1}^{k} \widehat{\beta}_j^2 + 2\sum_{j=1}^{k-1}\sum_{j=i+1}^{k} \widehat{\beta}_j \widehat{\beta}_l \widehat{\rho}_{j,l} \tag{8}$$

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

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

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

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

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

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

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

$$\phi_j = \frac{\Delta_j}{k!}, \tag{9}$$

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

C.2.1 Example with two inputs

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

TABLE C.2: Example of Shapley-Owen Decomposition

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

C.2.2 Unpacking the Shapley-value with two inputs

To better understand what the Shapley-value for each parental input comprises, we express it as a function of regression coefficients, variances, and covariances in the two-input case. Let ϕ_1 be one parent's Shapley value—i.e., the contribution that the parent's input makes to the overall R^2 when regressing child outcomes on both parents' inputs. Applying equation (9), we have

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

Further, using equation (7), we have

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

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

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

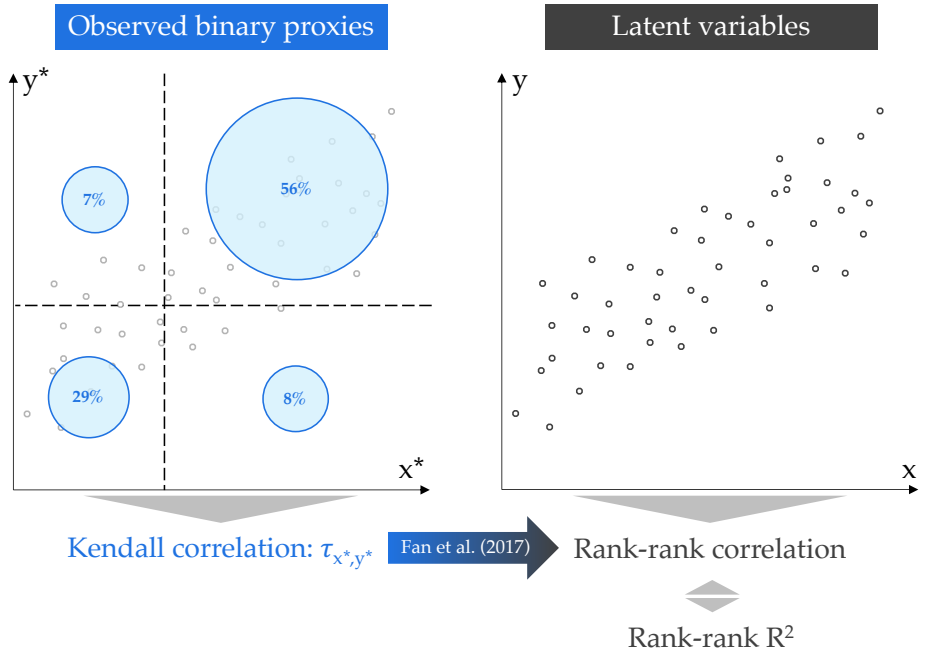
For rank-rank regressions, we have

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

C.3 Semi-parametric latent variable method

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

FIGURE C.1: Illustrating the Semi-Parametric Latent Variable Method



Notes: This figure illustrates the semi-parametric latent variable method, recovering rank-rank mobility (R^2) in latent variables from observed binary proxies. Assuming that the underlying latent variables are drawn from a joint Gaussian copula distribution, pairwise rank-rank correlations can be identified from Kendall's correlation between the observed binary proxies using the bridging function in (12). Rank-rank regressions can be identified from the pairwise correlation matrix using equations and (13) and (14).

We assume that the dependent and independent variables are drawn from a joint Gaussian copula distribution. That is, we assume that there exists a set of unknown monotonic transformations f_y, $f_1, \cdot\cdot, f_k$ such that $f_y(y_i), f_1(x_{1i}), f_k(x_{ki}) \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = \mathbb{1}$. 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. Note that the normality assumption does not impose that the latent variables of interest (e.g., human capital) are jointly normally distributed. Rather, it requires that there exists some monotonic transformation of the latent variables that is jointly normally distributed.

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 $\tilde{x}_{1i}$ which is one if x_{1i} is above an arbitrary cut-off and zero otherwise. Third, that only observe binary proxies of each variable are observed.

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

tween Kendall's rank correlation coefficient and Σ_{12} is

$$\Sigma_{12} = \sin\left(\frac{\pi}{2}\tau_{12}\right). \tag{10}$$

Therefore, our estimate $\hat{\Sigma}_{12}$ is the sample equivalent of equation (10).

