Spurious Mobility in Imperfectly Linked Historical Data

Ross Mattheis[*]

Job Market Paper

This version: November 5, 2024

Most Recent Draft

Abstract

When was the United States a land of opportunity? This paper revisits the history of intergenerational mobility in the US, accounting for the impact of imperfectly linked census data. Incorrectly linked observations typically attenuate ordinary least squares (OLS) estimates, such as the association of income ranks among fathers and sons. This attenuation exaggerates levels of mobility, as mobility is inversely related to the strength of the relationship between parents' and children's outcomes. I address bias due to imperfectly linked data from the perspective of nonclassical measurement error and propose a class of models for misclassification—error in discrete data—that rely on a repeated, conditionally independent measure of the misclassified variable. A natural source for such a repeated measure can be found by linking observations into an additional sample. In a validation exercise, the proposed estimator reduces bias by 50-90% relative to OLS, with a larger reduction in bias on more severely misclassified samples. After correcting for misclassification error, estimates of the rank-rank slope of occupation status for White men born between 1832 and 1910 are 50-100% higher than OLS estimates, depending on the cohort. Revised estimates suggest a U-shape pattern for intergenerational mobility in US history. Individuals born before the Civil War experienced levels of mobility comparable to the present, while those born between the Civil War and WWI—who entered the workforce during the highest levels of inequality in the US before the present—experienced lower levels of mobility than in any region in the US today.

[*] Department of Economics, Harvard University. Email: mattheis@g.harvard.edu. I am especially grateful to my advisors, Claudia Goldin, Larry Katz, Ed Glaeser, and Isaiah Andrews, for their kindness, guidance, and generous support. I thank Alex Albright, Yong Cai, Kevin Chen, Raj Chetty, Lindsey Currier, Melissa Dell, Michael Droste, James Feigenbaum, Martin Fiszbein, Grant Goehring, Bob Margo, Nathan Nunn, Dev Patel, Deniz Sanin, Ursina Schaede, Rahul Singh, Jesse Shapiro, Neil Shepard, Marco Tabellini, Elie Tamer, Davide Viviano, Chris Walker, David Yang, and Noam Yuchtman for lending their brilliance and insight. I am grateful for financial support through the Thomas Cochran Fellowship in Business and Economic History. All errors are mine.

The main danger of making such comparisons [of constructed historical data and more nearly accurate modern data] may be to overestimate how much the economy has changed.

(Romer, 1986, p. 32) *"Spurious Volatility in Historical Unemployment Data"*

1 Introduction

When, if ever, did the United States economy deliver on the promise of opportunity? The recent large-scale digitization and linking of historical census data provide social scientists a powerful tool to address questions like this in American history (Ruggles et al., 2018; Abramitzky et al., 2021a). Work using this new resource has found exceptionally high rates of intergenerational mobility among White men born before the Civil War, and relatively stable levels of mobility thereafter (Long and Ferrie, 2013; Song et al., 2019). This pattern may be surprising given the sweeping changes in the economy over the last 150 years, including technological advancement (Gordon, 2016), the amassing of capital (Gallman and Rhode, 2020), growth in organizational complexity (Chandler, 1977), urbanization (Glaeser, 2012), changes in the returns to and levels of human capital (Goldin and Katz, 2008), waves of immigration (Abramitzky and Boustan, 2022), and the rise, fall, and return of inequality (Lindert and Williamson, 2016; Piketty, 2014). Yet comparing levels of mobility between the past and present is complicated by the fact that historical data are often less reliable than modern data. Linked historical census data are imperfect, with linked samples covering only 25-50% of the population and false positive rates—in which linked observations do not correspond to the same individual—of 15-40% (Bailey et al., 2020a). Both sources of error in record linkage may lead to bias in subsequent estimates, including the parent-child associations of status often used to measure mobility.

In this paper, I propose a methodology to address biases due to imperfectly linked data, and I find estimates that suggest a qualitatively different history of intergenerational mobility in the United States. Previous estimates of the rank-rank slope are substantially attenuated relative to the truth: standard estimates are attenuated by 50% for cohorts born in the mid-19th century and by 30% for those born at the turn of the 20th century relative to my estimates. Consequently, I find levels of mobility among cohorts of White men born before the Civil War comparable to the present and exceptionally low levels of mobility among cohorts born between the Civil War and World War One. Trends in mobility follow a U-shape pattern, with levels of mobility falling among cohorts born in mid-19th century, lowest among those born around the turn of the 20th century, and rising to modern levels. This new pattern reveals a previously obscured negative association between mobility and inequality across time in the US, consistent with the "Gatsby curve" documented across modern economies (Corak, 2013; Durlauf et al., 2022).

Imperfectly linked data can bias downstream estimates in two ways: linked samples may be *nonrepresentative* if less than the full population is linked, and linked samples are *contaminated* if some of the linked observations do not correspond to the same individual.[1] Nonrepresentativeness in linked data is well studied (Bailey et al., 2020b) and similar to the challenge of estimation and inference with nonrepresentative samples across the social sciences (Beresovsky et al., 2024).[2] While bias from nonrepresentativeness in linked samples is an important concern, this paper focuses on the impact of linkage errors in the final sample. To the extent that there is a trade-off between constructing larger, more representative linked samples and linked samples with lower rates of false positive links (Abramitzky et al., 2021a), my proposed methodology reduces the "marginal cost" of false positive links, potentially allowing researchers to choose more representative samples.

Contamination from incorrect links affects analysis in a way that may be counter-intuitive. It is widely understood that classical measurement error—additive and independent of the unobserved true value—attenuates estimates towards zero when present in the regressor, while similar errors in the outcome do not lead to bias. In contrast, when false positive links are independent, the resulting errors in variables are highly nonclassical and lead to attenuation bias whether errors are in the regressor or the outcome.[3]

To illustrate this source of bias, consider the following example. Suppose a researcher would like to estimate the return to schooling where educational attainment is observed in one source, income is observed in another, and—lacking a unique identifier—the researcher links observations using information like name and age. If names are not strongly associated with earnings, the income observed for an incorrect link will not depend on whether the linked individual has nine or eighteen years of education. Consequently, measurement error in income will be negatively associated with the level of educational attainment and estimates of the return to schooling will be attenuated.

To address bias from contamination in imperfectly linked data, I propose a class of misclassification models that build on results in the nonclassical measurement error literature. Measurement

[1] I discuss a simple bias decomposition that more concretely illustrates the role of each source of bias due to imperfectly linked data in OLS estimates in Appendix A.1.

[2] If selection into the linked sample is based on observed characteristics, researchers can re-weight the linked sample to mitigate bias from nonrepresentativeness, as is common across the empirical social sciences. For example, it is often challenging to link Black Americans across historical US censuses, and coverage of Black Americans remains a challenge in modern censuses (Sabety and Spitzer, 2023). In a more extreme example, women have generally been excluded from linked samples due to the convention of changing surnames at marriage. Recent work including women in large-scale samples has relied on additional sources sources including marriage records, social security applications, and genealogical information (Bailey et al., 2023; Althoff et al., 2024; Buckles et al., 2023).

[3] For incorrectly linked data, the difference between error in one variable or another is largely semantic. For any linked data connecting two sources, the researcher may choose to define the variables in either source as correctly observed while variables in the other source may be contaminated by errors in record linkage.

error in a discrete variable like occupation is called *misclassification* and is inherently nonclassical.[4] I consider a setting in which we observe a discrete regressor and two noisy measures of a discrete latent outcome. For example, the regressor may be father's occupation while the outcome is son's occupation and noisy measures come from observations in two, potentially imperfect, linked samples. When the regressor and the observed outcomes are independent conditional on the true value of the outcome, we can recover the true relationship between regressor and the unobserved, correctly measured outcome.[5] I then introduce a set of parametric misclassification models that build on the particular pattern of errors that result from record linkage errors. Adding structure to the model of misclassification allows me to consider more granular variables, such as narrower categories of occupations, that would otherwise pose challenges to estimation and inference due to the high dimensionality of the nonparametric model.

The proposed approach can be applied in any setting in which a repeated measurement satisfying the identifying assumptions is available. This approach does not require access to the information underlying the generation of linked data, such as the names, addresses, or other identifying information that is often restricted for reasons of privacy. Additionally, the estimator does not rely on a generative model of linked data, which requires identifying and estimating a vast number of cumbersome quantities, like the probability that an individual named Francis Sinatra appears as "Frank" in the census, or a Francis Fitzgerald is recorded as "F Scott".

The likelihood in the model is a simple function of a tabulation of the data, counting each combination of values for the regressor and the noisy measures of the latent outcome. Consequently, the model lends itself naturally to estimation by maximum likelihood. To measure intergenerational mobility, I first estimate the joint distribution of father's and son's occupations in a misclassification model. I then plug in these estimates and a ranking of occupations to arrive at an estimate of the rank-rank slope of father's and son's occupation status.[6] I implement the

[4] Misclassification error is necessarily nonclassical. Recall that classical measurement error refers to errors that are additive and independent of covariates. To illustrate why misclassification is always nonclassical, consider a binary latent variable X with observed value $\widetilde{X}$. Measurement error $\epsilon := \widetilde{X} - X$ takes values in $\{-1, 0\}$ conditional on $X = 1$ and in $\{0, 1\}$ conditional on $X = 0$. Clearly then, measurement error in $\widetilde{X}$ is only classical if it is always zero.

[5] Hu (2008) provides sufficient conditions to nonparametrically recover the distribution of regressor and the latent outcome as well as the conditional distributions of the misclassified variables given the latent truth. In the setting considered in this paper, three assumptions are sufficient. First, each noisy measure and the regressor should be mutually independent conditional on the latent outcome and an optional set of controls. While measurement error is immediately nonclassical in this setting because all variables are discrete, the continuous analogue of this condition is also a relaxation of classical measurement error, since the noisy measures can depend arbitrarily on the value of the latent outcome. Second, the noisy measures should be more likely than not to report the correct value. The last assumption simply requires that the distribution of values of the regressor, such as father's occupations, not be identical for any two values of the latent outcome, such as son's occupations.

[6] My approach is not limited to estimating the rank-rank slope, the interpretation of which has been a recent topic of debate (Chernozhukov et al., 2024). The last step in the estimation procedure can be substituted with any function of the joint distribution of the regressor and latent outcome.

estimator through a new R package created for this purpose.[7]

I evaluate the maximum likelihood estimator for the misclassification model in a validation exercise that mirrors the challenge of historical record linkage. Specifically, I add synthetic noise to multiple copies of the complete count 1940 census before linking subsets of those copies together with standard algorithms (Bailey et al., 2020b). Consistent with the bias decomposition in (10), I find that OLS estimates are attenuated roughly in proportion to the share of false-positive links in each sample. Estimates accounting for misclassification due to imperfectly linked data reduce bias relative to OLS estimates by between 50% and 90%, with proportionally larger improvements for more highly contaminated linked data. Importantly, the validation exercise does not impose the identifying assumptions in the misclassification model. The validation results suggest that the proposed approach can substantially reduce bias due to imperfect record linkage, even if the identifying assumptions are mildly violated.

Finally, I apply my methodology to study intergenerational mobility among White men born between 1833 and 1910. Flexible estimates of occupational misclassification suggest that errors in record linkage substantially bias the observed distribution of fathers' and sons' occupations. When I impose more structure on the model of misclassification, I find estimates of false positive linkage rates around 40%, slightly higher than estimates of false positive rates coming from high quality ground truth samples (Bailey et al., 2020a).

After correcting for misclassification due to imperfectly linked data, estimates of the rank-rank slope of occupation status for fathers and sons are substantially higher than OLS estimates throughout the period.[8] To benchmark levels of mobility based on rank-rank slopes, we can compare them to estimates across commuting zones in the modern US, the vast majority of which fell between 0.24 and 0.4 (Chetty et al., 2014a).[9] I find rank-rank slopes of around 0.3 among cohorts born before the Civil War, compared to OLS estimates of around 0.15 in the same samples. Rank-rank slopes of 0.3 roughly match levels in the US today, whereas a rank-rank slope of 0.15 is considerably lower than the most mobile localities. Revised estimates of the rank-rank slope rise into the early twentieth century, roughly in parallel with OLS estimates. The relative increase in the period is smaller than OLS, potentially reflecting increases in the

[7] The package is available at https://github.com/ramattheis/misclassifyr/.

[8] I structure the model of misclassification for these estimates so that the latent outcome corresponds to son's occupation at a point in time. Consequently, estimates of the rank-rank slope in this paper do not directly address concerns related the well-studied life-cycle bias in estimates of intergenerational mobility Lee and Solon (2009). Concerns about life-cycle bias are partially mitigated by observing sons' outcomes later in adulthood and by using occupation to define the rank rather than income, which may be thought to be relatively less noisy across time. The methodology in this paper could be extended to explicitly incorporate life cycle patterns in outcomes that may otherwise bias estimates of mobility, though this is left to future work.

[9] Two differences complicate the direct comparison of my estimates of the rank-rank slope with those in Chetty et al. (2014a). First, I use rankings based on occupations rather than averages of income, as individual-level income isn't observed in national censuses before 1940. Second, I define occupation ranks in the population rather than in the sample.

quality of census data later in the period. For cohorts born around the turn of the twentieth century, I find estimates of the rank-rank slope of nearly 0.5, reflecting levels of mobility lower than anywhere in the US today.

Connecting the revised estimates in this paper with modern estimates unaffected by high levels of record linkage error, I find a U-shaped pattern of intergenerational mobility over the last two centuries. My estimates suggest—contrary to famous examples of upward mobility during this period, like Walter Chrysler or William Boeing—that individuals who entered the workforce during the high levels of inequality in the Gilded Age, the Roaring '20s, or the Great Depression experienced exceptionally low levels of relative mobility.

Related Literature This paper contributes to two broad literatures: the measurement of intergenerational mobility, and the challenge of learning when information about individuals is scattered across sources.

First, this paper contributes to the literature on the measurement of historical mobility in the United States (Ferrie, 2005; Long and Ferrie, 2013; Olivetti and Paserman, 2015; Feigenbaum, 2018; Song et al., 2019; Collins and Wanamaker, 2022; Ward, 2023; Buckles et al., 2023). Compared to the majority of the literature, I find substantially lower levels of intergenerational mobility among cohorts born between 1830 and 1910. In particular, my estimates are inconsistent with previous work which found exceptionally high levels of intergenerational mobility in the mid 19th century relative to today, though it still may be true that levels of mobility were high in the US in that period relative to those in European countries (Long and Ferrie, 2013; Feigenbaum, 2018; Song et al., 2019).

The closest related paper is Ward (2023), which studies the role of racial gaps and measurement error in occupations in intergenerational mobility in US history. Whereas Ward (2023) focuses on the gap between White and Black Americans and assumes classical measurement error in occupations, this paper considers how nonclassical measurement error due to record linkage affects estimates of intergenerational mobility, primarily among the White population.[10] Consequently, direct comparisons of estimates of the rank-rank slope between those in Ward (2023) and in this paper are difficult.

This paper also relates to the much larger theoretical and empirical literature on the measurement and causes of intergenerational mobility (Lee and Solon, 2009; Chetty et al., 2014b, 2020; Ray and Genicot, 2023; Jácome et al., 2021). The estimates in this paper provide historical context for modern levels of intergenerational mobility and debates about the factors contributing to upward mobility. In particular, the U-shape pattern of estimates found in this paper is consistent

[10] My proposed methodology can be extended to include measurement error in the regressor, though I do not do so in this paper. There is undoubtedly some error in the occupations in historical census data. However, the quantity and source of the error is debatable and an important open question for US economic historians.

with falling rank-rank slopes reported in Jácome et al. (2021) for cohorts born between 1910 and 1970.

Beyond work concerning intergenerational mobility, this paper contributes to the large and varied literature on record linkage. My proposed methodology is applicable to the large and rapidly growing literature on historical census record linkage.[11] More broadly, the measurement-error approach to estimation with imperfectly linked data is complimentary to approaches that rely on generative models of record linkage (Han and Lahiri, 2019), or attempt estimation without explicitly linking records (Olivetti and Paserman, 2015; Balabdaoui et al., 2021; Pananjady and Samworth, 2022; D'Haultfœuille et al., 2023; Santavirta and Stuhler, 2024). Last, the misclassification models discussed in this paper may be useful in any setting with a repeated measure, regardless of how the linked data are generated, so long as the identifying assumptions are satisfied (Betancourt et al., 2022).

2 Models of Misclassification and Mobility

Researchers often hope to study the relationship between two variables—such as the economic status of parents and children in adulthood—which are observed in separate sources. If records are linked imperfectly across sources, then the data observed by the researcher reflect two data generating processes: the joint distribution of the unobserved, true values of the variables and the distribution of the observed, mismeasured data given the latent truth due to imperfections in record linkage. I refer to the former component of the problem as the "economic" part of the model and the latter as the "misclassification" component. In this section, I present a menu of models for each component of the problem. I start with the most flexible approach, building on the identification result in Hu (2008). For each model, I allow for fully flexible dependence on a vector of controls, which may include information used in the linking process. I then introduce additional structure to the problem which takes advantage of contextual knowledge about the economic question of interest and the particular structure of errors due to imperfectly linked data. Last, I discuss a maximum likelihood approach to estimation and inference.[12]

2.1 Nonparametric models of misclassification

The researcher observes N independent and identically distributed draws of four discrete variables: a regressor X, two noisy measures Y_1 and Y_2 of a latent outcome Y^*, and linkage controls

[11] For recent reviews of work using linked historical census data in economic history and social science more broadly, see Ruggles et al. (2018); Bailey et al. (2020a); Abramitzky et al. (2021a).

