# A Machine Learning Approach to Improving Occupational Income Scores*

Martin Saavedra
Department of Economics
Oberlin College

Tate Twinam
Interdisciplinary Arts and Sciences
University of Washington Bothell

November 4, 2018

**Abstract**

Historical studies of labor markets frequently suffer from a lack of data on individual income. The occupational income score (OCCSCORE) is often used as an alternative measure of labor market outcomes. We consider the consequences of using OCCSCORE when researchers are interested in earnings regressions. Using modern Census data, we find that the use of OCCSCORE biases results towards zero and can result in statistically significant coefficients of the wrong sign. We use a machine learning approach to construct a new adjusted score based on industry, occupation, and demographics. Our alternative score reduces bias and errors of sign in both modern and historical contexts. We illustrate our approach by estimating earnings gaps in the 1915 Iowa State Census and intergenerational mobility elasticities using linked data from the 1850-1930 Censuses.

**JEL codes**: C21, J71, N32
**Keywords**: OCCSCORE, occupational income score, LIDO score, machine learning, lasso, non-classical measurement error, occupation, earnings gaps

*We are grateful for feedback from Ran Abramitzky, Brian Beach, Hoyt Bleakley, Melissa Dell, James Feigenbaum, Ekaterina Jardim, Bruce Sacerdote, James Siodla, Randall Walsh, and Eugene White, as well as seminar participants at the University of Michigan, Harvard University, the College of William and Mary, Marquette University, and Clemson University. We are also grateful for comments from participants at the following conferences: WEAI, the Northeast Ohio Economics Workshop, SEA, the Pacific Northwest Labor Day conference at UW Seattle, and the North American Summer Meeting of the Econometric Society. Grant Goehring provided excellent research assistance. All errors are our own.

# 1 Introduction

Before 1940, data on individual wages and education are not available in the U.S. Census. Consequently, occupation is often the only measure of labor market outcomes available to economic historians. Occupation is a categorical variable; however, many economists use occupational indexes as continuous measures of historical labor market outcomes. One popular example is the 1950 occupational income score (OCCSCORE), which is the median income of an occupation in 1950. Originally developed by Sobek (1995) as a continuous measure of occupational earnings potential, Sobek acknowledged that "Although the income score is derived from individual-level data, it should not be interpreted as actual income." Since then, occupational income scores have been used to examine labor market outcomes going as far back as 1850, and studies using this approach have been published in numerous top journals in economics and other fields.[1]

Although occupational income scores are a reasonable proxy for occupational status, it is unclear how much bias this measurement error induces if researchers are primarily interested in earnings rather than occupation. Additionally, it is unclear if 1950 occupational income scores are good measures of income when examining Censuses several decades before 1950. While this potential bias has been acknowledged in the literature, only a handful of attempts have been made to quantify and diagnose its impact on inferences.[2] In this study, we attempt to measure this bias directly and examine how much it can be mitigated through adjustments to occupational income scores based on demographic and geographic variables available in all U.S. Censuses dating back to 1850.

We first develop a formal model of the measurement error problem posed by occupational income scores. The model allows us to determine when attenuation bias will occur and to explicitly quantify its magnitude. We then take this model to the data to estimate the

[1] See section 2 for examples.

[2] Abramitzky, Boustan and Eriksson (2014) point out that using OCCSCORE allows one to measure native-immigrant earnings convergence between occupations but not within occupation. Using 1970 and 1980 Census data, they estimate that using OCCSCORE captures at least 30% of total earnings convergence between these two groups.

OCCSCORE-induced bias. Because it is difficult to make historical data better, we analyze the performance of occupational income scores by making modern data worse.[3] We generate 2000-based occupational income scores and examine how well they predict income in the decades between 1950 and 2000. We then use this index to examine race, gender, and geographic earnings disparities from 1950 through 2000 and compare these to the true gaps estimated using actual earnings data. Finally, we use cross-validated lasso regressions to construct new data-driven adjusted OCCSCOREs (based on industry, occupation, race, sex, age, and state). We compare estimated earnings gaps based on our lasso-adjusted industry, demographic, and occupation (LIDO) scores to those generated using OCCSCORE and true earnings.

We find that although OCCSCORE is correlated with income even for Censuses five decades removed from the base year, earnings gaps are attenuated when using OCCSCORE as a proxy for income. Assuming the researcher is interested in an earnings regression, the use of OCCSCORE can result in statistically significant coefficients of the wrong sign up to 20 percent of the time in our modern data, particularly for variables indicating state of residency (often used in difference-in-differences analysis exploiting state-level variation in treatments).[4] This is even the case in earnings regressions where the sample is restricted to white males only. We find that adjusting OCCSCORE by race, sex, age, industry, and geography–adjustments that few papers in the literature have made–reduces this bias.

To examine the performance of our LIDO scores in a historical context, we exploit a rare source of pre-1950 earnings data: the 1915 Iowa State Census.[5] Estimated race and gender earnings gaps in 1915 Iowa using true earnings data are sizable and negative; howe-

[3]Our approach is in the spirit of Romer (1986), who shows that excess volatility in unemployment time series during the pre-war era is an artifact of the interpolation methods used before the Current Population Survey. Applying the same interpolation methods to unemployment data during the post-war period results in similar levels of volatility.

[4]This finding is similar to that in Bertrand, Duflo and Mullainathan (2004), who find that difference-in-differences models that do not account for serial correlation in the error terms can result in statistically significant estimates of placebo treatment effects 40% of the time.

[5]The 1915 Iowa State Census data was digitized by Claudia Goldin and Lawrence Katz (Goldin and Katz, 2010).

ver, when using standard OCCSCORE as a proxy for earnings, the racial earnings gap is underestimated by almost half and the gender earnings gap is statistically significant and of the wrong sign. Our adjusted OCCSCORE yields race and gender earnings gaps close to the true values. Finally, we conduct an analysis of OCCSCORE-induced bias in measures of intergenerational income transmission. This analysis is based on father-son pairs linked from the 1880 decennial Census to the 1850, 1860, 1900, 1910, 1920, and 1930 decennial Censuses. In this setting, we find that standard OCCSCOREs and our alternative scores perform similarly for white males, because measurement errors for fathers and sons are likely to be correlated. However, transmission coefficients are attenuated for black men. We conclude with recommendations for future research in economic history.

## 2 Previous Literature

The occupational income score (OCCSCORE) was developed by Matthew Sobek and IPUMS for the purpose of representing "the material rewards accruing to persons in different occupations" (Sobek, 1995). It provides a continuous alternative to coarse occupational groupings and is more comparable to earnings regressions that are common in modern labor economics. To understand how researchers use this variable, we searched for papers containing either "OCCSCORE" or "Occupational Income Score" in top general interest journals and top field journals in labor economics and economic history. This search yielded the 25 papers listed in Table 1.[6]

Most of the articles have been published within the last decade, with a median publication year of 2014. Sixteen use the log of occupational income score as a dependent variable, and consequently, we focus our empirical analysis on the log of occupational income score. Of these 25 papers, only four adjust occupational income scores by any demographic variables.[7]

[6]This search included articles in the following journals: *American Economic Journal: Applied Economics*; *American Economic Journal: Economic Policy*; *American Economic Review*; *Explorations in Economic History*; *Journal of Economic History*; *Journal of Human Resources*; *Journal of Labor Economics*; *Quarterly Journal of Economics*; *Review of Economics and Statistics*, and *The Review of Economic Studies*.

[7]The occupational earnings measure used by Collins and Wanamaker (2014) varies by race and region;

Table 1: Published Studies using Occupational Income Scores

| Article | Description | Adjusted | Log |
|---|---|---|---|
| Collins (2000) | Occupational mobility of blacks during the 1940s. | Both | Yes |
| Minns (2000) | SES growth of immigrants relative to natives. | No | Yes |
| Angrist (2002) | Effect of sex ratios on marriage markets and labor force participation. | Yes | Yes |
| Chin (2005) | Effects of incarceration in internment camps on labor-market outcomes. | No | No |
| Sacerdote (2005) | Intergenerational effects of slavery. | No | No |
| Bleakley (2007) | Hookworm eradication during school-age on human capital. | No | Yes |
| Bleakley and Lange (2009) | Quantity-quality childbearing, hookworm, and the returns for schooling. | No | Yes |
| Bleakley (2010) | Childhood exposure to malaria and adult SES | No | Yes |
| Abramitzky et al. (2012) | Returns to migration and self-selection. | No | Yes |
| Lee (2013) | Repeal of Sunday closing laws and years of schooling. | No | Yes |
| Aaronson et al. (2014) | The Rosenwald School initiative and quality-quantity childbearing. | No | Yes |
| Abramitzky et al. (2014) | Occupational advancement among European immigrants during the Age of Mass Migration | No | No |
| Collins and Wanamaker (2014) | The returns to migration and self-selection for blacks during the Great Migration. | Yes | Yes |
| Cook et al. (2014) | Distinctively black names and socioeconomic status. | No | No |
| Stephens and Yang (2014) | Sensitivity of prior estimates for the returns to schooling to region-specific birth year effects. | No | No |
| Collins and Wanamaker (2015) | Self-selection of inter-regional and intra-regional migration. | No | Yes |
| Lleras-Muney and Shertzer (2015) | The effect of English-only statutes on immigrant children literacy, years of schooling, and occupations. | No | Yes |
| Olivetti and Paserman (2015) | Creates pseudo-links to estimate father-son and father-daughter elasticities for the intergenerational transmission of SES. | No | Yes |
| Saavedra (2015) | The effect of school-age incarceration in internment camps on adult outcomes. | No | No |
| Cook et al. (2016) | 19th century blacks with distinctively black names live longer. | No | No |
| Massey (2016) | U.S. immigrant quota affected the selection of immigrants. | No | Yes |
| Bleakley and Ferrie (2016) | The effects of a Georgia land lottery on human capital investment. Uses OCCSCORE to measure returns to literacy. | No | No |
| Lee and Lin (2017) | How natural amenities affect neighborhood income ranks. | No | No |
| Saavedra (2017) | Early-life exposure to yellow fever affected occupational status. | No | No |
| Ward (2017) | Self-selection of return migrants. | No | Yes |
| Carruthers and Wanamaker (2017) | Effect of differential school quality to the black-white income gap. | Yes | Yes |

Typically, these papers analyze historical Census data for which income or wage data are not available. Some of the papers, however, present occupational income score along with wage/income data as an alternative measure of socioeconomic status (see Stephens and Yang (2014) or Chin (2005)). Some papers attempt to reduce bias by limiting the sample to a particular demographic group, typically white males (see Bleakley (2010)). These papers often examine intergenerational mobility, racial and ethnic SES gaps, migrant selection, and the effects of schooling or health interventions.