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

$$\tau_{12} = 4\Phi_2\left(\Delta_2, 0, \frac{\Sigma_{12}}{\sqrt{2}}\right) - 2\Phi\left(\Delta_2\right) \tag{11}$$

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 $\hat{\Delta}_2 = \Phi^{-1}(1 - \bar{x}_1)$ where $\bar{x}_1 \equiv \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{1i}$. Because equation (11) is strictly increasing in Σ_{12} (see (Fan et al., 2017) for the proof), Σ_{12} is identified as the unique root of equation (11) where τ_{12} and Δ_2 are replaced with their finite sample analogues.

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

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

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 (12) 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,

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{R}}_x\right)^{-1}\widehat{\boldsymbol{R}}_{xy} \tag{13}$$

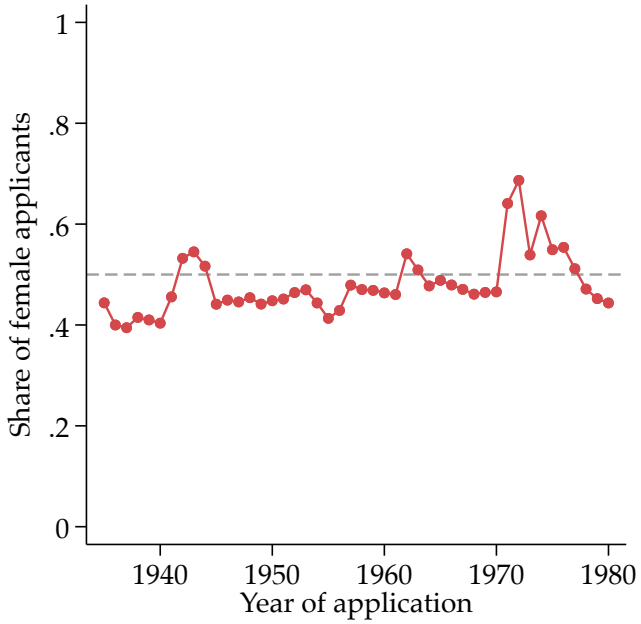
where $\widehat{\boldsymbol{R}}_x$ is a $k \times k$ rank-rank correlation matrix of the independent variables and $\widehat{\boldsymbol{R}}_{xy}$ is a $k \times 1$ vector of rank-correlations between the independent variable and dependent variable. $\hat{\alpha}$ is then computed as $\bar{y} - \widehat{\boldsymbol{\beta}}'\bar{\boldsymbol{x}}$.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

$$R^2 = \widehat{\boldsymbol{R}}'_{xy}\left(\widehat{\boldsymbol{R}}_x\right)^{-1}\widehat{\boldsymbol{R}}_{xy}. \tag{14}$$

Equations (13) and (14) are numerically equivalent to the rank-rank coefficient vector and R^2 in the case without latent variables (for a proof, see e.g., O'Neill (2021) and impose that the marginal distributions of the variables are identical). From equations (13) and (14), we also see the relation between the slope coefficient and R^2 and in the univariate case discussed in Appendix C.1.1: $\hat{\beta} = \sqrt{R^2}$.

D. Data Appendix

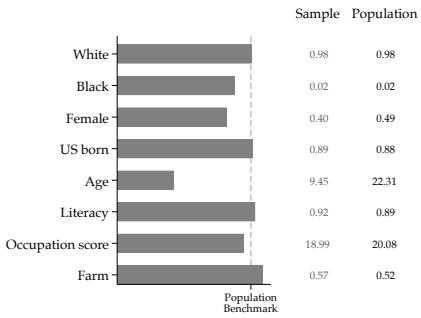
FIGURE D.1: Share of Female Applicants



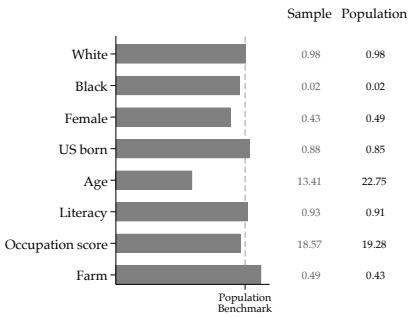
Notes: This figure shows the share of SSN applicants who are female by year of application.

