Essays on Intergenerational Mobility and Inequality in Economic History

Essays on Intergenerational Mobility and Inequality in Economic History

A dissertation presented by

James Joseph Feigenbaum

to

The Department of Economics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Economics

Harvard University
Cambridge, Massachusetts

April 2016

Dissertation Advisors:
Professor Claudia Goldin
Professor Lawrence Katz

Author:
James Joseph Feigenbaum

Essays on Intergenerational Mobility and Inequality in Economic History

ABSTRACT

This dissertation explores intergenerational mobility and inequality in the early twentieth century. The first chapter asks whether economic downturns increase or decrease mobility. I estimate the effect of the Great Depression on mobility, linking a sample of fathers before the Depression to their sons in 1940. I find that the Great Depression lowered intergenerational mobility for sons growing up in cities hit by large downturns. The effects are driven by differential, selective migration: the sons of richer fathers are able to move to better destinations. The second chapter compares historic rates of intergenerational mobility to today. Based on a sample matched from the Iowa 1915 State Census to the 1940 Federal Census, I argue that there was more mobility in the early twentieth century than is found in contemporary data, whether measured using intergenerational elasticities, rank-rank correlations, educational persistence, or occupational status measures. In the third chapter, I detail the machine learning method used to create the linked census samples used in chapters 1 and 2. I use a supervised learning approach to record linkage, training a matching algorithm on hand-linked historical data which is able to efficiently and accurately find links in noisy in historical data.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I am indebted to my committee members for their guidance and support. To my main advisors, Claudia Goldin and Larry Katz, I owe an enormous thanks. From my first year as a student in their History and Human Capital course to today, they have spent many hours advising and encouraging me. Every question I ask, any data I collect, and every project I work on, I begin by asking whether Claudia and Larry will think it is good economics. The writing in this dissertation owes much to Claudia's red pen and her deciphering of Larry's marginalia. Larry's advice on points big and small shaped this dissertation, from my footnotes, axis labels, and table captions, to the structure of the chapters and my empirical strategy. Richard Hornbeck never stopped asking me what else I was working on, pushing me to keep coming up with new ideas and questions. Ed Glaeser made sure I wrote a paper that could appeal to any audience, and his questions during my presentations ensured I had a back-up slide for every claim.

My education in economics began as an undergraduate at Wesleyan University. Without the incredible teaching and advising I benefited from at Wesleyan, I doubt I would be completing this dissertation today. My two undergraduate advisors, Richard Grossman and Cameron Shelton, deserve special thanks. I received my first taste of economic history as a research assistant for Richard Grossman, transcribing early twentieth century German bank balance sheets. Cameron Shelton agreed to advise my senior thesis during our first meeting at his office hours, and he guided my first foray into independent research.

I have been very lucky work with a number of excellent coauthors and am proud to call them all my friends. A special thanks to Andrew Hall, Alex Fouirnaies, Chris Muller, Martin Rotemberg, Jamie Lee, and Filippo Mezzanotti, who all read drafts of these chapters and picked up the slack on our projects while I worked on my job market paper. And I assure all of you that the next draft of whatever it was that I promised you last fall is almost finished.

For detailed comments and suggestions on the chapters of this dissertation, I also thank Natalie Bau, Miles Corak, Ellora Derenoncourt, Daniel Gross, Christoph Hafemeister, Simon Jaeger, Sandy Jencks, Robert Margo, Suresh Naidu, Nathan Nunn, Evan Roberts, Gary Solon,

and Bryce Millett Steinberg, as well as seminar participants at the Harvard Economic History Tea, the Harvard Labor Workshop, NBER DAE Summer Institute, the Berkeley Demography Conference on Census Linking, the University of Michigan H2D2 Seminar, the Cliometrics World Congress, and in the Harvard Multidisciplinary Program in Inequality & Social Policy. I am also appreciative for institutional support from Gia Petrakis, Brenda Piquet, and Pam Metz. I thank Viroopa Volla, Andrew Creamer, Justin Meretab, and especially Alex Velez-Green for excellent research assistance.

Finally, I thank my wife, Alexandra Steinlight, who is my best critic, best editor, and best friend.

INTRODUCTION

This dissertation explores the historical roots of economic inequality and mobility and looks to the past to expand our understanding of contemporary economic questions. In particular, I focus on intergenerational mobility in the early twentieth century United States. I examine the effects of economic conditions on mobility, compare mobility rates across time, and develop new tools to construct linked longitudinal microdata across generations.

Do severe economic downturns increase intergenerational economic mobility by breaking links between generations, or do they instead reduce mobility by limiting opportunity for the young? I answer this question in the first chapter, estimating rates of intergenerational mobility during the Great Depression in American cities that experienced downturns of varying severity. I create two new historical samples, digitizing and transcribing archival data on individual earnings and linking fathers to sons before and after the Depression. To build these longitudinal samples, I develop a machine learning approach to census matching that enables me to link individuals accurately and efficiently between censuses in the absence of unique identification numbers. I find that the Great Depression lowered intergenerational mobility for sons growing up in cities hit by large downturns. These results are not driven by place-specific mobility differences: for the generation before the Depression, mobility between 1900 and 1920 is unrelated to future downturn intensity. Differential directed migration is a key mechanism to explain my results. Although sons fled distressed cities at similar rates, the sons of richer fathers migrated to locations that had suffered less severe Depression effects. The differences in rates of intergenerational mobility for sons in the most and least Depression-affected cities are comparable to the differences between the United States and Sweden today.

In the second chapter, I ask whether intergenerational economic mobility was higher in the early twentieth century in the United States than today. I combine two historical data sources to estimate mobility between 1915 and 1940. I match fathers from the Iowa State Census of 1915 to

their sons in the 1940 Federal Census, the first state and federal censuses with data on income and years of education. In my sample, I estimate a lower intergenerational elasticity of income than is found in contemporary studies of the United States, suggesting higher levels of income mobility. Mobility measured with relative income ranks also shows higher mobility historically. Intergenerational mobility of education is higher in my sample than in contemporary measures as well. In addition to providing evidence that there was more mobility historically than today, I document comparably high levels of mobility across a variety of measures based on earnings, occupation, and education.

The final chapter details the machine learning approach to record linkage used to construct the samples I analyze in the first two chapters. Thanks to the availability of new historical census sources, economic historians are becoming big data genealogists. Linking individuals over time and between databases has opened up new avenues for research into intergenerational mobility, the long run effects of early life conditions, assimilation, discrimination, and the returns to education. To take advantage of these new research opportunities, scholars need to be able to accurately and efficiently match historical records and produce an unbiased dataset of links for analysis. I detail a standard and transparent census matching technique for constructing linked samples that can be replicated across a variety of cases. The procedure applies insights from machine learning classification and text comparison to record linkage of historical data. My method teaches an algorithm to replicate how a well trained and consistent researcher would create a linked sample across sources. I begin by extracting a subset of possible matches for each record, and then use training data to tune a matching algorithm that attempts to minimize both false positives and false negatives, taking into account the inherent noise in historical records. To make the procedure precise, I trace its application to an example from my own work, linking children from the 1915 Iowa State Census to their adult-selves in the 1940 Federal Census. In addition, I provide guidance on a number of practical questions, including how large the training data needs to be relative to the sample.

1. INTERGENERATIONAL MOBILITY DURING THE GREAT DEPRESSION

1.1 *Introduction*

Do economic downturns expand or contract intergenerational economic mobility? The effects of macroeconomic conditions on the stakes of the lottery of birth are unclear. A significant disruption of the economy may diminish or even render irrelevant inequities of opportunity bestowed by the previous generation, decoupling the fates of children from their parents. Alternatively, poorer families may be less able to endure a downturn, and children who might have climbed the income ladder in normal times—perhaps with more education or by making savvier migration choices—would instead emerge from a crisis no better off than their parents. To answer this question, I estimate the effects on intergenerational mobility of the largest economic cataclysm in American history, the Great Depression. The Depression presents an ideal natural experiment for studying the impact of a downturn on the transmission of economic status for three primary reasons. First, with a quarter of the labor force unemployed, the magnitude of the Depression dwarfs the recessions of the postwar era; if downturns do alter mobility rates, this should be most observable in the largest one. Second, fortunes large and small were destroyed in the 1929 stock market crash as well as in many local real estate crises and bank failures across the United States, creating geographic variation that enables me to compare intergenerational mobility across cities affected unevenly by the Depression. Third, the passage of time allows me to observe the downturn-exposed children as adults in 1940, which is not yet possible for the children of the recent Great Recession.

I measure mobility in two historical datasets that I construct, linking parents before the Depression to their children as young adults in 1940. Historical microdata reporting earnings as well as names and ages—necessary to link records between sources—are exceedingly rare before

1940.[1] I add to the stock of these datasets by digitizing and transcribing the 1918-1919 Bureau of Labor Statistics (BLS) cost of living survey that includes names, ages, addresses, income, and occupation for 12,817 families in 99 cities. I link the parents in the survey ahead to their children in 1940, creating a new dataset of intergenerational earnings. I supplement this data by building a second linked sample of parents and children containing individual occupations and average occupational earnings, drawn from the 1920 and 1940 US Censuses of Population.

To create these intergenerational linked samples, I develop a machine learning approach to record matching across historical microdata.[2] Linking historical data—without unique identification numbers—is difficult and imprecise, relying on demographic information like name, age, and place of birth. Manual linking by a trained researcher yields accurate and comprehensive matches, but at the cost of time and replicability. Algorithmic approaches have been developed in the historical literature, but their rigid rules are often quite inefficient—many records go unmatched—and inaccurate in the face of messy historical data. My technique uses supervised learning to train an algorithm to replicate the process of manually matching individual records across sources. I am thus able to increase the speed, accuracy, and consistency of creating historical linked samples.

I find that economic mobility was lower in cities with more severe downturns during the Depression: both the 1940 earnings and occupations of sons growing up in these cities are more closely linked to their father's outcomes than are those of sons in less negatively affected cities. I measure relative intergenerational mobility in three ways: the elasticity of the son's earnings with respect to the father's earnings; the coefficient from a regression of the son's position in the earnings distribution in 1940 on the father's position in 1920; and the elasticity of the son's occupation score to the father's occupation score.[3] All three measures show similar reductions in

[1] In the United States, there are two such sources: the 1915 Iowa State Census, which includes records for only the state of Iowa (Goldin and Katz, 2000), and published lists of personal income tax payers in New York City in 1923 and 1924 (Marcin, 2014). The 1940 Federal Census is the first census to include data on earnings and completed years of education. The original manuscripts with names and addresses were released in 2012 following the standard 72-year privacy period.

[2] The use of machine learning techniques in the economics literature has increased recently both in traditional prediction tasks (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015) and in causal policy evaluation (Athey, 2015).

[3] I focus on relative intergenerational mobility, estimated via the slope of a regression of the sons' outcomes on their fathers' outcomes, rather than absolute mobility, which indicates the expected position of a child born to a parent at

mobility caused by the Great Depression. The effects are of similar magnitude for all sons in my sample, regardless of whether they were in grade school in 1929, of high school age, or already in the labor force when the Depression hit.[4]

To explore whether the Depression effects on mobility are causal—and not driven by pre-existing city-level differences in mobility that happen to correlate with Depression severity—I perform a parallel analysis on intergenerational mobility from 1900 to 1920. Implicitly, this is a differences-in-differences framework, comparing sons in cities before and after the Depression that experienced Depression downturns of varying magnitudes. For this earlier generation, I find no *ex ante* differences in mobility between cities that later experienced larger or smaller Depression downturns. This result suggests that it is unlikely that low levels of mobility drove local Depression severity, or that mobility and Depression severity are correlated outside the Depression generation. I also show that Depression severity does not predict intergenerational mobility in the late 20th century.

How did the Depression decrease mobility in cities that experienced more severe shocks? Migration is a key mechanism. Local Depression severity drove out-migration, as sons fled distressed cities for better opportunities elsewhere. However, not all sons migrated at the same rates or to the same places. Rather, migration varied by father's earnings: the sons of richer fathers were able to make better migration decisions, moving to cities and regions that had suffered less severe Depression downturns, reducing estimated economic mobility. Formal education plays no role in explaining my estimated effects; while sons from cities with more severe downturns did accumulate more years of schooling than sons from other cities, the Depression did not affect the link between a father's earnings and a son's education. I also find that the Depression increased income inequality in cities, a change that may have made climbing the ladder of economic status more difficult. Finally, I show that local variation in New Deal spending did not affect mobility.

I motivate my empirical analysis with a model of intergenerational mobility based on Solon (2004) and in the spirit of Becker and Tomes (1979), but augmented with migration. Parents invest

a given place in the earnings distribution. In Section 1.4, I document the effects of variation in Depression severity on absolute mobility as well.

[4] I estimate the relative effect of the Great Depression, exploiting variation in severity across cities. I refer to this local, differential cross-city impact of the Depression as the Great Depression effect. However, I am unable to directly assess the aggregate effect of the Depression on intergenerational mobility.

in their children's human capital, and parents with higher income invest more heavily. Parents with higher income also endow their children with larger social networks, enabling them to migrate to regions with higher match-specific returns. Both education and migration could explain why a large macroeconomic downturn with spatial variation could decrease intergenerational mobility, but my empirical results suggest that migration is the primary mechanism.

The estimated effects of the Depression on mobility are economically large. The differences in intergenerational mobility rates between sons in cities with more severe and less severe downturns are comparable to the differences in mobility between the United States and Sweden today, the least and most economically mobile countries in the OECD (Corak, 2013). The differences I find are also comparable to the differences in mobility between Charlotte, NC and Salt Lake City, UT, two American cities that are currently at opposite ends of the mobility spectrum (Chetty, Hendren, Kline, and Saez, 2014). The Depression calcified the mobility ladder for the generation of children unlucky enough to be born in cities with especially severe downturns.

This historical study offers a lens on intergenerational mobility today and in the future. Mobility appears to have remained stable for the last three decades (Chetty, Hendren, Kline, Saez, and Turner, 2014; Lee and Solon, 2009). Has the Great Recession upset that stability? The children of the Great Recession are not yet old enough for researchers to observe their lifetime earnings or occupation choices. Both the Great Recession and Great Depression were periods of large economic and financial market disruption and uncertainty, accompanied by large increases in unemployment. The results of my study, that mobility decreased in cities hardest hit by the Depression, may predict that scholars will find less mobility in the years to come, particularly in those regions that suffered the worst downturns during the Great Recession.

By comparing mobility across cities, I am able to overcome three challenges to measuring historical intergenerational mobility. These potential issues could bias overall mobility estimates, but because they do not vary across the cities in my sample or correlate with local Depression severity—claims for which I will give evidence throughout the paper—they will not determine my estimates of the Depression effect on mobility. First, a pair of historical challenges. I observe only a single year of earnings or occupations for fathers and sons. In addition, any intergenerational matching procedure across different data sources, mine included, will introduce some share of false matches into the sample. Second, the sample of fathers drawn from the BLS survey

is not entirely nationally representative. The BLS targeted married families in 99 urban areas, most of whose male heads worked in industrial occupations, and all families are drawn from the middle of the earnings distribution, missing the very rich and the very poor. Third, I rely on the 1940 Federal Census because it is the first and only national survey with both earnings data and the necessary biographical details to enable intergenerational linking, but there are some concerns with measuring mobility using the 1940 census. Only labor earnings were enumerated in the census, not capital or self-employment income. Earnings and occupations are observed as of 1939; while the economy had recovered somewhat from the Depression and the 1937 Recession, the unemployment rate did not return to pre-1929 levels until at least 1941. The sons in my sample are still relatively young in 1940, and samples with sons who have not yet reached their permanent earnings levels tend to produce artificially high rates of mobility (Corak, 2006; Grawe, 2004; Mazumder, 2015). However, as I document, none of these challenges varies across cities in my sample and will be differenced out when I examine the effects of relative Depression severity across cities.

The paper proceeds as follows. In Section 1.2, I detail the historical data I collect, digitize, and link. I also document the machine learning based census matching procedure I develop to facilitate linking large samples. In Section 1.3, I motivate my analysis with a model of intergenerational mobility in the face of large and spatially varied macroeconomic downturns, allowing for investment in human capital and endogenous geographic mobility. I present my main results in Section 1.4: for the sons growing up in cities with the most severe Depression downturns, intergenerational mobility was greatly reduced. I present evidence that these Depression effects are specific to the 1920 to 1940 period, showing no relationship between local Depression severity and mobility in the generation preceding the Depression (1900 to 1920) or in the current period (1980-82 to 2011-12). In Section 1.5, I explore several possible mechanisms that might drive the main results, showing that differential directed geographic migration best explains the decrease in intergenerational mobility in Depression-affected cities. I also show that the Great Depression increased earnings inequality. I explore the effects of New Deal on mobility in Section 1.6. Section 1.7 concludes.

1.2 *Data and Census Record Linking*

I draw on several historical data sources, including new archival microdata I collected for this project. I have developed a matching algorithm to facilitate linking parents and children across censuses and other data sources, applying insights from the machine learning literature. The algorithm learns the implicit rules that a careful and well-trained researcher uses to match records across historical samples and replicates these decisions for the full dataset, increasing the speed, accuracy, and consistency of the process. The generated matched samples are both large and unbiased. I combine these data with measures of local variation in Great Depression severity based on retail sales from Fishback, Kantor, and Wallis (2003).

1.2.1 *Earnings Data*

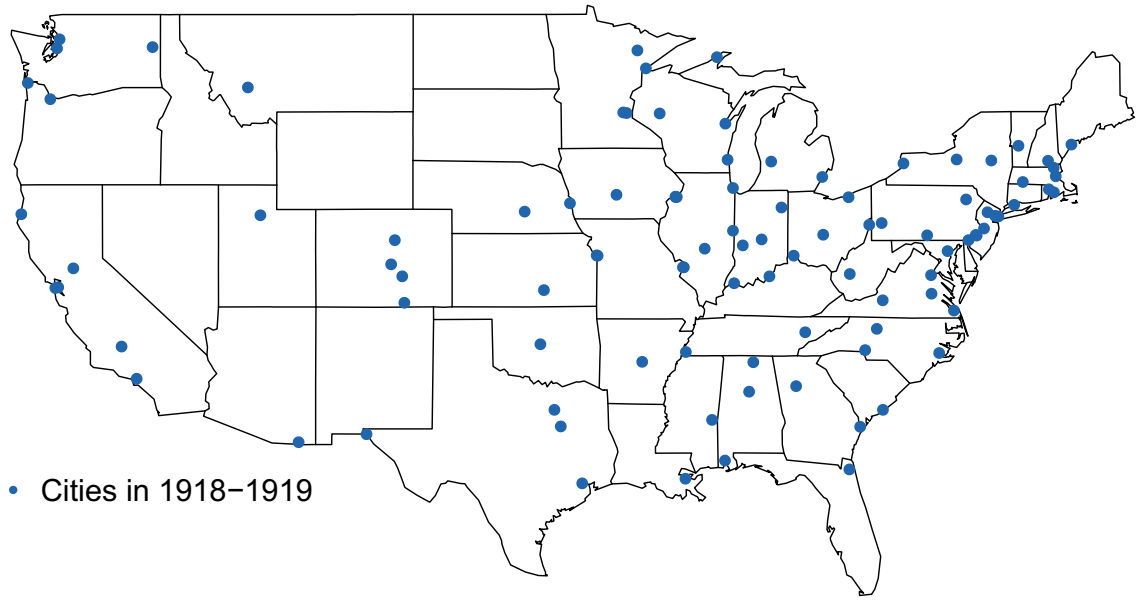
I combine two data sources to measure the earnings of fathers before the Great Depression and their sons after it. For the fathers, the data come from a 1918-1919 Bureau of Labor Statistics (BLS) survey that provides one of the earliest national samples to include both earnings and—crucially for record linkage—names and ages. I digitize and transcribe the names for the complete sample from the original surveys stored at the National Archives in College Park, MD. For the sons, I rely on the complete count 1940 census.

In 1918 and 1919, the BLS conducted one of the first cost of living surveys in the United States.[5] The BLS surveyed 12,817 families in 99 industrial and mining cities between July 31, 1918 and February 28, 1919. Though the primary focus of this survey was estimating the cost of living across urban areas during the First World War, the BLS also collected detailed data on earnings and labor supply.[6] These economic variables are instrumental to my study, as are the names and addresses of the original survey respondents. I have collected and digitized the information from each original BLS survey response deposited at the National Archives and linked the names and

[5] The survey has been used by economic historians, including Dora Costa (1997, 2000, 1999), Shawn Kantor and Price Fishback (1996), Carolyn Moehling (2001, 2005), Martha Olney (1998), and Evan Roberts (2003).

[6] Most of the variables originally collected by the BLS, including earnings and occupation, were transcribed and digitized by the Interuniversity Consortium for Political and Social Research (ICPSR) in the 1980s. Those data are available via the ICPSR dataverse at as ICPSR study 8299. However, the ICPSR files do not contain transcribed names or addresses of the respondents. According to Peter Granda, an associate director of ICPSR, these names may have been transcribed when the survey was first digitized in the early 1980s, but those files were never used for privacy reasons and have since been lost.

Figure 1.1: 99 Industrial Cities included in the BLS Sample, 1918-1919



The cities range in population from New York City (5.6 million in 1920) to Calumet, MI (2,390 in 1920).

addresses to the rest of the survey data.[7] After linking the BLS respondents to their sons in 1940, this detailed data enables me to construct estimates of intergenerational mobility of earnings and occupation.

In Figure 1.1, I map the cities included in the BLS survey. The largest cities in the sample are New York (516 observations), Boston (405), and Chicago (348). However, many smaller urban areas were also included, for example Bisbee, AZ (population 9,205 in 1920 with 80 observations in the BLS sample) and Calumet, MI (population 2,390, 73 observations). The sample is 93% white. The African American sample is concentrated in a few cities, and only Baltimore, New Orleans, and St. Louis have more than 75 sampled black families. I restrict my analysis to white families.[8]

What types of respondents were targeted by the survey? The BLS sampled only married

[7] I traveled to the National Archives to access the original survey responses and photographed the first page of each survey, capturing the names, addresses, and other demographic information of each respondent family. Each survey included an identification number that made linking between my transcriptions and the original ICPSR data exact. The original survey responses are deposited in 93 boxes at the National Archives in RG 257, accession NN-373-183.