This list underestimates how many researchers use occupational income scores or similar measures. Other papers in these journals may have used average or median income/wages by occupation as a dependent variable but do not refer to the variable as an occupational income score. Some authors, recognizing the issues we address here, have constructed occupational income scores adjusted for relevant covariates. For example, Bailey and Collins (2006) construct average occupational wages across sex-race-industry-region cells in 1940 to analyze the wage gains of black women between 1910 and 1940. Occupational income scores are also used in other fields, especially sociology. A Google Scholar search for research articles containing "OCCSCORE" or "occupational income score" currently yields 291 articles. Many of these articles are working papers that will eventually be published in top economics journals. For example, four NBER working papers since 2017 contain the phrase "OCCSCORE" or "Occupational Income Score."

# 3 Theory

In this section, we construct a stylized model of labor market outcomes which allows for sorting across occupations and heterogeneous earnings within occupation. We use the model to illustrate the bias induced when occupational earnings are used as a proxy for true earnings, and we show how this bias can be mitigated through the use of an adjusted occupa-

Angrist (2002) varies by age and sex; and Collins (2000) presents results using both an unadjusted and a race-adjusted OCCSCORE.

tional income score. Finally, we discuss a data-driven approach to constructing an adjusted OCCSCORE that is broadly applicable to pre-1950 Censuses.

## 3.1 Measurement Error

Consider a linear model with classical measurement error (CME) in the dependent variable. The researcher is interested in $y_i = \alpha + \beta X_i + \epsilon_i$, but instead of observing $y_i$, the researcher observes $\tilde{y}_i$, which equals $y_i$ plus a measurement error term $e_i$. In the CME model, $y_i$ and $X_i$ are uncorrelated with $e_i$, implying that $e_i$ is by definition correlated with the observed value $\tilde{y}_i$. Thus, a regression of the mismeasured $\tilde{y}_i$ on $X_i$ is equivalent to a regression of $y_i$ on $X_i$ with an error term of $\epsilon_i - e_i$. Since $e_i$ is uncorrelated with the true $y_i$ and $X_i$, regressing the observed value on $X_i$ is equivalent to adding variance to the error term of the regression. For this reason, CME in the dependent variable affects the precision of the regression estimates, but does not lead to bias.

Unfortunately, the CME model is a poor description of occupational income scores. Suppose the true $y_i$ is income. Without knowledge of $y_i$, the researcher replaces $y_i$ with their best guess of income given occupation. Perhaps this measure is mean or median earnings for a given occupation. Since the reported $y_i$ is the researcher's best guess of income, the measurement error must be uncorrelated with the reported value, and by definition correlated with the true value. Thus, the opposite of the CME model holds.

Occupational income scores are better described using the Optimal Prediction Error (OPE) model of Hyslop and Imbens (2001). In the OPE model, if researchers use their best guess of $y_i$ (income) given a noisy signal (occupation), then estimates of $\beta$ are often biased towards zero. Hyslop and Imbens (2001) refer to this as OPE(1). However, if researchers instead use their best guess of $y_i$ given both the noisy signal and relevant predictors $X_i$ (referred to as an OPE(2) model), then estimates of $\beta$ are unbiased. Consequently, instead of using occupational income scores, researchers should develop occupational income scores that are conditional on $X_i$.

In the subsection below, we develop a modified OPE model specifically for occupational income scores. In the model, researchers observe occupation, a vector of relevant covariates $X_i$, and, from a separate data source, mean earnings by occupation. Changes in $X_i$ can affect earnings through two channels: by shifting individuals from lower to higher paying occupations, by increasing earnings within a given occupation, or both. We then find conditions equivalent to the OPE(1) and OPE(2) from Hyslop and Imbens (2001). Lastly, it is not always possible to predict income conditional on both occupation and $X_i$, particularly when $X_i$ is not a standard Census variable. We show the conditions under which adjusting by predictors are correlated with $X_i$ will result in estimates that are less biased than the unadjusted occupational income score. We refer to this model as an OPE(3) model.

## 3.2 An Optimal Prediction Error Model with Occupational Income Scores

A researcher is interested in estimating $y_i = \alpha + \beta X_i + \epsilon_i$, where $y_i$ is the income of individual $i$, $X_i$ is a relevant predictor of interest, and $\epsilon_i \perp\!\!\!\perp X_i$. The researcher does not observe $y_i$, but observes both $X_i$ and occupation $j$. From a separate source, the researcher also observes the occupational income score of occupation $j$, which is $\mathrm{E}(y \mid \mathrm{occ} = j)$.[8] $X_i$ could increase income through two channels, either shifting the marginal worker into a higher paying occupation or increasing the earnings of a worker within a given occupation.

We start by modeling the first process. We assume that individuals are paid their marginal product, and that there exists a minimum marginal product an individual must produce to enter an occupation. For example, a surgeon must perform surgeries at some minimally acceptable level to legally practice. Suppose there is a continuum of occupations, and let $O_j$ be the minimum marginal product of occupation $j$, which can be thought of as the starting or entry-level salary. $X_i$ then shifts individuals into occupations with either higher or lower

[8]The IPUMS OCCSCORE uses median earnings by occupation instead of mean earnings. We use mean earnings to simplify the model.

starting salaries through the following data generating process:

$$O_{ij} = \delta_0 + \delta_1 X_i + \eta_i. \tag{1}$$

$O_{ij}$ denotes that individual $i$ is employed in occupation $j$ and earns $O_j$; since by construction $O_{ij} = O_j$ for all workers, we omit the $i$ subscript in the rest of the discussion. The parameter $\delta_1$ captures the extent to which $X_i$ shifts workers across occupations. The nuisance term $\eta_i$ captures that some workers enter occupations with higher or lower entry-level earnings than $X_i$ would predict, perhaps because of unobserved preferences, ability, or luck.

We then model earnings above the baseline pay, which we refer to as excess earnings, as a separate process. Let the excess earnings of individual $i$ be given by

$$y_i - O_j = \gamma_0 + \gamma_1 X_i + \nu_i. \tag{2}$$

$\gamma_1$ reflects the extent to which $X_i$ affects within-occupation earnings. For each worker $i$ in occupation $j$, total earnings $y_i$ equals the baseline pay $O_j$ plus excess earnings. Thus, $\delta_0 + \gamma_0 = \alpha$, $\delta_1 + \gamma_1 = \beta$, and $\eta_i + \nu_i = \epsilon_i$. We assume that $\eta_i, \nu_i \perp\!\!\!\perp X_i$. The assumption that $X_i$ is independent of the error terms $\eta_i$ and $\nu_i$ is a weak assumption given that we already assume $X_i$ is independent of $\epsilon_i$.

We consider two variations of this model. In the first, the error term influencing occupational sorting ($\eta_i$) is independent of that affecting excess earnings within an occupation ($\nu_i$). This simplifies the analysis and allows for some sharper conclusions. However, this assumption contradicts much established thought among labor economists. In a standard Roy (1951)-type model, individuals choose an occupation to maximize their expected earnings. If idiosyncratic individual factors affect occupational sorting, it seems likely they would influence excess earnings within occupation, and vice versa. If this is the case, it is unlikely that $\eta_i$ and $\nu_i$ are independent. We consider both the general case of dependent errors and the special case of independence.

### 3.2.1 OPE(1): $DV = \mathrm{E}(y \mid \mathrm{occ} = j)$

Consider the situation where $y_i$ is not observable, but the researcher observes the occupational income score of occupation $j$, which is $\mathrm{E}(y \mid \mathrm{occ} = j)$. In this case,

$$\operatorname{plim} \hat{\beta} = \frac{\operatorname{Cov}(\mathrm{E}(y_i \mid O_j), X_i)}{\operatorname{Var}(X_i)} = \delta_1 + \gamma_1 \frac{\operatorname{Cov}(\mathrm{E}(X_i \mid O_j), X_i)}{\operatorname{Var}(X_i)} + \frac{\operatorname{Cov}(\mathrm{E}(\nu_i \mid O_j), X_i)}{\operatorname{Var}(X_i)}. \quad (3)$$

Since $\frac{O_j - \delta_0}{\delta_1} = X_i + \frac{\eta_i}{\delta_1}$, we find that

$$\mathrm{E}(X_i \mid O_j) = \phi_0 \mu_X + \phi_1 \frac{O_j - \delta_0}{\delta_1} \quad (4)$$

where

$$\phi_0 = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \delta_1^2 \sigma_X^2}, \quad (5)$$

$$\phi_1 = \frac{\delta_1^2 \sigma_X^2}{\sigma_\eta^2 + \delta_1^2 \sigma_X^2}. \quad (6)$$

Similarly, we find that

$$\mathrm{E}(\nu_i \mid O_j) = \xi_0 + \xi_1 O_j \quad (7)$$

where

$$\xi_0 = -\xi_1 (\delta_0 + \delta_1 \mu_X), \quad (8)$$

$$\xi_1 = \frac{\sigma_{\eta\nu}}{\sigma_\eta^2 + \delta_1^2 \sigma_X^2}, \quad (9)$$

and $\sigma_{\eta\nu} = \operatorname{Cov}(\eta_i, \nu_i)$.

Some algebra will show that

$$\operatorname{plim} \hat{\beta} = \delta_1 + \gamma_1 \phi_1 + \delta_1 \xi_1. \quad (10)$$

If $\sigma_{\eta\nu} = 0$, this reduces to

$$\text{plim}\,\hat{\beta} = \delta_1 + \gamma_1\phi_1. \tag{11}$$

Since $\phi_1 \in (0,1)$, $\hat{\beta}$ is biased towards zero if $\text{sgn}(\delta_1) = \text{sgn}(\gamma_1)$.[9] If $\sigma_{\eta\nu} \neq 0$, then the direction of bias is unclear, as it depends on the sign of $\sigma_{\eta\nu}$, $\gamma_1$, and $\delta_1$, as well as the relative magnitudes of all of the terms in equation (10).[10]

### 3.2.2 OPE(2): $DV = \text{E}(y_i \mid \text{occ} = j, X_i)$

Suppose the researcher can observe, from another source, $\text{E}(y_i \mid \text{occ} = j, X_i)$, and uses this as the dependent variable. This is analogous to using a demographically-adjusted occupational income score (such as the LIDO score, which we develop below) when the predictor of interest is one of the demographic variables used to construct the adjustment (e.g., a gender or race indicator).