[12] I discuss details about the implementation in Appendix D.

W.[13] The number of unique values taken by the regressor and the observed and unobserved outcome variables is assumed to be the same and is denoted J.[14] In the context of intergenerational mobility, X is the father's occupation and Y_1 and Y_2 are two measures of a son's occupation. While Y_1 and Y_2 are interchangeable in the nonparametric model, I will refer to Y_1 as the primary outcome and Y_2 as the instrument or repeated measure. The roles of Y_1 and Y_2 are differentiated in the parametric models of misclassification below.

Without imposing some restrictions on measurement error, almost any distribution of the data X, Y^* can be rationalized (Molinari, 2008). Below, I discuss a set of assumptions sufficient for point identification of the economic and misclassification components of the model. Point identification for a broader class of misclassification models was established in Hu (2008), and I connect the assumptions and notation in this paper to those in Hu (2008) in appendix section C.

First, I assume that the regressor X and the noisy measures Y_1, Y_2 are mutually independent conditional on the latent truth Y^*.

Assumption 1 *The data are mutually independent conditional on the latent outcome:*

$$X \perp Y_1 \perp Y_2 \quad | \quad Y^*, W$$

Note that the scalar analogue of assumption 1 is weaker than that of classical measurement error, since Y_1 and Y_2 may still depend on Y^* arbitrarily so long as they are mutually independent. When measuring intergenerational mobility with linked census data, assumption 1 implies that, given knowledge of the son's actual occupation in adulthood, we would not be able to better predict the son's occupation observed in one census if given knowledge of the father's occupation or the son's occupation in another census.

There are a few important threats to this assumption when using linked data. First, this assumption may be violated if information about the regressor and the outcomes are used in constructing the sample. If, for example, a researcher or genealogist discriminated between candidate links based on the observed occupations in two census years or based on the occupations of sons and fathers in adulthood, this may introduce dependence in Y_1, Y_2, and X even after conditioning on the truth. While it is standard practice to exclude such information when linking

[13] I refer to X as the regressor and Y^* as the outcome as the primary object of interest in this application is a regression coefficient for Y^* on X. The roles of the outcome, instrument, and regressor are mostly interchangeable, though, and can be considered arbitrary discrete variables. The choice of the term "linkage controls" for the variable W reflects that information in W is assumed to be related to the linkage process, such as place of birth, age, race, and name commonness for individuals. The primary role of W is to allow flexibility in the relationship among the other variables in the model, as it does in Assumption 1.

[14] Identification in the model does not require the regressor X to have support on the same number of values as the latent outcome Y^* Hu (2023). However, for consistency with the original result in Hu (2008) and because the number of unique values is the same for X and Y^* in my main application, I assume that X and Y^* share a common number of dimensions.

across censuses in most academic work, it is important to keep this assumption in mind while interpreting results that rely on relatively less well understood approaches to record linkage, such as genealogical links.

Assumption 1 may also be violated if it is more likely than random for mistaken links to find the same individual. This concern is more serious if the typical underlying cause of record linkage errors is permanent changes like permanent name changes, deaths, or emigration. Alternatively, violations of this assumption are less likely if errors in record linkage are due to idiosyncratic features like choices of the census interviewer or errors in the digitization of the archival records. For example, Ghosh et al. (2024) show that the quality of handwriting in the original archival data causally affects the rate of record linkage in standard samples.

Last, Assumption 1 requires that the propensity of linkage outcomes—i.e. whether an individual is correctly or mistakenly linked—be conditionally unassociated with X and Y^*. For example, if black Americans or recent immigrants are more likely than white natives to be incorrectly linked, then errors in the observed occupation in one census year would predict errors in another. This concern highlights the importance of conditioning on information, like place of birth and race, that is likely to predict linkage outcomes.

Under Assumption 1, the conditional distribution of X, Y_1, Y_2 given W is:

$$Pr(X, Y_1, Y_2|W) = \sum_{y^* \leq J} Pr(Y_1|Y^* = y^*, W) Pr(Y_2|Y^* = y^*, W) Pr(Y^* = y^*, X|W) \tag{1}$$

In the context of measuring intergenerational mobility, the summand in this expression is the product of the probability of an occupational transition from father to son $Pr(Y^* = y^*, X|W)$ and the misclassification probability of the son's occupation in each observation in adulthood $Pr(Y_1|Y^* = y^*, W)$ and $Pr(Y_2|Y^* = y^*, W)$. The probability of observing a particular combination of occupations for a father-son pair, then, sums this probability over the possible values of the son's latent, actual occupation.

It will be useful to introduce some notation. For simplicity, I will omit the control variable W from all notation below with the understanding that all probabilities implicitly condition on W. Let $\pi_{i,j} := Pr(Y^* = i, X = j)$, and let the matrix Π collect the full joint distribution of X and Y^*:

$$\Pi := \begin{pmatrix} \pi_{1,1} & \cdots & \pi_{1,J} \\ \vdots & \ddots & \vdots \\ \pi_{J,1} & \cdots & \pi_{J,J} \end{pmatrix}$$

When X and Y^* are the occupations of fathers and sons, the matrix Π encodes the probability of all transitions from father's occupation to son's occupation.

Similarly, let the probability of conditional probability of observing $Y_1 = j$ and $Y_2 = k$ given the latent outcome $Y^* = i$ is denoted $\delta_{1,j,i} := Pr(Y_1 = j|Y^* = i)$ and $\delta_{2,k,i} := Pr(Y_2 = k|Y^* = i)$. The probability of any misclassification is summarized in the matrices $\Delta^{(1)}$ and $\Delta^{(2)}$ where

$$\Delta^{(1)} := \begin{pmatrix} \delta_{1,1,1} & \cdots & \delta_{1,1,J} \\ \vdots & \ddots & \vdots \\ \delta_{1,J,1} & \cdots & \delta_{1,J,J} \end{pmatrix} \qquad \Delta^{(2)} := \begin{pmatrix} \delta_{2,1,1} & \cdots & \delta_{2,1,J} \\ \vdots & \ddots & \vdots \\ \delta_{2,J,1} & \cdots & \delta_{2,J,J} \end{pmatrix}$$

Let θ collect all the elements of Δ and Π. Last, let $\mathbf{n}$ be a tabulation of the data, where $n_{q,r,s} := \sum_{i \leq N} \mathbf{1}\{x_i = q, y_{1,i} = r, y_{2,i} = s\}$.[15]

I restrict attention to cases in which the distribution of the regressor varies for each value of the outcome Y^*.

Assumption 2 *There are no two values of the outcome Y^* for which the distribution of X conditional on Y^* is identical: $\Pi_{i,\cdot} \not\propto \Pi_{j,\cdot}$ for all $i \neq j$.*

In the context of intergenerational mobility, Assumption 2 states that there are no two occupations (among sons) whose fathers share the same distribution of occupations. This assumption is satisfied, for example, if in pairwise comparisons, sons and fathers are more likely to have the same occupation.

I also restrict attention to cases in which misclassification is not extremely severe:

Assumption 3 *For any value of the latent outcome, correct classification is the more common than not: $\delta_{1,i,i} > 1/2$, $\delta_{2,i,i} > 1/2$, $\sum_{j \neq i} \delta_{1,i,j} < 1/2$, $\sum_{j \neq i} \delta_{1,j,i} < 1/2$, $\sum_{j \neq i} \delta_{2,i,j} < 1/2$, $\sum_{j \neq i} \delta_{2,j,i} < 1/2$.*

In linear algebra, this condition is known as diagonal dominance in rows and columns. While false-positive error rates in linked historical data are often high, most cases are thought to have false-positive rates far below 50% (Bailey et al., 2020a).

Likelihood The probability of a particular realization of the triple $(X, Y_1, Y2) = (i, j, k)$ is, following equation (1):

$$Pr(X = i, Y_1 = k, Y_2 = l) = \sum_{j \leq J} \delta_{1,k,j} \delta_{2,l,j} \pi_{j,i}$$

Since draws are i.i.d., the likelihood of the data $\mathbf{n}$ is a mixture of multinomials summing over the

[15] For simplicity of the likelihood expression, the index of $\mathbf{n}$ is increasing first in the values of X, then Y_1, and last Y_2: $\mathbf{n} := (n_{1,1,1}, n_{2,1,1}, ..., n_{J,1,1}, n_{1,2,1}, ..., n_{1,J,1}, n_{1,1,2}, ..., n_{J,J,J})$.

latent outcome Y^*:

$$Pr(\mathbf{n}|N,\theta) = c(\mathbf{n}) \prod_{i,k,l\leq J} \left(\sum_{j\leq J} \pi_{j,i}\delta_{1,k,j}\delta_{2,l,j} \right)^{n_{i,k,l}}$$

where the normalizing constant $c(\mathbf{n}) := \frac{N!}{\prod_{i,k,l\leq J} n_{i,k,l}!}$ does not depend on the parameters θ. The log likelihood is:

$$\ell(\mathbf{n}|N,\theta) - \log c(\mathbf{n}) = \sum_{i,k,l\leq J} n_{i,k,l} \log \sum_{j\leq J} \delta_{1,k,j}\delta_{2,l,j}\pi_{j,i} = \mathbf{n} \cdot \log(\text{vec}(\Pi^\top \Delta^{(1)} \overline{\Delta^{(2)}})) \tag{2}$$

where vec(M) flattens the matrix M column-wise and $\overline{\Delta^{(2)}}$ is a $J \times J^2$ matrix with the rows of $\Delta^{(2)}$ stretched across the diagonals.

As a special case of Hu (2008), Assumptions 1, 2, and 3 together imply that the nonparametric joint distribution Π and the misclassification matrices $\Delta^{(1)}$ and $\Delta^{(2)}$ are point identified. I discuss the connection between Hu (2008) and these assumptions in appendix section C. As the parametric models below are nested by the nonparametric model, they are also point identified.

2.2 Parametric models of Δ and Π

The flexibility of the nonparametric model comes at the cost of high-dimensionality, as the number of free parameters in Π, $\Delta^{(1)}$, and $\Delta^{(2)}$ is quadratic in the number of unique values taken by Y^*, J. Estimation and inference with the nonparametric model becomes infeasible when the dimension of the data is large, as it would be for richer definitions of occupation or when estimating rates of migration between counties.[16] Consequently, studying richer outcomes will require imposing additional restrictions on patterns of misclassification or the underlying economic model. Fortunately, misclassification from imperfect record linkage has a clear and simple pattern whenever the record linkage errors are not strongly associated with the variables of interest. Below, I list parametric alternatives to the nonparametric approach above.

Record linkage measurement error Errors in variables due to imperfect record linkage have a particular structure when record linkage errors are not strongly associated with the variables of interest. When one record is incorrectly linked to another, we can think of the characteristics as a random draw from the distribution of entities with similar linkage information.

To make this intuition more concrete, suppose that misclassification process for Y^* is as follows:

[16] Concretely, the number of free parameters in the nonparametric model is $3J^2 - 2J - 1$. For context, this means that measuring occupational transitions with all "occ1950" categories in IPUMS would require estimating over 200,000 free parameters, and measuring county-level migration would require estimating over 20 million free parameters.

for $Y^* = y$, the probability of a successful link is $1-\alpha(y)$, in which case $Y = Y^*$; with probability $\alpha(y)$ the link is a failure and Y is drawn independently from the distribution of Y^* in the population.[17] This form of misclassification error results in the following conditional distribution for the observed outcome Y on the latent truth Y^*:

$$Pr(Y_j = y'|Y_j^* = y) = \begin{cases} (1-\alpha) + \alpha Pr(Y^* = y) & \text{for } y' = y \\ \alpha Pr(Y^* = y') & \text{for } y' \neq y \end{cases} \tag{3}$$

Stacking the conditional probabilities across the values of Y^*, we can express the misclassification distribution as the sum of two terms: a diagonal matrix weighted by $(1-\alpha)$ corresponding to the successfully linked observations, and a matrix with each column equal to the marginal distribution of Y^*:

$$\Delta^{RL} := (1-\alpha)\mathbf{I}_J + \alpha \begin{pmatrix} Pr(Y^* = y_1) & \cdots & Pr(Y^* = y_J) \\ \vdots & & \vdots \\ Pr(Y^* = y_1) & \cdots & Pr(Y^* = y_J) \end{pmatrix} \tag{4}$$

where $\mathbf{I}_J$ is a $J \times J$ identity matrix and the matrix in the second term has constant columns.

Equation (4) suggests a few different approaches to modeling $\Delta^{(1)}$ and $\Delta^{(2)}$ in practice. First, we can allow the rate of record linkage error and the marginal distribution of Y^* to be flexible:

$$\Delta_1^{RL} := \text{diag}(1-\boldsymbol{\alpha}) + \boldsymbol{\alpha}\boldsymbol{\rho}^\top \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_J)^\top$ reflects different rates of record linkage for latent values of Y^* and $\boldsymbol{\rho} = (\rho_1, ..., \rho_J)^\top$. Equation (5) provides a relatively flexible model of misclassification due to record linkage error. In some cases, such as when using full censuses, we may take the marginal distribution of Y^* as known.[18] In these cases, we may assert that $\rho = \rho^*$ and hold it fixed when estimating the misclassification matrices:

$$\Delta_2^{RL} := \text{diag}(1-\boldsymbol{\alpha}) + \boldsymbol{\alpha}\boldsymbol{\rho}^{*\top} \tag{6}$$

Last, if we assume the rate of record linkage error is independent of Y^*, we arrive at a model of misclassification with just one free variable:

$$\Delta_3^{RL} := (1-\alpha)\mathbf{I}_J + \alpha\mathbf{1}_j\boldsymbol{\rho}^{*\top} \tag{7}$$

I compare the pattern of misclassification errors in equation (7) to the observed distribution of

[17] This pattern of errors can be generated by a model of record linkage in which rates of record linkage error are independent conditional on Y^*.

[18] More specifically, we may assume to know the marginal distribution of Y^* *within* the linked sample. Selection into the linked sample could shift the distribution of Y^* from what is observed in a sample cross-section.

errors in a naturalistic exercise in appendix section A.2. Without imposing the independence of record linkage errors and outcomes, I find that equation (7) closely approximates the observed misclassification distribution in an exercise where record linkage errors are the only source of misclassification, as shown in figure A.1.

Mixed models of measurement error Contextual knowledge may suggest that the first and second measure of the outcome have a different misclassification structure. In these cases, we can choose different models for $\Delta^{(1)}$ and $\Delta^{(2)}$. For example, suppose that we hope to measure migration across locations l between two periods 0 and 1. Location in the "origin" period l_0 is observed without error, and the location in the "destination" period l_1 differs only from the true location l_1^* only due to imperfect record linkage. To overcome error from imperfect record linkage, we can follow individuals to their location at another point in time, l_2. Differences between l_2 and the latent location l_1^* reflect both record linkage errors (i.e. the difference between l_2 and l_2^*) and mobility between the two periods (i.e. the difference between l_1^* and l_2^*). Assuming that record linkage errors are independent across periods, it makes sense in this context to use a record-linkage model for misclassification in l_1 and a more flexible model for l_2. In the context of estimating intergenerational mobility, occupations change over time, and we may choose a model of misclassification assuming record linkage is the primary form of misclassification, such as Δ_3^{RL} in equation (7) for $\Delta^{(1)}$, and a flexible model for misclassification in the second observation of occupation, $\Delta^{(2)}$.

Stacked models of measurement error Additional variables may be useful for estimating misclassification error. For example, if the location of residence for an individual is not used in the record linkage process, it can be used to improve estimates of misclassification error due to imperfect record linkage. In these cases, we can "stack" repeated tabulations of the underlying data with alternative definitions of the outcome and regressor and estimate a (partially) *shared* model of misclassification.

Parametric models of Π The most flexible model considered in this paper puts zero restrictions on the joint distribution of X and Y^*. While flexibility in this relationship allows for fully data-driven analysis of the relationship between X and Y^*, there may be natural economic or other contextual restrictions on this relationship. Imposing parametric models for the joint distribution of X and Y^* can lead to more efficient estimation by ruling out irrelevant portions of the parameter space of Π. Alternatively, we are often interested in functions of Π or model parameters that may generate joint distributions of X and Y^* rather than Π itself. The misclassification component of the model can be viewed as a "wrapper" to address errors due to imperfect record linkage while directly estimating parametric models of interest. For example, we can explicitly model the relationship between child and parents occupational status, as in Becker and Tomes (1979).

2.3 Estimation and Inference

I estimate the distribution of the economic data Π and the misclassification matrices $\Delta^{(1)}$ and $\Delta^{(2)}$ via maximum likelihood. Because all of the variables are discrete in this setting, the likelihood is a simple function of the tabulation of the data, as shown in equation (2).