[8] In future work, I hope to exploit different datasets, including linked census samples and World War II enlistment surveys, to explore intergenerational mobility among African Americans, especially in light of the high levels of geographic mobility during the Great Migration. Unfortunately, the BLS sample is not well suited to studying African Americans or other racial, ethnic, or religious minorities.

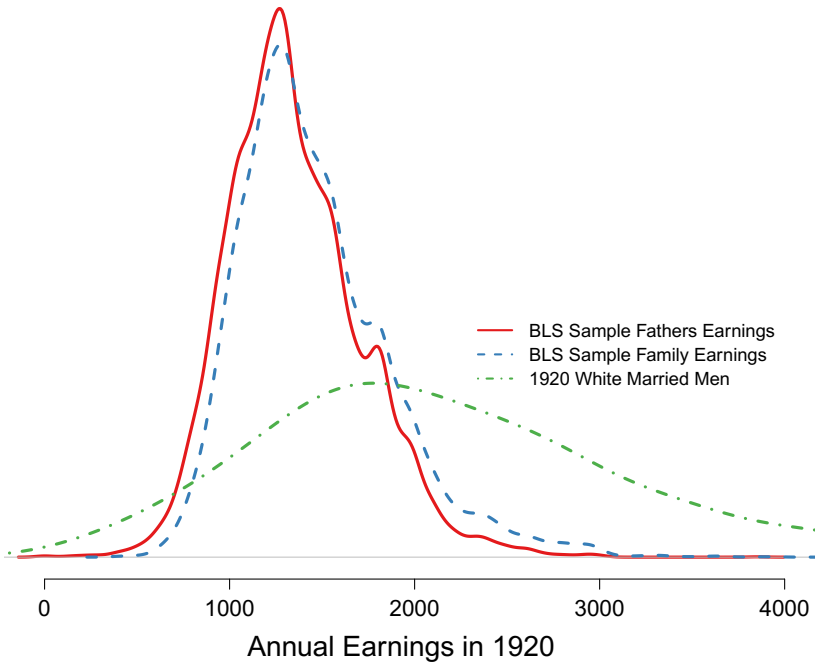
couples with both spouses living in the same house, and all families had to include at least one child. The surveys were conducted in respondents' homes by local female enumerators, and the sample was limited to English speakers who had resided in the "community" for more than one year. The BLS intended to sample only wage earners of any annual income and salaried workers earning less than $2000 per year. Of course, enumerators did not know respondents' incomes before asking, so the sample was selected imperfectly; some people with higher or lower incomes are included. This fact is apparent in the distribution of income graphed in Figure 1.2. The families sampled by the BLS were concentrated between the 30th and 70th percentile of the full 1920 earnings distribution.[9] In 94% of families, the husband worked more than 40 weeks a year; only 2% of wives worked more than 40 weeks, and only 8.5% had positive annual earnings. The richer fathers in the sample often worked for the railroads as machinists, foremen, or carpenters; other rich fathers were foremen in machine shops, foundries, and mines. The poorest fathers were mainly factory and mill laborers.[10]

The BLS survey comprises a representative but slightly younger sample of white middle income families living in large and small urban areas around 1920. I compare the demographics of the BLS sample with an urban sample of the 1920 census, drawn from the IPUMS 1% census extract, in Table 1.1. I restrict the census sample to white families in cities with married heads of household, both spouses present, and at least one child, replicating the BLS demographic sampling frame. The families sampled by the BLS tended to be younger: both the fathers and the mothers sampled were about five years younger than the comparable census sample. The children in the BLS sample are also three to four years younger on average. In terms of family size, however, the two samples are similar. The average family in both samples had 2.5 children. Because I can match only sons and not daughters forward into the future censuses, I focus on them. On average, there are 1.25 sons in each family in the BLS sample and 1.29 in the census

[9] A complete earnings distribution for 1920 does not exist. I estimate the distribution by drawing on the 1940 earnings distribution from the complete 100% sample of the 1940 census, weighting observations by occupation and industry shares from the 1920 census and converting 1940 dollars to 1920 dollars using the CPI. I detail the construction of the 1920 distribution in Appendix A.3. I also compare the families surveyed by the BLS in Davenport, IA and Des Moines, IA to the earnings distributions in those cities according to the 1915 Iowa State Census in Figure A.13. I use the CPI to convert the distribution from 1940 to 1920 rather than mean nominal earnings because I also convert the BLS earnings—collected across 8 months from 1918 to 1919—to 1920 dollars and need a monthly deflator. In practice, the choice of deflators does not change Figure 1.2 significantly.

[10] I present a word cloud illustrating common occupation strings by earnings in Figure A.4.

Figure 1.2: Distribution of Earnings in the BLS Sample



The BLS survey designers intended to sample only families earning less than $2000 per year, but excluding families on relief or charity; 80.8% of the sample earned between $1000 and $2000 annually, 91.8% between $800 and $2000. All earnings are measured in 1920 dollars. The complete earnings distribution for 1920 is calculated using the 1940 earnings distribution, weighted by occupation and industry shares from the 1920 census.

Table 1.1: The BLS Families are Younger but the Same Size as Families in the 1920 Census

	BLS Sample	IPUMS 1920 Sample	
		All Cities	BLS Sample Cities
Father's Age	36.94	42.09	41.89
	(8.52)	(11.78)	(11.54)
Mother's Age	33.32	38.04	37.92
	(7.86)	(11.23)	(11.04)
Number of Children	2.53	2.52	2.53
	(1.50)	(1.64)	(1.64)
Number of Sons	1.25	1.29	1.29
	(1.08)	(1.17)	(1.16)
Age of Oldest Child	9.27	13.42	13.29
	(6.06)	(9.25)	(9.18)
Age of Youngest Child	4.68	8.44	8.34
	(4.64)	(8.63)	(8.49)
Homeowner	0.27	0.40	0.34
	(0.44)	(0.49)	(0.48)
Observations	11946	67864	36416

The families sampled by the BLS are younger than families in the IPUMS 1920 sample, but the sizes of families are similar. They are less likely to own their own home, though this difference is smaller when comparing across the same set of cities. I restrict the census sample to white families in cities with married heads, both spouses present, and at least one child, replicating the BLS demographic sampling frame. The BLS sample is the complete BLS sample less 871 non-white families and 54 records missing from or illegible in the National Archives. The table presents means with standard deviations in parentheses.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample

sample.[11]

I convert earnings in the BLS sample to 1920 dollars to account for changes in the price level during the months of the survey using monthly urban consumer CPI data from the BLS. The respondents in my data were surveyed at different points over an eight month period, from July 1918 to February 1919. Earnings were reported to enumerators relative to the twelve months preceding the survey. These years were a period of rapid price changes, surely part of the impetus for the BLS to collect cost of living data in the first place. In Figure A.5 in the Appendix, I plot monthly inflation in these years.

After transcribing the BLS survey, I match families to the 1920 census to recover full census information on the sons: full name, state of birth, and year of birth. This facilitates matching

[11] Of the families with at least one son, the average number of sons is 1.7, with a maximum of 8, in the BLS sample. In the census sample, the conditional average is 1.7 as well. In Figure A.6, I plot histograms of the father's age in 1920 and the number of sons in each family in the BLS sample.

the sons into the 1940 census. In 1940, I observe annual earnings, weeks and hours worked, and occupation, as well as years of education, for each son in my sample. The earnings data in 1940 is restricted to labor earnings; business and self-employment income, as well as other capital income like farm-owner earnings, are not collected. Given that my sample is drawn exclusively from city families, the missing farmer income data is not relevant. However, some of the sons in my sample may have had business or capital earnings in 1940. Though the census did not record the exact amount, enumerators did ask whether or not the respondent earned $50 or more in non-wage or salary income during the year.[12] Only 12% of the sons in my sample reported such income in 1940, a reasonable share given that their fathers were all wage and salary earners in 1920.[13]

1.2.2 *Occupation Score Data*

In addition to the newly digitized BLS survey, I also create a matched sample of fathers and sons from the IPUMS 1% sample of the 1920 census to the 1940 census.[14] While this additional sample lacks data on fathers' incomes before the Depression because income was not reported in any US Census of Population until 1940, it enables me to calculate occupation score mobility and assess my main results on a much larger sample.[15]

I ensure comparability between the BLS sample and the IPUMS sample in two ways. First, I limit the sample to the sons of married fathers living in the same household in 1920. Second, I focus on sons living in one of the 99 cities sampled by the BLS.

My identification strategy requires that individuals in cities hit with larger Great Depression shocks were no less (or more) economically mobile prior to the Depression. To assess this claim,

[12] In the 1950 census, capital and labor earnings were reported. I impute capital income in 1940 for my sample, using the 1950 data and age, education, occupation codes, industry codes, and geographic location. See Appendix A.4. My main results are all robust to using imputed data for capital income earners.

[13] The earnings in the 1940 census are also top-coded at $5000. However, this restriction is not relevant for me, as only 16 of the sons in my sample report earning $5000 or more. I code earnings as $5000 in 1940 for these 16 sons.

[14] I use the IPUMS sample because it includes coded versions of important covariates—like occupation—which have not been digitized in the complete 1920 sample from Ancestry.com.

[15] Occupation scores are commonly used in historical research with data that lacks income information (Abramitzky, Boustan, and Eriksson, 2013a). The occupation scores are the median earnings of all workers in a given occupation in 1950. In Figure A.7, I show that occupation scores and earnings correlate strongly in my samples in both 1920 and 1940.

I also construct a linked sample beginning with the 1900 IPUMS 6% sample, matching fathers and sons in 1900 ahead to 1920.[16] By observing mobility *before* the Depression, I can test whether or not some locations are inherently less mobile.[17] In addition, I draw on data on recent intergenerational mobility from Chetty, Hendren, Kline, and Saez (2014) to show that mobility in the 1990s and 2000s is neither correlated with Great Depression severity, nor with my measures of mobility from 1920 to 1940.[18]

1.2.3 *Linking Census Microdata with a Machine Learning Approach*

Without a linked intergenerational sample, it is difficult to estimate economic mobility accurately. Constructing historical linked samples, however, requires matching across censuses and other sources without the use of unique identification codes such as tax payer IDs or social security numbers. I have developed an automated process for census matching that increases the speed, accuracy, and consistency of matching historical samples, and eliminates the need for hiring (costly) research assistants.[19] Starting with a small sample of training data with records that I have identified as matches or non-matches, the algorithm learns what features predict matches and then generates large and accurately matched samples.[20]

I begin the linking procedure with a list of sons from the 1920 census. I observe first and last

[16] Technically there is no 6% sample in IPUMS but a 1% sample of 1900 and a 5% sample of 1900, which do not overlap. I use an aggregation of both samples and refer to it throughout as a 6% sample.

[17] Implicitly, this sets up a difference-in-differences framework, where I compare mobility before and after the Depression, across cities with more severe and less severe Depression shocks.

[18] This test rules out a very long-run persistent effect of the Great Depression on mobility through the late 20th century. A natural question is how long the Depression's effect on mobility persisted. As more longitudinal data on mobility in the mid-century are constructed (after future census releases), these answers will come into better focus.

[19] In Feigenbaum (2015), I provide more details on the procedure, as well as tests of its accuracy and efficiency performed with cross validation on a test set. I highlight a comparison of my matching algorithm with methods that have been used previously by economists linking historical records. I show that my method yields a higher match rate overall. In addition, when comparing methods against data manually linked by a trained researcher, I show that my method identifies 90% of the links a researcher makes (efficiency), and that 90% of the links made by the algorithm were made by a researcher (accuracy); these rates are all evaluated on the test set, not the training set used to fit the model. In comparison, the popular soundex-based method will make only 60% of the links made manually, and only 75% of the links made by the soundex-based method would have been made manually. Exact matching fares far worse on efficiency (20% to 30% depending on the age rules) but without a significant decrease in accuracy (86%).

[20] The procedure I use to train my algorithm makes this an application of supervised learning (Kuhn and Johnson, 2013). The idea is straightforward: the researcher uses a labeled dataset and fits various models to predict the known labels. However, a portion of that labeled dataset is held back as the test set. To avoid overfitting, the candidate models are tested on the test set and a variety of out-of-sample prediction metrics can be compared. For more on machine learning and econometrics, see Varian (2014); Athey (2015); Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015).

names, years of birth, and states of birth. I then merge the data to the complete 1940 census, limiting the set of possible matches for any given son in 1920 to be the men in 1940 who meet the following criteria: born in the same state, born in the same year ± 3 years, and have a Jaro-Winkler string distance in first and last names of less than 0.25.[21] This yields a very large set of possible matches; on average each son in 1920 has 240 possible matches. John Malone, born in 1914 in New York, has 1346 possible matches, the maximum in my dataset. He has both a common first name and a common last name, both are similar to other common first and last names, and he was born in the largest state.

I then build a training data set of matches, manually identifying which—if any—of the possible matches are correct links for each son. I build this training data using only 1500 sons in 1920; across a variety of different census linking examples, the algorithm reaches peak precision and accuracy very quickly, with only 15% of the full data (Feigenbaum, 2015). I then use this training data to fit a probit model which generates a score, π_{ij}, indicating how likely it is that a given record i in the original dataset matches with a given record j in the target dataset.[22] However, the scores do not account for the fact that each original record can match only (and at most) once. Thus, I require each match to be (1) the best match for a given son, (2) a sufficiently good match ($\pi_{ij} > \gamma_1$), and (3) a sufficiently better match than the next best match for the son $\frac{\pi_{ij}}{\pi_{ij'}} > \gamma_2 \; \forall j'$.[23] With the model and selection rules trained on a subsample of my data, it is straightforward to construct a full matched sample on the entire data.

[21] The Jaro-Winkler algorithm is used to compute the distance between two strings. The Jaro-Winkler distance between Feigenbaum and Fiegenbaum is 0.03, while the distance between Feigenbaum and Teigenbaum is 0.07, an example of two transcription errors I have seen for my own name in census records that could be possible matches. Meanwhile Feinstein, which has a Jaro-Winkler distance of 0.27 from Feigenbaum, would not be in the set of possible matches because the string distance is too large. The Jaro-Winkler algorithm is popular in computing distances between proper nouns because it penalizes differing letters early in a string more heavily than letters later in the string. Jaro-Winkler has been used for record linkage in Mill (2012) and Nix and Qian (2015). For more on the string comparison algorithm, see Winkler (2006).

[22] I model links with a probit rather than any other machine learning algorithms for two reasons. First, unlike random forests or support vector matrices, probits are a standard empirical tool in the social sciences. Second, while other methods (in particular random forests) outperform the probit on the training dataset, the probit's fit on the testing data (the data that was held back from the initial fitting) is as good or better. I compare my method to other possible machine learning algorithms in Feigenbaum (2015) on the basis of the true positive rate (sensitivity) and the positive predictive value (precision).

[23] The parameters γ_1 and γ_2 are chosen to maximize the efficiency (defined as the true positive rate, the share of true matches identified by the algorithm) and the accuracy (defined as positive predictive value, the share of matches identified that are correct) on the training set via cross validation. The true positive rate corresponds to type I errors (rejecting good matches) and the positive predictive value corresponds to type II errors (accepting bad matches).

My linking procedure is very accurate. Across a variety of different samples and settings, the algorithm has type I and type II error rates of approximately 10% (Feigenbaum, 2015). Thus, the algorithm can identify at least 90% of the matches that would be made in a careful manual linking process by a trained researcher, and at least 90% of the matches made by the algorithm would have been made by the researcher. Still, some links between fathers and sons will be missed, and other links will be inaccurate. How will imperfect matching affect my results? Random mismatches will downwards bias estimated intergenerational mobility. If the outcome of interest were a persistence parameter on its own, one might imagine scaling up the estimated coefficient by the matching error rate.[24] However, in this paper, my main question depends not on the level of the persistence parameter, but rather on the variation of the parameter against with other covariates; in particular, how mobility changes with local Depression severity. If, as I document in the next section, inaccurate matches or biases in the linking procedure do not correlate with these other covariates of interest, then my results are unlikely to be driven by selection bias induced by the matching process.

1.2.4 *Matched Samples*

The match rates for my two main samples—the sons from the BLS survey and the sons of the IPUMS 1920 sample—using my automated linking procedure are quite high, resulting in large samples to analyze intergenerational mobility.

I match 56% of the sons in the BLS sample, in line with or better than past record linkages in the early 20th century.[25] I begin with the 12,871 family observations in the BLS survey. Some

[24] Such a procedure may be overly conservative. Incorrect matches are not likely to be random because they will be to false sons with names (and ages and states of birth) very similar to the correct sons. Given the strong evidence for the socio-economic content of names, both in the contemporary period (Fryer and Levitt, 2004) and the early 20th century (Olivetti and Paserman, 2015) the mismatches may be relatively good proxies for the true outcomes of sons. As I argue in Feigenbaum (2014), a better way to compare persistence parameters estimated with noisy, possibly mismatched historical data with current data may be to increase the noise and mismatches in the modern estimates to reflect biases in the historical data, similar to the approach Romer (1986) takes to industrial production data.

[25] Other record linking attempts in this period include Parman (2011), who links fathers and sons within the Iowa 1915 census with a 50% match rate. Boustan, Kahn, and Rhode (2012) link the 1920 and 1930 censuses to study migration in response to natural disasters and match 24% of individuals. Collins and Wanamaker (2014) match men from the 1910 to 1930 censuses with a 21% match rate. Hilger (2015), aiming to estimate intergenerational mobility of education, links children aged 10 to 17 from the 1930 census to the 1940 census, using complete samples of each; however, as he requires exact and unique matches on first name, last name, year of birth, state of birth, race, and sex, his match rate is only 14%. Mill and Stein (2012) report a range of match rates that vary with the strictness of their matching procedure, ranging from 11% to 34%. The Parman (2011) match rate is comparable to mine because

of the records have been lost in the National Archives and others are not legible, while another 3,277 of the families surveyed did not have a son. I also exclude the 871 non-white fathers. That leaves me with 6,685 fathers linked from the BLS to the 1920 census, with 11,195 sons for which to search.[26] My matching algorithm links 6,269 sons ahead to 1940, accounting for 4,385 fathers, resulting in a 56% match rate.

The match rate for my urban sons sample from IPUMS is also high, and the sample size is much larger. I begin with the 1% sample of the 1920 census and limit it to sons living in urban locations in 1920 with married fathers present in their household to replicate the sampling frame used in the BLS survey. This yields 110,339 sons to search for, with 64,078 unique fathers. I find 51,699 sons using my matching algorithm, a match rate of 46.9%. From this larger sample of urban sons in 1920, I can create a subsample limited to just the cities included in the BLS survey. This subsample includes 45,698 sons in 1920 and I find 20,283 of them in 1940, a match rate of 44.4%.[27]

1.2.5 *Biases in the Linking Procedure*

Which sons do I match into the 1940 census? While my matching procedure is not able to link every son in my sample, I show in this subsection that the matches are not a selected subsample, so my results will not be biased by the data construction.

There are two main reasons sons in 1920 should not be found in 1940: death or emigration. However, mortality risk should be low for my sample, as I observe sons in childhood and search for them in adulthood before middle age.[28] In addition, I search for the sons in 1940, a year before

that sample was linked manually, the standard my machine learning approach replicates. The other rates are lower because they rely on more restrictive linking techniques that are less efficient.

[26] The average father in the full sample has 1.3 sons; conditional on having at least one son, the average father has 1.67 sons. In the analysis that follows, I will cluster at least at the family level to account for multiple sons with the same father. Most clustering will be more conservative (at the city level) when I compare across cities with different Great Depression severity.

[27] The IPUMS match rate is approximately 10 percentage points lower than the BLS match rate. The primary explanation is lower quality transcription of names in the IPUMS sample. The name transcription I undertook in the BLS surveys was done with the explicit goal of accurately capturing the respondent names, while names were just one of many variables transcribed in the IPUMS data collection process. It is possible names were the least important variable collected by IPUMS, relative to demographically and economically vital variables like race, age, sex, and occupation. In addition, the sample frame of the BLS survey—stable community members, neither on public assistance, nor too rich—could create a list of sons who are more likely to be matched successfully.

[28] While under-1 mortality in 1920 was 98.1 per 1000 for white males, childhood mortality was only 2.7 per 1000 for

Pearl Harbor, America's entry into World War II, and any combat fatalities.[29] It is difficult for the sons to move out of the sample because I am matching into the entire 1940 Federal Census, and emigration rates are quite low in this period.

Transcription errors are the most likely cause of an unmatched son from 1920 to 1940. These transcription errors could be caused by the historical census enumerators misrecording a respondent's name or by today's data entry workers incorrectly digitizing the enumerator's original entry. The errors could occur in either the 1920 or the 1940 data. I quantify transcription quality that might make records more difficult to ultimately match in three ways: name string characteristics, enumerator effects, and family effects.

First, I use name commonness and length. More common names are likely easier to transcribe, but at the same time harder to match because common names will have more possible close matches. I measure commonness by counting the number of men in the 1% 1920 IPUMS sample with the same first (last) name. In addition, longer names have more letters, presenting more opportunities for transcription error.

Second, I make use of variation in enumerator quality. The BLS survey was originally recorded by 360 agents working in each of the 99 cities. Each agent enumerated an average of 30 to 40 surveys and worked in 4 to 5 different cities. These agents not only asked the survey respondents questions, but they also filled out the 11-page survey forms by hand. The quality of handwriting is extremely variable from enumerator to enumerator. While some printed in clear block print, most wrote in cursive with varying degrees of legibility, leading to variation in match rates from 30% to 80% between the best and worst enumerators.[30] To assess how im-

white males aged 5 to 14. Mortality rates for white males aged 1 to 4 were slightly higher at 9.8. The low rates for the cohort in my sample persisted from 1920 to 1940 (Linder and Grove, 1947).