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

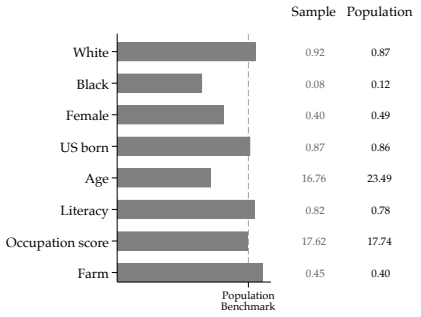
(A) 1850



(B) 1860

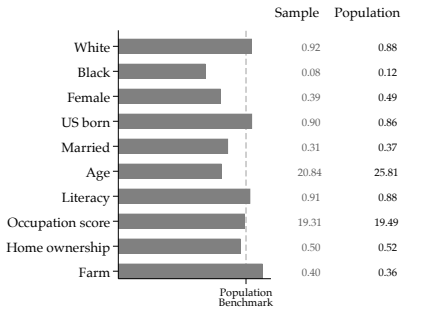


(C) 1870

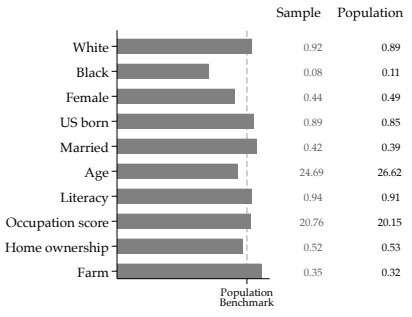


(D) 1880

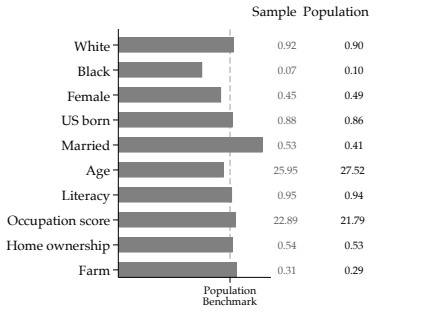
(E) 1900



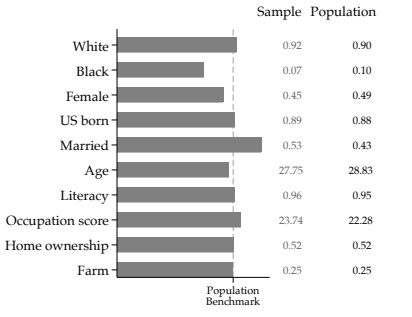
(F) 1910



(G) 1920

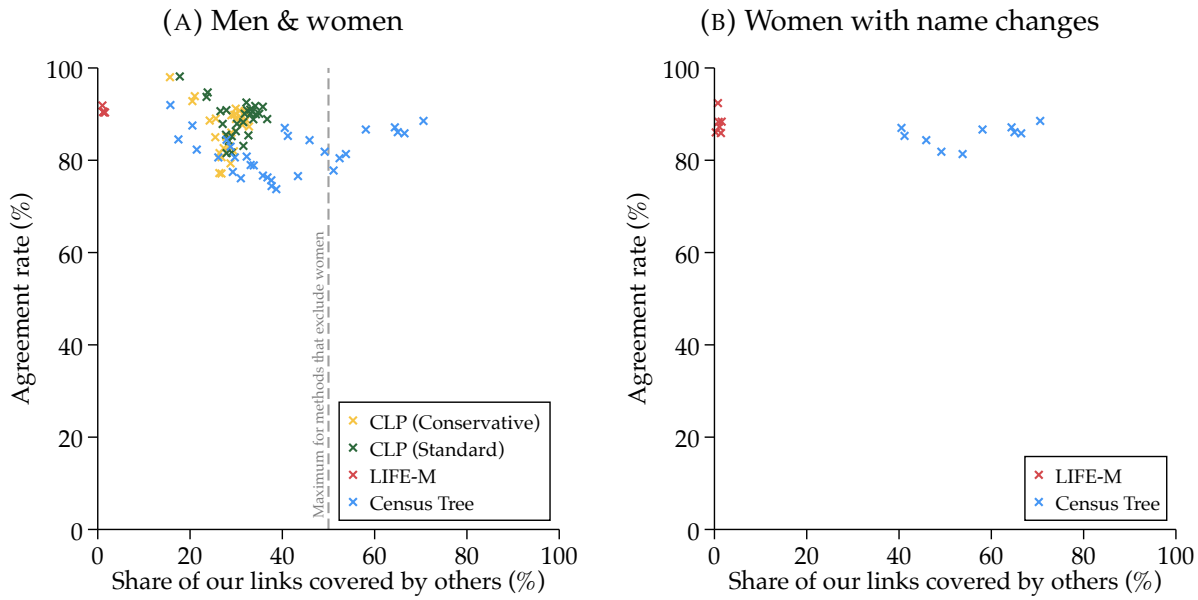


(H) 1930



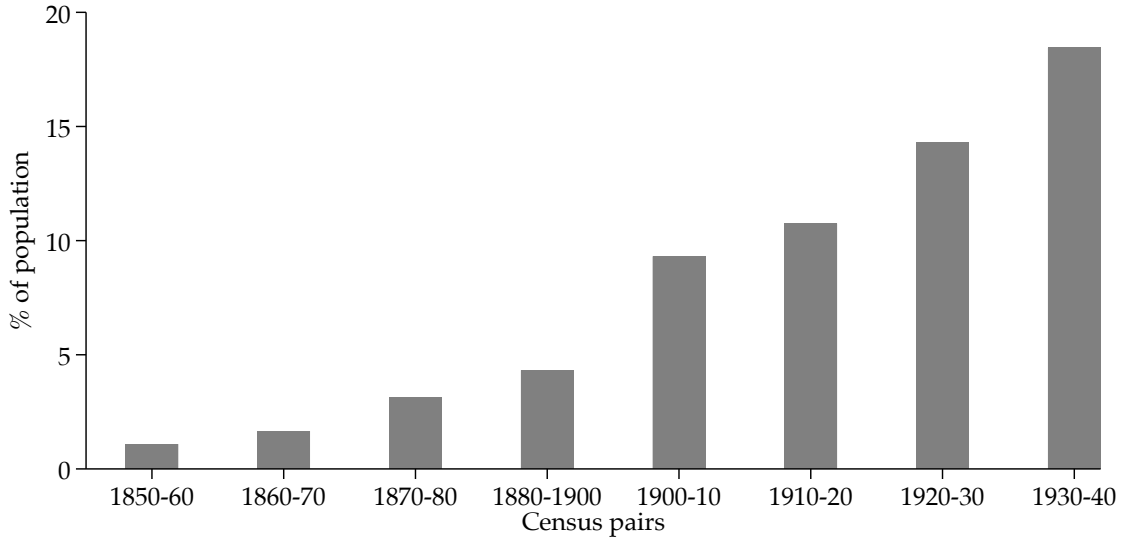
Notes: This figure shows the representativeness of characteristics among individuals who we successfully assign an SSN compared to the full population in each census before 1940. The sample is exceptionally representative compared to existing panels, most notably with respect to sex and race. Because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

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



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

FIGURE D.4: Fraction of US Population Linked in Our New Panel



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

D.1 Linking Procedure

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

First stage: Applicant SSN $\leftrightarrow$ census.