Then, we find that

$$\begin{aligned}
\hat{\beta} &= \frac{\text{Cov}\left(\text{E}\left(y_i \mid O_j, X_i\right), X_i\right)}{\text{Var}\left(X_i\right)} && (12)\\
&= \frac{\text{Cov}\left(\alpha + \beta X_i + \text{E}\left(\epsilon_i \mid O_j, X_i\right), X_i\right)}{\text{Var}\left(X_i\right)} && (13)\\
&= \beta\frac{\text{Cov}\left(X_i, X_i\right)}{\text{Var}\left(X_i\right)} = \beta. && (14)
\end{aligned}$$

Thus, $\hat{\beta}$ is unbiased.[11]

[9] This sign restriction implies that if $X_i$ increases income across occupations, it also increases income within occupations. This assumption will almost certainly hold for demographic groups that have historically earned less in labor markets (such as women and African-Americans). This assumption likely holds for human capital interventions that increase ability. It is conceivable that there are cases for which $\gamma_1$ and $\delta_1$ have different signs. For example, suppose an intervention increased the probability that a college graduate continues to law school, but had no other labor market consequences. This intervention would likely increase income across occupations ($\delta_1 > 0$), but not within occupation ($\gamma_1 = 0$), since almost all lawyers have law degrees.

[10] As we document empirically below, attenuation bias appears to be the most commonly observed outcome.

[11] This holds for arbitrary dependence between $\eta_i$ and $\nu_i$.

### 3.2.3 OPE(3) $DV = \mathrm{E}\left(y_i \mid \mathrm{occ} = j, Z_i\right)$

Suppose now that the researcher is interested in $y_i = \alpha + \beta X_i + \epsilon_i$, but cannot observe $\mathrm{E}\left(y_i \mid O_j, X_i\right)$. However, the variable of interest $X_i$ is correlated with another variable $Z_i$, and the researcher does observe $\mathrm{E}\left(y_i \mid \mathrm{occ} = j, Z_i\right)$. Further, assume that $Z_i$ is correlated with $X_i$ such that $Z_i = \lambda_0 + \lambda_1 X_i + \psi_i$, where $\psi_i$ is independent of $X_i$, $\eta_i$, and $\nu_i$. For example, suppose $X_i$ measures early-life malaria exposure during the late nineteenth or early twentieth century. Because the LIDO score, or any other index of occupational earnings, does not take into account early-life malaria exposure, the researcher does not observe $\mathrm{E}\left(y_i \mid \mathrm{occ} = j, X_i\right)$. However, the LIDO score does take into account geographic variables, such as state of residency, that will be correlated with early-life malaria exposure.

A regression of $\mathrm{E}\left(y_i \mid O_j, Z_i\right)$ on $X_i$ gives us estimates

$$\operatorname{plim} \hat{\beta} = \frac{\operatorname{Cov}\left(\mathrm{E}\left(y_i \mid O_j, Z_i\right), X_i\right)}{\operatorname{Var}\left(X_i\right)} = \delta_1 + \gamma_1 \frac{\operatorname{Cov}\left(\mathrm{E}\left(X_i \mid O_j, Z_i\right), X_i\right)}{\operatorname{Var}\left(X_i\right)} + \frac{\operatorname{Cov}\left(\mathrm{E}\left(\nu_i \mid O_j, Z_i\right), X_i\right)}{\operatorname{Var}\left(X_i\right)}. \tag{15}$$

Given $O_j$ and $Z_i$, we now have two noisy measures of $X_i$ since $\frac{O_j - \delta_0}{\delta_1} = X_i + \frac{\eta_i}{\delta_1}$ and $\frac{Z_i - \lambda_0}{\lambda_1} = X_i + \frac{\psi_i}{\lambda_1}$. Therefore, we find that

$$\mathrm{E}\left(X_i \mid O_j, Z_i\right) = \theta_0 \mu_X + \theta_1 \frac{O_j - \delta_0}{\delta_1} + \theta_2 \frac{Z_i - \lambda_0}{\lambda_1} \tag{16}$$

where

$$\theta_0 = \frac{\sigma_\eta^2 \sigma_\psi^2}{\sigma_\eta^2 \sigma_\psi^2 + \delta_1^2 \sigma_X^2 \sigma_\psi^2 + \lambda_1^2 \sigma_X^2 \sigma_\eta^2}, \tag{17}$$

$$\theta_1 = \frac{\delta_1^2 \sigma_X^2 \sigma_\psi^2}{\sigma_\eta^2 \sigma_\psi^2 + \delta_1^2 \sigma_X^2 \sigma_\psi^2 + \lambda_1^2 \sigma_X^2 \sigma_\eta^2}, \tag{18}$$

$$\theta_2 = \frac{\lambda_1^2 \sigma_X^2 \sigma_\eta^2}{\sigma_\eta^2 \sigma_\psi^2 + \delta_1^2 \sigma_X^2 \sigma_\psi^2 + \lambda_1^2 \sigma_X^2 \sigma_\eta^2}. \tag{19}$$

Turning to the final term,

$$\mathrm{E}\left(\nu_i \mid O_j, Z_i\right) = \zeta_0 + \zeta_1 O_j + \zeta_2 Z_i, \tag{20}$$

and the coefficients can be obtained using the standard Frisch-Waugh-Lovell approach:

$$\begin{aligned}
\zeta_0 &= -\zeta_1\left(\delta_0 + \delta_1\mu_X\right) - \zeta_2\left(\lambda_0 + \lambda_1\mu_X\right), && (21)\\
\zeta_1 &= \frac{\sigma_{\eta\nu}\left(\lambda_1^2\sigma_X^2 + \sigma_\psi^2\right)}{\sigma_\eta^2\sigma_\psi^2 + \delta_1^2\sigma_X^2\sigma_\psi^2 + \lambda_1^2\sigma_X^2\sigma_\eta^2}, && (22)\\
\zeta_2 &= \frac{-\sigma_{\eta\nu}\delta_1\lambda_1\sigma_X^2}{\sigma_\eta^2\sigma_\psi^2 + \delta_1^2\sigma_X^2\sigma_\psi^2 + \lambda_1^2\sigma_X^2\sigma_\eta^2}. && (23)
\end{aligned}$$

Some algebra will then reveal that

$$\operatorname{plim}\hat{\beta} = \delta_1 + \gamma_1\left(\theta_1 + \theta_2\right) + \delta_1\zeta_1 + \lambda_1\zeta_2. \tag{24}$$

Comparing (24) to (10) highlights the advantages of using the adjusted occupational income score. Since $\theta_1 + \theta_2 > \phi_1$ when $\lambda_1 \neq 0$, the bias on $\gamma_1$ is lower in this case. Noting that

$$\delta_1\xi_1 = \delta_1\frac{\sigma_{\eta\nu}}{\sigma_\eta^2 + \delta_1^2\sigma_X^2} = \delta_1\frac{\sigma_{\eta\nu}\sigma_\psi^2}{\sigma_\eta^2\sigma_\psi^2 + \delta_1^2\sigma_X^2\sigma_\psi^2} \tag{25}$$

and

$$\delta_1\zeta_1 + \lambda_1\zeta_2 = \delta_1\frac{\sigma_{\eta\nu}\sigma_\psi^2}{\sigma_\eta^2\sigma_\psi^2 + \delta_1^2\sigma_X^2\sigma_\psi^2 + \lambda_1^2\sigma_X^2\sigma_\eta^2}, \tag{26}$$

it is clear that the extra bias present when $\sigma_{\eta\nu} \neq 0$ is also mitigated in this case. If $\lambda_1 = 0$, i.e., if $Z_i$ and $X_i$ are independent, then conditioning on $Z_i$ provides no information about $y_i$, so $\theta_2 = 0$, $\theta_1 = \phi_1$, and $\delta_1\zeta_1 + \lambda_1\zeta_2 = \delta_1\xi_1$. In this case, (24) reduces to (10). As $\sigma_\psi^2$ goes to zero, $Z_i$ becomes collinear with $X_i$, and the bias goes to zero (as in (14)).

## 3.3 Constructing an Adjusted Occupational Income Score

The above model suggests that a score derived from average incomes conditioned on both occupation and a suite of common explanatory variables should provide estimates closer to those of a true earnings regression. There are a number of ways to approach this problem. The most important commonly-available variables influencing labor market outcomes are industry, occupation, sex, race/ethnicity, age, and geographic location.[12] Any adjusted OCCSCORE should account for differences along these lines. While other variables may no doubt be important, we focus on these because they are consistently available across decennial Censuses, meaning that our adjusted score will be widely applicable. However, researchers examining more particular questions or using non-Census data sources can use the method we propose below to construct scores adjusted by any relevant variables of interest.

The simplest and most general approach would be to adjust OCCSCORE in a fully nonparametric manner. For example, one could take an individual's OCCSCORE to be the median or mean income in a given base year for that individual's occupation within cells defined by their sex, age, race, state, and industry.[13] The advantage of this measure is that it allows for arbitrary interactions between all of the adjustment variables. For example, the age-earnings profile may differ flexibly between men and women in a given occupation, or the wage gap between races in a given occupation may vary between regions. The disadvantage of this approach is that stratifying on so many variables may result in small or empty cells, leading to excessively variable or missing OCCSCOREs for many individuals. Constructing new OCCSCOREs based on the (relatively small) 1% sample of the 1950 Census exacerbates

[12] Another common demographic variable that could be used is an indicator for foreign-born status. However, we caution that this may be misleading. The composition of the foreign-born population in the U.S. in 1950 differs dramatically from that in earlier years such as 1900 and 1850 in both racial/ethnic makeup and human capital. Thus, adjusting on this variable in a given base year may lead to inaccurate results when applied to other years.

[13] For example, Angrist (2002) constructs age- and sex-specific OCCSCOREs based on median income within cells. Collins and Wanamaker (2014) compute income scores by occupation and region specifically for black men.

this problem.

An alternative that avoids this problem involves a less flexible parametric approach. For example, one could regress income in a given base year on a series of occupation, sex, age, race, and geographic state indicator variables. The fitted coefficients could then be used to generate an adjusted OCCSCORE for each possible individual. This strategy is computationally simple and generates an adjusted OCCSCORE for all individuals. However, it likely misses many important interactions. For example, there is little reason to believe that early $20^{\text{th}}$ century earnings gaps between whites and blacks did not differ by region, nor does it seem likely that the age-earnings profile was the same across all occupations.