[29] Mortality is also a potential problem for any matching procedure: individuals who die between the collection of the initial records and the second census wave should not be matched. With my algorithm, the rates of ghost-matching are very low. For a sample of individuals I know to be deceased by the 1940 census, only 5% are "located" in the 1940 census (Feigenbaum, 2015).

[30] In Figure A.8 of the Appendix, I present the signatures of four enumerators, two with match rates of approximately 80% and two with match rates of only 35%. Though not obvious in their signatures, the enumerators with high match rates tended to write in block print or very clean, simple cursive; the enumerators with very low match rates instead used sweeping, ornamental cursive, making the letters much more difficult to decipher for the researcher. The BLS enumerators are not the only agents whose handwriting might determine the match rates. Both the 1920 and 1940 censuses were, of course, recorded by enumerators, also with large variation in penmanship. However, because the Federal censuses were such large undertakings, very few of my sample records were enumerated by the same census agent, and I do not analyze match rates based on census enumerator.

portant enumerators are to the matching process, I calculate the leave-one-out match rates for each record, which indicates the probability of matching the other people surveyed by that same enumerator.[31]

Third, I make use of the 3,088 sets of brothers included in my sample (including 1,889 pairs). Within families (sets of brothers), I calculate the leave-one-out match rate for each record, quantifying the probability a given son's brothers will be matched.

Which features of name strings predict a successful match from the BLS data into the 1940 census? I present these results in Table 1.2. Sons with more common names, both first and last, are less likely to be matched. Even with information on state of birth and year of birth, there will be far more possible matches for a common name like "James Smith" that—to avoid false positives—the matching algorithm will not identify as a match. Records with longer first names are more likely to be matched, perhaps because this correlates with commonness. I also find that the BLS survey enumerator, likely driven by the clarity of her handwriting, is a strong predictor of which records are ultimately matched; the coefficient in column (3) suggests that shifting from the enumerator with the lowest match rate—Marcia G. Brown (30%)—to the enumerator with the best—Lucille Henry (81%)—increases the probability of matching a son from the BLS sample to the 1940 census by 16 percentage points. In columns (4) and (5), I restrict to the sample of sons with brothers. Sons with a matched brother are 16.9 points more likely to be matched themselves, 15.5 when conditioning on name features and enumerator effects.

The matching rates in my sample are unrelated to any of the important economic variables I consider in my study. In Figure 1.3, I present binned scatter plots comparing match rates with father's age, father's income, local Great Depression severity measured by the decline in retail sales, and finally the interaction of father's income and Depression severity. In all plots, I control for first and last name string characteristics.[32] In none of the cases do the match rates vary systematically across covariates. In regressions corresponding to each sub-figure, I cannot reject that the variable of interest has no effect on the probability of making a match. This suggests that the matching procedure, while imperfect, is not introducing bias into the linked census samples.

[31] In Table A.3, I show that my main results are robust to cutting the data based on enumerator match rates.

[32] Specifically, I include controls for name commonness, length, and enumerator leave-one-out match rates. In the plot of match rates against the interaction term, I directly control for father's income and Depression severity as well.

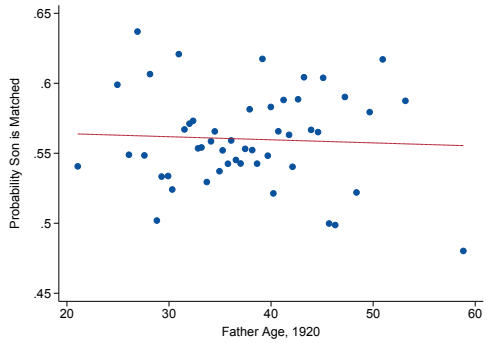
Table 1.2: Probability of Matching a Son from the BLS Survey to the 1940 Federal Census

	(1)	(2)	(3)	(4)	(5)	(6)
Name commonness, first name	-0.014**			-0.005		-0.007
	(0.006)			(0.006)		(0.007)
Name commonness, last name	-0.096**			-0.089**		-0.073*
	(0.038)			(0.038)		(0.041)
String length, first name		0.013***		0.012***		0.010***
		(0.003)		(0.003)		(0.004)
String length, last name		0.006*		0.005*		0.004
		(0.003)		(0.003)		(0.003)
Enumerator Match Rate Leave Out Mean			0.319***	0.317***		0.219***
			(0.047)	(0.048)		(0.054)
Brothers Match Rate Leave Out Mean					0.169***	0.155***
					(0.018)	(0.019)
Match Rate	0.560	0.560	0.560	0.560	0.558	0.558
Observations	11195	11195	11195	11195	7912	7912
Clusters	6371	6371	6371	6371	3088	3088
Adjusted R^2	0.001	0.002	0.005	0.007	0.021	0.024

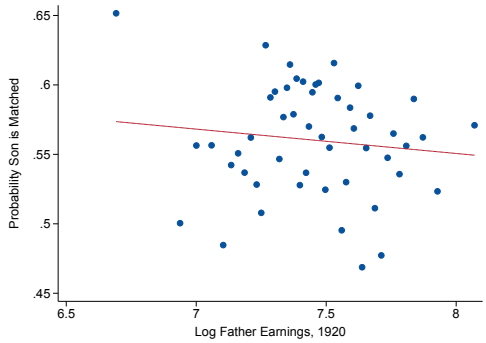
Linear probability model with an indicator variable for a successful match as the outcome. Standard errors are clustered by family in the BLS data. Results are consistent using a probit or logit model as well. Name commonness is measured as the share of 100 men in the 1920 IPUMS sample with the same first or last name. Name length is the number of characters in the first or last name.

Source: BLS Cost of Living Survey 1918-1918

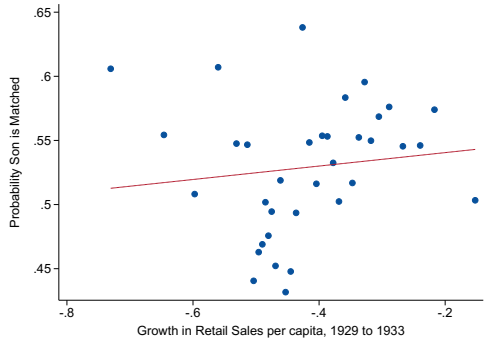
Figure 1.3: Match rates for BLS sample sons into the 1940 census do not correlate with important covariates



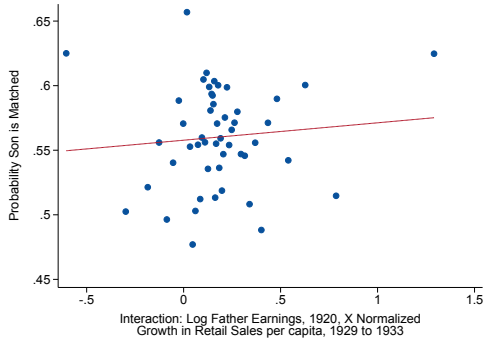
(a) Father's Age in 1918-1919



(b) Log Father's Income in 1918-1919



(c) Decline in retail sales



(d) Interaction

Match rates for sons do not correlate with father's age (a) or income (b). In addition, sons experiencing more severe Great Depression downturns in their city of origin are no more or less likely to be found in the 1940 census (c). Finally, and most importantly for my study, the match rates do not correlate with the interaction of Great Depression severity and father's income (d). All binned scatter plots control for first and last name commonness, length, and enumerator leave-one-out match rates. In the plot in (d), direct controls for father's income and Depression severity are included. The p-values for the main (x-axis) variable coefficients are 0.723, 0.343, 0.470, and 0.450 in separate regressions.

1.2.6 *Great Depression Severity*

To assess the effects of the Great Depression on mobility—the father-son measures just described—I require variation in Great Depression severity across cities. To measure local Depression severity in this paper, I use the decline in retail sales per capita from 1929 to 1933. The Great Depression was felt throughout the country, but there was local variation in its severity. For example, per capita retail sales fell in Davenport, IA by nearly 50%, while the decline in Des Moines, IA was only 25%.

The path of the unemployment rate in the early twentieth century underscores the massive disruption of the Great Depression. Between 1900 and 1930, unemployment exceeded 10% only once (1921). From 1931 to 1940, unemployment never fell below 14%, hitting 25% in 1933 at the nadir of the Depression (Bordo, Goldin, and White, 1998).[33] While national unemployment statistics are available throughout the Great Depression period, local rates, either at the county or city-level, are not. Instead, I use retail sales as an alternative, local measure of Depression severity.

I draw data at the county level from the Census of Retail Sales, originally digitized by Fishback, Horrace, and Kantor (2005). These censuses, taken biennially, allow me to measure retail sales per capita during the Depression era. Retail sales have been used frequently to measure local Depression severity and variation (Fishback, Horrace, and Kantor, 2005; Fishback, Haines, and Kantor, 2007; Fishback, Johnson, and Kantor, 2010); the measures were also some of the only indicators of economic conditions available to New Deal policymakers and administrators during the 1930s (Fishback, Horrace, and Kantor, 2005). As my main measure of Depression downturn, I calculate the log difference of retail sales per capita in 1933, at the nadir, and in 1929, before the Depression.[34] I merge the counties in the retail sales data to the cities in my linked intergenerational mobility samples.

The economic downturn in specific cities in my sample paints a vivid picture of the severity and variation of the Great Depression. In Figure A.1, I map severity in each city in my sample

[33] Real GNP fell by more than one-third during the Depression (Bordo, Goldin, and White, 1998). The Dow Jones lost 80% of its value from 1929 to 1932, and corporate profits were $2.7 billion in the red (Tyack, Lowe, and Hansot, 1984).

[34] Specifically, I define Depression severity as $\log\left(\frac{retailsales_{1933}}{population_{1933}}\right) - \log\left(\frac{retailsales_{1929}}{population_{1929}}\right)$.

according to the change in retail sales per capita. Portland, OR is the median city, with per capita retail sales declining by 0.41 log points from 1929 to 1933. Everett, WA suffered from a local downturn one standard deviation worse than Portland; retail sales fell in Everett by 0.55 log points. The downturn in Manchester, NH was relatively small, one standard deviation better than Portland, and yet per capita retail sales still fell by 0.29 log points.

In Appendix A.2, I explore several determinants of local Great Depression severity and review the literature on the Depression's geographic variation. I find cities specializing in durable manufacturing and in the extraction of raw materials used in construction and manufacturing to have suffered especially large downturns. In addition, the Depression was milder in the South.[35]

1.3 *A Model of Mobility in a Depression*

How do large macroeconomic downturns affect intergenerational mobility? To fix ideas about what factors may drive mobility, I consider a model based on Solon (2004), augmented with endogenous migration and macroeconomic shocks.[36] The model suggests that downturns will tend to lower intergenerational mobility, but could do so through two key channels: human capital and geographic migration.

Consider fathers and sons who live for two periods with each family indexed by i. In the first period, $t-1$, fathers supply labor inelastically and each earn incomes $Y_{i,t-1}$. Fathers can spend income on either consumption, $C_{i,t-1}$, or investment in the child's human capital, $I_{i,t-1}$. The sons work in the second period and have income Y_{it}. Families are liquidity constrained: a father cannot borrow against his son's future earnings in the first period to finance additional consumption or investment. The budget constraint for fathers is:

$$Y_{i,t-1} = C_{i,t-1} + I_{i,t-1} \tag{1.1}$$

Investments in human capital are transformed into the son's human capital, h_{it}, through a simple

[35] The southern United States has especially low mobility, both historically (Olivetti and Paserman, 2015) and today (Chetty, Hendren, Kline, and Saez, 2014). Given the region-wide mild downturn, this could explain a spurious finding of Depression severity leading to more mobility. However, I find the opposite result. Further, I make use of city fixed effects, which subsume regional fixed effects. My main results are also robust to including interactions of father's income and census region indicators, a control that allows for different mobility elasticities in different regions. For these robustness checks, see Figure 1.7.

[36] I abstract away from government policy in the form of taxation or public human capital investment. The Solon model is a modification of the classic Becker and Tomes (1979) model of intergenerational mobility.

log production function:

$$h_{it} = \theta \log(I_{i,t-1}) + e_{it} \tag{1.2}$$

where e_{it} is the child's (genetic or cultural) endowment. The parameter θ governs the efficiency of human capital production. The endowment is inherited from the father:

$$e_{it} = \delta + \lambda e_{i,t-1} + v_{it} \tag{1.3}$$

where the heritability factor is $\lambda \in (0, 1)$ and v_{it} is white-noise error. The earnings function is a standard Mincerian return to human capital, augmented by a son-city match specific return:

$$\log(Y_{it}) = \mu + ph_{it} + m_{it} \tag{1.4}$$

with p as the earnings return to human capital. m_{it} is the son-city match return, representing a good (or bad) economic fit between the child and his city of residence in adulthood. It is difficult for a son to predict whether he will be a good fit for a given city or not. Historically, connections like kin and ethnic networks were vital to job search and placement (Lebergott, 1964; Rosenbloom, 2002). Sons with richer fathers are more likely to have more connections in more cities and thus be able to learn more about city options. As the number of cities about which the son has information increases, the expected match quality will increase: in the limit, a son who could sample the entire distribution of cities will know which city match is the best match for him. Richer sons may also be able to afford to migrate multiple times, trying out more cities while searching for the best match. In addition, sons with higher earning fathers receive greater human capital investment (as the model will predict) and more education might make a migrant better at finding city matches. Motivated by these links between parental income and match quality, I model the match return as a function of parental income in a reduced-form way:

$$m_{it} = \omega \log(Y_{i,t-1}) \tag{1.5}$$

where ω corresponds to the match quality return to parent's earnings.[37] Fathers have a stan-

[37] I model the city match as a function of father's income directly, but it may be that match quality is also a function of human capital investment. Wozniak (2010) shows that the highly educated are more mobile in response to labor market shocks, and Malamud and Wozniak (2014) provide positive causal estimates of the effect of college education on migration. If these factors held during the Depression generation, that suggests that investment by fathers in sons could also increase expected city match quality. As I will show in equation (1.8), optimal investment is increasing in parental income and the model predictions are not affected by modeling city match as a function of income, investment, or both.

dard Cobb-Douglas utility function with preferences for both consumption and child's income, weighted by the father's degree of altruism, α:

$$U_i = (1-\alpha)\log(C_{i,t-1}) + \alpha\log(Y_{it}) \tag{1.6}$$

Plugging the budget constraint in equation (1.1), the human capital production function in equation (1.2), the location match in equation (1.5), and the income function in equation (1.4) into the father's utility, the parent chooses the amount of investment in his child, $I_{i,t-1}$, to maximize:

$$U_i = (1-\alpha)\log(Y_{i,t-1} - I_{i,t-1}) + \alpha\mu + \alpha p\theta\log(I_{i,t-1}) + \alpha pe_{it} + \alpha\omega\log(Y_{i,t-1}) \tag{1.7}$$

The optimal amount of investment in the child is

$$I^*_{i,t-1} = Y_{i,t-1} \times \frac{\alpha p\theta}{1-\alpha(1-p\theta)} \tag{1.8}$$

increasing in parental resources, altruism, and the returns to human capital. Plugging the optimal choice into the human capital and income functions yields a function that suggests a regression of the son's income on the father's income (in logs):

$$\log(Y_{it}) = \mu^* + (p\theta + \omega)\log(Y_{i,t-1}) + pe_{it} \tag{1.9}$$

where $\mu^* = \mu + p\theta\log(\frac{\alpha p\theta}{1-\alpha(1-p\theta)})$. The coefficient on $\log(Y_{i,t-1})$, which I denote as β, is a persistence parameter; $1-\beta$ indicates the amount of mobility. A larger β suggests a stronger link between the income of children and parents and thus less mobility. However, the true population β will not be exactly $p\theta + \omega$ because the "error" term is correlated with $\log(Y_{i,t-1})$. In fact, as pointed out by Solon (2004), this is the regression of an autoregressive variable (income) with an error term (the endowment) that is autoregressive as well. Working out the covariance and variance terms, the estimated β is

$$\hat{\beta} = \frac{p\theta + \omega + \lambda}{1 + (p\theta + \omega)\lambda} \tag{1.10}$$

The four model parameters affect the degree of intergenerational mobility in the following ways: the persistence parameter, β, will be higher, which implies less intergenerational mobility, when (1) inheritance is stronger (λ); (2) the returns to human capital are higher (p); (3) the human capital investment efficiency is greater (θ); and (4) the location-specific match is greater in parent income (ω). These terms help direct my exploration of possible mechanisms that might be affected by the Great Depression and ultimately determine the degree of intergenerational

mobility across my sample. Although it seems unlikely that the Depression changed the persistence of inherited traits, it may have changed the returns to human capital or the productivity of investment in human capital, as well as the ability of parents to help their children make good geographic matches.

The framework of the model also enables me to show that my results are not driven by measurement error. The Great Depression may induce measurement error in income, either father's or son's. With my historical data, I only observe a single year of income.[38] In cities with more severe Depression shocks, that single year of observed income may be a worse proxy for permanent income, the real Y_i in the model above.[39] When I introduce measurement error in either the father's or the son's income into the framework above, I find that the familiar attenuation bias from classical errors in variables applies, even though income in each generation is autoregressive, as is the endowment error term.[40] The probability limit of β is decreasing in the variance of measurement error. If the Depression increased the difficulty of observing accurate income data for affected fathers or sons, this would tend to overstate mobility (understating persistence), the opposite of what I find.

Differential variance in income between cities most and least affected by the Great Depression does not explain my results either. To show this, I consider the Corak (2004) decomposition of the persistence parameter β. Let ρ be the correlation between the log of father's income and the log of son's income. Let σ_{t-1} and σ_t be the standard deviation of log fathers' and log sons' income. Omitting, family subscripts i, it is straightforward to rewrite

$$\beta = \frac{Cov(\log(Y_{t-1}), \log(Y_t))}{Var(\log(Y_{t-1}))} = \rho \frac{\sigma_t}{\sigma_{t-1}}$$

This implies that even if the correlation between father's and son's outcomes is the same across cities, the persistence parameters may still be different. This variation in persistence could be driven solely by changes in the relative standard deviations of income in the two generations. In

[38] Historical single-year income data make estimating precise permanent persistence parameters more difficult. However, to the extent that I am concerned with *differences* in mobility estimates across cities and all the mobility estimates are based on the same imperfect data sources, these issues should not affect my relative results.

[39] Recall that I observe father's income in 1918 or 1919, a decade before the Depression. Though this income measure will not be changed by the Depression, if the Depression significantly changes the father's permanent income, this measure a decade earlier will be less accurate in the cities with more severe downturns.

[40] See Appendix A.5. The attenuation bias is somewhat complicated by the λ term.

particular, if the Great Depression induced additional variation in income for sons in cities where the shocks were especially bad, β will be higher in those cities, implying less mobility. It does not appear that this mechanical relationship explains my results of less mobility in cities hit with more severe downturns, because neither the variance in log earnings of sons nor the variance in log earnings of fathers varies across cities with above or below median Depression shocks.[41]

1.4 *Empirical Strategy and Results*

In this section, I describe my empirical strategy to measure mobility and then present the main results of the paper: that Great Depression severity reduced intergenerational economic mobility. I also argue for the causal nature of my findings by documenting that mobility and future Depression severity by city were not related in the generation before the Depression.

1.4.1 *How to Measure Mobility*

Intergenerational mobility is the relationship between the outcomes in one generation and the outcomes of the following generation. In this paper, I focus on relative mobility in earnings and occupation, estimating persistence parameters that correspond to the slope coefficients in a regression of sons' outcomes on fathers' outcomes.

Let Y_i be the outcome of interest: log earnings, rank, or occupation score. Outcomes for fathers are Y^f, outcomes for sons Y^s. The intergenerational persistence parameter is the β estimated by

$$Y_i^s = \alpha + \beta \cdot Y_i^f + \epsilon_i \tag{1.11}$$

Larger estimates of β mean a tighter link between father and son and thus less mobility. A society with no relationship between parents' and children's outcomes would have complete mobility and a persistence parameter of $\beta = 0$. Conversely, a perfectly immobile society, with relative income preserved across generations, would have a persistence parameter of $\beta = 1$.[42]

[41] The variance in 1940 log earnings of sons is 0.6, both for sons growing up in cities hit with above median Depression downturns and below median downturns. The variance in 1920 log earnings of fathers is 0.1 across cities. The variance is much lower for fathers than sons due to the BLS sample construction.

[42] In theory, β is not bounded between 0 and 1. However, such extremes are unlikely. $\beta < 0$ only if relatively richer parents had relatively poorer children and $\beta > 1$ only when incomes "regress away from the mean" (Mulligan, 1997).

When focusing on income mobility, there are two traditional measures in the literature. The first, known as Intergenerational Elasticity (IGE), is estimated by logging both the father's income and the son's income. The IGE allows for a straightforward interpretation of intergenerational persistence: a 1 percent increase in the father's income is expected to increase the son's income by β. With logged incomes, $1 - \beta$ captures regression to the mean in percentage terms (Mulligan, 1997) and can be used to compute how many generations a rich or poor family would take to converge to the average (Mazumder, 2015). However, the log-log relationship between father's and son's income is just one functional form. In practice, the implied linear relationship between logs of income does not always hold (Chetty, Hendren, Kline, and Saez, 2014).[43] A second specification of mobility considers instead the relative income rank of both fathers and sons, relative to their respective cohorts (Dahl and DeLeire, 2008; Chetty, Hendren, Kline, and Saez, 2014). Thus, Y in equation (1.11) is the income percentile, and β is a rank parameter or the rank-rank coefficient. In my sample, this entails calculating the sons' earnings ranks in the national 1940 earnings distribution and the fathers' earnings ranks in the 1920 earnings distribution.[44]

Occupation score mobility is an alternative measure of economic status that can be used to estimate intergenerational mobility. In my sample, I observe occupations for both generations in the census and in the BLS survey. This allows me to compute occupational score mobility, where the occupation score is the median earnings by occupation. While occupation scores crudely ignore any variation within occupations, they may be better suited to deal with life-cycle bias in mobility estimates. If occupations with higher status or lifetime earnings feature steeper earnings trajectories, observing a younger sample will lead to biased overestimates of mobility. However, such life-cycle induced bias will not be a problem for occupation score based mobility measures. Using occupation scores may also smooth out noise in annual earnings data and uncover a more accurate measure of economic standing when working with historical data.