- *Preparing SSN data*: We use a digitized version of the Social Security Number application data from the National Archives and Records Administration (NARA) known as the Numerical Identification Files (NUMIDENT). We harmonize the application, death and claims files to capture all the available information of each SSN record. These data include each applicant's name, age, race, place of birth, and the maiden names of their parents. We recode certain variables to align with census data, for example, we ensure codes for countries of birth, race and sex are consistent across the SSN and Census. Additionally, we apply the ABE name cleaning method to names of applicants and their parents resulting in an "exact" and a NYSIIS cleaned version of all names (Abramitzky et al., 2021a)[9].

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

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

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

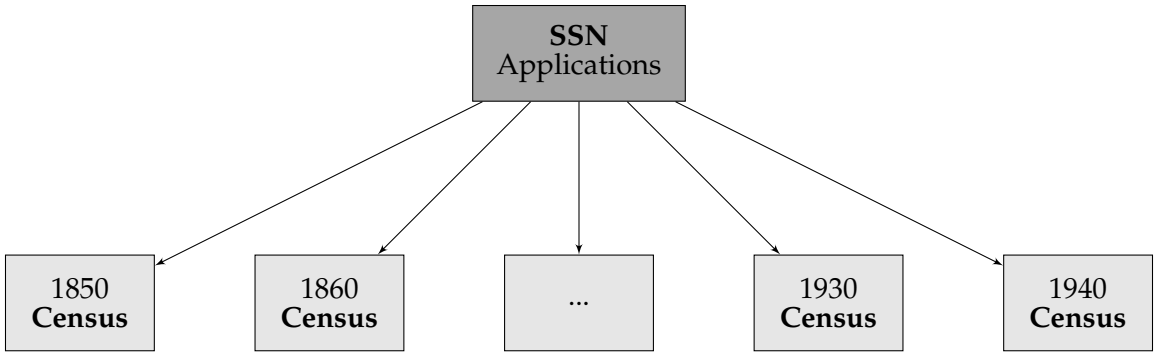
[9]The use of the NYSIIS phonetic algorithm helps in matching names with minor spelling differences, as mentioned in Abramitzky et al. (2021a)