Our approach bridges these alternatives and aims to balance the need for a rich model of income determinants with the limitations imposed by the small number of observations available for some occupations. For a given base year, we compute a set of lasso-adjusted industry, demographic, and occupation (LIDO) scores as follows. For each Census-classified industry, we regress log income on a set of demographic covariates for all individuals between the ages of 20 and 70 employed with positive earnings in that industry.[14] We use the lasso algorithm, which solves the standard least squares problem subject to a constraint on the sum of the absolute values of the model coefficients (Tibshirani, 1996). This regularization approach controls the complexity of the model based on the importance of the predictors and the size and composition of the sample.

We allow for the following regressors: indicators for all occupations within the given industry, a polynomial for age, indicators for sex, race, and state of residence, and interactions between (1) sex and race, (2) sex and region, (3) occupation and sex, (4) occupation and an indicator for white, (5) Census region and an indicator for white, and (6) Census region and an indicator for black. In 1950, this results in a maximum of 654 possible covariates for the industry with the largest number of represented occupations (educational services). In general, the number of possible covariates is large relative to the sample size for each

[14]Industries follow the 1950 Census Bureau industrial classification system. Stratifying by industry eases the computation burden of the algorithm.

industry, and in some cases it may exceed the number of observations. The lasso algorithm shrinks coefficients depending on their relative importance, with the constraint forcing the coefficients on the least relevant predictors to zero. The sparsity induced by the lasso depends on the choice of tuning parameter $\lambda$ for each particular industry (described further below).

The set of potential predictors allows for occupational income to depend on a wide range of factors. Income for a given occupation can vary depending on the particular industry in which an individual works; it can also vary in flexible ways with race, sex, and geographic region. The age profile of earnings can also differ by industry. This generates income scores that more closely reflect reality than the interaction-free regression approach described above. It also avoids the small-cell overfitting problem that arises from the fully nonparametric approach; the lasso retains only the most relevant predictors of income differences, and the size of the model is scaled depending on the number of observations in each industry.

The extent to which the lasso generates a sparse model depends on the choice of tuning parameter $\lambda$, which reflects the stringency of the constraint. Since the importance of different demographic factors likely varies by industry, a one-size-fits-all choice would be inappropriate. We instead use 10-fold cross-validation to select a $\lambda$ that minimizes out-of-sample mean squared error for each industry.

# 4 Results

In this section, we document the performance of the LIDO score relative to OCCSCORE using modern Census data. Initially, we examine the stability of occupational earnings over time. We then estimate earnings regressions between 1950 and 2000 and measure the extent to which OCCSCORE causes errors of sign and magnitude in earnings regressions, and the extent to which the LIDO score ameliorates this problem.

## 4.1 Persistence of Occupational Income

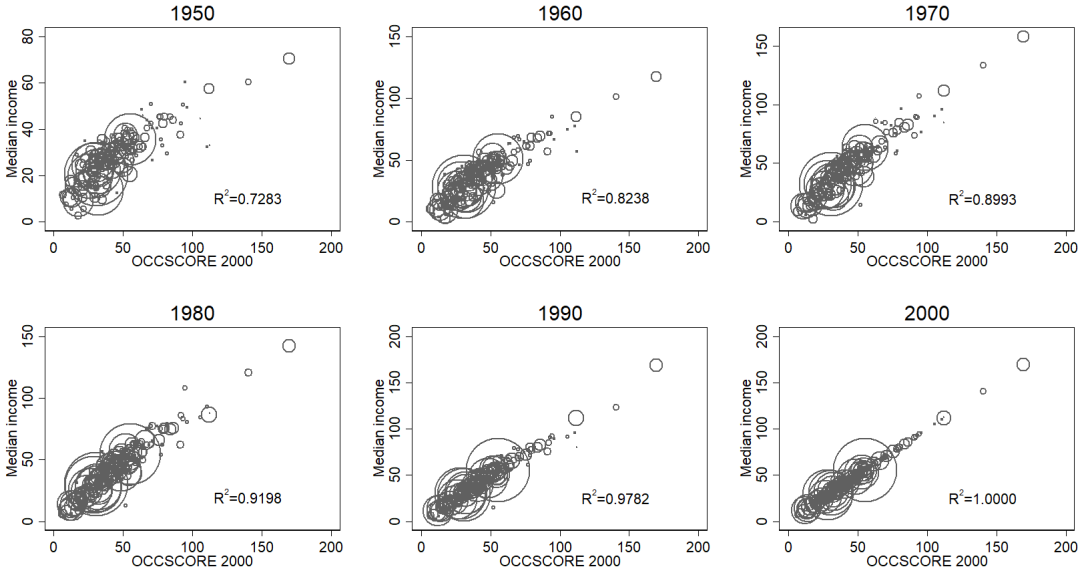
The occupational income score is a weighted average of the median earnings for males and females for each occupational category in 1950. This variable is likely a reasonable proxy for earnings in 1950, but it is unclear whether the relative earnings of occupations are sufficiently stable for it to remain an accurate proxy for income in earlier decades. If they are not, then even an adjusted version of OCCSCORE may perform poorly.

We test whether median earnings of an occupation accurately predict median earnings in the decades before the base year. We do this by constructing a 2000-based OCCSCORE and testing how well it predicts median earnings from past Censuses. If the 2000-based OCCSCORE successfully proxies for median earnings in 1950, then the 1950 OCCSCORE may be a reasonable proxy for median occupational earnings in 1900.

The results of this exercise can be seen in Figure 1. Each circle is an occupation weighted by the size of the occupational cell. The 2000 OCCSCORE perfectly predicts median earnings in 2000 by construction. For each decade removed from 2000, the $R^2$ decreases, implying that OCCSCORE is becoming worse as a proxy for median earnings. Even 50 years removed from the base year, $R^2 = 0.73$, implying that OCCSCORE remains a strong proxy for median earnings.

To provide further evidence, we examine changes in the rank correlation of median occupational income between 1950 and 2000. Measured by Spearman's rank correlation coefficient, the correlation between occupational rankings in 1990 and 2000 is 0.97. While this declines over time, it does so gradually. Between 1950 and 2000, the correlation is 0.81, which is still very high. While we cannot examine how this correlation changes nationally in the decades before 1950, we can examine the correlation between occupational income in 1950 and that in 1915 Iowa using state Census data (described in section 5.1). Using this information, we find that the rank correlation between 1950 and 1915 occupational earnings in Iowa is 0.7. This analysis provides some validation for the use of occupational income scores, since it demonstrates that the earnings hierarchy of occupations is likely to be reasonably

Figure 1: 2000 occupational income score and median income




**Notes:** Median earnings for each occupation are from the 1% sample of the U.S. Census. The 2000-based OCCSCOREs are the median earnings from the 2000 Census for individuals in each occupation. The size of each circle corresponds to the number of individuals in the occupational category during that Census year. Median earnings and OCCSCORE are measured in hundreds of 1950 dollars.

stable over time.

## 4.2 Errors of Magnitude

In this section, we analyze the magnitude of bias induced when using OCCSCORE and LIDO score as a proxy for income in an earnings regression. For the moment, we focus on estimating earnings gaps by race, gender, and state of residence. Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) introduce the Type M error rate as the expected value of an estimate divided by the true parameter value, conditional on the estimate being statistically different from zero. In this context, the true earnings gap is the earnings gap found using actual income data, and the estimated earnings gap is the gap using a proxy for income (either OCCSCORE or LIDO score).[15]

[15] Although we do not formally take the expectation of the earnings gaps, the large sample size of the Census ensures that the standard errors are small and the estimated coefficients will be close to their expectations.

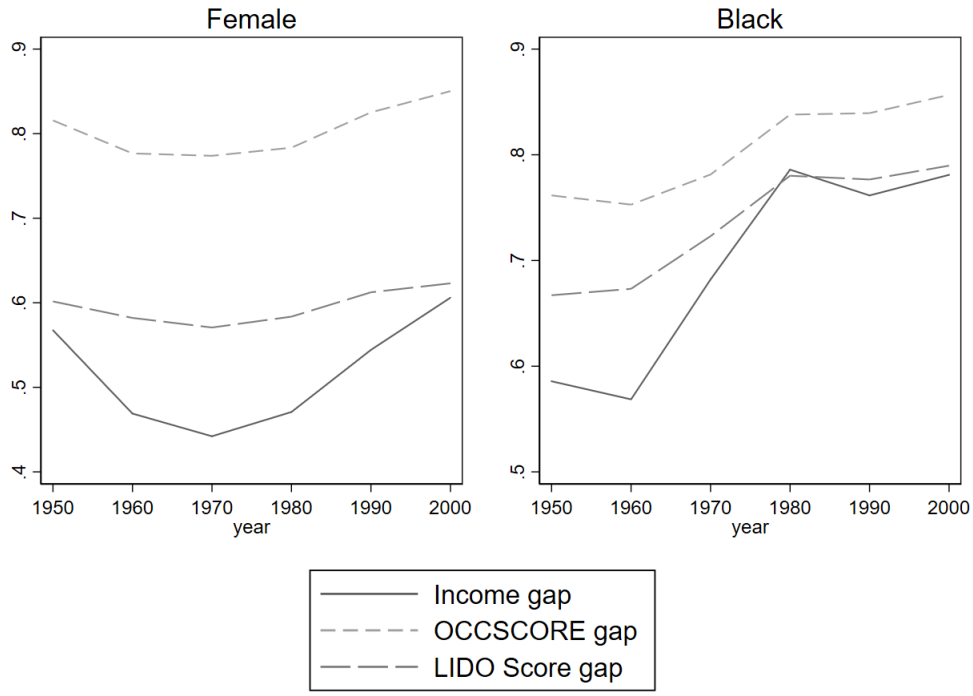
While many economic historians use 1950-based occupational income score as a proxy for occupational status, some interpret it as a proxy earnings, and then estimate models using data from pre-1950 Census years. We cannot directly test whether OCCSCORE produces coefficients similar to earnings regressions using pre-1950 national Census data. However, we can make modern data worse, so that the modern data suffers from the same problems as historical data. Here, we generate a 2000-based occupational income score and compare estimated racial and gender gaps from 1950-1990 with the true earnings gaps a researcher would have obtained by using actual earnings instead of the proxy.