There are three other potential measures of mobility that I do not focus on in this paper. First,

[43] Another limitation of the log-log method generally is that the IGE parameter is affected by both the intergenerational correlation of income and the variances of income within each generation (Pfeffer and Killewald, 2015).

[44] Mazumder (2015) highlights one reason why researchers might prefer rank parameters to IGE estimates when comparing mobility across cities, counties, or commuting zones: the IGE in a given city signals the regression to the mean in that city, while rank-rank coefficients are based on national income distributions.

the intergenerational correlation of education between generations could also describe mobility.[45] However, I am limited by the data available on completed years of schooling. I observe education only in 1940; thus, I know schooling for only the sons in my sample and the possibly selected subset of fathers found in 1940.[46] Second, wealth mobility would also be a valuable metric, but I do not observe summary measures of wealth for either generation.[47] Third, Mulligan (1997) focuses on consumption as a primary measure of intergenerational mobility. Though the fathers in my sample are drawn from a BLS cost of living survey with incredibly rich detail on consumption, the census data in 1940 are not well suited for measuring the consumption of the sample sons.[48]

1.4.2 *Intergenerational Mobility in the Early Twentieth Century*

How much intergenerational mobility was there in the early twentieth century? Across a variety of mobility measures and in both of my samples, I find more mobility historically than is typically found today. I also document that the relationship between father's and son's outcomes in my sample is linear in log earnings, earnings rank, and log occupation score.[49]

---

[45] I calculate education mobility for the fathers and sons of Iowa from 1915 to 1940 (Feigenbaum, 2014); Hilger (2015) uses education to calculate mobility from 1940 to the present.

[46] In addition, without years of schooling for the 1900 to 1920 sample, I would be unable to calculate educational mobility for the pre-Great Depression period.

[47] I observe housing value (for homeowners) and rent (for renters) in the 1920 and 1940 censuses. It is unclear how well these variables proxy for wealth historically, particularly for some sons who are not household heads in 1940.

[48] In the historical intergenerational mobility literature, two other non-regression alternatives to mobility measurement have been used: Altham statistics (Altham and Ferrie, 2007; Long and Ferrie, 2015, 2007b) and rare surnames (Clark, 2014). Altham statistics are useful when data on income are unavailable. Rather than forcing cardinal comparisons of occupational data, the Altham statistics are measures based on occupational transition tables that indicate how many sons are likely to enter occupations different from their fathers. However, these statistics require slotting occupations into a few categories: often the four groups are farmers, white-collar workers, skilled laborers, and unskilled laborers. For my sample of urban fathers and sons, the farmer category is nearly empty, and accurately assigning many of the other occupations is difficult. Further, as the number of categories in an occupation transition table increases, the measure becomes less informative, as almost all sons will be in different categories from their fathers. I take advantage of the income and detailed occupation measures in my data instead of using transition tables. Rather than link one generation to the next, Clark (2014) uses rare surnames and documents high levels of persistence in socio-economic status. Throughout a variety of periods and countries, Clark (2014) finds that rare surnames with high status in the initial period tend to be relatively overrepresented in later periods in high income or high status professions. While this technique yields interesting results for very long run mobility, it is not well suited to the task of comparing cities across the country in a one-generation period.

[49] On intergenerational mobility today, see Chetty, Hendren, Kline, and Saez (2014); Chetty, Hendren, Kline, Saez, and Turner (2014), Solon (1999), Corak (2006). On mobility in the early twentieth century, see Parman (2011). In my own work, I document higher rates of income mobility in Iowa in the early twentieth century than today (Feigenbaum, 2014).

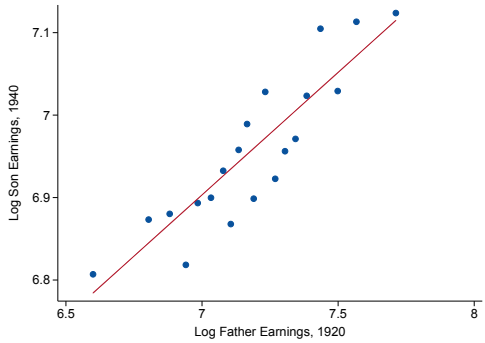
However, the claim that my samples show more mobility historically than prevails today should be tempered somewhat by five limitations of my historical data. First, my sample is built by linking fathers and sons across censuses, possibly imperfectly, which may downwards bias estimated coefficients. Second, I observe income and occupation in only one year for fathers and sons. To the extent that a single year of income or occupation is a noisy signal of permanent status, the estimated persistence parameters will be biased towards zero. Third, the sons in my sample range in age from 20 to 40, but many are in their late 20s. Mobility can appear spuriously high in samples with especially young sons and old fathers (Corak, 2006; Grawe, 2004; Mazumder, 2015). Fourth, my BLS sample was restricted by the original enumerators to collect data only on families in the middle of the income distribution in 1920. Solon (1989) argues that homogeneous samples will also bias the persistence parameters downwards, and Chetty, Hendren, Kline, Saez, and Turner (2014) show how estimated mobility rates vary depending on the income distribution of the parents in the sample. Finally, my sample is restricted to white families in urban areas in 1920 and only father-son links, so any estimates of mobility are specific to that demographic group.

The measures of mobility in my two samples are approximately linear, as documented by Figure 1.4. In each binned scatter plot, I pool fathers into 20 bins, each representing five percent of the data, ranked by the father's 1920 outcome. The linearity is at least partially driven by the original BLS plans to survey only middle-income wage and salary workers in urban areas. However, it enables me to represent relative mobility or persistence with one simple term, the β from equation (1.11).

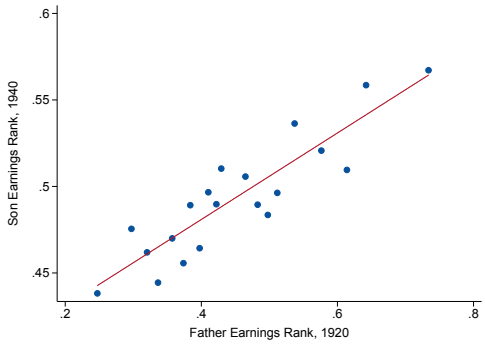
In Table 1.3, I present estimates of several mobility measures across my two main linked samples. In all cases I regress son's outcomes on father's outcomes, controlling for quartic polynomials in father's age and son's age following Lee and Solon (2009). I cluster my standard errors at the family level because brothers in the sample will have the same father with the same data. I use state fixed effects and then city fixed effects, based on the son's city of residence in the BLS sample in 1918-1919, to control for any state-level or city-level common shocks to long run outcomes.

Consider first the IGE estimates in columns (1) and (2), calculated by regressing the son's log earnings in 1940 on the father's log earnings in 1918. The IGE parameter is 0.275 with state fixed

Figure 1.4: Mobility measures are approximately linear in my samples

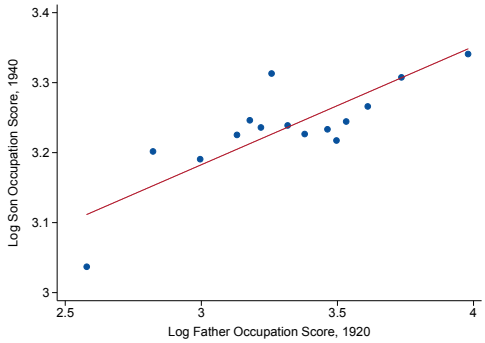


(a) Intergenerational Elasticity of Earnings, BLS Sample



(b) Rank-Rank Parameter, BLS Sample

(c) Occupation Score Elasticity, BLS Sample



(d) Occupation Score Elasticity, IPUMS Sample

Binned scatter plots of mobility in my samples. In all cases, these relationships are approximately linear. I measure mobility in three ways in the BLS sample: (a) intergenerational elasticity (log of son's earnings regressed on log of father's earnings), (b) rank-rank (position in the national earnings distribution in 1940 of the son regressed on the father's earnings position in the national 1920 distribution), and (c) occupation score elasticity (log of the son's occupation score on the log of the father's occupation score). I also measure mobility in the larger IPUMS sample that links the 1920 census to the 1940 census in (d), relying on occupation score elasticities, as earnings and income were not collected in the 1920 census. The binned scatter plot pools fathers into 20 bins, each representing five percent of the data, ranked by the father's 1920 outcome. Points are plotted at the mean for father's and son's outcomes within each bin.

effects or 0.272 with city fixed effects. In both cases, the results imply much greater mobility than is often found in contemporary American data, where common IGE estimates are often greater than 0.4 or 0.45 (Corak, 2006).[50]

As pointed out by Chetty, Hendren, Kline, and Saez (2014), the IGE estimates assume a linear relationship between log incomes across generations. That assumption is not always satisfied in contemporary data (Nybom and Stuhler, 2014a). While the linear approximation fits my data well, I supplement my IGE analysis with mobility estimates based on income ranks. Calculating each son's position in the national income distribution in 1940 is straightforward with the full 1940 census with earnings data. Calculating ranks for the fathers in 1918 is more difficult because the full income distribution is unknown. I calculate ranks in three ways, each yielding similar results. In each case I start with the national 1940 income distribution, based on the 1% IPUMS sample of the 1940 census. First, I adjust for changes in the price level between 1918 and 1940. Second, I adjust for inflation and changing occupation shares by weighting each observation by the share of men with that occupation in the 1920 census. Third, I adjust for inflation and changing occupation and industry shares by weighting each observation by the share of men with that occupation and industry in the 1920 census. In columns (3) and (4) of Table 1.3, I present results using the first method.[51] Similar to the IGE parameters, I find significantly more mobility for my historic sample than is found in current data using the rank-rank method (Chetty, Hendren, Kline, and Saez, 2014; Chetty, Hendren, Kline, Saez, and Turner, 2014; Mazumder, 2015).

Occupation score mobility, presented in the final four columns of Table 1.3, provides an alternative to income mobility measures. Occupation scores are calculated as the median earnings within an occupation code. While the scores eliminate variation in outcomes within an occupation, there are reasons to prefer the scores as a measure of mobility. For one, a single year of

[50] Both Mazumder (2015) and Mazumder (2005), meanwhile, find much higher IGEs, greater than 0.6. Chetty, Hendren, Kline, and Saez (2014) also estimate an IGE of 0.45, but argue that the parameter is highly sensitive to the treatment of children or parents with very small or very large incomes. They also suggest that in contemporary data the conditional relationship between parent's log income and child's log income is non-linear. However, for the historical BLS sample I consider in this paper—a sample drawn from a set of fathers with a condensed distribution of income—I cannot reject linearity, as seen in Figure 1.4 (a).

[51] See Appendix A.3 for more detail on all three methods of calculating the father's earnings rank, including a plot of the empirical cumulative distribution functions.

Table 1.3: Estimated Rates of Intergenerational Mobility, 1920-1940

	IGE		Rank-Rank		OccScore (BLS)		OccScore (IPUMS)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Father Earnings 1920	0.275***	0.272***						
	(0.042)	(0.045)						
Father Earnings Rank 1920			0.217***	0.214***				
			(0.036)	(0.038)				
Log Father Occupation Score 1920					0.202***	0.215***	0.154***	0.144***
					(0.041)	(0.042)	(0.011)	(0.011)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4730	4730	4952	4952	4819	4819	13216	13216
Clusters	3572	3572	3708	3708	3437	3437	10509	10509
Adjusted R^2	0.180	0.187	0.150	0.152	0.048	0.056	0.072	0.076

All standard errors are clustered at the family level to account for multiple observations of the same father across brothers. Columns (1) and (2) give the IGE for the BLS sample where the dependent variable is log son earnings in 1940. Columns (3) and (4) present rank-rank mobility estimates following Dahl and DeLeire (2008) and Chetty, Hendren, Kline, and Saez (2014), where the dependent variable is son's earnings rank in 1940. Columns (5) and (6) present occupation score mobility in the BLS sample, where the dependent variable is the log son occupation score in 1940. Columns (7) and (8) present occupation score mobility in the IPUMS sample, where the dependent variable is the log son occupation score in 1940. All regressions include controls for quartics in father's age (in 1920) and son's age (in 1940). The odd columns add state fixed effects, based on the state where the father and son lived in 1920. The even columns substitute city fixed effects for state fixed effects, also based on 1920 residence. Sons with only capital or self-employment earnings, which were not recorded in the 1940 census, are excluded from both the IGE and rank-rank analysis. In addition, sons with no labor earnings in 1940 are excluded from the IGE analysis, but are included in the rank-rank analysis.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample; 1940 Complete Count Census.

occupation is likely a much less noisy proxy for lifetime average earnings (or for socio-economic class or status) than a single year of income. Further, occupations should suffer from a smaller life-cycle bias than contemporaneous income, a complication that is known to drive down income mobility estimates when sons are observed early in their working lives (Corak, 2006). Finally, occupations may be generally more accurate measures of socio-economic status than income. More vital to my study, occupation scores are necessary to estimate mobility in the IPUMS matched sample as the 1920 census lacks income data. I present occupation score elasticities in columns (5) through (8), regressing the log of son's occupation score in 1940 on the father's occupation score in 1920, for the BLS sample and then the IPUMS sample. Occupation score persistence in the BLS sample is slightly larger than occupation score mobility in the IPUMS sample, but both rates are comparable.

The father and son ages at which income or occupations are observed often complicate the intergenerational mobility estimates. For example, many early IGE studies have been criticized for observing sons who were too young and fathers who were too old, which biases parameter estimates down and implies spuriously large levels of mobility (Corak, 2006; Grawe, 2004; Mazumder, 2015). I show in Appendix Figure A.9 that my mobility estimates are all relatively stable across sons ages.

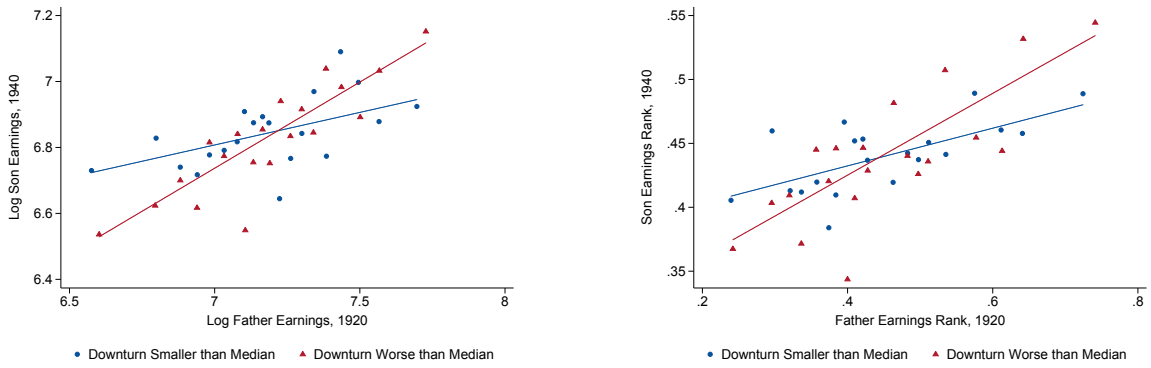
The results in Table 1.3 suggest that mobility was higher in the early twentieth century than it is today; in the next section, I explore what role the Great Depression played in driving this result.

1.4.3 The Great Depression Decreased Economic Mobility

At the nadir of the Depression in 1933, more than one-quarter of the American work force was unemployed, and others saw their hours or wages fall and their life savings and home values erode. Did this disruption break the economic links between fathers and sons, driving mobility higher in cities that experienced more severe downturns? Or did the Depression underscore differences in outcomes for children in different parts of the income distribution, reducing upward mobility for sons with poorer parents more than it reduced mobility for middle class sons?

To summarize the results in this section, I graph two binned scatter plots of father's earnings

Figure 1.5: Intergenerational mobility of earnings is lower in cities with more severe Great Depression downturns



(a) Intergenerational Elasticity of Earnings

(b) Rank-Rank Mobility

The steeper slope between the father's log earnings and son's log earnings in cities with downturns more severe than the median implies less mobility. Cities are split into groups with above or below median Great Depression severity, as measured by the decline in per capita retail sales between 1929 and 1933. The binned scatter plot pools fathers into 20 bins, each representing five percent of the data. Points are plotted at the mean for father's and son's log earnings within each bin.

against son's earnings in my matched BLS sample in Figure 1.5. In (a), I plot log earnings, and in (b), earnings rank. Each point represents one-twentieth of the data in each category, plotted at the mean for father's and son's outcomes within the bin; these binned scatter plots present the raw earnings data, without controlling for father or son ages or any fixed effects. The relationship between father's and son's earnings is steeper in cities where the downturn was larger than the median.[52] A steeper curve implies a stronger link between father's and son's outcomes and thus less mobility: the sons growing up in cities where the Great Depression was worse have much less mobility than sons in cities with relatively mild downturns.

Figure 1.5 also indicates a surprising fact about the effect of the Depression on absolute mobility, defined by Chetty, Hendren, Kline, and Saez (2014) as the expected adult outcome for a son born to a father at a given rank. For both the IGE and the rank-rank, Figure 1.5 shows that the best fit lines overlap around the middle of father's distribution, suggesting that a son born to a father in the middle of the distribution has the same expected earnings in 1940 whether he grew up in a city hit by a severe or mild downturn. Absolute mobility, however, does respond

[52] Cutting Depression severity at the median produces a convenient graphical presentation, but as I show in Table 1.4, the results hold with a more continuous measure of Depression severity.

to the Great Depression for the sons of both richer and poorer fathers. The typical son in the bottom half of the father's earnings distribution did worse in cities with more severe downturns, but absolute mobility increases for sons from the top half of the distribution. The best fit lines in Figure 1.5 provide a simple measure of absolute mobility. For example, a son born to a father at the 25th percentile of earnings could expect to be in the 41st percentile in 1940 if he grew up in a city with a downturn less severe than the median, but in the 38th percentile in a city with a more severe Depression. For sons born at the 75th percentile, the expected rankings are 48th percentile from the less severe cities and 54th percentile from the more severe cities. In Section 1.5, I will explore why it might be the case that a locally severe downturn is so harmful to the sons of poorer fathers and beneficial to the sons of richer fathers.

The graphical results in Figure 1.5 are only illustrative, and thus I turn to a fuller regression analysis, controlling for covariates at both the father-son and city-level. To determine the direction of the Great Depression effect on mobility, I run regressions of the form:

$$Y_{i,son} = \beta_0 + \beta_1 \times Y_{i,father} + \beta_2 \times Y_{i,father} \times GD_{city} + \gamma_{city} + \epsilon_i \tag{1.12}$$

where $Y_{i,son}$ is the son's outcome in 1940 and $Y_{i,father}$ is the father's outcome in the 1918-19 BLS sample or the 1920 IPUMS sample. The outcomes are either log earnings, position in the earnings distribution, or occupation score. GD_{city} is the severity of the Great Depression in the son's city of residence in childhood. I cluster standard errors at the city level, to reflect that each son growing up in a city is subject to common city-level shocks and the same observed city-level Depression severity measure.

Two different parameterizations of severity are used, both based on my underlying measure of the decline in per capita retail sales between 1929 and 1933. In the first, I normalize the sales growth, subtracting the mean in my sample and dividing by the standard deviation. In the second, I generate an indicator variable that takes a one in cities with a downturn worse than median and a zero in cities with a downturn more mild than median. In both cases, the β_1 parameter can be easily interpreted. When the severity is normalized, β_1 is the persistence parameter in a city with a mean Great Depression downturn; when severity is measured with an indicator variable, β_1 is the persistence in a city with a more mild than median downturn. The key variable of interest, however, is β_2, the interaction term. When severity is normalized, β_2 will

imply the addition or reduction in persistence with a one standard deviation worse downturn. This is the difference between the downturn in Portland, OR (retail sales fell by 0.41 log points from 1929 to 1933) and Everett, WA (retail sales fell by 0.55 log points). When severity is an indicator for worse than median local conditions, β_2 will imply the change in persistence if the downturn was worse than median.

I find that in cities with more severe Depression downturns, intergenerational mobility is lower, as presented in Tables 1.4 and 1.5.[53] In the first table, I examine mobility measured by the IGE and the rank-rank coefficient. Consider the estimated IGE parameter: it is 0.280 in column (2) with city fixed effects. A city with a downturn one standard deviation worse than average, however, is expected to have a persistence parameter of 0.388, indicating significantly less mobility. In columns (3) and (4), I use the coarser parameterization of Depression severity, identifying cities with downturns worse than the median city, Portland, OR. The point estimates suggest similarly dramatic effects of a worse than median Depression on mobility. For cities hit with severe downturns, the predicted persistence is 0.184 to 0.197 points higher, roughly the difference between the United States and Sweden in contemporary IGE estimates (Corak, 2013). The rank-rank results in columns (5) through (8) depict similarly dramatic effects of the Depression on mobility.[54] Today, the difference in rank-rank mobility between one of the least mobile cities in America (Charlotte, NC) and one of the most (Salt Lake City, UT) is 0.133 (Chetty, Hendren, Kline, and Saez, 2014). This difference is less than the difference in rank-rank mobility between a city hit with a one standard deviation more severe local Depression and a one standard deviation more mild shock, or the difference between a city hit with an above median severity downturn versus not.[55]

[53] In both tables, the odd columns include state fixed effects and direct controls for city-level Great Depression severity, while the even columns replicate the previous column with the addition of city fixed effects.

[54] The samples included vary slightly between the total number of sons matched from 1920 to 1940, as well as between the IGE and the rank measures of mobility. I drop all sons with $0 in earnings in the 1940 census from the IGE estimates because I cannot log 0. In the rank-rank estimates, sons with no earnings are included because I am able to calculate their position in the earnings distribution. However, in both samples I exclude the sons with no reported labor earnings in 1940 but who do report capital or self-employment earnings, as their ranking, based on no earnings, would be very misleading. In Appendix A.4, I show that my results are robust to imputing capital income for sons with no labor earnings, using the 1950 census. Finally, I drop 516 sons with illegible or mistranscribed earnings in the 1940 census.