While $\Delta^{(1)}$, $\Delta^{(2)}$, and especially Π may be of interest to researchers on their own, the target is often a functional $\beta(\cdot)$ of the joint distribution of X and Y^*. For example, the rank-rank slope of occupations for fathers and sons is a function of Π where X is the father's occupation, Y^* is the latent true occupation of the son in adulthood, and the $\beta(\cdot)$ corresponds to the linear regression coefficient of the ranks of occupations for sons regressed on the ranks of occupations for fathers. Building on this observation, I use a simple plug-in estimator, which depends on an estimate for the joint distribution $\widehat{\Pi}$ and two vectors corresponding to the scalar values associated with each category for X and Y^*. As a slight abuse of terminology, I will sometimes refer to plug-in estimates $\beta(\widehat{\Pi})$ as maximum likelihood estimates of β.

I estimate the variance of the parameters in the model by the inverse Fisher information. To obtain confidence intervals for $\beta(\widehat{\Pi})$, I use the delta method.[19]

3 Validation

To assess the performance of an estimator on imperfectly linked data, we would ideally like to compare estimates to those that we would get using perfectly linked data. The closest available data to such a ground truth in the context of historical censuses are the high-quality linked samples made available to the public by the LIFE-M project (Bailey et al., 2020b) and the Census Tree (Price et al., 2021). While these samples are believed to have much lower false positive rates than standard approaches, neither sample is perfectly representative of the full population, and the absence of a linked pair in either data set does not imply that the link is a false-positive[20] Consequently, it would be unclear if differences between OLS estimates using

[19] In practice, I often find that the Fisher information is singular to numerical precision. The MLE appears to be stable across multiple starting values for the optimization algorithm, suggesting that the source of the singularity is a failure of point identification. Instead, the issue is likely because parameters in the nonparametric model of Π are close to the boundary—i.e. zero. In these cases, I use the Moore-Penrose pseudoinverse of the Fisher information to estimate the variance of Π and downstream parameters. One more reliable alternative for constructing confidence sets for parameter estimates would be the Monte-Carlo confidence algorithms discussed in Chen et al. (2018). The `misclassifyr` package allows for constructing such confidence sets, though this approach to inference is more computationally demanding.

[20] The LIFE-M sample successfully links about one half of records, and the genealogical links in the Census Tree sample represent less than 20% of the full census.

LIFE-M or the Census Tree and estimates from the proposed methodology on standard linked data from equivalently defined populations would be due to bias in the proposed estimator or selection bias in either linked sample. For this reason, I use a synthetic ground truth for the validation exercise, in which sample selection is held constant and the rate of data corruption from false positive links is known. Specifically, I use common algorithms to construct linked samples across synthetically corrupted copies of the 1940 census. I estimate returns to schooling with a simple linear Mincer regression (8) in each linked sample via OLS and maximum likelihood estimates for models of misclassification.

$$\underbrace{Y_i^*}_{\text{wage income}} = \gamma + \beta \underbrace{X_i}_{\text{years of education}} + \epsilon_i \tag{8}$$

I then compare OLS and MLE estimates with the OLS estimates that would be obtained on the correctly linked subset of the data. By conditioning on the linked sample, I shut down the possibility of selection bias in the exercise.

I generate corrupted copies of the 1940 census as follows. I start with all men age 30 to 49 in the full count 1940 census with observed years of educational attainment and wage income. From this population, I independently draw three samples: a 20% sample which I denote as A and two 50% samples, B_1 and B_2. Missing observations in B_1 and B_2 reflect unrecoverable mutations in the linkage information—such as changes in variables used to "block" candidate links, like the recorded place of birth, or swapping of middle and first names after childhood—as well as emigration and death. The scale of these changes reflects the difficulty of linking historical census data. The LIFE-M project, for example, is unable to link about 50% of the initial sample despite extensive effort and the use of additional information from auxiliary sources.

I then add naturalistic, independent synthetic noise to names and ages in each sample, calibrated to the degree of noise found in linked historical census data. Specifically, I add -1, 0, or 1 from each individual's age with probability 0.1, 0.8, and 0.1 respectively. I then round ages for 20% of each sample to the nearest age ending in five or zero to reflect age heaping. I swap names for common "nicknames" observed in linked historical data, such as "Wm" for "William" or "Jon" for "Jonathan". Finally, I garble letters in last names, either deleting or swapping a letter, and swap or add middle initials with probability 20%. Table E.1 compares distances in names and ages in one draw of the validation exercise to those in the LIFE-M sample.

While the synthetic noise in the validation exercise is drawn independently, this does not imply that linkage errors are independent in the B_1 and B_2 samples.[21] The frequency of name, age, and birth place in the population lead to a range of successful linkage rates which may be associated

[21] To illustrate the influence of increased dependence between the measurement error in Y_1 and Y_2, one could introduce a non-zero correlation in the missingness of samples B_1 and B_2.

with educational attainment and earnings. Indeed, I find that linkage errors are correlated in the B_1 and B_2 samples.

After generating the corrupted samples, I link sample A to B_1 and B_2 using five linkage algorithms: standard and conservative versions of the deterministic algorithm (ABE) in Abramitzky et al. (2014); standard and conservative versions of the unsupervised approach in Abramitzky et al. (2019) that estimates a Felligi-Sunter (FS) model of record linkage; and a "Representative" approach which estimates FS probabilities and takes the best candidates without thresholding on a minimum linking probability or a minimum difference between the linking probability of the best and second-best candidates.

For each linked sample, I estimate β in the Mincer regression (8) via OLS and via maximum likelihood for misclassification models. I report estimates for two models of misclassification: a nonparametric model of the misclassification and a record linkage model of misclassification, specifically Δ_1^{RL} defined in equation (5). I bin wage income into eight groups: zero wage income, increments of $500 up to $3000, and between $3000 and the 1940 top-code of $5000. While estimating models of misclassification, I condition on a coarsened definition of birthplace and age rounded to the nearest ten years, for a total of 18 covariate cells. Note that the role of the linking controls is not to project out variation in X, in the sense of Frisch-Waugh-Lovell theorem. Instead, conditioning allows estimates of Π, $\Delta^{(1)}$, and $\Delta^{(2)}$ to vary across values of the linkage controls. I take a weighted sum over estimates of Π across covariate cells to compute an estimate of β for the full population.

The influence of record linkage error and the adjustment provided in the estimates of the misclassification model are illustrated in Figure 1, which plots results from one draw of the validation exercise. The bottom right matrix shows the difference in the empirical distribution of wages and educational attainment in the full linked sample and its correctly linked subset. Relative to the correctly linked subset of the sample, the observed distribution is missing mass around the diagonal of the matrix with excess mass further off-diagonal. This bias reflects that the distribution of wages does not depend strongly on educational attainment for incorrectly linked pairs of observations. The top right matrix shows the difference between the maximum likelihood estimate of the joint distribution of wages and educational attainment from the misclassification model and the empirical distribution in the linked sample. This difference largely mirrors the bias in the empirical distribution in the linked sample, as the MLE adds mass near the diagonal and subtracts mass in the off diagonal, correcting for the bias from record linkage errors. The matrix in the bottom left shows the remaining error in the MLE relative to the empirical distribution in the correctly linked sample. While the MLE does not perfectly fit the actual joint distribution, the clear pattern of bias in the on/off diagonal is no longer seen.

The performance of the maximum likelihood estimator for two models of misclassification is summarized in Table 1. The first row of the table shows the average proportion of links that were correct across draws of the validation for the particular linkage algorithm used. The conditions are ordered in terms of the difficulty of the inference problem, with higher false positive rates in the sample when moving from left to right. The median bias in the maximum likelihood estimates for each model of misclassification is reduced relative to OLS across all conditions. In the first two columns, the false positive rate is quite low, at four or five percentage points. This level of accuracy in the linked sample is likely a better approximation of links in higher quality modern data. The bias in OLS estimates is correspondingly small, at about −\$3 or −\$4 per year of education, relative to the OLS estimate in the full population of \$91.40 per year of education. The median bias of the MLE is about half the magnitude of the bias in OLS at about −\$1 or −\$2 per year of education for both the nonparametric and record linkage models of misclassification.

The third and fourth columns report results for linked samples that are a better approximation of the challenges posed by imperfectly linked historical censuses, with false positive rates of 13pp and 21pp respectively. The higher false positive rates lead to larger biases in OLS estimates, at about 10% and 15% of the population OLS coefficient respectively. Bias in the MLE for both models remains at about the same low level in Condition 3, now around 90% smaller than in the OLS estimates. The median bias starts to increase in the MLE for the fully nonparametric model, increasing in magnitude to −\$5 per year of educational attainment. The last column reports estimates in the most challenging inference environment, with false-positive rates of 44pp on average. Condition 5 is a more challenging problem than in most settings with historically linked data, though it may better reflect links with limited information or links among difficult subpopulations. In this setting, OLS is severely biased, with attenuation of over 30%. Bias is also larger in MLE estimates for both misclassification models, with attenuation of about 5% and 13% in the record linkage and nonparametric models respectively. The false positive rate in condition 5 is sufficiently high that Assumption 3 nearly fails. Correspondingly, it appears that the MLE for models of misclassification, especially the most flexible models, begin to suffer. It is notable, though, that the direction of the bias appears to be consistently in line with that in OLS, and the magnitude of the bias remains a fraction of that in OLS.

4 Intergenerational Mobility 1850-1920

In this section, I present revised estimates of intergenerational mobility allowing for misclassification in sons' occupation due to record linkage errors. Maximum likelihood estimates of misclassification suggest that sons' occupations in adulthood are measured with substantial error, with the latent and observed occupation agreeing about 60% of the time, averaging across years and occupations. Consequently, I find that OLS estimates are substantially attenuated relative to maximum likelihood estimates, with levels of attenuation varying from approximately

50% for cohorts born in the mid nineteenth century to approximately 30% among cohorts born around the turn of the twentieth century.

4.1 Data sources and sample construction

The main analysis depends on three pieces of data: I use complete count census data to observe occupations and construct father-son pairs; I use existing links between censuses to infer the occupations of sons in adulthood; and I use census data along with supplementary sources to assign rankings to occupations over time. To facilitate replicability, the main analysis depends only on publicly available data and can be implemented with a new R package.

I use IPUMS complete count data for censuses between 1850 and 1940. I use three sources of crosswalks for linked census data: the Census Linking Project Abramitzky et al. (2020), the Census Tree (Price et al., 2021), and the IPUMS Multigenerational Longitudinal Panel project Helgertz et al. (2023).[22] For each census year between 1850 and 1910, I subset to father-son pairs in which the son is below age 18. The misclassification model relies on repeated observations of the outcome, so I restrict the sample to father-son pairs in which the son is linked into two censuses as an adult. I consider only one group of three census years for each base year for ease of interpretation. When possible, I use census years two and three decades after the base year for the instrument and main outcome, when sons are 20-37 and 30-47 years old, respectively. The base years 1860 and 1870 measure outcomes for sons at different ages due to the destruction of the 1890 census. I weight final samples to adjust for selection bias on observables in record linkage. Specifically, I estimate propensity to be included in the linked sample via logistic regression with LASSO penalization on father's age, race, birth place, and occupation.

It is important to know which variables enter the linkage algorithm. Any information that is predictive of misclassification error—and consequently record linkage errors—should be included as a linkage control. Consequently, the methodology proposed in this paper may perform poorly when variables that are used in the analysis are also used to construct the linked sample. There is tension, then, between obtaining (presumably) higher quality links based on a richer set of information and retaining enough observations within each covariate cell for estimation.

To reduce the dimension of the problem, I use the "mesooccupation" coding of occupations from Song et al. (2019). This categorization is composed of nine groups: lower manual occupations, such as operatives and laborers; service workers, such as janitors and deliverymen; farmers, fisherman, and other primary occupations; craft occupations, such as blacksmiths, bricklayers, and bakers; sales occupations, such as insurance and real estate agents; clerical occupations,

[22] Complete count US census data may be downloaded at https://usa.ipums.org/usa/. Crosswalks for linked census data may be downloaded at https://censuslinkingproject.org/ for the Census Linking Project, https://www.censustree.org/ for the Census Tree, and https://usa.ipums.org/usa/mlp/mlp.shtml for the IPUMS Multigenerational Linked Panel.

such as postal clerks and bookkeepers; managers and officials; other professions, such as school teachers and religious workers; and classical professions, such as doctors, lawyers, engineers, and professors.[23]

The primary estimand is the rank-rank slope of father and son's occupations. For any estimate of the joint distribution of father's and son's occupations and for each ranking of occupations, we can compute a corresponding rank-rank slope of status. I rank occupations using estimates of earnings and human capital based on census data and other sources, summarized in Table B.1. One important decision in constructing occupational status rankings is whether you allow the ranking to depend on covariates like race, region, and birth year. In reality, the average earnings or educational attainment may vary considerably across places, between groups, or through time and we may wish for measures of occupational status to reflect this variation (Saavedra and Twinam, 2020; Song et al., 2019). Allowing measures of occupational status to depend on race, for example, substantially increases the rank-rank slope in father-son occupational status due to the large racial gap in outcomes between the White and Black populations in the United States (Ward, 2023), as can be seen in Figure B.1. In my view, the choice of occupational ranking is a choice in estimand rather than different approaches to estimating a common target. Interpreting any rank-rank slope as the "true" association would require writing down a model that relates the observed data to the preferred estimand or bluntly asserting that the occupational ranking is correct.

The most notable source of missing coverage in the US Census is the exclusion of slaves, who were recorded in separate schedules in 1850 and 1860.[24] Consequently, linked census data based on the 1850 and 1860 censuses miss the vast majority of the non-White individuals and levels of intergenerational mobility in these samples will be higher than it is in the full population. Following Ward (2023), I impute father-son pairs among slaves in the 1850 and 1860 censuses under the assumption that almost all Black individuals born in the US in the antebellum period would be the children of slaves and that outcomes like the average income, the average years of schooling, or the proportion literate among slaves is zero. Additionally, to approximate a sample in which sons are observed twice in adulthood, I use links between Black adults in the 1870 and 1880 censuses and 1880 and 1900 censuses who were age 0-17 in 1850 and 1860 respectively. These assumptions ignore the (relatively small) free black population in the US before the Civil War and any variation in status among the enslaved, which certainly was non-zero (Fogel and Engerman, 1974).[25] The scale of errors due to these assumptions, though, may be small relative to linking

[23] The original categorization in Song et al. (2019) includes a separate category for members of armed forces. Because this is a relatively small and inconsistently sized occupational group over time, I subsume members of armed forces into the managers and officials category, which matches most closely on median educational attainment and total income.

[24] The digitized Slave schedules are available here: https://usa.ipums.org/usa/slavepums/documentation/about.html.

[25] There were large differences, for example, in the experience of slaves coerced into plantation labor and those in

the full free black population in the 1850 and 1860 censuses. For consistency in the population across years, I report estimates either based on the White population or the full population in the census with the addition of imputed Freedmen. I include the imputed slave populations for all measures of occupation rank.

4.2 Occupation Misclassification

While we are primarily interested in the joint distribution of the correctly measured data, the approach proposed in this paper produces estimates of the misclassification error as a bi-product. In the context of estimating intergenerational mobility, this means that we will estimate the probability that a farmer is misclassified as a doctor in addition to the probability the son of a farmer becomes a doctor.

Figure 2 reports maximum likelihood estimates for a nonparametric model of misclassification in occupations in the 1910-1940 linkage. The figure appears to show the record linkage pattern of misclassification error, as discussed in section 2.2, with a heavy mass on the diagonal and off-diagonal mass roughly proportional to the marginal density of sons' occupations. Because the model is fully flexible, this pattern was not guaranteed. Values in the diagonal of the distribution are consistent with false positive rates slightly over 30% in the linked census data, which is at the high end of the range of false positive linkage rates found when evaluating linked historical censuses against a ground truth (Bailey et al., 2020a).

At the same time, there are clear deviations from the record linkage pattern of misclassification, likely reflecting occupational changes between the two census years in which the sons' occupations were observed, 1930 and 1940. In particular, there appears to be a large degree of mobility in the bottom of the matrix, in which the MLE suggests that 24% of individuals whose latent occupation is as a service worker are observed in a lower manual occupation, and 16% of individuals who are lower manual workers are observed in a craft occupation in 1940. Further up the occupation distribution, 18% of sales workers in the MLE are observed in an managerial or official occupation. This pattern of estimates reflects, in part, ambiguity in the definition of the sons' latent occupation.

While the nonparametric model of misclassification presented in section 2.1 is point identified under Assumptions 1, 2, and 3, the interpretation of the latent outcome is somewhat ambiguous when measured at two points in time. To make the make the mapping from the latent outcome to the observed data more concrete, we can use a mixed model of misclassification error in which the only source of misclassification in one of the two measures is assumed to be record linkage error. Under this model, we can interpret the latent outcome as being tied to a point in time. For example, when linking sons observed in childhood in 1910 to their outcomes in 1930 and

urban settings engaged in domestic service or craftwork.