Figure 2 graphs the implied earnings gaps using the three regressions. The first specification regresses the log of earnings on a set of dummies for state of residence, sex, race, and nativity. In addition to these dummy variables, the regression includes age and age squared. We run the regressions separately for every Census year from 1950 to 2000. Because we assume researchers would have used earnings instead of occupational income scores if earnings data were available, we treat these coefficients as the true parameters that researchers would like to estimate. The second regression uses the log of the 2000-based OCCSCORE instead of log earnings as the dependent variable. The dependent variable for the last regression is the log of the 2000-based LIDO score. For each regression, we restrict the sample to adults ages 25-65 who were in the labor force.

As expected, earnings gaps have declined for blacks since the 1950s and for women since the 1970s, and this is reflected in all three models. For all years, the coefficients on sex and race are of the same sign and statistically significant in all specifications, but the coefficients from the OCCSCORE specification suffer from attenuation bias. Using the LIDO score as the dependent variable reduces this bias, but does not eliminate it. The earnings gap estimated using the LIDO score more closely mirrors the true earnings gap than the OCCSCORE estimates.

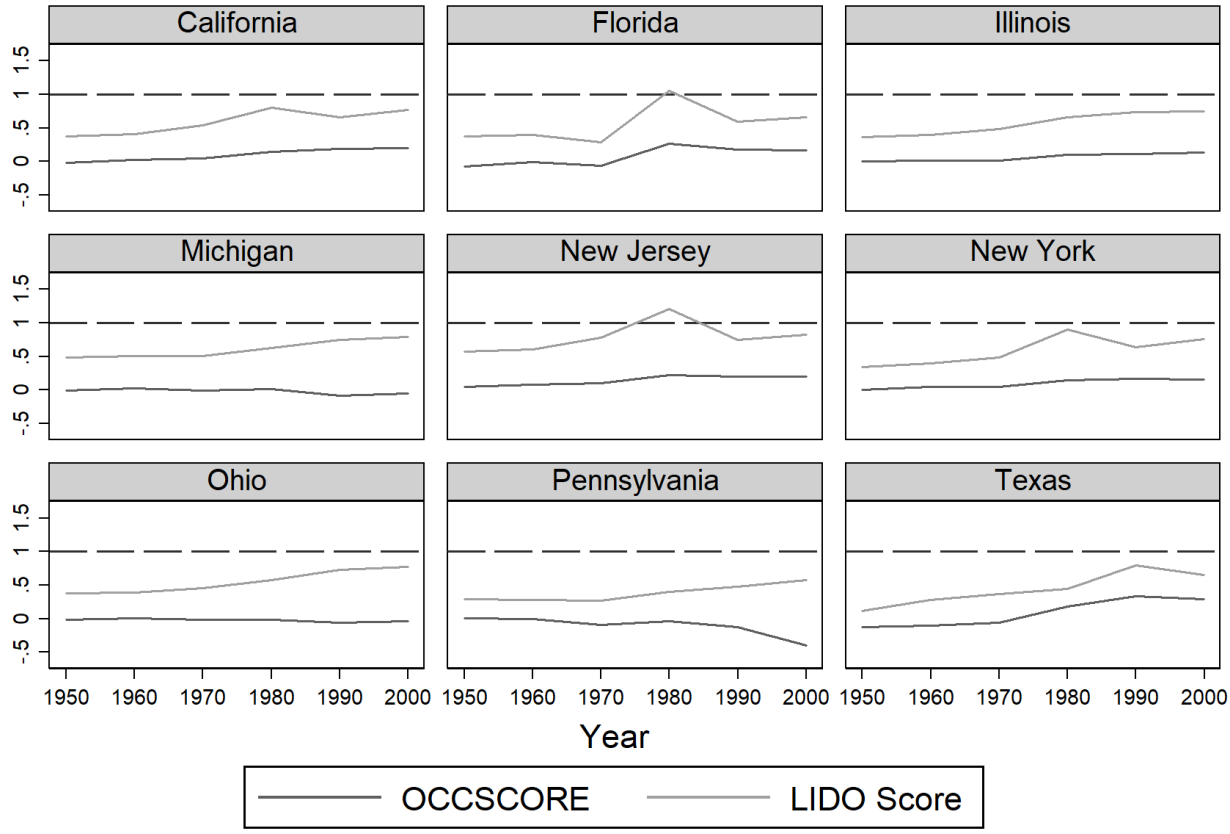
Our estimates of the female/male earnings ratio are similar to the extant literature (Goldin, 1990, p. 62). The female/male earnings ratio declined between 1950-1960, after which

Figure 2: Earnings ratios using earnings, OCCSCORE, and LIDO score




**Notes:** The data are from IPUMS (see Ruggles, Genadek, Goeken, Grover and Sobek (2015)). The graph displays the implied female/male and black/white income ratios from six earnings regressions. Note that the gaps are conditional on age, age squared, a dummy variable for U.S.-born, and state of residency. The female earnings gap is conditional on race, and the black/white earnings gap is conditional on sex. OCCSCORE uses a 2000-based occupational income score, whereas LIDO score is a 2000-based occupational income score constructed as described above. The sample is restricted to those between ages 25 and 65 who were in the labor force.

Figure 3: Ratios of estimated to true state fixed effects




**Notes:** The graph displays the ratio of estimated to "true" state fixed effects from six earnings regressions. OCCSCORE uses a 2000-based occupational income score, whereas LIDO score is a 2000-based occupational income score constructed as described above. The sample is restricted to those between ages 25 and 65 who were in the labor force.

the gender gap slowly narrowed. Margo (2016) provides Census estimates of the black/white earnings gap that are similar to ours. Black income increased relative to whites during the 1960s and 1970s, but the ratio has not narrowed significantly since the 1980s. Smith (1984) estimates the black/white income gap by assigning each individual the average income of a race by sex by age group cell from the 1970 Census. These estimates, produced at least a decade before the IPUMS OCCSCORE variable was regularly in use, are in essence an adjusted OCCSCORE.

Lastly, given the importance of geographic variables in difference-in-differences models, we analyze where geographic occupational gaps mirror geographic earnings gaps. Earnings can vary across space due to differences in the distribution of occupations or differences in

within-occupation earnings. While the LIDO score measures both, OCCSCORE captures only the former. Thus, if there are substantial within-occupation earnings differences across states, the use of OCCSCORE will fail to reflect the true spatial heterogeneity of earnings. Figure 3 plots the ratio of state fixed effects estimated using OCCSCORE and LIDO score to the "true" values estimated using actual income. These estimates are conditional on the same covariates as above. The Figure includes the nine most populous states in 2000 and eight most populous in 1950, spanning the East and West Coast, Rustbelt, and Sunbelt.

The graph illustrates the attenuation bias that occurs when OCCSCORE is used. Most of the spatial heterogeneity in earnings disappears completely, falsely indicating little difference in earnings across states. This can even occur in the base year where the OCCSCORE was constructed. When the LIDO score is used instead, estimates are much closer to their true values. While the performance drops off farther back in time, in many cases the 2000-based LIDO score gives more accurate results in 1950 than the 2000-based OCCSCORE does in 2000. Interestingly, the relative improvement generated by using the LIDO score does not appear to differ substantially by region. The 2000-based LIDO score leads to improved estimates over the entire time span for states that have seen dramatic industrial decline since 1950; it leads to similar improvements for states that have experienced much more positive economic trends.

## 4.3 Errors of Sign

If researchers are primarily concerned with the direction of an effect rather than its magnitude, the previous results suggest that some qualitative conclusions may not be seriously affected by the use of occupational income scores. The use of OCCSCORE as the dependent variable did not result in sign changes for gender and racial earnings differences. However, this result does not generalize to regressors with signs less predictable than race and gender indicators. Here, we show that OCCSCORE can result in errors of sign, or Type S errors. A Type S error occurs when the true population parameter is non-zero and the estimate is

statistically significant and of the wrong sign (see Gelman and Carlin (2014) and Gelman and Tuerlinckx (2000)).

In this section, we consider two models. We regress log income on 176 dummy explanatory variables: a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. Then, we estimate the model with a 2000-based occupational income score as the dependent variable and then again with a 2000-based LIDO score. Standard errors are clustered at the state level.

We then compare how often these models give conflicting results compared to the "true model" in which the dependent variable is log income. Many researchers are satisfied to use estimators that are biased towards zero since the sign of the estimator is likely to be the same as that of the parameter they are trying to estimate. A more serious problem occurs when researchers find spurious results that are statistically significant and of the wrong sign. For this reason, we say that the two models conflict for a particular coefficient if they produce opposite signs, but the estimates are statistical significant in both models.

The results from this exercise are shown in Table 2.[16] Estimates that are significant in both the earnings and OCCSCORE regressions are of conflicting signs 4% of the time in the base year, and the problem worsens as one gets farther from the base year. By 1950, 20% of statistically significant coefficients have the wrong sign when using OCCSCORE in place of earnings. The variables that are particularly affected are state and age. This finding is troubling for difference-in-differences estimates in which the treatment variable is often an explicit function of state of residency and/or birth cohort. In 1970, 33% of the age coefficients are incorrectly signed; in 1950, 52% of the state coefficients are incorrectly signed. This problem is reduced when using the LIDO score as the dependent variable. From 1980-2000, none of the coefficients are statistically significant and of the wrong sign. The numbers for 1950-1970 are only 4%, 1%, and 2%, respectively.

[16]The results are similar if we drop those in agriculture, an industry in which measuring income is particularly difficult (see Steckel (1991)).