[55] In Table 1.4, I measure earnings without adjusting for differences in local prices. In Table A.1, I show that my main results are robust to accounting for variation in local price levels. I use median city rents in 1920 to calculate fathers' real earnings and median county rents in 1940 to calculate sons' real earnings.

Table 1.4: Great Depression Severity Decreases Intergenerational Mobility: IGE and Rank-Rank Measures

	IGE				Rank-Rank			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Father Earnings 1920	0.285***	0.280***	0.190***	0.181***				
	(0.041)	(0.043)	(0.054)	(0.056)				
Log Father Earnings 1920 X GD Normalized Severity	0.087**	0.108**						
	(0.039)	(0.042)						
Log Father Earnings 1920 X GD Above Median Severity			0.184**	0.197**				
			(0.081)	(0.082)				
Father Earnings Rank 1920					0.219***	0.213***	0.131***	0.124**
					(0.031)	(0.033)	(0.047)	(0.051)
Father Earnings Rank 1920 X GD Normalized Severity					0.078**	0.100***		
					(0.033)	(0.036)		
Father Earnings Rank 1920 X GD Above Median Severity							0.167**	0.175***
							(0.065)	(0.066)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4730	4730	4730	4730	4952	4952	4952	4952
Clusters	99	99	99	99	99	99	99	99
Adjusted R^2	0.181	0.188	0.181	0.188	0.151	0.153	0.151	0.153

Estimates of intergenerational mobility based on a linked sample from the BLS survey of urban families in 1918-1919 to the 1940 Federal census. Each column is a regression of the son's outcome in 1940 on the father's corresponding outcome in 1918-1919, a measure of Great Depression severity in the city of residence in 1918-1919, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. In the odd columns, I include state fixed effects and direct controls for Great Depression severity (normalized in columns 1 and 5, above or below median in columns 3 and 7) but omit the point estimates from the table. In the even columns, these controls are absorbed by the city fixed effects. All fixed effects are based on the city of residence in 1918-1919. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. IGE is the intergenerational elasticity of income, and the dependent variable is log son earnings in 1940. Rank-rank mobility compares the son's position in the earnings distribution in 1940 to the father's position in 1918-1919 and the dependent variable is son earnings rank in 1940.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample; 1940 Complete Count Census.

Based on results from both my BLS sample and the larger IPUMS-based sample, the Great Depression also decreased occupational mobility. The correlation between occupation scores in two generations is a coarser measure of economic status, as it varies only across occupations and not within them. In columns (1) through (4) of Table 1.5, I measure occupation score mobility in the sample of fathers and sons matched between the BLS survey and the 1940 census. Though the results are somewhat less precise, I find similarly that cities with more severe Great Depression downturns have much less intergenerational mobility. Turning to my larger sample of fathers and sons living in the BLS cities in 1920, based on the IPUMS 1920 sample, in columns (5) through (8) of Table 1.5, I find large Great Depression effects on mobility as well.[56]

Comparing the Depression effects to the overall persistence parameter illustrates the magnitude of the effects. In columns (1) and (2), as well as columns (5) and (6), I show that the intergenerational occupation score elasticity is roughly one-third larger in a city with a one standard deviation worse Depression decline. Similarly, when I measure severity as worse than the median downturn or not, as in columns (3) and (4) and columns (7) and (8), I find the occupation score elasticity increases by almost 50 percent.

The Great Depression effects on mobility do not vary across age ranges, as I demonstrate in Figure 1.6. Here, I replicate the specification from column (2) of Table 1.4 and column (6) of Table 1.5, but allow the Depression interaction to vary by son's age in 1940. In the BLS sample, the point estimates are consistently between 0.101 and 0.125, with 95% confidence intervals from roughly 0.014 to 0.212. Though the confidence intervals are slightly smaller in the second panel, based on data from the IPUMS 1920 to 1940 sample, they do overlap with zero for several of the son's ages. Nonetheless, the point estimates are similarly stable.

Although there are five key limitations of my data that make it difficult to compare overall historical mobility rates to estimates for recent decades, none of these drawbacks will bias my estimates of the Depression effects on mobility. First, as I documented in Section 1.2.5, imperfections in the matching procedure do not vary systematically across families or cities in ways that could affect the estimated Depression effect. In addition, I show in Table A.3 that my results are

[56] The Depression effect on occupational mobility is not particular to the cities sampled by the BLS. Using all urban sons in the 1920 IPUMS census sample matched ahead to 1940, I show in Table A.2 similar declines in mobility in cities with more severe Depression downturns.

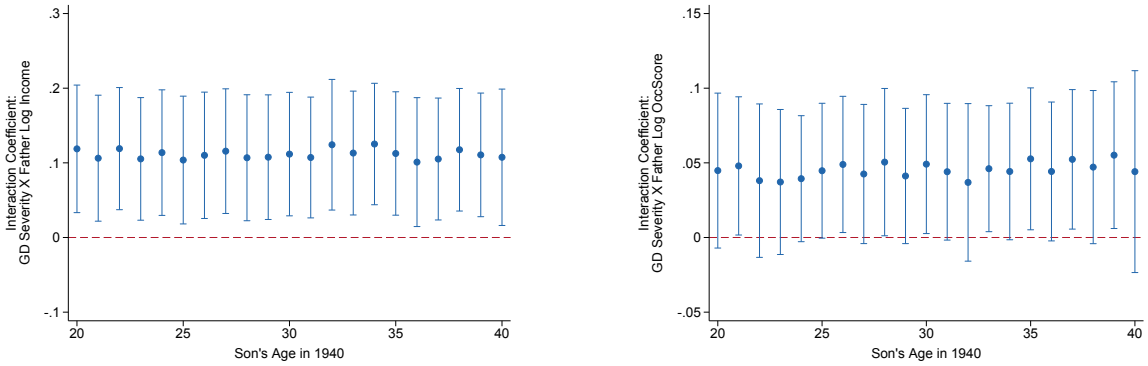
Table 1.5: Great Depression Severity Decreases Intergenerational Mobility: Occupation Score Elasticity Measures

	Son OccScore (BLS)				Son OccScore (IPUMS)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Father Occupation Score 1920	0.182***	0.198***	0.119**	0.131***	0.161***	0.157***	0.128***	0.125***
	(0.041)	(0.043)	(0.052)	(0.047)	(0.013)	(0.014)	(0.017)	(0.017)
Log Father Occupation Score 1920 X GD Normalized Severity	0.071**	0.067*			0.050**	0.045*		
	(0.029)	(0.034)			(0.020)	(0.023)		
Log Father Occupation Score 1920 X GD Above Median Severity			0.124	0.128			0.067**	0.055*
			(0.080)	(0.079)			(0.028)	(0.028)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4819	4819	4819	4819	11555	11555	11555	11555
Clusters	99	99	99	99	92	92	92	92
Adjusted R^2	0.049	0.057	0.048	0.057	0.075	0.081	0.079	0.081

Estimates of intergenerational mobility in columns (1) through (4) are based on a linked sample from the BLS survey of urban families in 1918-1919 to the 1940 Federal census. Mobility estimates in columns (5) through (8) are based on a linked sample from the IPUMS 1% sample of the 1920 census to the full 1940 census. Each column is a regression of the son's log occupation score in 1940 on the father's log occupation score in 1918-1919 or 1920, a measure of Great Depression severity in the city of residence in 1918-1919 or 1920, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. In the odd columns, I include state fixed effects and direct controls for Great Depression severity (normalized in columns 1 and 5, above or below median in columns 3 and 7) but omit the point estimates from the table. In the even columns, these controls are absorbed by the city fixed effects. All fixed effects are based on the city of residence in 1918-1919 or 1920. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. Occupation scores are calculated as the national median income for men in the occupation.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample; 1940 Complete Count Census.

Figure 1.6: The effect of Great Depression severity on intergenerational mobility is stable across sons by age



(a) BLS Sample **(b)** IPUMS 1920-1940 Sample

In the BLS sample, the effect varies between 0.11 and 0.14, roughly half the estimated intergenerational mobility elasticity in a city with an average Great Depression downturn. This implies that sons in a city with a local Depression one standard deviation more severe than average had mobility rates of 0.37 to 0.41. In the IPUMS sample, I calculate mobility rates using occupation scores and find similarly stable effects between 0.04 and 0.06, roughly one-third the overall estimated occupation score elasticity. The IPUMS results are slightly less precise, likely reflecting the noise in measuring occupations and assigning occupation scores. Great Depression severity is measured with normalized retail sales growth per capita from 1929 to 1933, where higher severity implies less growth.

robust to various sample restrictions based on enumerator quality, one of the stronger predictors of linking records from 1920 to 1940. Second, I only observe income or occupations in one year for both the fathers and sons in my sample. Overall, this might downwards bias the IGE or rank-rank parameters, suggesting more mobility than there actually was. However, I only observe one year of income for all fathers and sons in my data. This downward bias will affect mobility for all cities, whether the Depression downturn was severe or not. Third, the ages of my sons are somewhat younger relative to other mobility samples, but this does not drive the differential effects of the Depression, as I show in Figure 1.6. Fourth, while the BLS sample was restricted to the middle of the earnings distribution, these restrictions were consistent across cities. And finally, while my conclusions are necessarily limited to the demographic group I observe in the BLS—white families living in cities—I argue that this sample is particularly relevant for understanding the effects of the Great Depression. 1920 marked the first year America was a majority urban nation.

1.4.4 *Falsification: Cities with Severe Downturns Do Not Always Have Lower Mobility*

In the previous section, I showed that the intergenerational links between fathers and sons were stronger in cities that suffered larger Great Depression declines. However, Depression shocks are not randomly assigned throughout the country. In this section, I turn to two additional data sources on mobility—a linked sample from 1900 to 1920, as well as data on intergenerational mobility in the late 20th century—to show that Depression severity does not predict mobility for either the generation before the Depression or for the contemporary period.

By studying the generation before the Great Depression, I can uncover city variation in mobility that is unrelated to large macroeconomic downturns. If mobility in this period were related, either positively or negatively, with Depression shocks, that would suggest that the finding presented in the main section does not reveal the effects of the Depression on mobility. One concern would be that places with especially low levels of mobility in all periods would be more likely to be hit by Great Depression shocks. This could be because low mobility causes regional severity in the downturn or simply that the two variables are correlated. For example, certain industries, which may be regionally concentrated, could reduce economic mobility and could have been particularly susceptible to employment loss in the Depression. Implicitly, this is a differences-in-differences framework, comparing cities before and after the Depression with larger and smaller Depression downturns. The identifying assumption, as in a diff-and-diff, is that no other determinants of mobility are changing over this period differentially across cities in a way that is correlated with the downturn.[57]

To estimate mobility before the Depression, I create an additional matched sample that links fathers from the 1900 census to their sons in 1920. I begin with the IPUMS 1900 6% sample, limiting to only families living in one of the cities in the BLS survey with both a father and at least one son present in the household. There are 70,561 such families, with a total of 125,078 sons. I then search for these sons in the complete 1920 census, following the linking procedure used throughout this paper. I locate 58,600 sons for a match rate of 46.9%. However, neither

[57] New Deal spending might be one obvious candidate. However, as I document in Section 1.6, particularly in Table A.5, the New Deal had no significant effects on mobility. If anything, certain forms of New Deal aid might be expected to increase mobility and, given that the spending was targeted at cities with the most severe downturns, this would bias me against finding the Depression and mobility effects I observe in Table 1.3.

the 1900 nor the 1920 census includes income or earnings information. Instead, I make use of occupation scores as a measure of economic status.[58] In the previous section, I found that the generation coming of age during and after the Great Depression had less intergenerational mobility in cities hit by more severe downturns. Does this same pattern emerge in the previous generation?

I present the "effects" of the Great Depression on mobility between 1900 and 1920 in Table 1.6. Mobility is generally high in this sample, which I document in the first row of the table. The estimated elasticity of a son's occupation score in 1920 with respect to his father's occupation score in 1900 ranges from 0.110 to 0.115, lower than the corresponding estimates in Table 1.3.[59] However, mobility is no higher or lower for sons born in cities that would, thirty years later, experience more severe contractions during the Great Depression. In the third and fourth columns, I include an interaction of Great Depression severity and the log of the father's occupation score; in the fifth and sixth columns, the interaction is between an indicator for a worse than median downturn and the log of the father's occupation score. But in all cases, with or without city fixed effects, the interaction coefficients are small and statistically insignificant. Further, the corresponding estimated actual Great Depression effects from Table 1.4 for both the BLS sample of sons or the IPUMS sample are well above the confidence intervals for each of the interaction terms in Table 1.6.

Table 1.6 suggests that there are no pre-treatment differences in economic mobility rates across cities with varying levels of Depression downturns. This result buttresses my main claim that the Depression drove lower levels of mobility.

I also find that the Great Depression does not have a lasting effect on mobility in the cur-

[58] For the fathers in 1900, I use the occupation scores based on occupation codes created by IPUMS. However, for the sons in 1920, standardized occupation codes are not available in the complete count data. Instead, I draw on the raw occupation strings entered by census enumerators and transcribed by Ancestry.com and manually link these strings to occupation codes and strings, basing my links on the occupation coding in the 1% IPUMS sample of the 1920 census. Due to missing or unclear occupation strings, I only observe an occupation score for 42,393 matched father and son pairs, 72% of the matched pairs. I use this sample in Table 1.6.

[59] It is not surprising that there was more occupational mobility from 1900 to 1920 than I found between 1920 and 1940. Long and Ferrie (2007b) show higher levels of mobility for a linked sample of fathers and sons between 1860 and 1880 than between 1880 and 1900, suggesting that mobility rates decreased in the nineteenth and early twentieth century. The contemporary pattern of stable mobility rates documented by Lee and Solon (2009) and Chetty, Hendren, Kline, Saez, and Turner (2014) does not begin until the 1970s.

Table 1.6: Great Depression Severity Does Not Affect Intergenerational Mobility 1900 to 1920

	Log Son Occupation Score					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Father Occupation Score 1900	0.1117***	0.1101***	0.1117***	0.1101***	0.1153***	0.1144***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Log Father Occupation Score 1900 X GD Normalized Severity			0.0004	-0.0002		
			(0.008)	(0.008)		
Log Father Occupation Score 1900 X GD Median Severity					-0.0077	-0.0085
					(0.017)	(0.017)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes
Observations	42,393	42,393	42,393	42,393	42,393	42,393
Clusters	89	89	89	89	89	89
Adjusted R^2	0.028	0.030	0.028	0.029	0.029	0.029

Intergenerational mobility estimated on a matched sample of fathers from the 1900 census to sons in the 1920 census. The outcome variable throughout is the son's logged occupation score in 1920. The sample consists of sons in the IPUMS 6% sample of the 1900 census living in a city later surveyed by the BLS. Standard errors are clustered by city. In the odd columns, I include state fixed effects based on the state of residence in 1900 and a direct control for Depression severity. In the even columns, I add city fixed effects which absorb both the state fixed effect and the Depression severity measure.

Source: IPUMS 1900 6% Census Sample; 1920 Complete Count Census; Census of Retail Sales

rent period.[60] Drawing on county-level data on mobility from Chetty, Hendren, Kline, Saez, and Turner (2014); Chetty, Hendren, Kline, and Saez (2014), which measures mobility for the cohort of children born between 1980 and 1982 and their parents, I regress mobility against Great Depression severity. If the Depression did influence contemporary mobility, that might prompt concern that the observed severity effect on mobility in the previous section was driven by other fixed local factors that determined both mobility and severity. However, in Table 1.7, I show that there is no relationship between Great Depression severity and rates of economic mobility today, measured either by the slope of the rank-rank regressions or the expected adult outcomes for children born in the 25th percentile.[61] These results further strengthen my claim that the cities shocked by the Great Depression do not have fundamentally different rates of mobility from cities with milder downturns.[62]

[60] I describe this exercise in full detail in Appendix A.6.

[61] I merge my county-level measures of Great Depression severity to the county level measures of recent mobility. Cities comprised of multiple counties like New York and St. Louis enter the sample multiple times.

[62] The results also rule out very long run persistence of the Great Depression on mobility many generations later.

Table 1.7: Great Depression Severity Does Not Affect Intergenerational Mobility Today

	Rank-Rank Slope		Rank-Rank 25th Percentile	
	(1)	(2)	(3)	(4)
Great Depression Severity	-0.005	0.013	0.761	0.491
	(0.011)	(0.011)	(0.540)	(0.567)
Geographic Controls	No	Yes	No	Yes
Observations	107	107	107	107
Y Mean	0.35	0.35	40.52	40.52
Adjusted R^2	-0.007	0.220	0.004	0.202

Great Depression severity is the normalized growth (decline) in retail sales per capita from 1929 to 1933. County level data on Great Depression severity is matched to county level data on mobility in the contemporary period from Chetty, Hendren, Kline, Saez, and Turner (2014). There are 107 observations, rather than the 99 cities in my main sample because several cities include multiple counties in the contemporary data, including New York, St. Louis, and three independent cities in Virginia. In the first two columns, the outcome variable is the slope from a regression at the county level of children's rank in the income distribution on parent's rank in the income distribution, for children born between 1980 and 1982. In the second two columns, the outcome variable is the expected rank of children with parents at the 25th percentile of the national income distribution, calculated from a regression of children's rank on their parent's rank. The geographic controls row indicates the inclusion of linear controls for latitude and longitude.

Source: Census of Retail Sales; Chetty, Hendren, Kline, and Saez (2014)

The results in this section have underscored the causal argument I offered in the previous sections: cities with more severe downturns during the Great Depression had less mobility in that period, but not before or much later. It is unlikely, therefore, that the Depression effects on mobility are merely correlated with some underlying geographic fixed driver of mobility. The Depression reduced mobility for the unlucky sons growing up in the worst hit cities.

1.4.5 Robustness: City Heterogeneity Does Not Reduce the Depression Mobility Effect

In the previous section, I showed that Great Depression severity does not predict mobility for the generation before the Depression. However, it could be the case that other city-level heterogeneity, which correlates with the Great Depression, drove mobility between 1920 and 1940. In this section, I consider several such sources of variation—industry mix, regional fixed effects, age variation, population size, historical growth rates, and inequality—and show that my central finding is robust.

While individual economic outcomes are likely the result of many generations of family inputs (Long and Ferrie, 2015), given the high levels of geographic mobility in the US during the last century, these null results are not surprising and not the ideal test of long run economic shocks on multi-generation mobility. A better test would link people today to their grandfathers or great-grandfathers during the Depression and to use the location of the grandfather rather than the location of the child to measure Depression severity. However, due to privacy restrictions on census data, such multigenerational matches ending with final generations in the contemporary period are not yet possible.

I use the same empirical framework throughout this section. Let $Z_{i,city}$ be some candidate father- or city-level covariate that may be correlated with the Great Depression and driving my results. I run regressions of the form:

$$Y_{i,son} = \beta_0 + \beta_1 \times Y_{i,father} + \beta_2 \times Y_{i,father} \times GD_{city} + \beta_3 \times Y_{i,father} \times Z_{city} + \gamma_{city} + \epsilon_i \tag{1.13}$$

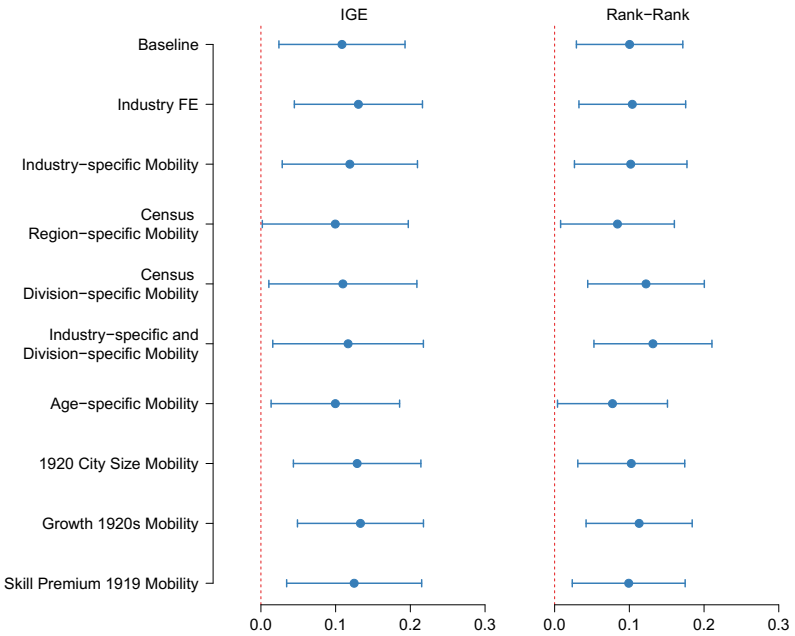
with variables as in equation (1.12).[63] The β_3 term allows mobility to differ according to $Z_{i,city}$. When the candidate covariate is scalar—population in 1920, 1920s growth rate—I recenter the variable around zero to simplify the interpretation of β_2, the effect of the Depression on intergenerational mobility. When the candidate covariate is a set of indicators—father's industry, census region or division, son's age—I saturate the regression and estimate a mobility rate for each indicator, dropping β_1. In both cases, the variable of interest remains β_2: the decrease in mobility due to a more severe Depression downturn. I plot β_2 for each candidate covariate in Figure 1.7, measuring mobility with both the intergenerational elasticity of income and the rank-rank parameter. The samples and controls in the baseline (first row) match exactly columns (2) and (6) of Table 1.4.

Industry-mix at the city-level was one of the strongest predictors of Depression severity, which I document in Appendix A.2. Industry may also determine the degree of intergenerational mobility. Certain fields or professions may be extremely rigid: miners' sons become miners and remain relatively poor, while the sons of mine electricians and carpenters may move on to more lucrative employment. Did cities with industries that were less mobile also have more severe Depression downturns, and could this explain my main finding? As I show in the second and third rows of Figure 1.7, it is unlikely. I first include fixed effects for each father's industry, allowing the son's outcome to differ based on father's industry. I then include interactions of these fixed effects with father's income, allowing mobility to vary by industry as well. In both cases, the Depression effect on mobility remains stable.[64]

[63] $Y_{i,son}$ is the son's outcome in 1940 and $Y_{i,father}$ is the father's outcome in the 1918-19 BLS sample. The outcomes are log earnings, position in the earnings distribution, or occupation score. GD_{city} is the severity of the Great Depression in the son's childhood city of residence. I cluster standard errors at the city level because sons in the same city are subject to common city-level shocks and the same observed city-level Depression severity.