1940, we can choose 1940 to be the year we wish to measure associations with the sons' latent occupation by enforcing that the misclassification matrix $\Delta^{(1)}$ reflects only measurement error due to imperfect record linkage, while a nonparametric model for $\Delta^{(2)}$ allows for misclassification for record linkage errors along with other sources of change in the outcome.

The mixed model of record linkage also allows for direct estimates of share of incorrectly linked observations. Figure 3 presents maximum likelihood estimates for α as defined in equation (7) for mixed models of misclassification in which $\Delta^{(1)}$ is assumed to have the form of Δ_3^{RL} and $\Delta^{(2)}$ is nonparametric. Maximum likelihood estimates of record linkage error are high for all census years, with false positive error rates near 40% throughout and as high as 50% for cohorts observed in childhood in the 1860 census. There does not appear to be a clear trend in the levels of record linkage error over time, which may reflect partially offsetting trends as both the quality of the underlying data improve and the composition of the linked sample shifts moving from earlier to later censuses.[26]

Across the three linked samples in Figure 3, the conservative version of the deterministic ABE algorithm in Abramitzky et al. (2014) has the lowest estimates of record linkage error, the standard variant of the ABE algorithm typically has the highest estimates of the record linkage error with estimates of false positive rates 5-10pp higher than the conservative samples, and the full Census Tree has estimates of false positive rates that typically fall between estimates of the other two samples. Consistent with this ordering of linked samples, OLS estimates of the rank-rank slope on the standard ABE linked samples are the lowest, estimates for the conservative ABE sample are consistently higher, and estimates in the full Census Tree sample are in between, as shown in Figure E.2.

Estimates of the false positive linkage rate vary significantly across values of the linkage controls. Figure E.1 plots maximum likelihood estimates of α within linkage control cells for the 1910-1940 linkage. Linkage error rates are substantially larger for the rural population than the urban population, with the exception of the rural population in the West, and higher among the foreign population than the native born.

The levels of the MLE for record linkage error are higher than prior estimates based on ground truth samples (Bailey et al., 2020a). It is possible—though subjectively unlikely—that ground truth samples themselves contain false positive errors. A fundamental challenge in historical record linkage is that we cannot know with certainty that a set of linked observations are correct, we can only assemble larger constellations of evidence in support of particular views of the world. Alternatively, it is possible that the subset of standard linked data that are not found in ground

[26] For example, the share of immigrants in the linked sample is substantially higher among the cohorts observed in childhood in 1900 and 1910 relative to those in 1850 and 1860. Estimates of linkage rates in the 1870 census appear to outliers relative to the other linked samples; this may be because it is the only initial census year in which the instrument is observed after (in 1910) the main outcome (in 1900). I am unable to keep the relative timing of the earlier and later years constant because the 1890 census is missing due to a fire in 1921.

truth linked samples have higher rates of linkage error than the overlapping samples. Last, other forms of measurement error may lead to inflated estimates of record linkage error. Errors in the recorded occupation—whether due to errors in the initial interview, the digitization of the archival text, or the IPUMS coding of raw occupation strings—could lead to higher estimates of α, especially if the distribution of misclassification errors is similar to the pattern generated by record linkage errors. There is some evidence in favor of the latter interpretation provided in Ward (2023), which found high rates of error in recorded occupations in two enumerations of St. Louis in 1880.

4.3 Intergenerational mobility

Figure 4 illustrates how estimates of the apparent relationship between father's and son's occupations change after correcting for misclassification from imperfect record linkage. The matrix on the left plots the maximum likelihood estimate of the conditional distribution of son's occupation in 1940 on father's occupation in 1910 (equal to $\widehat{\Pi}$ normalized to have columns sum to one) from a mixed model of misclassification in which measurement error in 1940 is assumed to have a record linkage structure, Δ_3^{RL}, and measurement error in the 1930 occupation is fully flexible. The matrix on the right plots the difference between the MLE and the empirical conditional distribution in the linked sample. As in the validation exercise, the MLE shifts mass away from far off-diagonal, where sons have the largest change from their father's status, and towards the diagonal, where similar occupations are shared across generations.

The maximum likelihood estimate increases the share of sons in the same occupation (or occupation group) as their father by seven, five, and six percentage points among the sons of lower manual workers, farmers, and classic professionals respectively—representing a 21%, 13%, and 27% increase over the observed shares. Consequently, estimates of the direct transmission of occupation are relatively high, with over a quarter of the sons of professionals (e.g., lawyers and doctors) becoming professionals, nearly a third of the sons of craftsmen becoming craftsmen, and over two-fifths of the sons of farmers staying in farming. Similarly, large jumps in upward or downward mobility are less common in the MLE. The share of sons of lower manual workers becoming managers and officials is 2 percentage points or 20% lower than observed, and the share of sons of professionals entering lower manual work 3 percentage points or 38% less.

Estimates of the joint distribution of occupations provide a detailed description of intergenerational mobility, but the primary estimand is the rank-rank slope of occupation status for fathers and sons. Specifically, I estimate β in the linear regression:

$$\underbrace{y_i^*}_{\text{Rank of son's occupation } Y^*} = \gamma + \beta \underbrace{x_i}_{\text{Rank of father's occupation } X} + \epsilon_i \tag{9}$$

where y^* is the population rank of the son's (latent) occupation Y^*, and x is the population rank of the father's occupation. I estimate β in equation (9) via OLS and via plug-in for the MLE of the joint distribution of fathers' and sons' occupations, $\widehat{\Pi}$.

Figure 5 plots OLS and MLE estimates of the rank-rank slope of occupation status, defined by the average educational attainment in the 1940 and 1950 censuses, for cohorts of White men born between 1832 and 1910. The level and trend of OLS estimates is consistent with prior work in the literature (Ferrie, 2005; Song et al., 2019). OLS estimates of the rank-rank slope are around 0.16 for cohorts observed in childhood in 1850 and 1860 censuses and rise to 0.3 for cohorts born between 1892 and 1910. The proportional change is striking, with levels of the rank-rank slope nearly doubling across the period. These estimates paint a picture of an exceptionally mobile economy in the mid nineteenth century US, with mobility gradually declining to levels similar to those observed in the US in modern data.

Estimates correcting for misclassification due to imperfect record linkage reveal a different story. Levels of rank-rank slope are substantially higher across the full period, with OLS estimates attenuated by about one third at the end of the period and about one half in the earliest censuses. The difference in the rank-rank slope between OLS and MLE estimates is roughly constant across the full period, while the relative bias declines considerably. While estimates of the rank-rank slope increase substantially over the second half of the nineteenth century, levels in the earliest censuses are not exceptionally mobile relative to modern economies. Instead, levels of mobility among cohorts born between the Civil War and World War One appear to have experienced low levels of mobility in the distribution of modern economy, and lower than in earlier and later periods of American History. Estimates of the rank-rank correlation for cohorts born in the 1970s are around 0.3, suggesting an inverse-U shape pattern in levels of intergenerational mobility in the US between the mid nineteenth century and the late twentieth.

The rank-rank slope of occupation status varies with the choice of misclassification model, though the qualitative conclusions do not. Figure E.3 plots the estimates of the rank-rank slope for OLS and MLE for four models of misclassification: a fully nonparametric model of misclassification, a flexible record linkage structure Δ_1^{RL}, a model of record linkage error with known marginal distributions of the outcome Δ_2^{RL}, and a mixed model of misclassification in which $\Delta^{(1)}$ is assumed to have the record linkage structure Δ_3^{RL} and $\Delta^{(2)}$ is fully flexible. There do not appear to be a clear patterns in the order of estimates across misclassification models. There is proportionately less variation in maximum likelihood estimates of the rank-rank slope across linked samples relative to OLS, as shown in Figure E.4. Accounting for variation in the false positive rate across linked samples, the MLE reduces the range of estimates relative average estimate, from about 20% on average across years for OLS to about 13% for MLE. The difference between OLS and MLE estimates is similar across cohorts when occupations are ranked by the average total income of workers holding that occupation in the 1940 and 1950 censuses, rather than average educational attainment.

5 Conclusion

This paper introduces a methodology for estimation using imperfectly linked data and applies it to provides new estimates of intergenerational mobility for cohorts born between 1832 and 1910. Building on the identification result in Hu (2008), I show how independent repeated measures of linked data can be used to estimate parameters even when imperfections in record linkage lead to highly nonclassical errors in observed variables. The methodology developed in this paper is not limited to the measurement of intergenerational mobility. And while the most direct applications include other topics that rely on linked historical censuses such as immigrant assimilation (Abramitzky et al., 2021b) or the causes of racial gaps (Althoff and Reichardt, 2024), this approach may also be useful in modern applications with (much) less than perfect links. As linked data across nonstandard sources become more common in economics—for example, linking Linkedin resumes to Github profiles (Gortmaker, 2024) or venture capital funding to technologies (Narain, 2024)—it will be important to consider the potential for biases due to imperfections in record linkage.

Revised estimates of intergenerational mobility suggest a different trajectory of mobility over the last 200 years. The rank-rank association for individuals born between 1830 and the Civil War is comparable to modern levels, rather than the exceptionally low levels reported in prior work Long and Ferrie (2013); Song et al. (2019). The association between fathers' and sons' occupation rank increases among cohorts born from the Civil War to the First World War, reaching levels just outside the range of mobility experienced in commuting zones in the US today (Chetty et al., 2014a).

While the levels of the rank-rank association found here for the cohorts born before the Civil War are comparable to the modern US, this does not necessarily imply that levels of mobility—for White men—were not *exceptional*. It is arguably more meaningful to compare levels of mobility to contemporary economies in Europe and the New World (Long and Ferrie, 2013; Pérez, 2019). It is not clear ex ante whether we should expect errors in record linkage in historical census data to be more or less common in the United States compared to Britain, Norway, or Argentina. Consequently, it may still be that the Americas provided high rates of intergenerational mobility for White men in the mid nineteenth century relative to Europe.

Additionally, the inverse-U shape pattern in the rank-rank association found in this paper reflects trends in *relative* mobility, rather than absolute mobility. The United States experienced phenomenal rates of growth over the last 200 years, with average incomes doubling five times between the birth of the earliest cohorts in this analysis and today (Wright, 2024). The quality and length of life improved alongside earnings (Gordon, 2016). Reconciling the low levels of relative intergenerational mobility during the turn of the twentieth century found here with the high levels of contemporary growth is a promising area for future work. Measuring upward mobility with confidence across US history would require information about earnings of occupations across time and place—potentially drawing from many disparate sources—which could then be combined with estimates of the transmission of occupation between fathers and sons presented here.

References

Abramitzky, R. and Boustan, L. (2022). *Streets of Gold: America's Untold Story of Immigrant Success*. Hachette Book Group, New York, NY.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Perez, S. (2021a). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

Abramitzky, R., Boustan, L., Eriksson, K., Pérez, S., and Rashid, M. (2020). Census linking project: Version 2.0.

Abramitzky, R., Boustan, L., Jacome, E., and Perez, S. (2021b). Intergenerational mobility of immigrants in the united states over two centuries. *American Economic Review*, 111(2):580–608.

Abramitzky, R., Boustan, L. P., and Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(3):467–506.

Abramitzky, R., Mill, R., and Perez, S. (2019). Linking individuals across historical sources: a fully automated approach. *Historical Methods*, pages 94–111.

Althoff, L., Gray, H. B., and Reichardt, H. (2024). The missing link(s): Women and intergenerational mobility. Working paper.

Althoff, L. and Reichardt, H. (2024). Jim crow and black economic progress after slavery. *The Quarterly Journal of Economics*, 139(4):2279–2330.

Bailey, M., Cole, C., Henderson, M., and Massey, C. (2020a). How well do automated linking methods perform? lessons from us historical data. *Journal of Economic Literature*, 58(4):997–1044.

Bailey, M., Cole, C., and Massey, C. (2020b). Simple strategies for improving inference with linked data: a case study of the 1850–1930 ipums linked representative historical samples. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):80–93.

Bailey, M., Lin, P., Mohammed, A. R. S., Mohnen, P., Murray, J., Zhang, M., and Prettyman, A. (2023). The creation of life-m: The longitudinal, intergenerational, family electronic micro-database project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 56(3):138–159.

Balabdaoui, F., Doss, C. R., and Durot, C. (2021). Unlinked monotone regression. *The Journal of Machine Learning Research*, 22(1):7766–7825.

Becker, G. S. and Tomes, N. (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy*, 87(6):1153–1189.

Beresovsky, V., Gershunskaya, J., and Savitsky, T. D. (2024). Review of quasi-randomization approaches for estimation from non-probability samples. Working paper.

Betancourt, B., Zanella, G., and Steorts, R. C. (2022). Random partition models for microclustering tasks. *Journal of the American Statistical Association*, 117(539):1215–1227.

Buckles, K., Price, J., Ward, Z., and Wilbert, H. (2023). Family trees and falling apples: Historical intergenerational mobility estimates for women and men. Working paper.

Chandler, A. D. (1977). *The Visible Hand: The Managerial Revolution in American Business.* Belknap Press, Cambridge, MA.

Chen, X., Christensen, T., and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, 86:1965–2018.

Chernozhukov, V., Fernández-Val, I., Meier, J., van Vuuren, A., and Vella, F. (2024). Conditional rank-rank regression. Unpublished working paper.

Chetty, R., Hendren, N., Jones, M. R., and Porter, S. R. (2020). Race and economic opportunity in the united states: An intergenerational perspective. *Quarterly Journal of Economics*, 135(2):711–783.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014a). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623.

Chetty, R., Hendren, N., Kline, P., Saez, E., and Turner, N. (2014b). Is the united states still a land of opportunity? recent trends in intergenerational mobility. *American Economic Review: Papers and Proceedings*, 104(5):141–147.

Collins, W. J. and Wanamaker, M. H. (2022). African american intergenerational economic mobility since 1880. *American Economic Journal: Applied Economics*, 14(3):84–117.

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27(3):79–102.

D'Haultfœuille, X., Gaillac, C., and Maurel, A. (2023). Partially linear models under data combination. Working paper.

Durlauf, S. N., Kourtellos, A., and Tan, C. M. (2022). The great gatsby curve. *Annual Review of Economics*, 14:571–605. First published as a Review in Advance on May 13, 2022.

Feigenbaum, J. (2018). Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940. *The Economic Journal*, 128:F446–F481.

Ferrie, J. P. (2005). History lessons: The end of american exceptionalism? mobility in the united states since 1850. *Journal of Economic Perspectives*, 19(3):199–215.

Fogel, R. W. and Engerman, S. L. (1974). *Time on the Cross: The Economics of American Negro Slavery*, volume 1. W. W. Norton & Company, New York.

Gallman, R. E. and Rhode, P. W. (2020). *Capital in the Nineteenth Century.* University of Chicago Press, Chicago, Illinois.

Ghosh, A., Hwang, S. I. M., and Squires, M. (2024). Links and legibility: Making sense of historical u.s. census automated linking methods. *Journal of Business & Economic Statistics*, 42(2):579–590.

Glaeser, E. (2012). *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. Penguin Books, New York, NY.

Goldin, C. and Katz, L. F. (2008). *The Race Between Education and Technology*. Belknap Press, Cambridge, MA.

Gordon, R. J. (2016). *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War*. Princeton University Press, Princeton, NJ.

Gortmaker, J. (2024). Open source software policy in industry equilibrium. Working paper.

Han, Y. and Lahiri, P. (2019). Statistical Analysis with Linked Data. *International Statistical Review*, 87(S1):S139–S157.

Helgertz, J., Ruggles, S., Warren, J. R., Fitch, C. A., Hacker, J. D., Nelson, M. A., Price, J. P., Roberts, E., and Sobek, M. (2023). Ipums multigenerational longitudinal panel: Version 1.1. IPUMS.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61.

Hu, Y. (2023). The econometrics of unobservables: Latent variable and measurement error models and their applications in empirical industrial organization and labor economics. Unpublished manuscript.

Jácome, E., Kuziemko, I., and Naidu, S. (2021). Mobility for all: Representative intergenerational mobility estimates over the 20th century. Working Paper 29289, National Bureau of Economic Research.

Lee, C.-I. and Solon, G. (2009). Trends in intergenerational income mobility. *The Review of Economics and Statistics*, 91(4):766–772.

Lindert, P. H. and Williamson, J. G. (2016). *Unequal Gains: American Growth and Inequality since 1700*. Princeton University Press, Princeton, NJ.

Long, J. and Ferrie, J. (2013). Intergenerational occupational mobility in Great Britain and the United States since 1850. *The American Economic Review*, 103(4):1109–1137.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Narain, N. (2024). How patient is venture capital? Working paper.

Olivetti, C. and Paserman, M. D. (2015). In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850-1940. *American Economic Review*, 105(8):2695–2724.

Pananjady, A. and Samworth, R. J. (2022). Isotonic regression with unknown permutations: Statistics, computation and adaptation. *The Annals of Statistics*, 50(1):324–350.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press, Cambridge, MA. Translated by Arthur Goldhammer.

Price, J., Buckles, K., Van Leeuwen, J., and Riley, I. (2021). Combining family history and machine learning to link historical records: The census tree data set. *Explorations in Economic History*, 80:101391.

Pérez, S. (2019). Intergenerational occupational mobility across three continents. *The Journal of Economic History*, 79(2):383–416.