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

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

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

FIGURE D.5: First & Second Linking Stages

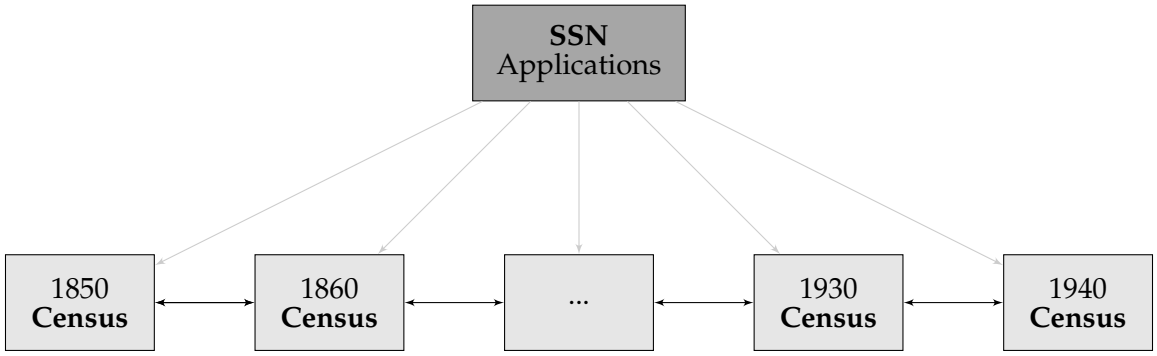


Notes: This figure shows the first and second step of our linking procedure—linking individuals' Social Security Numbers to their census records.

on their parents' birthplace (i.e., the SSN applicant's grandparents birthplace) which is available from the census records. For parents who are not SSN applicants themselves, we create a synthetic identifier similar to an SSN.

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

FIGURE D.6: Final Linking Stage



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

D.2 Sample Weight Construction

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

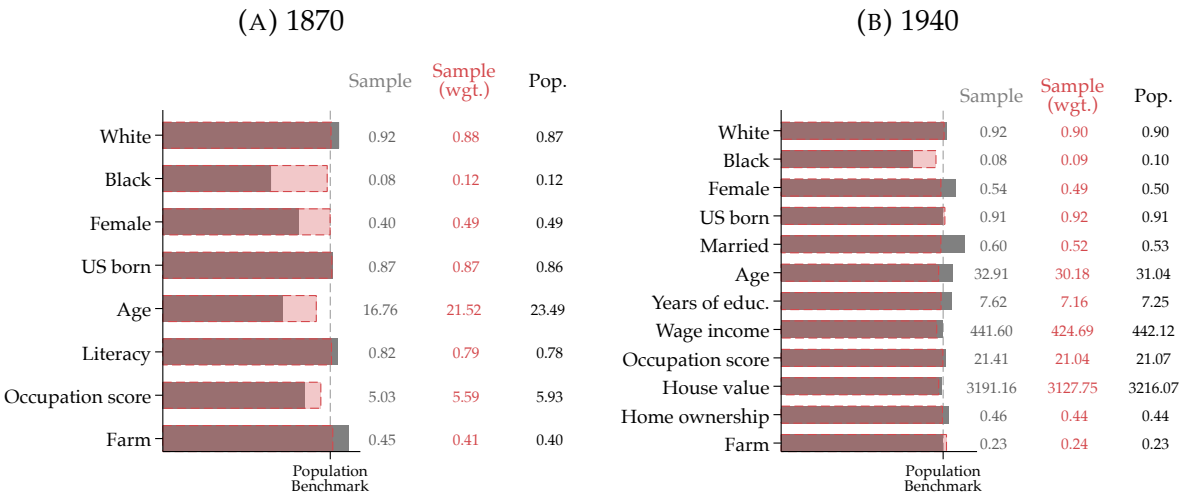
For each census between 1850 to 1940, we create indicator variables for whether (1) we have identified an individual's Social Security Number, (2–4) whether we have been able to measure the economic status of the individual's (2) mother, (3) father, or (4) both parents. Measuring parental economic status may itself involve census linking and does

not rely on observing parents in the same census wave.

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

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

FIGURE D.7: Sample Balance After Inverse Propensity Weighting (1870 & 1940)



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

Figure D.7 shows average sample characteristics after applying our new inverse propensity weights. The reweighted sample is almost perfectly representative of the full population in all dimensions, even those not targeted by our reweighting method. For example, wage income and occupational income scores match close to perfectly despite only having included coarse occupation and industry categories in our reweighting procedure. Similarly, housing wealth is not targeted but our reweighted sample closely mirrors the overall population.

REFERENCES

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

ABRAMITZKY, R., L. BOUSTAN, AND M. RASHID (2020): "Census Linking Project: Version 1.0," dataset: https://censuslinkingproject.org.

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

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

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

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

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.

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.