[64] I code industries that were initially recorded as strings in both the 1920 and 1940 censuses, as well as in the BLS sample using the IPUMS 1950 standardized industry list; I aggregate the industries to 14 broad categories, roughly corresponding to SIC 2-digit codes.

Figure 1.7: Estimated Depression effect on mobility, allowing mobility to vary with other city and individual covariates



The reduction in mobility from a more severe Depression downturn is robust to a variety of control specifications. I run my main specification—the son's earnings in 1940 on father's earnings, Depression severity, and the interaction—and include additional terms that allow mobility to vary along other dimensions. Above, I plot the primary coefficient of interest, the interaction of Great Depression severity with father's earnings (either logged or his rank in the overall distribution). In each row, I include a different robustness control, described along the y-axis. The samples and controls match columns (2) and (6) of Table 1.4, including city fixed effects.

The Depression exhibited some degree of regional variation (Rosenbloom and Sundstrom, 1999). Intergenerational mobility has a strong regional component, both historically and in the current period. Studying mobility in the early twentieth century using socioeconomic patterns in given names, Olivetti and Paserman (2015) find low mobility in the South and high mobility in the Midwest. These patterns persist today: Chetty, Hendren, Kline, and Saez (2014) also document low mobility in the South and high mobility in the plains states. If the Depression had struck regions that had lower rates of mobility, either coincidentally or because low mobility drove Depression severity, that could explain my results. In the previous section, I showed that this is not likely to be the case, because, at the city level, mobility from 1900 to 1920 is unrelated to Depression severity. Here, I push the analysis a step further. I include in my main specification an interaction of father's log earnings or earnings rank with census region fixed effects.[65] This allows there to be a different rate of mobility in each of the four census regions separate from the Depression affected mobility rate. As shown in the fourth row of Figure 1.7, the main results are generally unaffected by this control. Allowing mobility to vary at the nine census division levels does not affect the main results either (fifth row).[66] Finally, when I allow both industry-specific and census division specific mobility rates, the main results remain stable (sixth row).

One critique of many estimates of intergenerational mobility is life-cycle bias. Samples with either older fathers or younger sons tend to exhibit more mobility (Corak, 2006). The BLS city samples may have different age profiles, or the Great Depression may have affected the trajectory of earnings for sons over time. In Figure 1.6, I documented the stability of the Depression interaction effect across sons' ages in my sample. I confirm that finding in Figure 1.7 (seventh row), showing that the overall Depression effect does not vary when allowing for age-specific mobility rates.

Several other city-level covariates could influence mobility. The cities sampled by the BLS range in size dramatically from New York, NY to Calumet, MI. Smaller cities may have less mobility and may have been more susceptible to Depression downturns because they were less industrially diversified or because local labor markets were too small. There is also some evidence

[65] The census divides the 50 states into four census regions and nine census divisions.

[66] I do not include census region or division fixed effects directly because they are subsumed by city fixed effects.

that the Roaring 20s, the boom preceding the Depression, may have exacerbated the downturn.[67] Inequality could also have been correlated with the Depression, and there is a well known relationship between inequality and mobility today (Krueger, 2012; Corak, 2013). Unfortunately, inequality measures are not abundant in the early 20th century; instead, I turn to city-level data from the Census of Manufactures and calculate a worker-skill premium. For each city, I know the wages paid and the number of workers in three manufacturing skill categories: operators, clerks, and officers. I calculate the skill premium in each city as the ratio of wages per operator against wages per clerk and officer. As I show in the final three rows of Figure 1.7, allowing mobility to vary across any of these city-level dimensions does not affect the main finding.

Variations in military spending as the United States prepared to enter WWII could be another potential driver of mobility patterns. The wartime labor market could have compressed earnings variation, which might affect mobility overall. Alternatively, local defense spending could have affected sons' earnings in different ways across different cities. If these war effects were correlated with Depression severity, perhaps because spending was targeted to distressed areas, that could threaten my interpretation of the Depression effects. However, I find this to be unlikely. To begin with, I observe the earnings and occupations of sons in the 1940 census which refers to 1939 outcomes. Germany did not invade Poland until September 1939, and the US did not enter the war until after Japan attacked Pearl Harbor, on December 7, 1941. US defense spending remained at around 2% of GDP in 1939 and 1940, its average share throughout the 1930s. Cullen and Fishback (2006) argue that the economic mobilization for the war "started gradually"; monthly munitions output at the end of 1940 was at barely 2% of the peak in late 1943 (Krug, 1945). The massive production of the instruments of war did not begin until well after I observe the sons in my sample.

In the previous sections, I have shown that Great Depression severity reduced intergenerational mobility for sons growing up in cities with severe downturns. In the next section, I ask why.

[67] I explore determinants of local Depression severity in Appendix A.2 and find a weak connection between growth in the 1920s and the downturn, at least for the cities in the BLS sample.

1.5 *Mechanisms: How Did the Depression Lower Mobility?*

Sons growing up in cities with worse downturns during the Great Depression were "locked" into their fathers' outcomes. Guided by my model of intergenerational mobility, I explore several possible mechanisms: migration, education, and inequality. Ultimately, I find that differential geographic mobility by sons of rich and poor fathers offers the best explanation for my findings.

1.5.1 *Geographic Mobility*

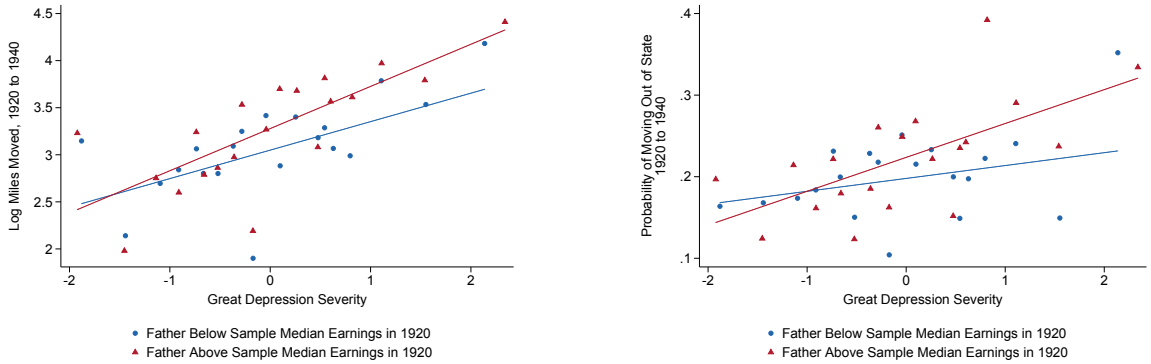
Since de Tocqueville (1839), observers have highlighted the high rates of geographic mobility in America. The sons in my sample are highly mobile, and this mobility differs by both father's income and Great Depression severity. In particular, I find that increased local Great Depression severity prompts all sons to migrate, but that the sons of richer fathers are more likely to make *better* migration choices, moving to cities and counties with less severe Depression downturns.

Historically, the young are among the most geographically mobile population groups. Counties with larger population shares between 10 and 19 years of age in 1930—a group comprising half of the sons in my sample—had more out-migration during the 1930s (Fishback, Horrace, and Kantor, 2006). The sons in my sample are no different, moving an average of 165 miles. Of those sons who leave their county of residence in 1918-19, they move an average of 340 miles. In my sample, 21% of the sons move to a different state between 1918-19 and 1940. Of those remaining in-state, 33% move counties.

Local Great Depression severity drove this migration. Overall, geographic mobility was lower during the Depression decade than during previous decades, but this decline in mobility is common during recessions.[68] However, across the country, I find that the sons in my sample growing up in cities suffering more severe Depression declines were more likely to migrate. Figure 1.8 shows that both the probability of moving to a new state and the distance moved from 1920 to 1940 increase for sons living in cities with larger downturns. Sons in cities with a Depression downturn one standard deviation worse than the mean were 3% more likely to

[68] See for example, the index of internal redistribution of population in Figure 1.2 of Rosenbloom (2002). There was less mobility in the 1930s than during the 1920s or 1940s, but low levels of migration during the 1890s and 1910s, decades with multiple macroeconomic contractions. During the first American depression in 1819, migration to the frontier declined (Lebergott, 1964).

Figure 1.8: Sons in cities with more severe Depression downturns were more likely to be move out of state and to move farther



(a) Distance Moved

(b) Moving Out of State

Relative Great Depression severity spurred out-migration of the sons in my sample between 1920 and 1940. Each figure presents a binned scatter plot, grouping the sons in the sample in Depression severity percentiles and plotting (a) the average log miles moved or (b) the share moving between states within each group. Sons are split in each plot by the earnings of their father in 1920; median earnings for fathers in the sample was $1300 in 1920 dollars. Sons of richer fathers appear to be slightly more likely to move or to move farther in response to Depression severity, but these differences are not statistically significant.

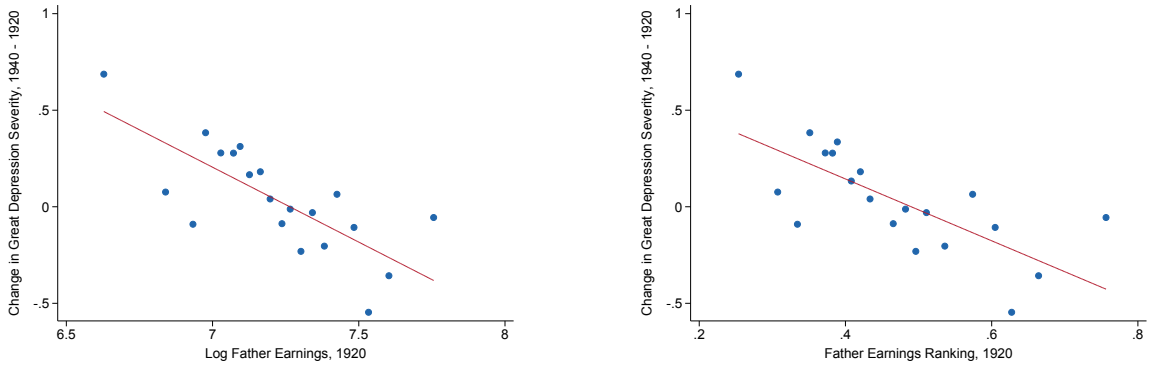
move out of state and moved 40% farther.[69] In Figure 1.8, I also explore whether this mobility effect differed by father's earnings. Splitting the sample of sons at the median of father's earning ($1300 in 1920 dollars), it appears that sons with richer fathers may be somewhat more mobile in response to increased Depression severity, but these differences are not statistically significant.[70]

However, the sons of richer fathers and the sons of poorer fathers did not move to the same sorts of places. In Figure 1.9, I show that sons from richer families made much better migration choices. I measure the quality of a migration by the difference in Great Depression downturn severity between the home county in 1920 and the destination county in 1940: a negative change in severity implies a move to a region with a less severe Depression. Among the sons moving out of state, the quality of the sons' moves increased with father's earnings in 1918-1919, measured

[69] The Depression pushed sons to move, but where did they go? In the appendix, I map the locations of all the sons in my sample in Figure A.2. Among the sons moving out of state, 19% of the sample ended up in California and another 7% in New York. Others moved to the industrial midwest, as 4 to 5% of sons went to Illinois, Michigan, Pennsylvania, Missouri, Ohio, and Indiana.

[70] In the regressions that correspond to the plots in Figure 1.8 (a), the semi-elasticity of distance moved with respect to Depression severity is 0.446 for sons with fathers earning above the median in 1920 and 0.300 for sons with fathers earnings below the median, but with overlapping confidence intervals. Similarly, for Figure 1.8 (b), the two slopes are 0.041 and 0.016, but are not statistically significantly different.

Figure 1.9: Sons of richer fathers moved to cities with less severe Depression downturns



(a) Log Father Earnings **(b)** Father Earnings Rank

The sons of richer fathers were more likely to make better migration choices, measured by the difference in Depression downturn severity in sending and receiving counties. Migration quality is measured by the difference in growth of retail sales per capita between 1929 and 1933 in the county of residence in 1940 and the county of residence in 1920. A negative value implies a move from a city with a higher level of Great Depression severity to a city with a lower level of Depression severity. Only sons moving out of state are included in this figure. Each figure presents a binned scatter plot, grouping the sons in the sample into fathers' earnings or ranking percentiles.

either as the log of earnings or the father's ranking in the earnings distribution.[71]

The differential in directed mobility can explain my main finding that the Great Depression decreased intergenerational economic mobility. To see this, I split my sample into movers and stayers and estimate intergenerational mobility on these selected samples in Table A.4. For both the IGE and rank-rank parameters, I find the Depression effect is concentrated within the migrant sample rather than the non-migrant sample.

In cities with more severe downturns, more sons migrated and thus had to make a decision about where to live. The sons with richer fathers were able to make better choices, moving to areas that experienced a less severe Depression shock. This explains the increase in their average outcomes, relative to the sons of richer fathers in cities with less severe downturns, in Figure 1.5. Even if the richer sons in mild Depression cities were just as likely or as able to make good migration choices, fewer of them did so, as the milder Depression compelled fewer to migrate in the first place. Conversely, Figure 1.9 shows that the sons of poorer fathers migrated to areas

[71] Expanding the sample to include all sons, not just those migrating between 1920 and 1940, reduces the estimated slopes in Figure 1.9, but does not change the main point that the sons of richer fathers were more likely to move to locations with a milder Depression downturn. In addition, controlling for 1920 severity or including 1920 city fixed effects does not break the relationship.

with more severe local Depression downturns, a choice that was likely to lower their eventual economic outcomes. Because the sons in cities with more severe Depression downturns were more likely to migrate, more of the poorer sons in these cities moved to cities with even worse local economic conditions. During the Great Recession, Yagan (2014) finds that richer workers bear a lower incidence rate from a local downturn thanks to migration. The same pattern played out in the Great Depression, with more geographic mobility and smarter geographic mobility protecting the sons of richer fathers in cities with worse downturns.

1.5.2 *Years of Education*

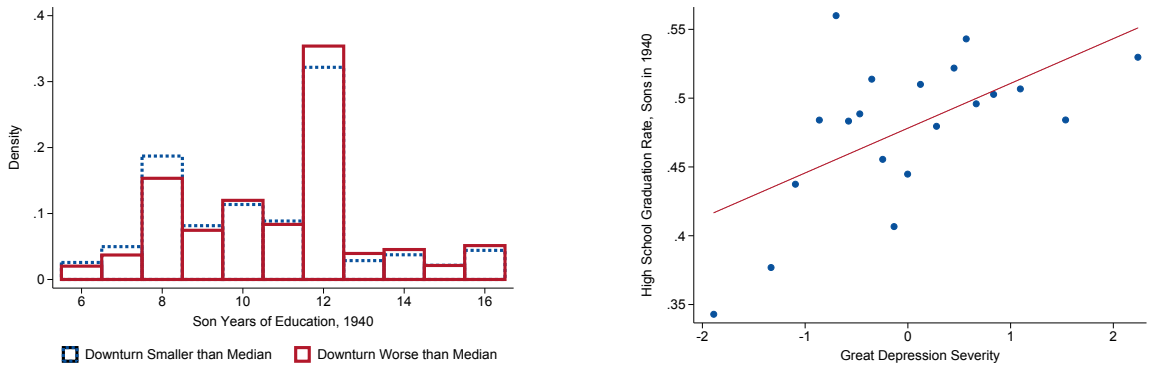
My model suggests that the Great Depression could have determined intergenerational mobility through a differential change in educational attainment. I find that in all cities the sons of richer fathers get more years of education, but that this education gap does not vary according to Great Depression severity. Therefore, education is not the mechanism through which the Depression affected mobility.

In cities with worse downturns, sons of poorer fathers may have been more likely to drop out of school earlier to enter the labor force, with income and employment reduced for their parents.[72] Had they remained in school, some of those sons would have climbed the education and income ladder: these Depression-induced drop outs would thus decrease local mobility relative to a city with a more mild downturn. Sociologists have speculated that, for the sons of poorer fathers, educational opportunities beyond high school may have been especially sensitive to the limits on resources during the Great Depression (Elder, 1999, p. 154).

However, the relationship between the Great Depression and education is much more complex. As I document in Figure 1.10, for sons growing up in cities with worse downturns, both total years of education and high school graduation rates actually increase. These changes are driven by a shift in the distribution of completed schooling from 8 to 12 years, as is clear in the histogram in Figure 1.10 (a). The high school graduation rate was only 45% in cities with mild Depressions, but was 51% in cities with downturns worse than median. These results echo the

[72] In a history of the working class in Depression era Chicago, Cohen (1991) recounts several cases of children serving as primary earners while parents were unemployed. Such stories are common in oral histories of the Depression generation (Terkel, 2013; Watkins, 2000).

Figure 1.10: Sons growing up in cities with more severe Depression downturns had more education in 1940



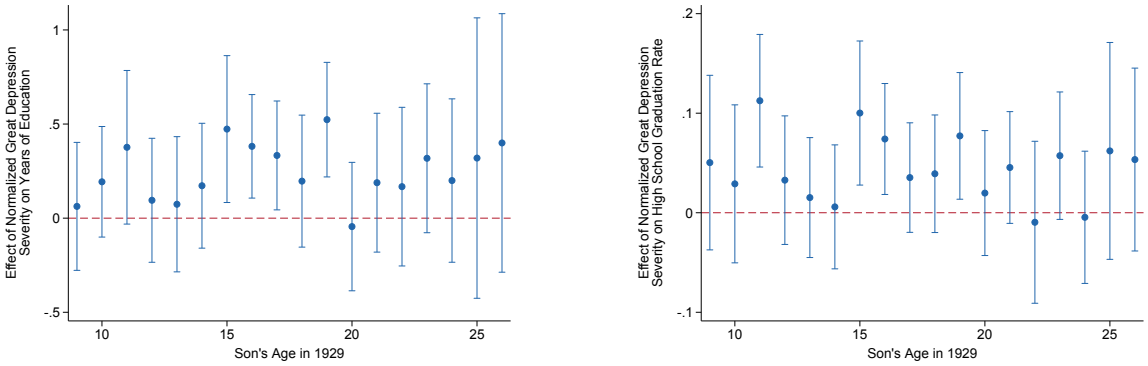
(a) Son's Years of Education in 1940 **(b)** Son's High School Graduation Rate in 1940

The Great Depression shifted sons from grade school graduates to high school graduates. In (a), I present a histogram of the completed years of schooling by Depression severity. Cities with downturns smaller than median are displayed in dashed blue bars; cities with downturns worse than median are in solid red bars. I omit the 3% of sons with 5 or fewer years of education from the histogram. In (b) the binned scatter plot demonstrates that high school graduation rates among sons in my matched sample increased in normalized Great Depression severity: the more severe the downturn, the higher the graduation rate.

findings in Goldin and Katz (1997) and Goldin (1998) on the positive effects of the Great Depression on education. Drawing on data at the state level, Goldin and Katz (1997) find that high school graduation rates increased the most in states with the largest increases in unemployment. In addition, states with large manufacturing bases also saw large increases in graduation rates during the Depression, likely because the New Deal era National Industry Recovery Act codes banned the employment of youths in manufacturing. Shanahan, Elder, and Miech (1997), analyzing different cohorts from the Stanford-Terman Study of Gifted Children, also find that the Depression increased educational attainment.

Did the Great Depression *cause* the increase in education suggested in Figure 1.10? Without comprehensive data on educational attainment before the Depression, that is difficult to prove, but the variation of the Depression effect within my sample is suggestive. In Figure 1.11, I plot coefficients from a pair of regressions of educational outcomes in my sample of sons on 1920 city-level Depression severity interacted with age in 1929; in (a) the outcome is completed years of education in 1940, and in (b) it is an indicator for high school graduation. It seems unlikely that sons who had already graduated or dropped out of school when the Depression hit in 1929 would be responsive to its severity. On the other hand, for sons at the ages on the margin of

Figure 1.11: Great Depression Severity Increased Education among Sons aged 15 to 19 in 1929



(a) Son's Years of Education in 1940 **(b)** Son's High School Graduation Rate in 1940

Both plots present coefficients from single regressions of (a) completed years of education in 1940 or (b) an indicator for high school graduation on the interaction of Depression severity with son's age in 1929. Standard errors are clustered at the city level. Controls include son's age fixed effects and a quartic in father's age.

working or staying in school, the Depression could have a large effect: if jobs were more scarce in cities with more severe downturns, then those sons would have a lower opportunity cost of education and be more likely to remain in school. The patterns in Figure 1.11 support these possibilities: for sons between 15 and 19 in 1929, a one standard deviation increase in Depression severity increases total years of education by 0.25 to 0.5 years and increases the probability of graduating from high school by 5 to 10 percentage points. However, sons at other ages, either younger or older, are no more likely to increase their educational attainment in response to Depression severity.

If the Depression-induced increase in education is differential according to father's income, that could explain my main finding. To test this, I regress measures of the sons' educational outcomes on the father's log income in 1920, an interaction of income and Depression severity, as well as age quartics and city fixed effects. In Table 1.8, I show that while sons of richer fathers have more years of education and are more likely to graduate from high school and college, there is no effect of Depression severity on these links. The relationship between father's income and son's education is stable regardless of Great Depression severity. However, education could still play an important role in my findings through its protective effects against downturns. In Figure 1.9, I showed the sons from more advantaged families moving to cities with less severe Depres-

Table 1.8: Fathers' Income and Sons' Educational Outcomes Across Depression Severity

	Years of Education		High School Graduate		College Graduate	
	(1)	(2)	(3)	(4)	(5)	(6)
Log Father Earnings, 1920	1.320***	1.318***	0.218***	0.218***	0.051***	0.051***
	(0.168)	(0.169)	(0.028)	(0.029)	(0.014)	(0.014)
Log Father Earnings 1920 X GD Normalized Severity		-0.034		-0.008		-0.003
		(0.125)		(0.023)		(0.012)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6129	6129	6129	6129	6129	6129
City Clusters	99	99	99	99	99	99
Adjusted R^2	0.095	0.095	0.083	0.083	0.013	0.013

Standard errors clustered at the 1920 city level. The sons of higher earning fathers have more years of education and are more likely to graduate from high school and college. However, the relationship between father's earnings and son's education is not significantly affected by local Great Depression severity. Sons have an average of 10.7 years of education in my sample. 48% of the sons are high school graduates and only 3% are college graduates.