Ray, D. and Genicot, G. (2023). Measuring upward mobility. *American Economic Review*, 113(11):3044–3089.

Romer, C. (1986). Spurious volatility in historical unemployment data. *Journal of Political Economy*, 94(1):1–37.

Ruggles, S., Fitch, C. A., and Roberts, E. (2018). Historical Census Record Linkage. *Annual Review of Sociology*, 44(1):19–37.

Ruggles, S., Nelson, M. A., Sobek, M., Fitch, C. A., Goeken, R., Hacker, J. D., Roberts, E., and Warren, J. R. (2024). Ipums 1940 census full count data: Version 4.0. IPUMS.

Saavedra, M. and Twinam, T. (2020). A machine learning approach to improving occupational income scores. *Explorations in Economic History*, 75:101304.

Sabety, A. and Spitzer, A. K.-L. (2023). Missing black men? the impact of non-reporting on estimates of labor market outcomes for black men. Working Paper.

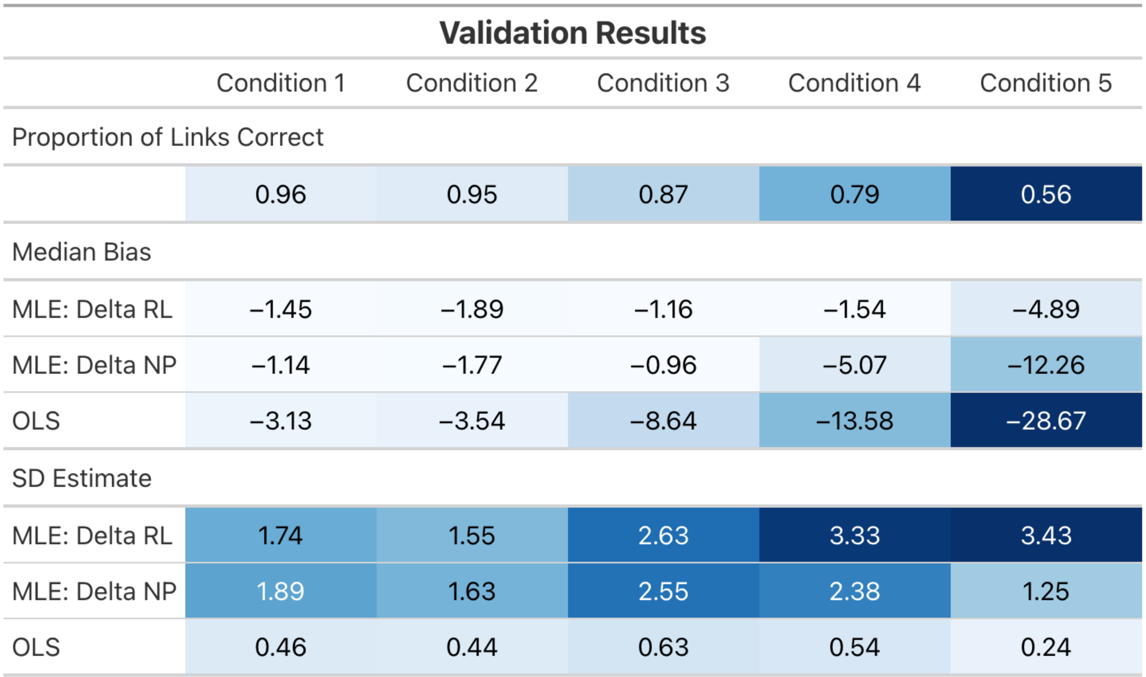
Santavirta, T. and Stuhler, J. (2024). Name-based estimators of intergenerational mobility. *The Economic Journal*, 134(663):2982–3016.

Song, X., Massey, C. G., Rolf, K. A., Ferrie, J. P., Rothbaum, J. L., and Xie, Y. (2019). Long-term decline in intergenerational mobility in the United States since the 1850s. *Proceedings of the National Academy of Sciences*, page 201905094.

Ward, Z. (2023). Intergenerational mobility in american history: Accounting for race and measurement error. *American Economic Review*, 113(12):3213–3248.

Wright, G. (2024). The antebellum US economy. In Diebolt, C. and Haupert, M., editors, *Handbook of Cliometrics*. Springer Nature Switzerland AG.

Tables and Figures



Validation Results

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
Proportion of Links Correct					
	0.96	0.95	0.87	0.79	0.56
Median Bias					
MLE: Delta RL	−1.45	−1.89	−1.16	−1.54	−4.89
MLE: Delta NP	−1.14	−1.77	−0.96	−5.07	−12.26
OLS	−3.13	−3.54	−8.64	−13.58	−28.67
SD Estimate					
MLE: Delta RL	1.74	1.55	2.63	3.33	3.43
MLE: Delta NP	1.89	1.63	2.55	2.38	1.25
OLS	0.46	0.44	0.63	0.54	0.24

Table 1: Validation results for the misclassification maximum likelihood estimator.

This table presents results from 100 draws of the validation exercise described in section 3. The columns of the table correspond to different linkage algorithms and rows report values of statistics for either the naive OLS estimator or the maximum likelihood estimator for the nonparametric misclassification model or the record linkage misclassification model Δ_1^{RL}. Condition 1 and Condition 2 are the conservative and standard versions of the thresholding rules used in Abramitzky et al. (2019) based on estimates of the probability of correct linkages in the Felligi-Sunter model; Condition 3 and Condition 4 correspond to the conservative and standard versions of the deterministic algorithm proposed in Abramitzky et al. (2014); and Condition 5 selects the candidate link with the highest estimated Felligi-Sunter probability without thresholding. The first row gives the average proportion of the linked sample for each condition that is correct. The next three rows report the median bias across draws for the OLS and MLE estimates. The last three rows report the standard deviation of the estimate for each estimator across validation draws.

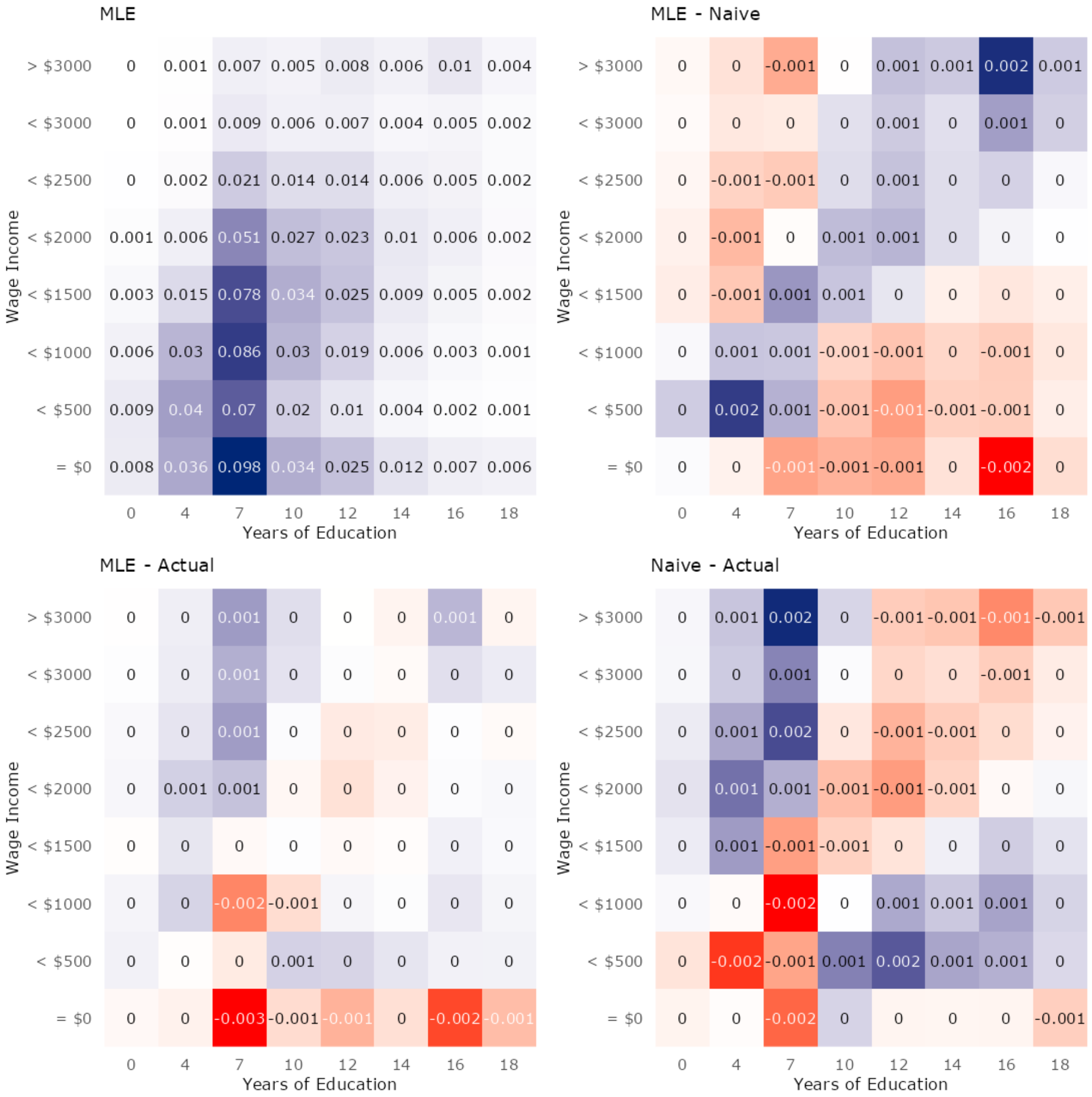


Figure 1: Bias in the distribution of education and income for naive and MLE estimates. This figure compares estimates of the distribution of wage income and years of educational attainment to the truth in one draw from the validation exercise, as described in Section 3. The "Naive" estimate is the sample mean in each cell of the joint distribution in the observed linked sample. "MLE" estimates are maximum likelihood estimates assuming an independent Record-Linkage structure for misclassification, specifically model Δ_2^{RL} defined in equation (6). The top left matrix plots maximum likelihood estimates of the joint distribution of education and wages. The top right matrix reports the difference between the maximum likelihood estimates and the means in the linked sample. The bottom matrices report the error in estimates for the joint distribution relative to the empirical distribution in the correctly linked subset of the data.

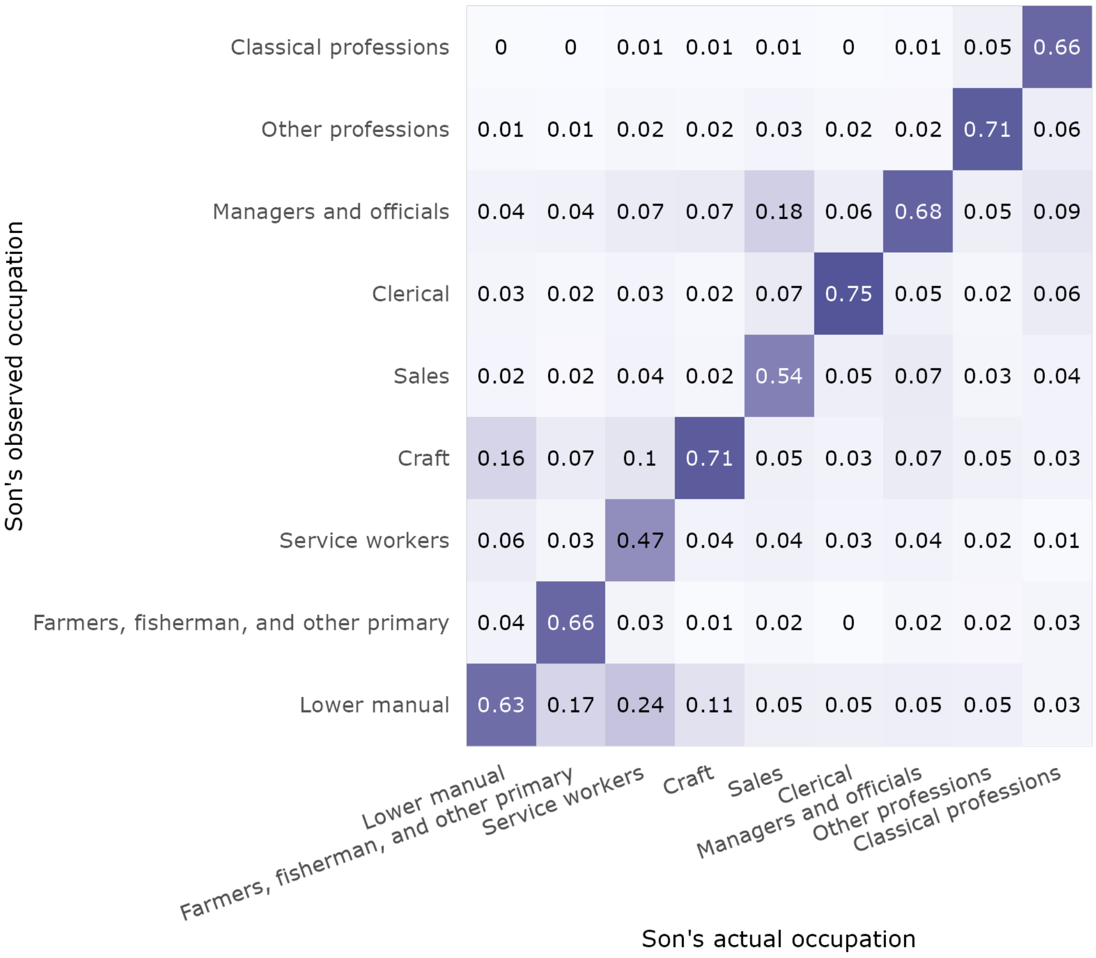


Figure 2: Nonparametric maximum likelihood estimates of the misclassification matrix, 1940. This figure reports maximum likelihood estimates of the misclassification matrix for son's occupations observed in the 1940 census based on a sample of linked census data 1910-1940 and 1910-1930. These estimates correspond to the nonparametric model of misclassification. Occupations are grouped into nine categories based on the "mesooccupation" classification in Song et al. (2019). The value and color of each entry correspond to the conditional probability in that cell.

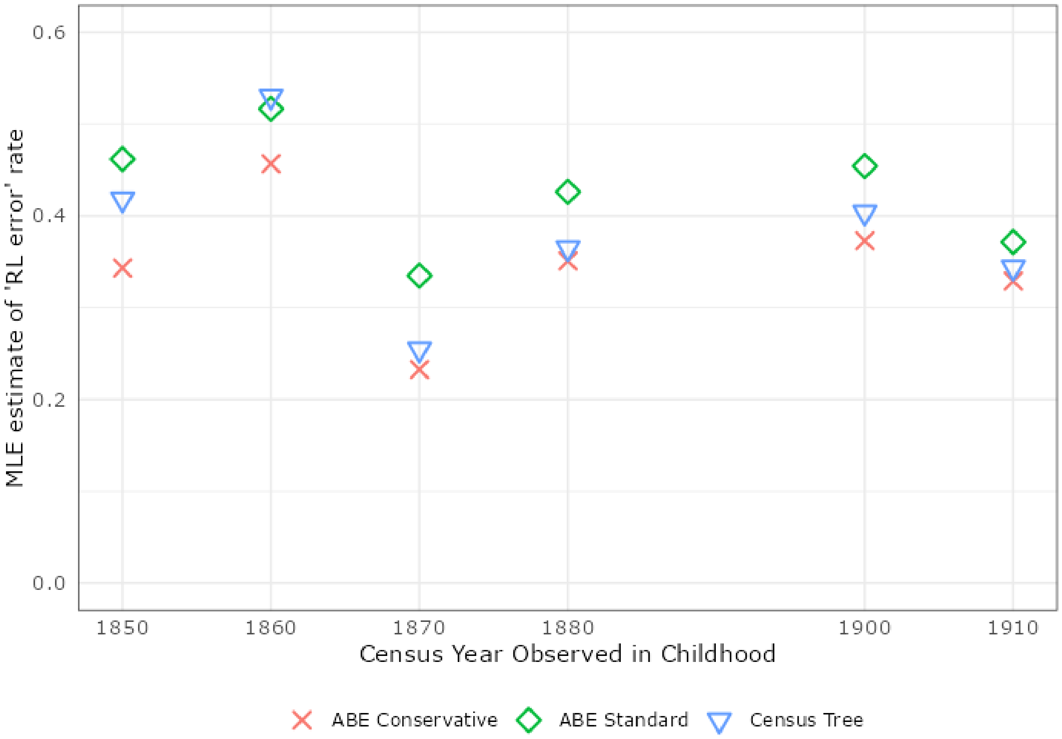


Figure 3: Estimates of record linkage error, 1850-1910.
This figure shows maximum likelihood estimates of "record linkage error" in linked census data between 1850 and 1910. This figure plots estimates of the record linkage error parameter, α, defined in equation (7). For each set of linked censuses, I estimate a mixed model of misclassification error in which the primary measurement is assumed to have a record linkage error structure, $\Delta^{(1)} = \Delta_3^{RL}$, and misclassification in the repeated measure is allowed to be flexible. Estimates are reported for three linked samples: ABE conservative and ABE standard are from the Census Linking Project (Abramitzky et al., 2020) and Census Tree is the full sample of links from the Census Tree (Price et al., 2021).