Table 2: Percent of significant coefficients with conflicting signs

| OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.06 | 0.00 | 0.33 | 0.17 | 0.11 | 0.14 |
| State | 0.52 | 0.32 | 0.05 | 0.00 | 0.00 | 0.00 |
| Birth place | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Race and sex | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Family and household | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mean | 0.20 | 0.10 | 0.14 | 0.06 | 0.03 | 0.04 |

| LIDO score | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.06 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 |
| State | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Birth place | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Race and sex | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Family | 0.10 | 0.04 | 0.04 | 0.00 | 0.00 | 0.00 |
| Mean | 0.04 | 0.01 | 0.02 | 0.00 | 0.00 | 0.00 |

**Notes:** Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

Table 3: Mean ratio of the estimated coefficient to the "true" coefficient

| OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.38 | 0.18 | -0.01 | 0.15 | 0.20 | 0.14 |
| State | 0.06 | 0.26 | 0.29 | 0.40 | 0.30 | 0.33 |
| Birth place | 1.07 | 0.74 | 0.66 | 0.73 | 0.93 | 0.76 |
| Race and sex | 0.56 | 0.55 | 0.59 | 0.49 | 0.51 | 0.55 |
| Family and household | 0.32 | 0.39 | 0.50 | 0.54 | 0.44 | 0.50 |
| Mean | 0.28 | 0.30 | 0.27 | 0.37 | 0.32 | 0.33 |

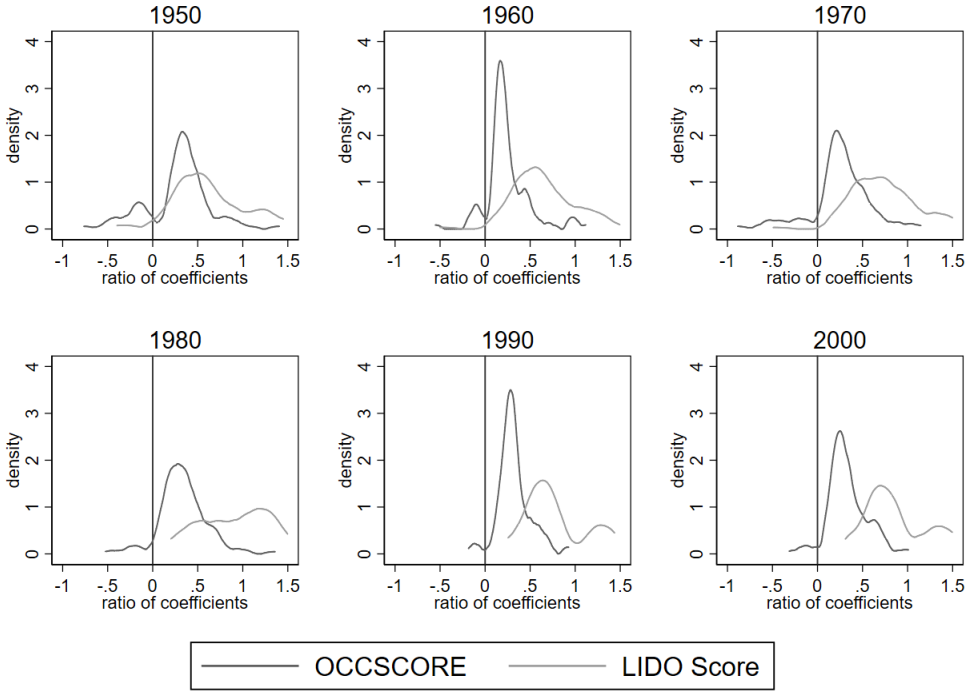
| LIDO OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 1.96 | 2.39 | 2.47 | 1.59 | 1.59 | 1.48 |
| State | 0.89 | 0.65 | 0.76 | 1.05 | 0.66 | 0.75 |
| Birth place | 1.06 | 0.69 | 0.63 | 0.74 | 0.90 | 0.91 |
| Race and sex | 0.94 | 0.89 | 1.07 | 0.90 | 0.90 | 0.97 |
| Family and household | 0.34 | 0.37 | 0.43 | 0.55 | 0.50 | 0.57 |
| Mean | 1.12 | 1.21 | 1.28 | 1.09 | 0.94 | 0.97 |

**Notes:** Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

Table 3 displays the mean ratios of the estimated coefficients to the "true" earnings regression coefficients. Ideally, these ratios would be close to 1 and would never be negative. The results suggest that the OCCSCORE coefficients are typically 27-37% of the earnings regression coefficients. The LIDO score coefficients are closer to being centered around 1 and depending on the year vary between 94-128% of the earnings regression coefficients. Variables that are unlikely to be correlated with the demographic and industry adjusting variables (such as household and family characteristics) produce similar estimates for both the OCCSCORE and LIDO score regressions.

Figure 4 graphs kernel density estimates of the ratio of the estimated and true coefficients. The density for LIDO score is closer to being centered around one (less attenuation bias) and has less weight to the left of zero (conflicting signs). Using OCCSCORE leads to less

Figure 4: Density of ratios of estimated to true coefficients using OCCSCORE and LIDO score




**Notes:** Data are from IPUMS (see Ruggles et al. (2015)). The figures display the density of the ratio of the estimated coefficients (using a 2000-based OCCSCORE) and the "true" coefficient using observable income. Each year contains 176 regression coefficients and includes a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The sample is restricted to those between ages 25 and 65 who were in the labor force.

accurate results the farther away from the base year we get, whereas estimates using the LIDO score are likely to be of the same sign even 50 years prior to the base year. Although the LIDO score is rarely of the wrong sign, it too suffers from some attenuation bias when the data are 50 years removed from the base year.

These results included non-whites and women, both of which would likely earn below median occupational earnings. For this reason, many papers using occupational income scores restrict the sample to only white males. In the Appendix, we repeat Tables 2 and 3 and Figure 4 while restricting the sample to only white males. We regress log earnings on 168 dummy variables for age, state of residency, state of birth, and family and household.

In Table A.1 we display the percent of coefficients that are statistically significant in both regressions that have conflicting signs. Restricting the sample to white males improves the performance of OCCSCORE relative to section 4.3; however, state coefficients are still of the wrong sign up to 50% when 50 years removed from the base years, 38% when 40 years, and 23% when 30 years removed from the base year. The state coefficients are never statistically significant using the LIDO score for any of the years. In Table A.2, we see that even for white males the OCCSCORE coefficients are approximately a third of the earnings regression coefficients, whereas the LIDO score coefficients are close to being center around 1. Lastly, in Figure A.1 we present kernel density estimates of the ratio of estimated coefficients to the true coefficients.

# 5 Applications

## 5.1 Earnings Gaps in the 1915 Iowa Census

The analysis in Section 4 shows that the LIDO score improves estimates of racial and gender earnings gaps in modern Census data. To assess whether this conclusion applies in a historical context, we exploit a rare source of pre-1950 income data, the 1915 Iowa State Census (Goldin and Katz, 2010).[17] This was the first Census in the U.S. to collect data on income prior to 1940. The sample contains records on 5.5% of the urban population drawn from three of Iowa's largest cities: Des Moines, Dubuque, and Davenport. It also contains 1.8% of the population of counties not containing a major city; the ten counties sampled span the geography of the state. This data allows us to compare racial and gender earnings gaps and the age-income profile estimated using OCCSCORE, LIDO score, and true earnings in a historical setting.

For the estimation, we restrict the sample to those between the ages of 20 and 70 and

[17]This data was recently used to examine intergenerational mobility by Feigenbaum (2018).

exclude those with missing occupation data or zero/missing earnings.[18] The Census reports occupation categories according to the 1940 scheme. We cross-walked these with the 1950 scheme to match individuals in 1915 to their 1950 OCCSCORE.[19] The final sample includes 15,201 individuals. We estimate the earnings gap between whites and blacks and men and women; approximately 1% of the sample is black (196 obs) and 14% of the sample is female (2,153 obs). We also estimate the age-earnings profile, urban-rural gap, and the native-foreign born gap.

In column (1) of the top panel of Table 4, we report the coefficients from a regression of log earnings on indicators for black, female, urban, and foreign-born, as well as a quadratic polynomial for age. Women and African-Americans earn less than white men and, as is typical, earnings increase with age but at a diminishing rate. In column (2), we replace log earnings with the log of the standard 1950 OCCSCORE. The black-white earnings gap coefficient declines by almost half. The gender earnings gap is positive. The age-earnings profile is attenuated, as is the gap between natives and immigrants. The urban premium is overestimated.

Moving to column (3), we replace the unadjusted OCCSCORE with 1950 LIDO score.[20] Using this approach, the earnings gap for women is similar to that estimated using true earnings. The estimated earnings gap for blacks is slightly larger than the true value but closer in magnitude than the standard OCCSCORE estimate. The age-earnings profile, while still attenuated, is closer to the correct value, as is the urban premium, which was not explicitly incorporated in the construction of the LIDO score. The OCCSCORE and LIDO score estimates of the immigrant penalty are similar.

The bottom panel of Table 4 repeats the above analysis excluding any individuals whose

[18]We also exclude those whose race is recorded as missing (19 observations) and those whose race is recorded as Mixed or Asian (5 observations).

[19]In some cases, the 1940 scheme aggregated some occupations; for example, bookkeepers, accountants, and cashiers fall into one occupation category in 1940 but are disaggregated into three separate categories in 1950. There are 7 occupation categories in 1940 (out of 194 total) that cannot be matched uniquely to a 1950 occupation. We exclude individuals in these categories.

[20]Because industry was not recorded in the Iowa State Census, we instead estimate our lasso-adjusted score by occupation (retaining all of the other predictors listed in section 3.3).

1950 occupational classification is "Farmers (owners and tenants)." Since farm income is unusually heterogeneous and farmers saw a substantial change in their occupational standing between 1850 and 1950, we examine our results excluding this category of earners. The LIDO score again yields estimates closer to the earnings regressions for the female gap and age-earnings profile, though it does not deliver a more accurate estimate of the black-white gap.[21]

## 5.2 Estimates of Intergenerational Mobility

Labor economists often measure intergenerational mobility by regressing a son's socioeconomic status on his father's socioeconomic status:

$$I_i^{\text{son}} = \beta_0 + \beta_1 I_i^{\text{father}} + u_i \tag{27}$$

where $I_i^{\text{son}}$ is the log income of a son observed during adulthood, and $I_i^{\text{father}}$ is the log income of a father observed while the son was a child. The transmission coefficient $\beta_1$ is an elasticity typically between 0 and 1, with 1 representing perfect immobility between generations and 0 representing perfect mobility. Historical evidence on occupational mobility across generations relies heavily on occupational income scores instead of income for two reasons. First, to obtain data on fathers' and sons' labor market outcomes in the Census, one needs to link across Census years, which is typically only possible using given and surnames. Names do not become publicly available in the Census until 72 years after the Census year, meaning occupations are the only available labor market outcomes for both fathers and sons. Second, estimates of how intergenerational mobility have changed over time require data spanning at least three generations, implying that such estimates must make use of historical data.

[21] Results are similar if we additionally exclude individuals whose occupational classification is "Managers, officials, and proprietors (n.e.c.)." This category is problematic due to the aggregation of small business proprietors and chief executives of large corporations.