Source: BLS Cost of Living Survey 1918-1918; 1940 Complete Count Census; Census of Retail Sales

sion downturns. These sons had more education (Table 1.8). Even if education differentials do not change with Depression severity, education could be one mechanism to explain differences in migration quality. More education might allow sons to find a better city match by opening up connections and information about different labor markets.

1.5.3 *Returns to Education*

The model of intergenerational mobility presented in section 1.3 makes clear the role of the return to skill in determining mobility rates. As p, the return to human capital parameter, increases, persistence rises and so mobility falls.[73] For the cities in my sample, the Great Depression does not appear to have increased the return to skills or human capital, suggesting that changes in the return to education do not explain the Great Depression effects on mobility. However, this analysis is necessarily limited by the fact that I cannot observe the returns to education *before* the Depression across cities.

[73] This prediction is borne out empirically in recent data as well. Comparing intergenerational mobility across counties today, Corak (2013) shows a strong positive correlation between the college premium and mobility. Aaronson and Mazumder (2008) document a correlation between the college premium and intergenerational mobility from 1940 to 2000 in the US. The recent increase or convexification in the return to education has likely driven part of the increase in inequality as well (Autor, Katz, and Kearney, 2008).

To test whether the Great Depression increased the return to skills in the US, I turn to the 1940 100% census. I observe annual earnings in 1939, completed years of education, age, and place of residence in 1940. Separately for each of the 3,071 counties in the country, I run a simple Mincerian returns to education regression, regressing log earnings on years of education and age dummies for all full-time employed white men between the ages of 16 and 65 in the county. The coefficient on education in each regression is the observed return to education; I map the returns in Figure A.3. Obviously, these estimated returns to education are descriptive and not causal. But they suggest which counties had relatively higher returns for skill and education.[74]

Nationwide, it does appear that Depression downturn severity, measured as usual by the decline in retail sales per capita from 1929 to 1933, have higher returns to education in 1940. The top panel of Figure 1.12 shows the correlation in a binned scatter plot. Each point represents one percentile of the Depression severity measure, comprising approximately 30 counties. However, this includes many small, rural counties. When I narrow my focus to the cities sampled by the BLS, the relationship disappears, as is evident in the bottom panel of Figure 1.12.[75] I find little evidence that the Great Depression was more (or less) severe in cities with high returns to education in 1940, suggesting that higher returns to schooling are unlikely to be the mechanism driving lower levels of mobility.[76] However, without good estimates of the return to education in these cities before the Depression, this conclusion is more speculative.

1.5.4 *Inequality*

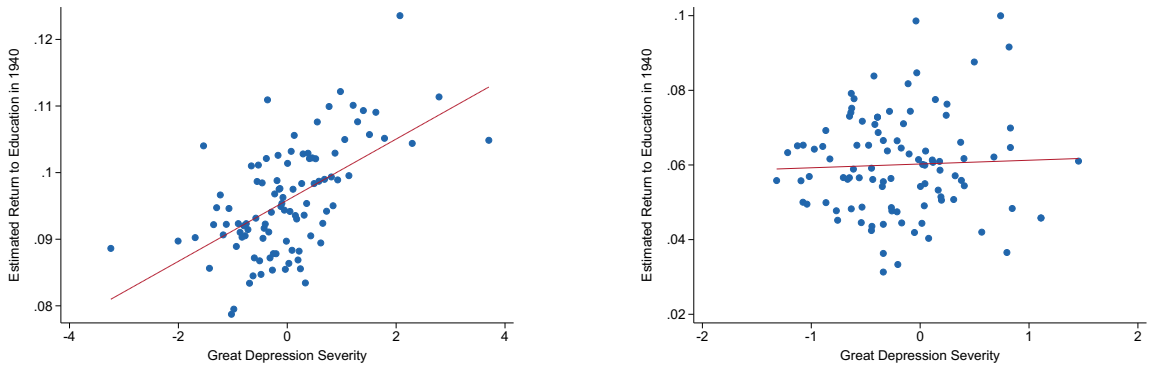
Intergenerational mobility and inequality are linked, both conceptually and empirically. Immobile societies are often highly unequal. I have shown here that the Depression hardened the link between parents and children's economic positions, but did it also change the relative economic distribution? I find that earnings inequality did increase in the cities hit hardest by the

[74] The returns to education correlate negatively with county population: an increase in county population by 1% correlates with a 1% smaller estimated return. It is also apparent from the map that the human capital returns were higher in the plains states—the Dakotas, Nebraska, and Kansas—as well as the in mid-South. The returns to education were extremely low in New England and New York, as well as in the industrial Midwest—Ohio, Illinois, Michigan, and Indiana—and on the West Coast.

[75] There is nothing particular about the BLS-sampled cities. The correlation between the estimated return to education and Depression severity is flat across all counties with populations over 25,000, for example.

[76] This finding tracks the time series evidence for inequality. Goldin and Margo (1992) document lower inequality in the decades following the Great Depression, a period of declining returns to education (Goldin and Katz, 2008).

Figure 1.12: Great Depression severity did not correlate with higher returns to education in 1940



(a) All Counties **(b)** BLS Sample Cities

The top panel shows a binned scatter plot, comparing Great Depression severity from 1929 to 1933 and the estimated return to education in 1940 across all counties. Counties hit by more severe downturns had substantially higher returns to an additional year of schooling in 1940. However, this relationship is driven by smaller, rural counties. The bottom panel limits the sample to the 99 industrial cities included in the BLS survey and shows no correlation between Depression severity and the return to education. Education returns are measured with county level regressions in the 1940 complete count census, regressing log annual earnings on years of education and complete age dummies for full time employed white men between 16 and 65. Great Depression severity is normalized retail sales growth per capita from 1929 to 1933, where higher severity implies less growth.

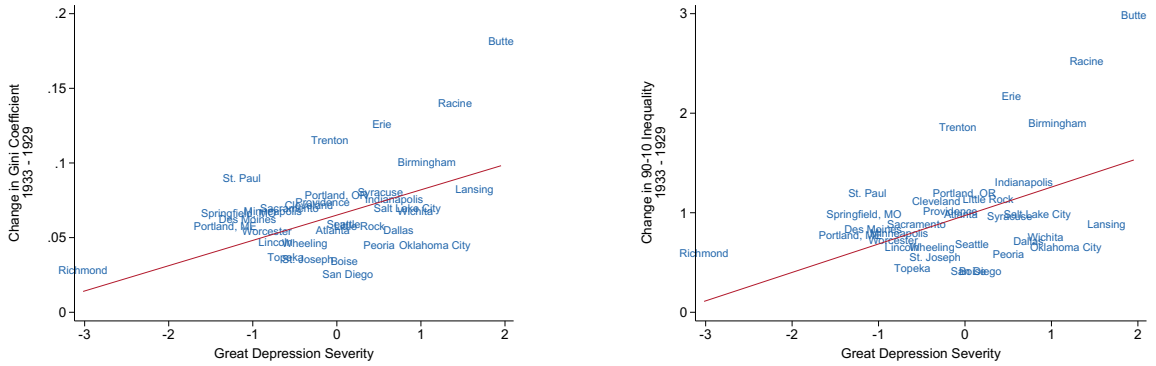
Depression, but that this increase in inequality is not nearly large enough to explain my mobility findings.

Data on inequality at the county or city level in the years before and after the Depression are rare. I use data from the only such contemporaneous survey I know of: David Wickens' Financial Survey of Urban Housing, which collected data on the income distribution in 1929 and 1933 for 33 American cities.[77] I recollect these data and calculate measures of inequality before (1929) and during (1933) the Great Depression. Cities range in size from Cleveland (1.2 million) to Boise (22,000). Only 21 of the 33 cities are in both the Wickens sample and BLS sample in 1918-1919.

Results from this small sample of cities suggest that the Depression increased inequality. In Figure 1.13, I plot the changes in Gini coefficients and in the difference between 90th and 10th percentile earnings against Depression severity at the city level: the worse the downturn, the larger the increase in inequality. The results hold for other inequality measures, including the 90-50 and 50-10 differences. In all cases, Depression severity is uncorrelated with inequality in

[77] This survey has been used in previous work on historical inequality by Grayson (2012) and Mendershausen (1946).

Figure 1.13: The Great Depression increased earnings inequality in cities between 1929 and 1933.



(a) Gini Coefficient

(b) 90th - 10th Percentile Earnings

Across the 33 cities included in the *Financial Survey of Urban Housing*, which collected data on the income distribution in 1929 and 1933, the change in inequality was correlated with the severity of the Great Depression. The relationship between Depression severity and the Gini coefficient and the 90-10 differential are plotted above, but both the changes in the 90-50 and in the 50-10 differential also correlate with Depression severity. There is no relationship between inequality in 1929 (by any measure) and Depression severity. Great Depression severity is normalized retail sales growth per capita from 1929 to 1933, where higher severity implies less growth.

1929 and positively correlated in 1933. If the Depression increased inequality and decreased mobility across cities, that follows the Great Gatsby curve identified across countries (Krueger, 2012; Corak, 2013) as well as contemporary patterns in mobility and inequality across US counties (Chetty, Hendren, Kline, and Saez, 2014).[78]

The magnitude of the effect of the Depression on inequality, however, is too small to explain my mobility results. Chetty, Hendren, Kline, and Saez (2014) regress intergenerational elasticities of income on Gini coefficients across countries, estimating a slope of 0.72. I estimate a 0.10 increase in IGE and a 0.02 increase in the Gini coefficient for a one standard deviation worse Depression shock. With the strong assumption that the Gatsby Curve is causal—and not inflated due to spurious correlation driven by public policy or social norms—and that the contemporary Gatsby Curve is relevant to the inequality-mobility relationship in the early 20th century, widening inequality only explains about 14% of the change of mobility.[79]

[78] The proposed causal effect of inequality on mobility within the US may not exist. Bloome (2015) exploits state and cohort variation within the US and finds "very little support for the hypothesis that inequality shapes mobility."

[79] The IGE in a city suffering a Depression downturn one standard deviation worse than the mean would be expected to increase by 0.10; the increase in the Gini coefficient would be 0.02, which when multiplied by the slope of the Gatsby Curve yields an expected IGE increase of $0.02 \times 0.72 = 0.0144$, or roughly 14% of the 0.10 direct effect of the Depression

1.6 *The New Deal and Mobility*

Franklin D. Roosevelt took office in March 1933 amidst the depths of the Depression. During and after his famed first hundred days, Roosevelt attacked the Depression from a number of angles. A bank holiday staved off a string of bank runs; leaving the interwar Gold Standard freed monetary policy from fixed-exchange shackles; and the Federal government began a new era of involvement in the economy, spending and loaning huge amounts of money through various programs, grants, and agencies. Did these New Deal programs—aimed at alleviating the economic misery of the Depression—also have an effect on intergenerational mobility?

I measure per capita New Deal spending at the county level with data drawn from Fishback, Kantor, and Wallis (2003). In addition to observing overall expenditures on grants and loans, I follow Fishback, Kantor, and Wallis (2003) in categorizing spending into three broad categories.[80] Relief spending comprised the first category, including Federal Emergency Relief Administration grants, Civil Works Administration grants, Works Progress Administration grants, and Public Assistance Grants via the Social Security Act. Spending was directed to areas most in need of relief, targeting cities and counties with the most severe downturns (Fishback, Kantor, and Wallis, 2003). In the cities in my sample, spending in the relief category accounts for roughly 43% of New Deal outlays in each city. In the second category is public works spending via Public Works Administration grants and loans, as well as grants from the Public Roads Administration and the Public Buildings Administration. The infrastructure programs endeavored to build useful projects, rather than immediately boosting employment during the Depression. In most of my sample cities, public works spending accounts for less than 20% of total New Deal grants and loans. Finally, a number of New Deal programs targeted the housing market with loans or the insurance of mortgages. This spending included loans from the Reconstruction Finance Corporation, from the Home Owners Loan Corporation, and grants and loans from the US Housing Administration.

---

on the IGE.

[80] A fourth category of New Deal grants and loans was spending directed to farmers and rural areas. However, while these New Deal outlays are important to the economic history of the Great Depression and New Deal, they are not relevant for my urban sample of fathers and sons. Only a handful of sons live in counties which received more than 10% of New Deal outlays in the form of farm assistance, and this spending was likely directed at the non-urban core of the counties. These programs include the AAA, FCA, FSA grants and loans, and the REA.

New Deal spending did not alter the relationship between the downturn and intergenerational mobility, as I show in Table A.5. In Panel A, I supplement my baseline IGE specification with interactions of normalized per capita New Deal spending and father's log income in 1920 in columns (2) and (3). Though the point estimates are negative—suggesting that New Deal spending may have in part counteracted the deleterious effects of the Depression on economic mobility—the results are imprecise. When I decompose New Deal spending into three constituent parts—relief spending, public works, and housing market assistance—I find weak evidence that one aspect of the New Deal, housing market support, may have reduced the negative mobility effects of the Depression. However, these results are only marginally significant and are more speculative. Further, when I examine the effects of the New Deal on rank-rank mobility in Panel B of Table A.5, I find no effects of the New Deal on mobility.

New Deal spending had long term positive effects in the cities and counties targeted: Fishback, Horrace, and Kantor (2005) show that a dollar of spending increased retail sales in 1939 by 44 cents. But the programs were slow to begin and did not ameliorate economic misery immediately. By the time the economic recovery began—fueled in part by New Deal programs—sons hit hardest by the downturn were likely already locked in to their fathers' relative economic position.

1.7 Conclusion

In this paper, I have explored the effects of the Great Depression on intergenerational mobility. Comparing across sons, I find that the Depression was bad for mobility. Despite—or perhaps because of—massive economic disruption, the economic slate was not wiped clean. Instead, my results suggest that sons growing up in cities with more severe downturns during the Great Depression had less economic mobility than sons from cities with milder shocks. These effects are not driven by pre-existing place-specific differences in mobility, as the mobility rates for the generations before and long after the Depression are unaffected by local downturn severity.

Differential directed migration is the primary mechanism to explain my results. The Depression pushed all sons to look elsewhere for better economic opportunities: geographic mobility increased with Depression downturn severity across my entire sample. However, the sons of richer fathers were able to make better choices, migrating to cities which had suffered far less se-

vere Depression downturns. Because the city-specific Depression effects were long lasting—cities with more severe downturns still had lower wages and higher unemployment in 1940—these directed movements were hugely beneficial to the sons making them.

I have shown the mobility effects of variation in Great Depression severity, but my empirical specification cannot directly address a different question: what would intergenerational mobility have been had the Great Depression not struck? Speculatively, my results may shed some light on this more dramatic counterfactual. My local findings may be an underestimate of the aggregate effects of downturns on mobility. The sons of richer fathers in my sample were able to migrate to "better" locations than their poorer peers, but their choices were constrained by the Depression: no American cities had robust wage growth or low levels of unemployment during the Depression-era. If such places existed, the effects of directed migration on son's outcomes may have been even more dramatic, which would make the mobility effect of the Depression that much stronger. However, the historical relationship between migration and macroeconomic growth complicates this interpretation. Overall migration rates during the Depression era were lower, as is usually true during periods of slow growth in American history (Rosenbloom, 2002; Lebergott, 1964). The lower geographic mobility rates in aggregate could have increased intergenerational economic mobility by reducing the share of sons with richer fathers benefiting from the gains of directed migration, pushing against the local effects I find. In this case, in contrast to my finding that local Depression severity reduced mobility, in aggregate it may have increased mobility.[81]

To create the longitudinal sample of fathers and sons necessary for my analysis, I developed a machine learning approach to census linking. With my method, I am able to quickly, accurately, and efficiently match large samples of people across censuses or into the census from other sources, like the BLS cost of living survey. Combining this method with many recently digitized historical microdata sets, it is possible to extend our understanding of the determinants of intergenerational economic mobility. The Depression reduced mobility, but what of the other major economic, social, and political events of the early twentieth century? Did the first wave

[81] It may also be possible that migration is more directed during downturns than during periods of growth, a claim that my census linking procedure could enable me to test by constructing longitudinal samples of fathers and sons across many different eras. A further complication is the unknown distributional effect of aggregate economic conditions on migration across richer and poorer sons.

of the Great Migration break intergenerational bonds among African American children born in this period—both for those who migrated and for those who remained in the South? What effect did the dramatic expansion of public schooling during the High School movement have on economic mobility? How did mobility change as the American population shifted from majority rural to majority urban?

2. A NEW OLD MEASURE OF INTERGENERATIONAL MOBILITY: IOWA 1915 TO 1940

2.1 *Introduction*

The history of income inequality throughout the twentieth century is well known (Piketty and Saez, 2003). Much less, however, is known about economic mobility in the United States in the first half of the twentieth century. How strong was the link between a child's outcomes in adulthood and the accident of his or her birth? And how does economic mobility in this earlier period compare to mobility today? How much more common were Horatio Alger's rags-to-riches heroes in the early twentieth century than in the early twenty-first? The concepts of inequality and intergenerational mobility are strongly linked, but inequality does not determine mobility, or vice versa, and so few clues are available in the inequality literature. The recent studies on trends in intergenerational mobility are unable to trace income mobility before the 1980s (Lee and Solon, 2009; Chetty, Hendren, Kline, Saez, and Turner, 2014). Moreover, historical analysis of economic mobility has long been limited by the available data sources. While the United States federal census began collecting information on respondents' occupations in 1850, the census did not include data on either years of educational attainment or annual income until 1940. Historians have collected detailed place-specific data on intergenerational mobility and transfers (see Thernstrom, 1964, 1973, on Newburyport, MA and Boston, for example), but these studies are often constrained by their inability to track those individuals who moved away from the original study site.

To measure economic mobility in the early twentieth century, I match fathers from the Iowa State Census of 1915 to their sons in the 1940 Federal Census, the first state and federal censuses with data on income and years of education. I estimate intergenerational mobility along three dimensions: income, education, and occupation. The estimates I present are the earliest intergen-

erational mobility estimates for both income and education in the United States.[1] I use these data to address the question of how intergenerational mobility changed both between 1915 and 1940 in the United States, as well as between 1915 and the present. In addition, because my sample includes intergenerational data on income, education, and occupation for the same individuals, I can determine whether these measures of intergenerational mobility all show consistent trends.

I summarize my primary results in Table 2.1. I find a lower intergenerational income elasticity during the first half of the twentieth century in the US than studies find in the second half of the century. This result implies that there was more mobility of income during my study period than there is today. I also measure intergenerational mobility using the rank-rank parameter (Chetty, Hendren, Kline, and Saez, 2014) and similarly find more mobility historically. Such differences between contemporary and historical mobility could be spurious, driven by measurement error or differences in sample construction. However, I show that the estimated differences between contemporary and historical mobility remain large after adjusting the contemporary sample to mirror the historical sample in measurement noise and demographic and geographic composition.

The results for education are broadly similar: there was more mobility in education in the early twentieth century than today as well. This is also the case for occupational mobility when measured with the standard transition matrices. My results indicate that mobility is higher for the early twentieth century than just after mid-century; I also find less occupational mobility during the twentieth century than others have found for the nineteenth century.[2]

Overall, the various intergenerational mobility measures point to one, main conclusion: there was more economic mobility in the early twentieth century than there is today.

The paper proceeds as follows. In the second section, I discuss the historical data that I

[1] My study builds on the earlier mobility work of Parman (2011), who also draws on the 1915 Iowa State Census to measure intergenerational mobility. However, data constraints limit the broad interpretability of his results. He was only able to match fathers to sons within 1915 Iowa. Thus his estimate is biased by omitting any sons who move out of the state. Even more problematic, the average age of the fathers in his sample is between 57 and 65 and his sons are between 25 and 30. Corak (2006) shows that parameters estimated with such old fathers and young sons are biased down to a large degree. These points are addressed in more detail later in this paper.

[2] This point contrasts somewhat with findings in the recent intergenerational mobility literature. Jencks and Tach (2006) suggest that intergenerational correlations of earnings and occupational rank are not good substitutes. They note, in particular, that in the US earnings correlation is higher than in other rich democracies but occupational rank correlation is low relative to such countries. For historical study, I find that occupational and income mobility measures are relatively similar.

Table 2.1: Intergenerational Mobility Results Summary

Intergenerational Mobility Measure	Estimates		Modern Source
	1915 to 1940	Modern	
Intergenerational Elasticity of Income	0.199 to 0.258	0.36 to 0.54	Lee and Solon (2009)
Income Rank-Rank Coefficient	0.169 to 0.219	0.307 to 0.317	Chetty et al. (2014)
Educational Persistence	0.206 to 0.264	0.46	Hertz et al. (2007)
Occupation Score Elasticity (1915 Basis)	0.162 to 0.265	.	
Occupation Score Elasticity (1950 Basis)	0.366 to 0.441	.	
Altham-Ferrie Occupation Transition Statistic	16.14	20.76	Ferrie (2005)

All measures of intergenerational mobility will be explained in detail in the text of this paper. Across all measures, higher estimates imply less mobility. The intergenerational elasticity of income is the regression coefficient on log of father's annual income, with son's annual income as the dependent variable. The income rank-rank coefficient is the regression coefficient on the father's income percentile, with the son's income percentile as the dependent variable. Educational persistence is the the regression coefficient on the father's years of education, with the son's years of education as the dependent variable. Occupation score elasticity is the regression coefficient on the father's occupation score, with the son's occupation score as the dependent variable, both in logs. The occupation scores are defined as the median income across all respondents in a given occupation in a given base year. The Altham-Ferrie occupation transition statistic relates the distance from a given occupation category transition matrix to the complete mobility matrix.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census; Lee and Solon (2009); Chetty, Hendren, Kline, and Saez (2014); Hertz, Jayasundera, Piraino, Selcuk, Smith, and Verashchagina (2007); Ferrie (2005)

draw on and my data collection and census-linking procedures. In section three, I review past measurements of intergenerational mobility, both in contemporary and historic samples. In particular, I focus on sources of measurement error that may bias the estimates of mobility up or down in historic data relative to recent data. In the fourth section, I present my estimates of intergenerational mobility in the early twentieth century for income, education, and occupation, and compare these results to estimates for the contemporary period. I also consider heterogeneity in the mobility parameters across my sample as well as geographic mobility. Section five concludes the paper.