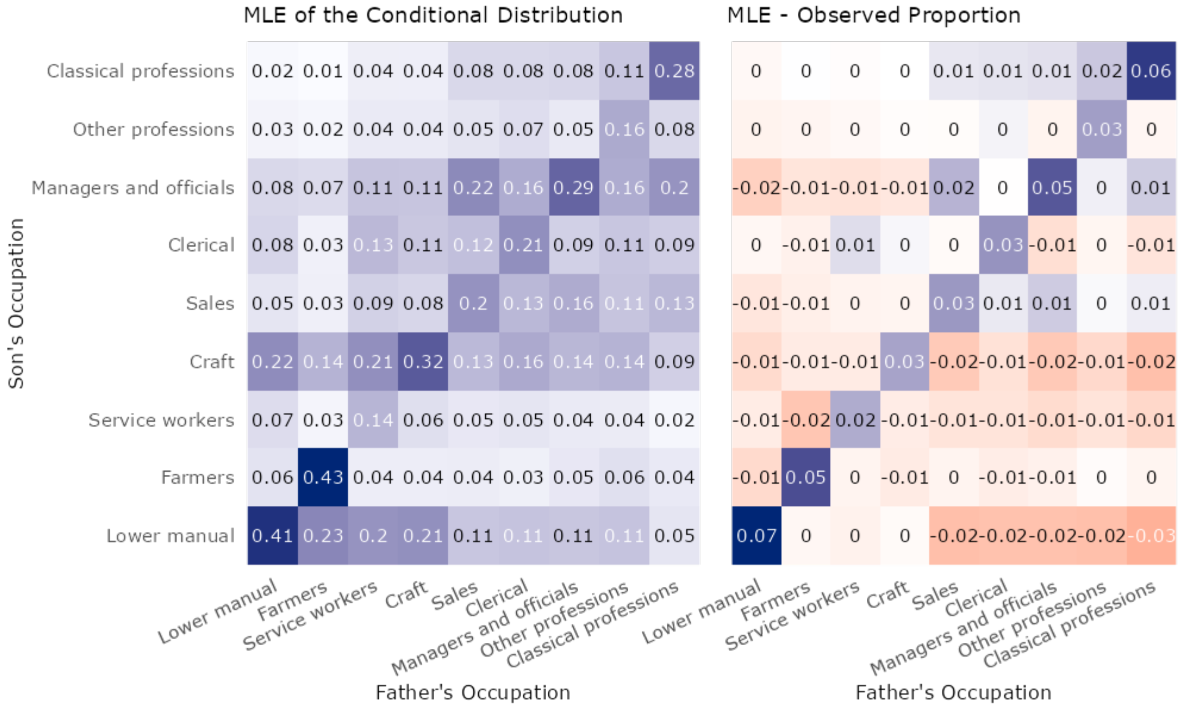


Figure 4: Estimated and observed conditional distribution of occupational status, 1910-1940. This figure presents maximum likelihood estimates of the distribution of sons' occupation conditional on their fathers' occupation in a model of misclassification and the difference between the MLE and the "naive" distribution in the linked sample. Estimates of the conditional distribution are from a nonparametric model of the joint distribution of father and son's occupations and a mixed model of misclassification: misclassification errors are assumed to have a record linkage structure in 1940, $\Delta^{(1)} = \Delta_3^{RL}$, and misclassification errors in 1930 are allowed to be flexible. The value and color of each entry correspond to the conditional probability or the difference in conditional probabilities in that cell. Linked data in this sample are from the standard variation of the ABE algorithm using exact names (Abramitzky et al., 2020). Occupations are grouped into nine categories based on the "mesooccupation" classification in Song et al. (2019).

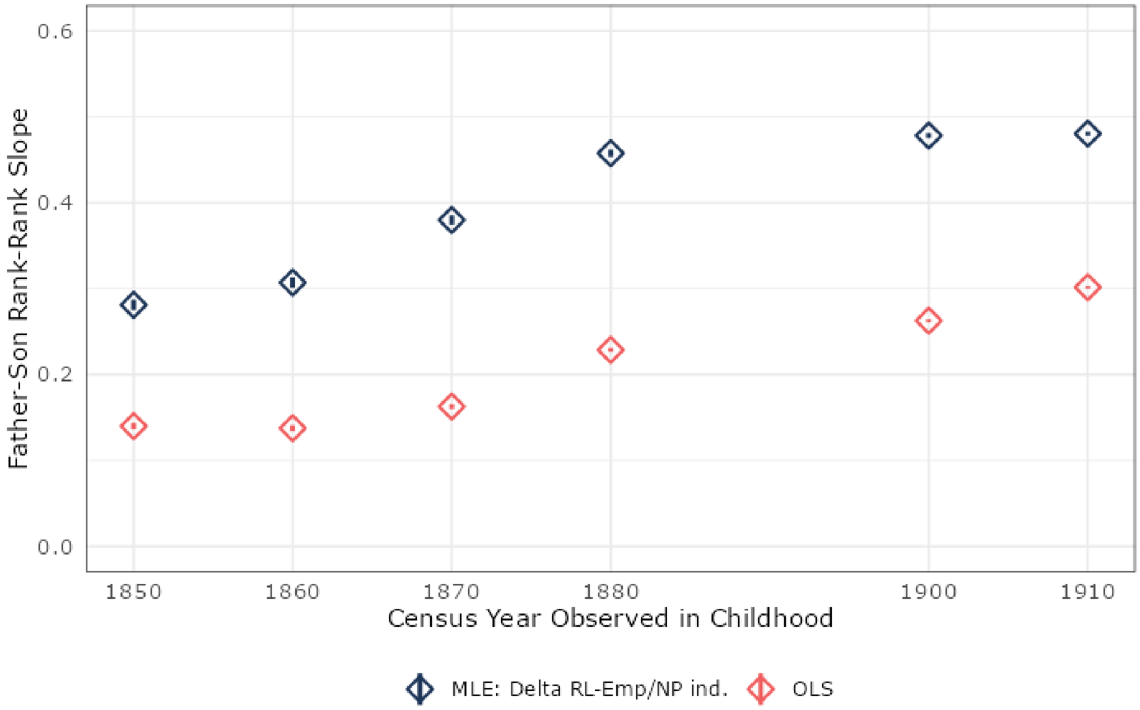


Figure 5: Intergenerational rank-rank slope of occupation status, 1850-1910.
This figure reports estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. OLS estimates are shown in orange and maximum likelihood estimates for a model of misclassification are shown in blue. Misclassification errors are assumed to have a record linkage structure in the year of the main outcome, $\Delta^{(1)} = \Delta_3^{RL}$, typically 30 years after observed in childhood, and misclassification is allowed to be fully flexible in the repeated measure. The vertical line through each estimate is a 95% confidence interval. Standard errors for the maximum likelihood estimates are computed via delta method. The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. This sample was generated using linked data from the Census Linking Project (Abramitzky et al., 2020). Occupations are grouped into nine categories based on the "mesooccupation" classification in Song et al. (2019).

Appendix A Illustration of Record Linkage Errors

A.1 Bias decomposition

Imperfections in record linkage may result in bias from *contamination* or *nonrepresentativeness* in the linked sample. To illustrate the role of each source of bias more concretely, consider the following simple bias decomposition. Suppose a researcher would like to regress an outcome y observed in data set B on a covariate x observed in data set A. Let λ^* be an indicator for the true linkage between data sets, and let λ be the linkage used by the researcher. That is, $\lambda^*_{ab} = 1$ means that the observations $a \in A$ and $b \in B$ correspond to the same individual, and $\lambda^*_{ab} = 0$ means that the observations do not. Similarly, $\lambda_{ab} = 1$ means that observations $a \in A$ and $b \in B$ are linked by some record linkage procedure while $\lambda_{ab} = 0$ means the observations are not linked. For simplicity, suppose that x and y are mean zero, x has unit variance and the true and observed linkages are independent of x and y. Then by the law of total probability, we can decompose the bias of the naive estimate:

$$\begin{aligned} E[\hat{\beta}] - \beta =& E[x_a y_b|\lambda_{ab} = 1] - E[x_a y_b|\lambda^*_{ab} = 1] \\ =& \underbrace{E[1 - \lambda^*_{ab}|\lambda_{ab} = 1]}_{\text{False Positive Rate}} \underbrace{(E[x_a y_b|\lambda^*_{ab} = 0, \lambda_{ab} = 1] - E[x_a y_b|\lambda^*_{ab} = 1, \lambda_{ab} = 1])}_{\text{Bias from Data Corruption}} + \\ & \underbrace{E[1 - \lambda_{ab}|\lambda^*_{ab} = 1]}_{\text{False Negative Rate}} \underbrace{(E[x_a y_b|\lambda^*_{ab} = 1, \lambda_{ab} = 1] - E[x_a y_b|\lambda^*_{ab} = 1, \lambda_{ab} = 0])}_{\text{Bias from Nonrepresentativeness}} \end{aligned} \tag{10}$$

The decomposition shows that the strength and direction of bias from imperfect record linkage is determined by four terms. The scale of bias from nonrepresentativeness depends on the false negative rate, and the direction depends on the difference in the covariance among actually linked individuals between observations linked or missed in λ. Intuitively, individuals with less common names, those who are illiterate or innumerate, and those who change names between censuses are likely nonrepresentative of the full population. Bias from data corruption depends on the covariance between x_a and y_b when observations a and b do not correspond to the same individual. Often, it's reasonable to assume that this term will be close to zero, or at least much closer to zero than the covariance among correctly linked observations. In the context of intergenerational mobility, this covariance depends on the strength of the association between linking information (e.g. names, age, and birthplace) and occupation status. Intuitively, the scale of this bias depends linearly on the fraction of false positives in the linkage λ. Consequently, OLS estimates of β in an otherwise representative linked sample with a false positive rate of α will be attenuated to zero, $E[\hat{\beta}] = (1 - \alpha)\beta$.

A.2 Illustration in the 1940 census

To illustrate errors due to record linkage in a more naturalistic setting, I generate a synthetic linked data set with zero correct links in the exercise below. Similar to the validation exercise described in section 3, I link synthetically corrupted copies to each-other using standard linking algorithms. In this exercise, though, I shift the birth years of all individuals in one copy by a decade. Consequently, links generated by this procedure are guaranteed not to correspond to the same individual. At the same time, the outcomes Y_1 and Y_2, may be substantially correlated if the information used to link is itself predictive of the latent outcome Y^*.

Figure A.1 compares the empirical distribution of years of education in a sample of entirely incorrect links to the predicted joint distribution of Y_1 and Y_2 from a record linkage misclassification model, Δ_3^{RL}, assuming that the false positive linkage rate is one. The matrix on the left shows that the relationship between the observed years of education is clearly not independent, despite the fact that the data contain no actual "links." However, the matrix on the right shows that conditioning the misclassification model on the same linking controls as the validation exercise largely matches the observed pattern of misclassification.

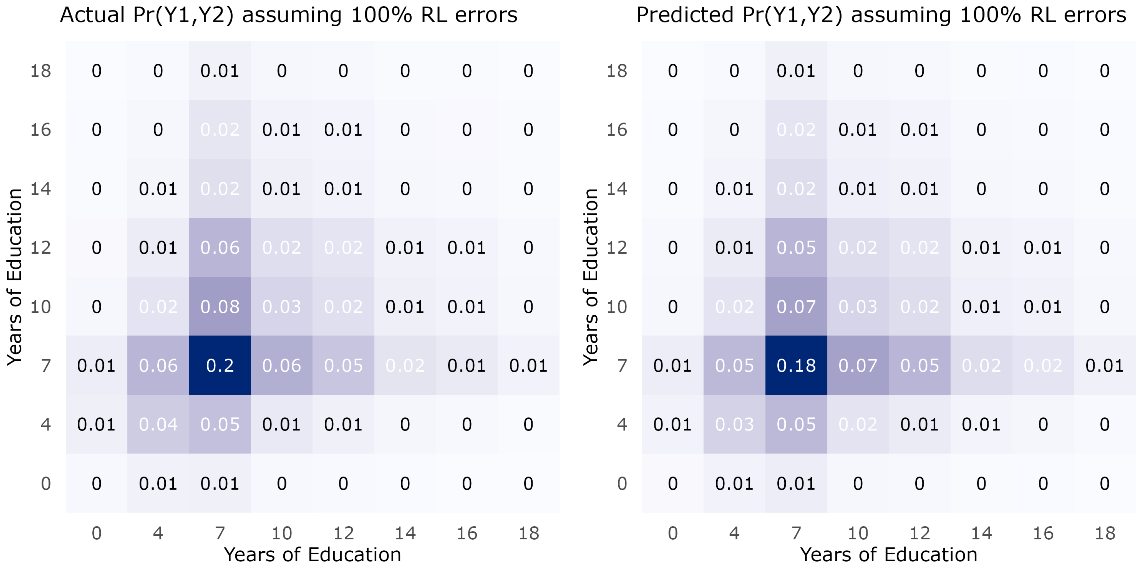


Figure A.1: Record linkage error illustration.

This figure compares the distribution of misclassification error in a sample with 100% record linkage errors and the predicted distribution in a model of record linkage misclassification. The matrix on the left shows the joint distribution of years of education in an exercise linking copies of the 1940 census in which successful linkage was impossible. The matrix on the right shows the predicted distribution of years of education on the same sample from a record linkage model of misclassification, Δ_3^{RL}, assuming that all links in the sample are incorrect, $\alpha = 1$.

Appendix B Measuring Occupation Status

Economic historians have proposed a long list of measures for inferring occupation status. To keep the estimand relatively consistent, each measure of occupation status I consider will be a rank. Specifically, for each outcome Z,

1. I first compute a predicted value (such as a sample average) for that outcome $\widehat{Z}_{k,c}$ within a demographic cell c for occupation k. The set of cells may be a singleton (i.e. the prediction is constant for the full population with occupation j) or it may capture all the unique combinations across multiple demographic factors, such as race, region, and cohort.

2. I then define the sample P_t of all men born in 10 year cohort t, observed at age 25-64 with observed occupations (i.e. the OCC1950 code is not equal to 979, 983, 984, 986, or 999). Let N_t be the sample size for each group P_t.

3. Last, I compute the sample percentile rank for each individual $i \in P_t$ with occupation $k(i)$ and demographic cell $c(i)$:

$$Y_{k(i),c(i),t} := \sum_{j \in P_t} \mathbf{1}\left(\widehat{Z}_{k(j),c(j)} \leq \widehat{Z}_{k(i),c(i)}\right) / N_t$$

I report estimates of mobility defined using various ways of ranking occupations. The definition of the population, outcome, and demographic cell are listed in table B.1. Note that even if the measure of the outcome is constant within a cohort, the percentile rank for that occupation may vary across time as the occupational distribution shifts. To illustrate: for the occupation rank **Education 1940**, Z is years of education in the 1940 census, demographic cells include all unique combinations of race, region, and birthplace. I then use empirical Bayes shrinkage assuming a normal prior and likelihood to predict the average years of education within demographic cells. Last, I construct the percentile rank for all individuals in each cohort between 1790 and 1950 using complete count censuses between 1850 and 1940, and various large samples between 1950 and 2017, following Ward (2023).

Variable Name	**Population**	**Outcome**	**Demographic Covariates**	**Prediction Type**
Inc. 1950 All	Men age 25-64, complete 1950 census	Total Income	Full Population	Sample Mean
Edu. 1940 All	Men age 25-64, complete 1940 and 1950 censuses	Years of Education	Full Population	Sample Mean
Inc. 1950	Men age 25-64, complete 1950 census	Total Income	Race, Region, Birthplace	Empirical Bayes
Edu. 1940	Men age 25-64, complete 1940 and 1950 censuses	Years of Education	Race, Region, Birthplace	Empirical Bayes
Song	Men age 25-64, complete censuses 1850-1940 and census / ACS samples 1950-2017	Literacy before 1880, Years of Education thereafter	Cohort	Sample Mean
Ward	Men age 25-64, complete censuses 1850-1940 and census / ACS samples 1950-2017	Literacy before 1880, Years of Education thereafter	Race, Region, Cohort	Sample Mean

Table B.1: Outcomes used to define occupation rankings.

This table collects definitions for the measures used to rank occupations and occupation × demographic group cells. Empirical Bayes prediction assumes a normal-normal likelihood and prior for outcome means across covariate cells. The variables **Ward** and **Song** are the average educational attainment and literacy within occupation before applying the rank transform as defined in Ward (2023).