As Solon (1989, 1992) has highlighted, measurement error in the dependent variable has the potential to bias intergenerational mobility estimates in favor of greater mobility. Let $e_i^{\text{son}} = I_i^{\text{son}} - \tilde{y}_i^{\text{son}}$ and $e_i^{\text{father}} = I_i^{\text{father}} - \tilde{y}_i^{\text{father}}$ be the measurement error from using an occupational index (either OCCSCORE or LIDO score) for the son and father, respectively. Then researchers estimate:

$$\tilde{y}_i^{\text{son}} = \beta_0 + \beta_1 \tilde{y}_i^{\text{father}} + \underbrace{\beta_1 e_i^{\text{father}} - e_i^{\text{son}} + u_i}_{\epsilon_i}. \tag{28}$$

This regression differs from the model in Section 3 since OCCSCORE appears on both the left-hand and right-hand side of the regression. The measurement errors $e_i$ are likely to be smaller if one uses a demographically-adjusted LIDO score instead of OCCSCORE, since racial, age, and regional differences in occupational earnings will not be captured in $e_i$. However, when using OCCSCORE, some of the measurement error is likely to cancel out since $e_i^{\text{son}}$ is positively correlated with $e_i^{\text{father}}$. Our estimate of the transmission coefficient will be biased if $\text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) \neq 0$. If there is little intergenerational mobility, in which case $\beta_1$ is close to 1, and if the son's measurement error is highly correlated with the father's measurement error, then the second term of covariance is close to zero. Alternatively, suppose $e_i^{\text{son}} = \tilde{\beta}_0 + \tilde{\beta}_1 e_i^{father} + \upsilon_i$, where $\upsilon_i$ is independent of all other variables. The transmission coefficient $\tilde{\beta}_1$ reflects that fathers who earn above average within their occupations are likely to have sons who earn above average within occupations. Then, $\text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) = \text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - \tilde{\beta}_1 e_i^{\text{father}})$. Thus, the bias from estimating equation 28 using OCCSCOREs will be small so long as the transmission in overall income from father to son is similar to the transmission of excess income within occupation. For these reasons, using occupational income scores instead of income should lead to a smaller amount bias in this context.

In Table 5, we provide estimates of intergenerational mobility using the IPUMS linked data sets. These data link the 1% samples of the 1850, 1860, and 1900-1930 Censuses to the

1880 complete count. The sample is restricted to those who during the first Census year were children of the household head, no older than 15 years old, and male. We regress the log of a son's OCCSCORE during the second Census year on the log of the father's OCCSCORE during the first Census year, and then repeat the regression using LIDO score. To make the estimates comparable, we restrict the sample to father-son pairs in which neither the father nor the son has a missing LIDO score. The resulting coefficients are elasticities with higher coefficients implying occupational immobility. Row 1 of columns (2) and (4) of Table 5 are replications of the estimates in row 7 of Table 3 of Olivetti and Paserman (2015), but restricting the sample to those with non-missing LIDO scores. We present estimates for whites using all samples, and for blacks using samples in which both the father and son are observed in the postbellum era.

The intergenerational mobility estimates for whites are similar for both measures. However, the OCCSCORE estimates suggest that blacks had twice the intergenerational mobility of whites ($\beta_1$ closer to zero). The LIDO score estimates suggest that black intergenerational mobility was much closer to white intergenerational mobility than previously thought. In fact, between 1880-1900 and 1880-1910, blacks had less intergenerational mobility than whites.

# 6 Conclusion

Using modern Census data, we find that median earnings within a given occupation are highly correlated over time. This implies that the occupational income score is a reasonable proxy for occupational status even when used with historical Census data. However, much of modern labor economics focuses on earnings regressions rather than occupational status, and we find that occupational income scores systematically underestimate income gaps due to race, gender, age, and location. Standard earnings regression covariates such as state of residency and state of birth indicators are attenuated and can be of the wrong sign up to

20 percent of the time, even when the regression is restricted to white males. We construct a new lasso-adjusted industry, demographic, and occupation (LIDO) score which flexibly accounts for differences in earnings across race, gender, age, state, occupation, and industry (variables available in every Census going back to 1850). Our alternative score reduces errors of magnitude and sign. We have made this alternative score available online, and encourage researchers to use a data-driven approach to construct their own alternative scores if the context demands.[22]

To examine the performance of the LIDO score in a historical context, we exploit the 1915 Iowa State Census, which collected data on both occupation and earnings. We find that estimated race and gender earnings gaps in 1915 Iowa using true earnings are sizable; however, when using standard OCCSCORE as a proxy, the racial earnings gap is attenuated by almost half and the gender earnings gap is incorrectly signed. Our LIDO score yields earnings gaps close to their true values. We also use the LIDO score to measure intergenerational income transmission. This analysis is based on father-son pairs linked across the 1850-1930 decennial Censuses. In this setting, we find that standard OCCSCOREs and LIDO scores perform similarly for white males because measurement errors for fathers and sons are likely to be correlated. However, transmission coefficients are attenuated for black men, suggesting a lower rate of intergenerational mobility than previously thought.

Our results suggest that future research in economic history and other fields should use LIDO scores if researchers are interested in earnings regressions. When should researchers continue to use standard occupational income scores? Unadjusted OCCSCOREs may be a reasonable proxy for total earnings over the life cycle. A recent college graduate may have a low adjusted OCCSCORE, since young professionals are below their peak occupational earnings, whereas the standard OCCSCORE would assign workers of all ages within the occupation the same earnings. Race and gender earnings gaps may also shrink over the life-cycle. Without linked Census data, it is impossible to know whether OCCSCOREs or

[22]It can be found at `http://www2.oberlin.edu/faculty/msaavedr/lido.html`.

LIDO scores provide a better proxy for lifetime wealth. Although linked Census data does exist, it does not cover the 1950-2000 periods in which earnings data are available.

Studies that are primarily interested in occupational status, rather than earnings, may wish to retain the standard OCCSCORE. The 1950 OCCSCORE provides a ranking of occupations by earnings, and we find substantial persistence in these occupational rankings. Even for such studies, labor economists using modern data almost invariably use earnings instead of using OCCSCORE, and using LIDO scores would make the historical literature more comparable to the modern labor economics literature. For these reasons, we recommend the LIDO score as a complement rather than a substitute to the occupational income score.

# References


Aaronson, Daniel, Fabian Lange, and Bhashkar Mazumder (2014). Fertility transitions along the extensive and intensive margins. *American Economic Review*, **104**(11), pp. 3701–3724.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, **102**(5), pp. 1832–1856.

——— (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, **122**(3), pp. 467–506.

Angrist, Josh (2002). How do sex ratios affect marriage and labor markets? Evidence from America's second generation. *Quarterly Journal of Economics*, pp. 997–1038.

Bailey, Martha J and William J Collins (2006). The wage gains of African-American women in the 1940s. *Journal of Economic History*, **66**(03), pp. 737–777.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, **119**(1), pp. 249–275.

Bleakley, Hoyt (2007). Disease and development: evidence from hookworm eradication in the American South. *Quarterly Journal of Economics*, **122**(1), p. 73.

——— (2010). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics*, **2**(2), pp. 1–45.

Bleakley, Hoyt and Joseph Ferrie (2016). Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *Quarterly Journal of Economics*, **131**(3), pp. 1455–1495.

Bleakley, Hoyt and Fabian Lange (2009). Chronic disease burden and the interaction of education, fertility, and growth. *Review of Economics and Statistics*, **91**(1), pp. 52–65.

Carruthers, Celeste K and Marianne H Wanamaker (2017). Separate and unequal in the labor market: human capital and the jim crow wage gap. *Journal of Labor Economics*, **35**(3), pp. 000–000.

Chin, Aimee (2005). Long-Run Labor Market Effects of Japanese American Internment during World War II on Working-Age Male Internees. *Journal of Labor Economics*, **23**(3), pp. 491–525.

Collins, William J (2000). African-American economic mobility in the 1940s: a portrait from the Palmer Survey. *Journal of Economic History*, **60**(03), pp. 756–781.

Collins, William J and Marianne H Wanamaker (2014). Selection and economic gains in the great migration of african americans: New evidence from linked census data. *American Economic Journal: Applied Economics*, **6**(1), pp. 220–252.

——— (2015). The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *Journal of Economic History*, **75**(04), pp. 947–992.

Cook, Lisa D, Trevon D Logan, and John M Parman (2014). Distinctively black names in the American past. *Explorations in Economic History*, **53**, pp. 64–82.

——— (2016). The mortality consequences of distinctively black names. *Explorations in Economic History*, **59**, pp. 114–125.

Feigenbaum, James J. (2018). Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940. *The Economic Journal*, **128**(612), pp. F446–F481.

Gelman, Andrew and John Carlin (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**(6), pp. 641–651.

Gelman, Andrew and Francis Tuerlinckx (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, **15**(3), pp. 373–390.

Goldin, Claudia (1990). *The gender gap: An economic history of American women*. New York: Cambridge University Press.

Goldin, Claudia and Lawrence Katz (2010). The 1915 Iowa State Census Project: ICPSR28501-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, `http://doi.org/10.3886/ICPSR28501.v1`.

Hyslop, Dean R and Guido W Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, **19**(4), pp. 475–481.

Lee, Dara N (2013). The impact of repealing Sunday closing laws on educational attainment. *Journal of Human Resources*, **48**(2), pp. 286–310.

Lee, Sanghoon and Jeffrey Lin (2017). Natural Amenities, Neighbourhood Dynamics, and Persistence in the Spatial Distribution of Income. *Review of Economic Studies*, p. rdx018.

Lleras-Muney, Adriana and Allison Shertzer (2015). Did the Americanization Movement Succeed? An Evaluation of the Effect of English-Only and Compulsory Schooling Laws on Immigrants. *American Economic Journal: Economic Policy*, **7**(3), pp. 258–290.

Margo, Robert A (2016). Obama, Katrina, and the Persistence of Racial Inequality. *Journal of Economic History*, **76**(02), pp. 301–341.

Massey, Catherine G (2016). Immigration quotas and immigrant selection. *Explorations in Economic History*, **60**, pp. 21–40.

Minns, Chris (2000). Income, cohort effects, and occupational mobility: a new look at immigration to the United States at the turn of the 20th century. *Explorations in Economic History*, **37**(4), pp. 326–350.

Olivetti, Claudia and M Daniele Paserman (2015). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1940. *American Economic Review*, **105**(8), pp. 2695–2724.

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

Roy, Andrew Donald (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, **3**(2), pp. 135–146.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015). Integrated Public Use Microdata Series Version 6.0, Machine-readable database. Minneapolis: University of Minnesota.

Saavedra, Martin (2015). School quality and educational attainment: Japanese American internment as a natural experiment. *Explorations in Economic History*, **57**, pp. 59–78.

——— (2017). Early-life disease exposure and occupational status: The impact of yellow fever during the 19th century. *Explorations in Economic History*, **64**, pp. 62–81.