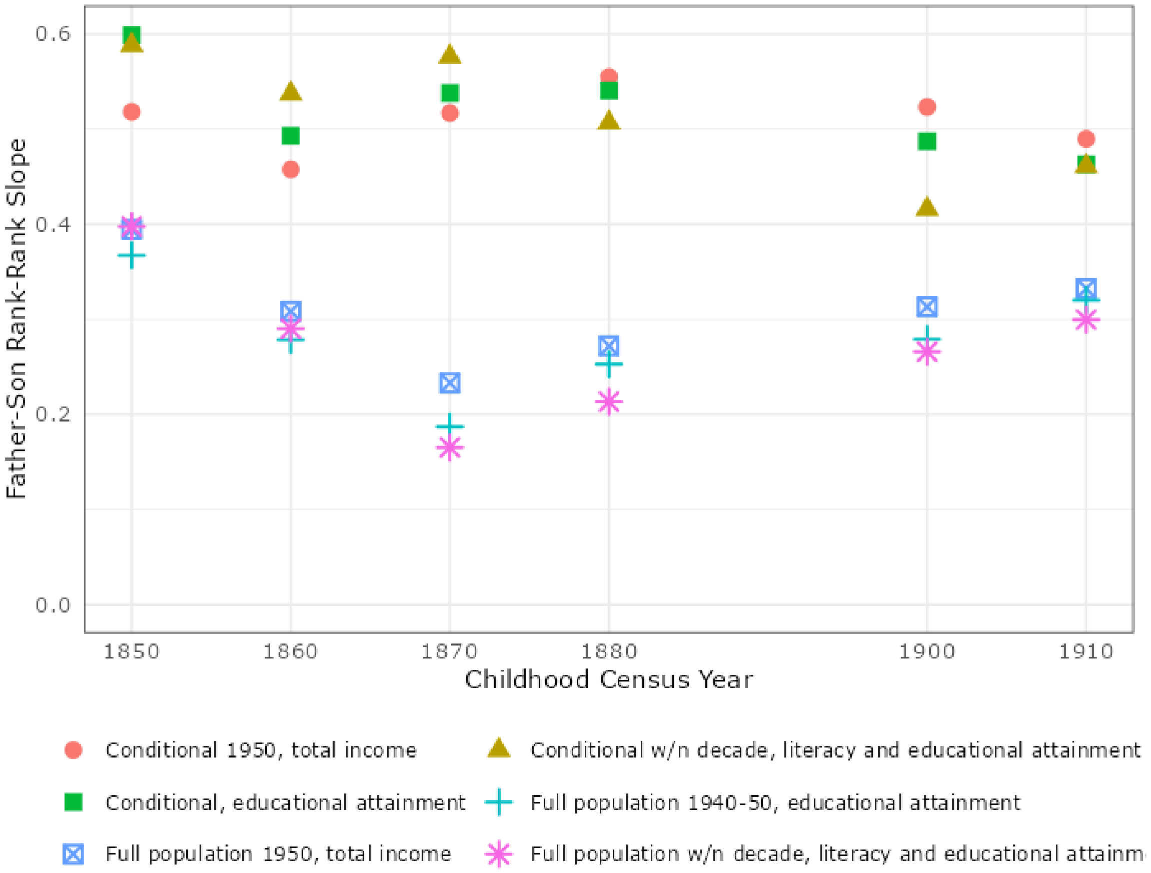


Figure B.1: OLS estimates of the rank-rank association by occupation ranking.

This figure reports OLS estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. The color and shape of each point corresponds to the definition of occupation status used to rank occupations, as described in Table B.1. The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. This sample was generated using linked data from the Census Linking Project (Abramitzky et al., 2020).

Appendix C Point identification of the nonparametric model

In this section, I connect the assumptions in section 2.1 to the identification result in Hu (2008), which I will refer to as Hu Theorem. First, I restate the Hu Theorem using the original notation. I then connect the assumptions and notation in this paper to that of Hu (2008).

C.1 Hu Theorem

The set up in Hu (2008) considers three variables: a dependent variable y, a latent true discrete variable x^* subject to misclassification error, and a vector of accurately measured independent variables w. The conditional density of y on x^* and w is denoted $f_{y|x^*w}(y|x^*, w)$. Researchers observe an i.i.d. sample of the variables $\{x, y, w, z\}$, where x is a misclassified measure of x^* and z is an instrumental variable satisfying two assumptions:

Assumption H 1 $f_{y|x^*xzw}(y|x^*, x, z, w) = f_{y|x^*w}(y|x^*, w)$.

Assumption H 2 $f_{x|x^*zw}(x|x^*, z, w) = f_{x|x^*w}(x|x^*, w)$.

Suppose that x, x^*, and z^* share the same support $\{1, ..., k\}$. The first additional assumption requires that the instrument z is non-trivially related to the regressor x^*.

Assumption H 3 *The matrix corresponding to the conditional distribution of the regressor on the instrument has full rank:*

$$\textbf{rank} \begin{pmatrix} f_{x^*|zw}(1|1, w) & \cdots & f_{x^*|zw}(k|1, w) \\ \vdots & \ddots & \vdots \\ f_{x^*|zw}(1|k, w) & \cdots & f_{x^*|zw}(k|k, w) \end{pmatrix} = k$$

Similarly, we need to assume that measurement error in x is not irrecoverably severe:

Assumption H 4 *The matrix of misclassification error, the conditional distribution of x given x^*, is invertible:*

$$\textbf{det} \begin{pmatrix} f_{x|x^*w}(1|1, w) & \cdots & f_{x|x^*w}(k|1, w) \\ \vdots & \ddots & \vdots \\ f_{x|x^*w}(1|k, w) & \cdots & f_{x|x^*w}(k|k, w) \end{pmatrix} \neq 0$$

Because the proof of Hu theorem relies on a spectral argument, we need an assumption that guarantees the uniqueness of eigenvalues for a particular matrix.

Assumption H 5 *There exists a function $\omega(\cdot)$ such that, for all $i \neq j$,*

$$E[\omega(y)|x^* = i, w] \neq E[\omega(y)|x^* = j, w]$$

Finally, Hu (2008) lists four additional assumptions, any of which in conjunction with assumptions H1-5, are sufficient for point identification of the nonparametric model. The first two additional assumption consider notions of monotonicity in the relationship between y and x^*.

Assumption H 6 *Given y and w, the conditional density $f_{y,x^*w}(y|x^*, w)$ is strictly increasing or decreasing in x^*.*

Assumption H 7 *There exists a function $\omega(\cdot)$ such that $E[\omega(y)|x^*, w]$ is strictly increasing in x^*.*

Alternatively, we can recover point identification by imposing additional assumptions on the distribution of misclassification errors.

Assumption H 8 *$Pr(x = 1|x^*, w)$ is strictly decreasing in x^* for $x^* \in \{1, ..., k\}$.*

Assumption H 9 *$Pr(x = i|x^* = i, w) > Pr(x = j|x^* = i, w)$ for $j \neq i$.*

We can now state the main result in Hu (2008).

Hu Theorem *If Assumptions H1-5 hold, and at least one of Assumption H6, H7, H8, or H9 hold, then the densities $f_{y|x^*w}$, $f_{x|x^*w}$, and $f_{x^*|zw}$ are point identified.*

C.2 Application to the nonparametric model of record linkage

Notation The set up in this paper differs slightly from that in Hu (2008). One immediately apparent difference is the source of measurement error: in Hu (2008), the regressor x^* is misclassified, whereas the misclassified variable is the outcome Y^* in this paper. While this difference appears substantial—as it is in the case of classical measurement error—the role of the instrument, the regressor, and the outcome are mostly interchangeable in the nonparametric model of misclassification (Hu, 2023). The mapping of variables between the two models is: the outcome y is the regressor X; the latent regressor x^* is the latent outcome Y^*; the noisy regressor x is the noisy outcome Y_1; the instrument z is the second outcome Y_2; and the controls w are labeled similarly W. These differences are primarily differences in labels alone. The model in this paper is simplified relative to the setting in (Hu, 2008) in one way: the regressor X is discrete, whereas the outcome y in Hu (2008) may be continuous. Consequently, all of the target distribution functions in this paper are finite dimensional while distribution functions involving y in Hu (2008) may be infinite dimensional.

Assumptions The first two Assumption H1 and H2 are together identical to 1. Assumption 2 implies Assumption 5. To see this, note that in the simpler setting with discrete X,

$$E[\omega(X)|Y^* = y_i, W] = \sum_{j \leq J} Pr(X = x_j|Y^* = y_i, W)\omega(x_j)$$

Stacking this expression across values of Y^*, the condition becomes: there exists a vector $\boldsymbol{\omega}$ such that $\Pi_{\cdot,i}\boldsymbol{\omega} \not\propto \Pi_{\cdot,j}\boldsymbol{\omega}$ for all $i \neq j$. It is sufficient, then, for no pair of rows in Π to be proportional.

Assumption 3 immediately implies Assumption H9, since $\delta_{ii} > 1/2 > \sum_{j \neq i} \delta_{ij} \geq \delta_{ik}$ for each $k \neq i$. Assumption 3 also implies Assumptions H3 and H4 by Gershgorin circle theorem, which implies that all strongly diagonal dominant matrices with non-zero diagonal elements are nonsingular, since zero lies outside all of the Gershgorin discs.

Thus, if Assumptions 1, 2, and 3 are satisfied, Hu Theorem applies, and the nonparametric model is point identified.

Appendix D Details on the misclassification estimator

I implement the maximum likelihood estimator for models of misclassification described in this paper through a new R package. The R package `misclassifyr` allows for more general models than the nonparametric set up in section 2. First, I allow the dimension of the regressor X to differ from the outcome, and I denote it K. Additionally, I consider a weaker form of independence, in which the noisy measures Y_1 and Y_2 are conditionally independent of the regressor but may be mutually dependent:

Assumption 4 *Measures are conditionally independent of the regressor:*

$$X \perp Y_1, Y_2 \quad | \quad Y^*, W$$

Under assumption 4, the conditional distribution of X, Y_1, Y_2 given W is:

$$Pr(X, Y_1, Y_2|W) = \sum_{y^* \leq J} Pr(Y_1, Y_2|Y^* = y^*, W) Pr(Y^* = y^*, X|W)$$

The relaxed form of independence requires slightly different notation for the misclassification matrices. Let he probability of observing $Y_1 = r$ and $Y_2 = s$ when the latent variable takes value $Y^* = q$ be denoted $\delta^{(q)}_{r,s} := Pr(Y_1 = r, Y_2 = s|Y^* = q)$. The probability of any misclassification is summarized in the matrix $\Delta = (\Delta_1, ..., \Delta_J)$ where

$$\Delta_s := \begin{pmatrix} \delta^{(1}_{1,s} & \cdots & \delta^{(1)}_{J,s} \\ \vdots & \ddots & \vdots \\ \delta^{(J)}_{1,s} & \cdots & \delta^{(J)}_{J,s} \end{pmatrix}$$

The submatrices Δ_s are indexed by the values of Y_2 for simplicity of the likelihood expression. I also consider a relaxed form of diagonal dominance:

Assumption 5 *Misclassification is diagonal dominant: $\delta^i_{i,k} > \delta^i_{j,k}$ and $\delta^i_{k,i} > \delta^i_{k,j}$ for all k and $j \neq i$.*

By Assumption 4, the probability of a particular realization of the triple $(X, Y_1, Y2) = (q, r, s)$ is:

$$Pr(X = r, Y_1 = r, Y_2 = s) = \sum_{t \leq J} Pr(X = r, Y^* = t) Pr(Y_1 = r, Y_2 = s|Y^* = t) = \pi_{t,q} \delta^{(t)}_{r,s}$$

The expression for the full log likelihood is now:

$$\ell(\mathbf{n}|N, \theta) - \log c(\mathbf{n}) = \sum_{q \leq K; r,s \leq J} n_{q,r,s} \log \sum_{t \leq J} \delta_{r,s}^{(t)} \pi_{t,q} = \mathbf{n} \cdot \log(\mathrm{vec}(\Pi^\top \Delta)) \tag{11}$$

where vec(M) flattens the matrix M column-wise. This expression naturally suggests estimating θ by choosing $\hat{\theta}$ to maximize $\mathbf{n} \cdot \log(\mathrm{vec}(\Pi^\top \Delta))$.

The relaxed, minimal assumptions allow researchers to explore a wider class of models which may not be point identified. Consequently, the package also builds-in tools for inference when point identification fails.

D.1 Inference

The discrete nature of the problem suggests estimation of Π and Δ via maximum likelihood. To improve the numerical stability of the optimizer, a transform the parameters of Π and Δ via the logistic link function.

I then use a plug-in estimator for β. I use the delta method to estimate of the variance of the misclassification estimator. To estimate the asymptotic covariance of the model parameters θ, I use the inverse of the Fisher Information: $\widehat{\Sigma}_\theta := I(\widehat{\theta})^{-1}$ where $I(\widehat{\theta}) := -\mathbf{H}(\widehat{\theta})$ and $\mathbf{H}(\theta)$ is the Hessian $\mathbf{H}(\widehat{\theta})_{ij} := \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j}\Big|_{\theta = \widehat{\theta}}$. I compute the covariances of Π and β via two applications of the delta method:

$$\widehat{\Sigma}_\Pi := \mathbf{J_g}(\widehat{\theta}) \widehat{\Sigma}_\theta \mathbf{J_g}(\widehat{\theta})^\top$$

where $\mathbf{J_g}$ is the Jacobian matrix of a map from the model parameters θ to the joint distribution Π, and

$$\widehat{\Sigma}_\beta := \nabla f(\widehat{\Pi}) \widehat{\Sigma}_\Pi \nabla f(\widehat{\Pi})^\top$$

where f is a functional that maps the joint distribution of Y^* and X to the regression coefficient β defined when Y^* and X have assigned scalar values.

Analytical confidence intervals through the delta method reflect sampling uncertainty in the tabulation $\mathbf{n}$. These confidence intervals may not have accurate size if (i) there is dependence in the true data generating process that is not reflected in the iid sampling process described in Section 2; (ii) if θ is near the boundary; or (iii) researchers consider models that do not satisfy the assumptions for point identification.

The true data generating process for linked data certainly violates independence across observations—for example, it's very common to require linked samples to have at most one link per observation in each data set. It's less clear whether dependence in this form is likely to lead to substantial undercoverage of confidence intervals based on iid assumptions.

To address the latter two concerns, the `misclassifyr` package includes the option to report highest posterior density sets (or projected highest posterior density sets) which have asymptotically exact coverage (or conservative coverage for subvector inference), as shown in Chen et al. (2018). Details on the implementation of the Gibbs sampler are in the package documentation.

Appendix E Additional figures and tables

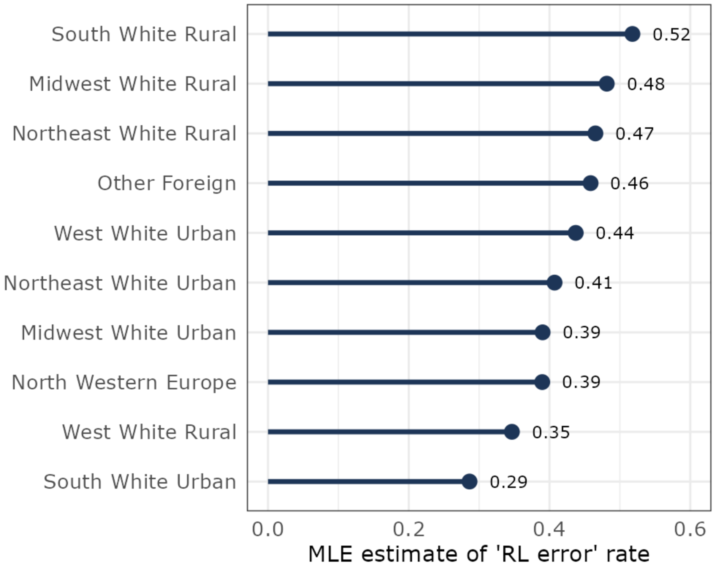


Figure E.1: Record linkage error estimates across control cells, 1940.

This figure shows the maximum likelihood estimates of the rate of record linkage error, α, in a mixed model of misclassification applied to a sample linking observations in 1910-1940 and 1910-1930, in which the misclassification in 1940 has a record linkage error structure, Δ_3^{RL}, and misclassification in 1930 is fully flexible.

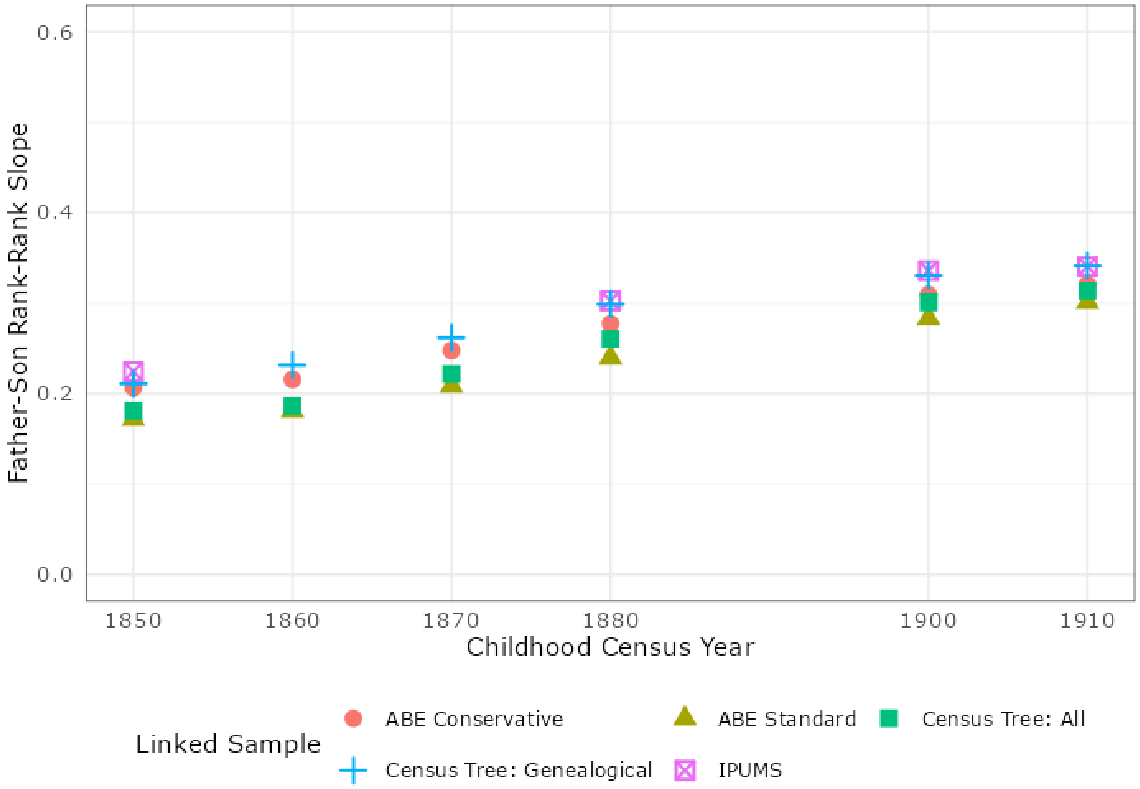


Figure E.2: OLS estimates of the rank-rank association for various linked samples. This figure reports OLS estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. The color and shape of each point corresponds to the source of the linked data, which includes links from the Census Linking Project (Abramitzky et al., 2020), the Census Tree (Price et al., 2021), and the Multigenerational Longitudinal Panel (Helgertz et al., 2023). The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. Occupation ranks are based on the average educational attainment within occupation in the complete count 1940 and 1950 censuses.

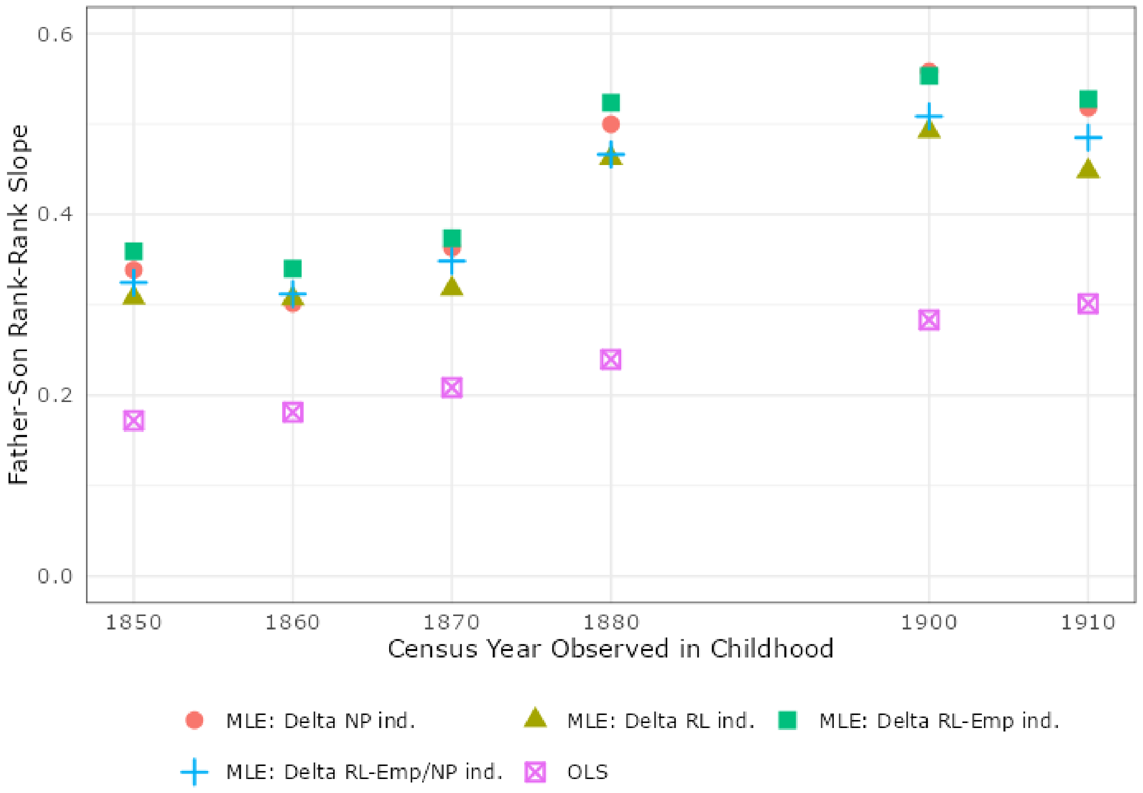


Figure E.3: Misclassification estimates of the rank-rank association by misclassification model.

This figure reports estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. The color and shape of points corresponds to the estimator used: OLS are shown in the pink crossed squares, and maximum likelihood estimates for various models of misclassification are shown. Orange circles correspond to fully nonparametric models of misclassification; olive triangles correspond to the record linkage model of misclassification Δ_1^{RL}; solid green squares correspond to Δ_2^{RL}; and the blue plus signs correspond to a mixed model of misclassification in which $\Delta^{(1)} = \Delta_3^{RL}$ and $\Delta^{(2)}$ is allowed to be fully flexible. The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. These samples were generated using the standard version of the ABE linked data using exact name matches from the Census Linking Project (Abramitzky et al., 2020). Occupations are grouped into nine categories based on the “mesooccupation” classification in Song et al. (2019).

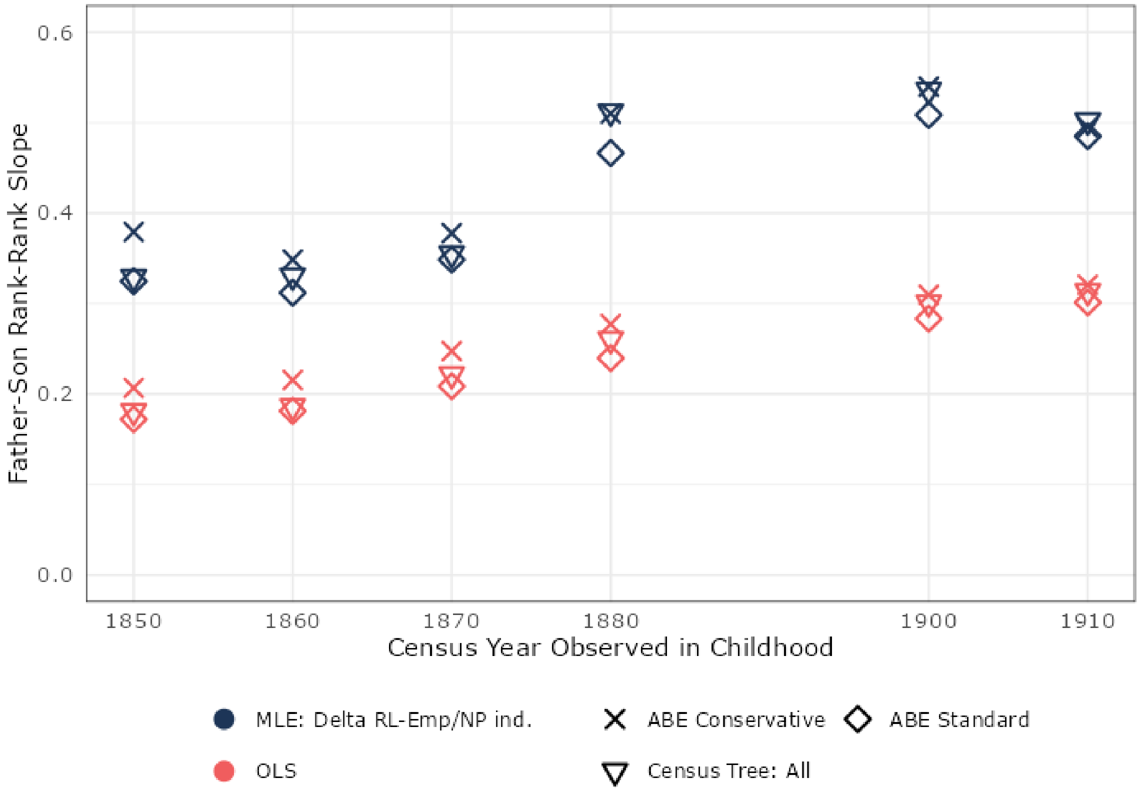


Figure E.4: Misclassification estimates of the rank-rank association by linked sample.

This figure reports estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. OLS estimates are shown in orange and maximum likelihood estimates for a model of misclassification are shown in blue. The shape of the points correspond to the linked sample used, which includes links from the Census Linking Project (Abramitzky et al., 2020) and the Census Tree (Price et al., 2021) Misclassification errors are assumed to have a record linkage structure in the year of the main outcome, $\Delta^{(1)} = \Delta_3^{RL}$, typically 30 years after observed in childhood, and misclassification is allowed to be fully flexible in the repeated measure. The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. Occupations are grouped into nine categories based on the "mesooccupation" classification in Song et al. (2019). Occupation rankings are defined by the average educational attainment within occupation in the complete count 1940 and 1950 censuses.

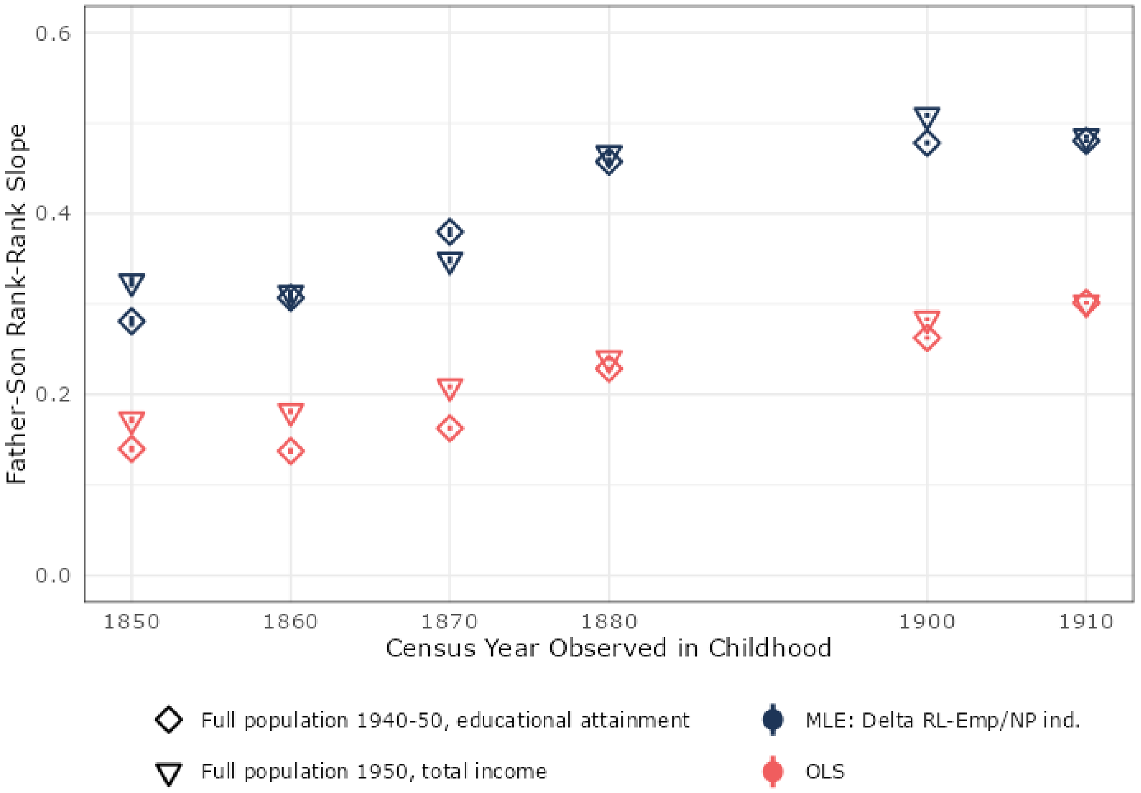
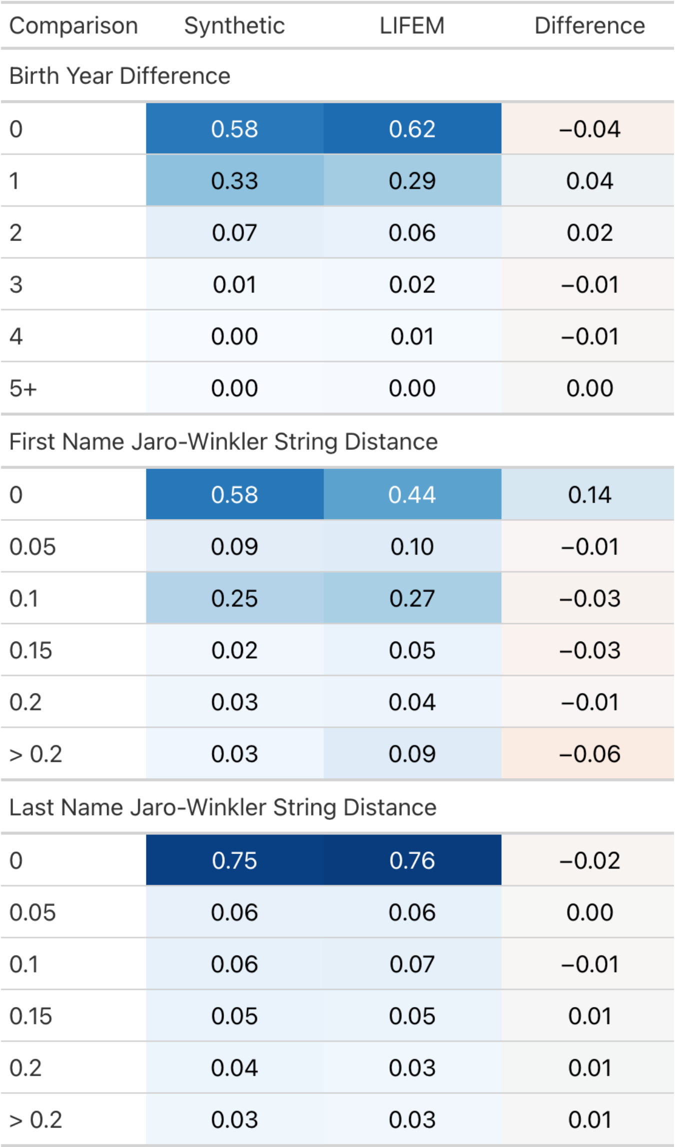


Figure E.5: Misclassification estimates of the rank-rank association by occupation ranking. This figure reports estimates of the intergenerational association of occupation rank among cohorts born between 1833 and 1910. OLS estimates are shown in orange and maximum likelihood estimates for a model of misclassification are shown in blue. The shape of the points correspond to the definition of the outcome used to define occupation ranks, as discussed in Appendix B. Misclassification errors are assumed to have a record linkage structure in the year of the main outcome, $\Delta^{(1)} = \Delta_3^{RL}$, typically 30 years after observed in childhood, and misclassification is allowed to be fully flexible in the repeated measure. The year corresponds to the census year in which father-son pairs are defined. Samples include sons age 0-17 in the base census year. To be included in the sample, sons must be observed over the age of twenty in two census years. These samples were generated using the standard version of the ABE linked data using exact name matches from the Census Linking Project (Abramitzky et al., 2020). Occupations are grouped into nine categories based on the "mesooccupation" classification in Song et al. (2019).



Comparison	Synthetic	LIFEM	Difference
Birth Year Difference			
0	0.58	0.62	−0.04
1	0.33	0.29	0.04
2	0.07	0.06	0.02
3	0.01	0.02	−0.01
4	0.00	0.01	−0.01
5+	0.00	0.00	0.00
First Name Jaro-Winkler String Distance			
0	0.58	0.44	0.14
0.05	0.09	0.10	−0.01
0.1	0.25	0.27	−0.03
0.15	0.02	0.05	−0.03
0.2	0.03	0.04	−0.01
> 0.2	0.03	0.09	−0.06
Last Name Jaro-Winkler String Distance			
0	0.75	0.76	−0.02
0.05	0.06	0.06	0.00
0.1	0.06	0.07	−0.01
0.15	0.05	0.05	0.01
0.2	0.04	0.03	0.01
> 0.2	0.03	0.03	0.01

Table E.1: Comparison distances in the LIFE-M sample and the validation exercise.

This table presents the distribution of comparison distances in synthetic ground truth linked samples and the LIFE-M sample. The first panel of the table reports the distribution of the absolute value of the difference in birth year for individuals in each sample. The second and third panels report the distribution of Jaro-Winkler string distances in the first and last names, respectively. The LIFE-M sample is linked between the 1910 and 1940 censuses. The synthetic ground truth is from one draw of the validation exercise. Census data are from the restricted complete count samples (Ruggles et al., 2024).