Sacerdote, Bruce (2005). Slavery and the intergenerational transmission of human capital. *Review of Economics and Statistics*, **87**(2), pp. 217–234.

Smith, James P (1984). Race and human capital. *American Economic Review*, **74**(4), pp. 685–698.

Sobek, Matthew (1995). The comparability of occupations and the generation of income scores. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **28**(1), pp. 47–51.

Solon, Gary (1989). Biases in the estimation of intergenerational earnings correlations. *Review of Economics and Statistics*, **71**(1), pp. 172–174.

——— (1992). Intergenerational income mobility in the United States. *American Economic Review*, **82**(3), pp. 393–408.

Steckel, Richard H (1991). The quality of census data for historical inquiry: A research agenda. *Social Science History*, **15**(4), pp. 579–599.

Stephens, Melvin and Dou-Yan Yang (2014). Compulsory education and the benefits of schooling. *American Economic Review*, **104**(6), pp. 1777–1792.

Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**(1), pp. 267–288.

Ward, Zachary (2017). Birds of passage: Return migration, self-selection and immigration quotas. *Explorations in Economic History*, **64**, pp. 37–52.

Table 4: Earnings in the 1915 Iowa State Census

| Full sample | Log of earnings | Log of 1950 OCCSCORE | LIDO score |
|---|---|---|---|
| | (1) | (2) | (3) |
| Black | -0.389*** | -0.243*** | -0.422*** |
| | (0.0378) | (0.0255) | (0.0232) |
| Female | -0.516*** | 0.020* | -0.507*** |
| | (0.0201) | (0.0106) | (0.0096) |
| Age | 0.216*** | 0.033*** | 0.078*** |
| | (0.0072) | (0.0045) | (0.0036) |
| Age$^2$ | -0.144*** | -0.008** | -0.069*** |
| | (0.0070) | (0.0037) | (0.0030) |
| Urban | 0.171*** | 0.332*** | 0.191*** |
| | (0.0106) | (0.0060) | (0.0049) |
| Foreign born | -0.146*** | -0.092*** | -0.088*** |
| | (0.0186) | (0.0099) | (0.0077) |
| Observations | 15,201 | 15,201 | 15,201 |
| $R^2$ | 0.152 | 0.125 | 0.316 |
| Excluding farmers | | | |
| Black | -0.375*** | -0.309*** | -0.438*** |
| | (0.0385) | (0.0211) | (0.0194) |
| Female | -0.470*** | -0.084*** | -0.504*** |
| | (0.0188) | (0.0118) | (0.0109) |
| Age | 0.164*** | 0.093*** | 0.150*** |
| | (0.0076) | (0.0055) | (0.0047) |
| Age$^2$ | -0.127*** | -0.034*** | -0.089*** |
| | (0.0078) | (0.0044) | (0.0040) |
| Urban | 0.316*** | 0.129*** | 0.122*** |
| | (0.0121) | (0.0077) | (0.0067) |
| Foreign born | -0.242*** | -0.118*** | -0.113*** |
| | (0.0194) | (0.0130) | (0.0111) |
| Observations | 11,982 | 11,982 | 11,982 |
| $R^2$ | 0.197 | 0.080 | 0.364 |

**Notes:** Linear regressions of earnings measures on indicators for black, female, urban, and foreign-born status as well as a quadratic polynomial in age. Top panel includes all individuals except those whose race is recorded as Missing, Mixed, or Asian (24 observations), those who are below the age of 20 or above the age of 70, those with missing occupation data, and those with zero or missing earnings. Bottom panel further excludes individuals whose 1950 occupational classification is "Farmers (owners and tenants)." *** p<0.01, ** p<0.05, * p<0.1

Table 5: Estimates of Intergenerational Mobility

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Panel A: Mobility among whites | | | | | | |
| Dependent variable: log of son's OCCSCORE | | | | | | |
| | 1850-1880 | 1860-1880 | 1880-1900 | 1880-1910 | 1880-1920 | 1880-1930 |
| Log of father's OCCSCORE | 0.402*** | 0.460*** | 0.544*** | 0.428*** | 0.403*** | 0.379*** |
| | (0.0219) | (0.0172) | (0.0127) | (0.0134) | (0.0150) | (0.0146) |
| $N$ | 2804 | 3945 | 8354 | 7345 | 5687 | 5524 |
| Dependent variable: log of son's LIDO score | | | | | | |
| | 1850-1880 | 1860-1880 | 1880-1900 | 1880-1910 | 1880-1920 | 1880-1930 |
| Log of father's LIDO score | 0.457*** | 0.439*** | 0.408*** | 0.367*** | 0.394*** | 0.387*** |
| | (0.0193) | (0.0154) | (0.0107) | (0.0105) | (0.0132) | (0.0138) |
| $N$ | 2804 | 3945 | 8354 | 7345 | 5687 | 5524 |
| Panel B: Mobility among blacks | | | | | | |
| Dependent variable: log of son's OCCSCORE | | | | | | |
| | 1850-1880 | 1860-1880 | 1880-1900 | 1880-1910 | 1880-1920 | 1880-1930 |
| Log of father's OCCSCORE | | | 0.174*** | 0.161** | 0.0825 | 0.184* |
| | | | (0.0514) | (0.0581) | (0.0698) | (0.0786) |
| $N$ | | | 520 | 375 | 250 | 204 |
| Dependent variable: log of son's LIDO score | | | | | | |
| | 1850-1880 | 1860-1880 | 1880-1900 | 1880-1910 | 1880-1920 | 1880-1930 |
| Log of father's LIDO score | | | 0.600*** | 0.541*** | 0.562*** | 0.506*** |
| | | | (0.0436) | (0.0600) | (0.0954) | (0.0969) |
| $N$ | | | 520 | 375 | 250 | 204 |

**Notes:** Data are from the IPUMS linked data files. These data are from 1% samples of the 1850, 1860, 1900, 1910, 1920 and 1930 Censuses linked to the 1880 complete count. The sample is restricted to those who during the first Census year were children of the household head, male, and no older than 15 years old. Standard errors are in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

# Appendix

Table A.1: Percent of significant coefficients with conflicting signs for white males only

| OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.03 | 0.00 | 0.13 | 0.00 | 0.00 | 0.00 |
| State | 0.50 | 0.38 | 0.23 | 0.03 | 0.00 | 0.00 |
| Birth place | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Family | 0.00 | 0.00 | 0.15 | 0.04 | 0.00 | 0.08 |
| Mean | 0.13 | 0.10 | 0.12 | 0.02 | 0.00 | 0.02 |

| LIDO score | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| State | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Birth place | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Family | 0.00 | 0.00 | 0.12 | 0.04 | 0.00 | 0.08 |
| Mean | 0.01 | 0.00 | 0.02 | 0.01 | 0.00 | 0.02 |

**Notes:** Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

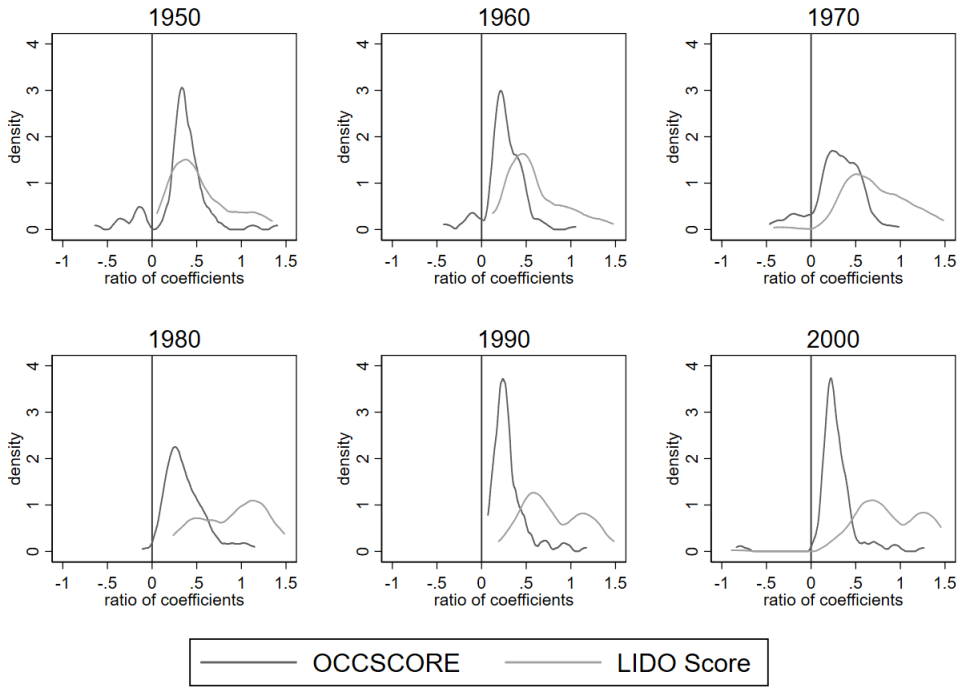
Table A.2: Mean ratio of the estimated coefficient to the "true" coefficient for white males only

| OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 0.41 | 0.21 | 0.18 | 0.25 | 0.20 | 0.18 |
| State | 0.10 | 0.04 | 0.17 | 0.38 | 0.27 | 0.30 |
| Birth place | 0.42 | 0.46 | 0.54 | 0.56 | 0.59 | 0.59 |
| Family | 0.33 | 0.33 | 0.28 | 0.39 | 0.59 | 0.36 |
| Mean | 0.32 | 0.25 | 0.29 | 0.35 | 0.33 | 0.28 |

| LIDO OCCSCORE | | | | | | |
|---|---|---|---|---|---|---|
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
| Age | 2.17 | 1.89 | 1.77 | 1.26 | 1.18 | 1.27 |
| State | 0.79 | 0.65 | 0.78 | 1.07 | 0.63 | 0.72 |
| Birth place | 0.39 | 0.47 | 0.51 | 0.62 | 0.79 | 0.68 |
| Family | 0.29 | 0.32 | 0.32 | 0.46 | 0.66 | 0.48 |
| Mean | 1.16 | 0.99 | 0.98 | 0.95 | 0.83 | 0.87 |

**Notes:** Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

Figure A.1: Density of ratios of estimated to true coefficients using OCCSCORE and LIDO score for White Males Only




**Notes:** Data are from IPUMS (see Ruggles et al. (2015)). The figures display the density of the ratio of the estimated coefficients (using a 2000-based OCCSCORE) and the "true" coefficient using observable income. Each year contains 185 regression coefficients and includes a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The sample is restricted to those between ages 25 and 65 who were in the labor force.