2.2 *Data and Record Linkage*

I draw my primary data for measuring intergenerational mobility in the United States early in the twentieth century from the 1915 Iowa State Census and the 1940 US Federal Census, both of which include measures of the earnings, education, and occupations of the respondents. I describe both data sources in this section, as well as the method used to link fathers in 1915 to their sons in 1940.

The 1915 Iowa Census covered all 2.3 million Iowa residents in 1914. It was the first American census of any kind to include data on both annual income and years of education in addition to more traditional census measures, and it also includes respondent name, age, place of residence,

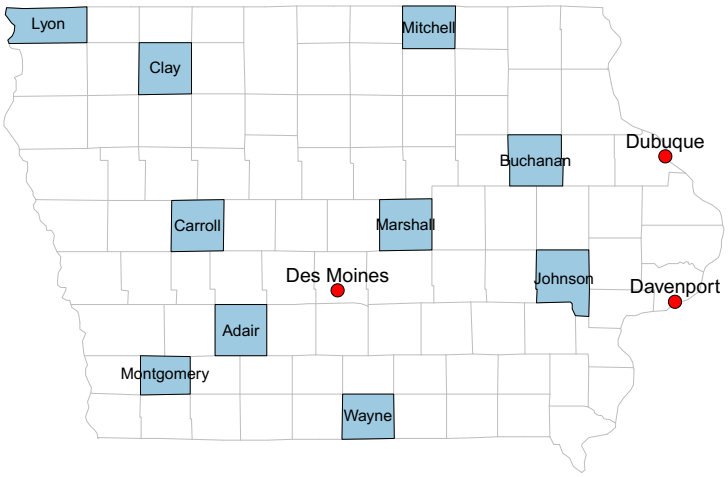


Figure 2.1: Map of Iowa 1915 Cities and County Sample

birthplace, marital status, race, and occupation. I use the Iowa State Census sample digitised by Claudia Goldin and Lawrence Katz for their work on the historical returns to education (Goldin and Katz, 2000, 2008). The Goldin-Katz sample includes 26,768 urban residents (5.5% of the total urban population of Iowa in 1915) and 33,305 rural residents (1.8% of the total rural population). Figure 2.1 presents a map of the counties and cities included in the Goldin-Katz sample. The three large Iowa cities sampled are Des Moines, Davenport, and Dubuque.[3] In 1915, the population of Des Moines was approximately 97,000 people, making it the 64th largest city in the country. Davenport and Dubuque were smaller, with approximately 46,000 and 39,000 people, respectively. The rural counties in the sample were selected by Goldin and Katz on the basis of both image and archive quality, as well as to provide a diverse geographic sample within the state, as shown in Figure 2.1.[4]

To construct my sample for census matching, I limit the Goldin-Katz sample to families with boys aged between 3 and 17 in 1915. These sons will be between 28 and 42 when I observe them again in 1940, which should reduce measurement issues due to life cycle variability in annual income. I restrict my analysis to sons in 1915, because name changes make it impossible to locate most daughters in the 1940 Census. This leaves me with a sample of 7,580 boys, 6,071 of

[3] The census manuscripts for Sioux City, one of the other large cities in Iowa, was unreadable and not collected by Goldin and Katz (Goldin and Katz, 2000).

[4] For more details on the construction of the Goldin-Katz sample, see Goldin and Katz (2000).

whom have fathers in their households and the requisite data on both the father's education and income. Each of those 7,580 observations is a son in 1915 Iowa.

To locate these sons in 1940, I utilise the 100% 1940 census sample deposited by Ancestry.com with the NBER. I collect the set of possible matches, using the son's first and last name, middle initial (when available), state of birth, and year of birth. Then, I train a record-linking algorithm and use the scores generated by that algorithm to identify the correct matches for each son from 1915 in the 1940 data.[5] Once the matched sons are identified, I record the pertinent data from the 1940 records. The 1940 Census was the first federal census to collect data on incomes, weeks of work, and years of education of the entire population.[6] Because it is a national sample, I do not have to worry about losing many sons to out-migration, which might otherwise bias my estimates.

My match rate is roughly 59%, which surpasses the rates of previous literature linking between censuses.[7] Table 2.2 shows my match rates for the rural and urban samples; match rates

[5] The machine learning approach, which I detail in Feigenbaum (2015), attempts to teach an algorithm how to replicate the careful hand-linking work of a researcher, but at scale and with extreme consistency. I generate the training data used to train the algorithm by manually comparing a subset of sons from 1915 to the set of possible links in 1940 and determining which records are in fact matches for the same person exact. The training dataset is built as a random sample of 30% of the full set of sons in Iowa in 1915. To be considered as a possible link, records must first match exactly on state of birth, be within 3 years difference in year of birth, and within 0.3 Jaro-Winkler string distance in both first and last names. Then, within that filtered list of possible matches, the records are double-entry matched to the 1940 census manually by trained research assistants. Records determined to match are marked as such, records without a clear and certain match in the 1940 census are marked as unmatched. With this corpus of links, I then train a match algorithm. The match algorithm is used to reduce between-researcher variability in match quality, to speed up the matching process, and ensure data replication. The method improves on previous efforts based on phonetic matching because typos and transcription errors will not cripple the matching. The matching algorithm uses Jaro-Winkler string distances in first and last names, exact matches on state of birth, absolute difference in year of birth, Soundex matches for first and last names, middle initial matches, matching first and last letters of first and last names, and other record-based variables to predict whether a record is a true or false match. The algorithm also factors in the match quality of other possible matches for the given record searched for, only making matches when a record is a significantly better match than other possibilities. Based on cross-validated out of sample predictions within my training data, the match algorithm has a true positive rate of nearly 90% and a positive prediction rate of 86%. In Table B.2 of the appendix, I detail the exact match algorithm used.

[6] Past federal censuses record contemporary school enrollment for each person (child), but not years of schooling completed for adults no longer in school. Earnings data was collected in 1940 only for wage and salary workers. The data collected are the "total amount of money wages or salary" but enumerators were instructed: "Do not include the earning of businessmen, farmers, or professional persons derived from business profits, sale of corps, or fees." The importance of this missing data will vary with the fraction of farmers and other business owners in my sample. It does not, however, affect farm labourers, whose earnings are reported the same as other occupations. Of my matched sample, 13.7% of the sons in 1940 are farm owners or operators without income. Initially, I drop these observations with missing earnings data in analyses on income data. However, in Appendix B.3, I impute earnings for farmers using the 1950 census, which did collect data on capital income and non-wage and salary earnings. Using these imputed earnings, I estimate even higher levels of mobility than in my main results.

[7] Parman (2011) reports match rates of just below 50% using hand matching. Guest, Landale, and McCann (1989)

are comparable between the two samples. My sample size of 4,478 father-son pairs is also comparable to many other projects measuring intergenerational mobility, both historically (Long and Ferrie, 2013) and recently (Lee and Solon, 2009).

Table 2.2: Sample Match Rates

	Rural	Urban	Total
Found	2657	1821	4478
	(60.36)	(57.30)	(59.08)
Total	4402	3178	7580
	(100.00)	(100.00)	(100.00)

Any project linking historical data is subject to possible bias due to the difficulty of making matches between datasets. I present an analysis of potential bias in the matching below, which suggests that the final sample does not suffer any crucial construction defects. Simple transcription errors are the most likely obstacle to linking between a son observed as a child in 1915 and as an adult in 1940. To test this, I calculate a number of string- and character-based statistics using the first and last names of the sons in my sample.

First, I determine the name commonness of both the first and last name, relative to all names in the pooled IPUMS sample of the 1910 and 1920 censuses.[8] A more common name is less likely to have a unique match in the 1940 Census, even after limiting the possible targets by state of birth and year of birth. Second, I calculate the length of each son's first and last name. Longer names are more likely to be incorrectly transcribed, but they are also more likely to be distinctive.

Third, I attempt to predict typographical errors using character similarity scores. Cognitive scientists and typographers have studied how likely certain letters are to be mistaken for one another or how similar two letters are visually. For example, readers are much more likely to mix up lower case p and q than they would be p for k. Further, some letters are more likely to be mis-transcribed than others: s is quite visually unique while l and n are both visually similar to other letters.[9] A name with a number of l's or n's in it is more likely to be mis-transcribed and

match at 39.4%. Other attempts at census linking using phonetic codes such as Soundex have lower match rates.

[8] The commonness statistic is measured as the share of 100 people in the pooled 1910 and 1920 sample with the same first (last) name. It ranges from 0.00118 (or roughly 1 person in 100,000 with the same name—these names are unique in my sample) to 1.72 for first names (John) or 1.02 for last names (Miller). Abramitzky, Boustan, and Eriksson (2012) use relative commonness as a predictor of census match success as well.

[9] l is likely to be confused with f and i for example, while n is similar to both h and m.

thus not matched when I search in the 1940 Census.[10] I use a matrix of letter visual similarity from Simpson, Mousikou, Montoya, and Defior (2013) to compute, for first and last names, a similarity score.[11]

Finally, I calculate a name's Scrabble score as an alternative measure of both name commonness and name simplicity.[12] Names with low Scrabble scores are likely to be made up of relatively common characters and are less likely to be changed or Americanised over time (Biavaschi, Giulietti, and Siddique, 2013). I use standardised z-scores for both the visual similarity scores and the Scrabble scores; the z-scores are based on the distribution of visual similarity scores and Scrabble scores within the pooled sample of my Iowa sons and the 1910 and 1920 censuses.

Table 2.3 presents the results from a series of linear probability models, predicting whether or not a son in 1915 is uniquely matched ahead to the 1940 Federal Census. Sons with more common last names are less likely to be matched, while first name commonness has a smaller, positive effect. Sons with longer first names or first names with higher similarity scores are more likely to be found, but both of these effects are quite small.[13] I include controls for all of these name string properties in all subsequent analysis.

More serious issues could be generated by differential matching rates according to father, son, or family characteristics in 1915. In my sample I find little evidence that such characteristics strongly affect the probability of matching. In Table 2.4, I present the estimated effects of a set of variables observed for fathers and sons in 1915 on the probability of positively locating the son in

[10] Recall matches are made using census indices transcribed by Ancestry.com and deposited with the NBER.

[11] Specifically, Simpson, Mousikou, Montoya, and Defior (2013) conduct surveys of college students and other native and non-native English readers to assess the similarity of letters on a 7 point scale, where 7 indicates exactly the same and 1 extremely different. For example, *i* and *l* have a similarity score of 6.13, while *w* and *t* have a similarity score of exactly 1. I take the highest (non-self) similarity score for each letter as a measure of a letter's likelihood of being mis-transcribed. Figure B.1 in the appendix graphs these scores for each letter. Then, I calculate the average of these scores for all letters in a given string (name). The scores from Simpson et al are based on both lower case and upper case letters in block type. As many of the Census files are in script, a visual similarity matrix for cursive letters would be ideal, but such a measure does not exist in the typography literature. As a robustness check, I also use a letter matrix of confusion probability from McGraw, Rehling, and Goldstone (1994) and find a high correlation between each letter's similarity score.

[12] Biavaschi, Giulietti, and Siddique (2013) introduce the use of Scrabble scores into the economic literature. They use this measure to predict name changes by immigrants to the United States during the early twentieth century. Scrabble point values were based, originally, on the frequency of letters on US newspaper front pages.

[13] With controls for commonness and length, the Scrabble scores do not seem to relate to match rates.

Table 2.3: Probability of Matching a Record from Iowa 1915 to the Federal Census 1940

	(1)	(2)	(3)	(4)	(5)
Name commonness, first	0.041**				0.056***
	(0.017)				(0.020)
Name commonness, last	−0.122***				−0.121***
	(0.039)				(0.039)
String length, first		0.013***			0.020***
		(0.004)			(0.004)
String length, last		−0.002			−0.002
		(0.004)			(0.004)
Normalized letter similarity, first			0.019***		0.024***
			(0.007)		(0.007)
Normalized letter similarity, last			0.006		0.005
			(0.007)		(0.007)
Normalized scrabble score, first				−0.001	−0.002
				(0.006)	(0.007)
Normalized scrabble score, last				0.009	0.008
				(0.006)	(0.006)
Observations	7580	7580	7580	7580	7580
Clusters	4731	4731	4731	4731	4731
Adjusted R^2	0.002	0.002	0.001	0.000	0.007

Linear probability model with an indicator variable for a successful match as the outcome. Standard errors are clustered by family. Results are consistent using a probit or logit model as well. Name commonness is measured as the share of 100 men in the 1910 and 1920 IPUMS sample with the same first or last name. Name length is the number of characters in the first or last name. Name similarity scores are based on character typology similarity from Simpson, Mousikou, Montoya, and Defior (2013).

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

Table 2.4: Effects of Family Covariates on the Probability of Matching Records from 1915 to 1940

			Predicted Match Rate with X at	
X	β	SE	25th Percentile	75th Percentile
Father Log Earnings	0.013	0.011	59.6	60.6
Father Education	0.004	0.002	59.2	60.0
Mother Education	0.003	0.003	59.8	60.3
Urban in 1915	-0.034	0.012	60.5	57.1
Son Born in IA	0.138	0.018	61.0	61.0
Father Foreign Born	-0.063	0.013	61.2	54.8

This table presents the coefficients from a series of linear probability regressions with X as the primary independent variable, controlling for first and last name commonness, length, letter similarity, and Scrabble score. As in Table 2.3, there are 7580 observations and 4731 clusters, clustering standard errors by family.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

the 1940 Census.[14] Each row in the table is a separate linear probability regression, reporting the coefficient of the listed X variable while controlling for first and last name commonness, length, letter similarity, and Scrabble score. I am slightly more likely to match sons who had higher income or more educated fathers (or mothers) in 1915, but these effects are both economically and statistically insignificant. For example, the probability of matching a son with a father at the 25th percentile of income is only 1 percentage point lower than matching a son with a father at the 75th percentile of income. Similarly small effects of both father's and mother's education can be seen as well. Confirming the results presented in Table 2.2, I am also less likely to match sons in the urban sample. I am also more likely to link sons born in Iowa, even after conditioning on name string characteristics.[15] All analysis undertaken in this paper will include controls for son's place of birth, place of residence in 1915, and, where appropriate, father's place of birth.

The first two columns of Table 2.5 present summary statistics for the fathers of children between 3 and 17 in the Goldin-Katz Iowa State Census sample. Observation counts are smaller than my full sample because fathers of multiple children or sons are not double counted. The fathers in the sample are restricted to fathers of sons between 3 and 17; fathers found are the father for whom sons were located in the 1940 Census through Ancestry.com. Average yearly earnings for the fathers are approximately $1000 in 1915 dollars. The average father had a half

[14] Results in this matching exercise are robust to alternative regression models, including logit and probit models. I use a simple linear probability model for ease of interpretation.

[15] 86% of the sons in my sample were born in Iowa so there is no difference between the 25th to 75th percentile for that covariate.

Table 2.5: Summary Statistics: Fathers in 1915 and Sons in 1915 and 1940

	Fathers		Sons	
	Fathers in Sample	Found Fathers	Sons in Sample	Sons Found
Yearly Earnings	1005.6	1007.1		1417.1
	(591.5)	(587.6)		(818.9)
Log Yearly Earnings	6.743	6.747		7.059
	(0.604)	(0.597)		(0.716)
Log Weekly Earnings	2.858	2.859		3.250
	(0.566)	(0.563)		(0.633)
Years of Education	8.491	8.507		10.40
	(2.837)	(2.803)		(3.068)
Age (1915)	41.92	41.73	9.740	9.605
	(9.262)	(9.307)	(4.350)	(4.357)
Born in Iowa	0.473	0.505	0.859	0.886
	(0.499)	(0.500)	(0.348)	(0.318)
Urban (1915)	0.452	0.436	0.419	0.415
	(0.498)	(0.496)	(0.493)	(0.493)
Observations	3713	2204	7580	3971

All summary statistics are based on those fathers and sons with complete data for all listed variables. This restriction reduces the number of observations from the count of all sons found, as presented in Table 2.2. The sample fathers include only men with sons between the ages of 3 and 17 in 1915. The found fathers are only those men with sons matched into the 1940 census. All sons includes any boys aged 3 to 17 in the Iowa sample in 1915; the found sons are only those boys linked from 1915 to 1940. For fathers, earnings, education, age, and urban status are measured in the 1915 Iowa State Census. For sons, earnings and education are measured in the 1940 Federal Census, while age and urban status are measured in the 1915 Iowa State Census.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

year more than a common school education (eight years) and was approximately 42 years old in 1915. Of the fathers in my sample, those fathers for whom I matched a son into the 1940 Census earned very slightly more, though not significantly so, measured either in levels, logs, and weekly earnings.

The final two columns of Table 2.5 present summary statistics for the Iowa sons in my sample. Only summary data for sons with complete information in the 1940 Federal Census is reported in the table, which lowers the number of observations in the final column to 3,971. The located sons earned more than $1400 in 1940, which is roughly the same in real terms as the average earnings for their fathers in 1915; the lingering effects of the Great Depression may have reduced any real income gains overall.[16] Also notable in the summary statistics is the fact that the sons

[16] I measure all dollar amounts in this paper in nominal terms. Because I use logged earnings in my regressions and income for all sons is measured in 1940 and for all fathers in 1915, any nominal to real conversions drop out into the unreported constant term.

Table 2.6: Summary Statistics: Sons in 1940

	Iowa Linked Sample	1940 IPUMS Sample	
	Sons Found	Unweighted	Weighted by State of Birth
Yearly Earnings	1417.1	1237.2	1256.1
	(818.9)	(889.4)	(898.3)
Log Yearly Earnings	7.059	6.819	6.850
	(0.716)	(0.887)	(0.863)
Log Weekly Earnings	3.250	3.102	3.099
	(0.633)	(0.758)	(0.758)
Years of Education	10.40	9.151	10.30
	(3.068)	(3.482)	(2.974)
Age (1940)	34.58	34.65	34.76
	(4.419)	(4.337)	(4.377)
Married	0.815	0.804	0.815
	(0.389)	(0.397)	(0.388)
Born in Iowa	0.886	0.0190	0.886
	(0.318)	(0.136)	(0.318)
Share White	0.994	0.902	0.992
	(0.0775)	(0.297)	(0.0891)
Observations	3971	112309	99105

All variables measured in 1940. After weighing the 1940 1% IPUMS sample of the census for state of birth, the matched sample of sons is comparable to the 1940 population of men between 28 and 42.

Sources: 1940 Federal Census

had on average two more years of schooling than their fathers. This is a striking example of the effect of the high school movement and the expansion of public education in Iowa, previously documented by Goldin and Katz (2008) and Parman (2011).

How does my sample compare to the rest of the US in 1915 or in 1940? In Table 2.6, I compare the sons in my sample to the national population in 1940, focusing on men aged 28 to 42. The second column is the IPUMS 1% sample of the 1940 census; the third column reweights the 1% sample by state of birth to match the states of birth in my Iowa sons sample. I find that my sample has slightly higher income than either the general population or the reweighted sample. After reweighting, however, my sample is representative in terms of education, age, marital status, and race.[17]

[17] In a later analysis of geographic mobility, I show in Table 2.12 that my sample's state of residence distribution in 1940 is roughly similar to the geographic distribution of all men born in Iowa between 1898 and 1912.

2.3 Past Estimates of Intergenerational Mobility

2.3.1 Intergenerational Income Mobility

The most frequent measure of intergenerational economic mobility used in the literature is the intergenerational elasticity of income (IGE), estimated by regressing the son's log income against his father's log income. Corak (2006), Solon (1999), and Black and Devereux (2011) present thorough reviews of the IGE literature.[18] These reviews all indicate a lack of historical data on intergenerational mobility: Corak (2006) documents 41 studies of the US IGE, none of which presents data before 1980. One aim of my project is to establish a correct measure of IGE well before the period previously studied in this literature.[19]

Corak's preferred measure of IGE in the US is 0.47,[20] in line with the reviews presented by both Solon and Black and Devereaux. Corak (2006) also documents large variations between US studies measuring the intergenerational elasticity of income.[21]

The IGE imposes a very particular functional form relationship between son's and father's

[18] The estimated elasticity of income between one generation and the next is commonly referred to as an IGE and I will use that abbreviation here.

[19] Aaronson and Mazumder (2008) take a different tack when measuring intergenerational income mobility. They use successive waves of the US federal census, from 1940 to 2000, and construct synthetic parents for observed individuals. They find low levels of mobility in 1940, but more mobility each decade until 1980. Mobility falls again in 1990 and 2000. However, the parents are constructed only using state of birth, age, and race; thus rather than regressing the son's income on the father's, they regress the son's income on the average income of same race men in the son's state of birth. While that is a possible proxy for father's income, it does not seem sufficiently detailed or granular to detect small shifts in the intergenerational transmission of income.

[20] Corak also gives lower and upper bounds of between 0.40 and 0.52. Mazumder (2015) argues for a parameter of 0.6

[21] See, for example, the first table in Corak's appendix. IGE estimates in the literature range from 0.09 to 0.61. Because of this variation scholars have focused on measuring changes over time in IGE within one consistent dataset. However, the results in this literature have also been rather inconsistent. Mayer and Lopoo (2005) use the PSID and collect a sample of 30 year olds, regressing son's income at age 30 on a three year average of father's income. They find a large and statistically significant downward trend in the IGE, suggesting that mobility has increased significantly in the last several decades. Levine and Mazumder (2002) present more mixed results in work using the NLS, GSS, and PSID. Levine and Mazumder observe sons between the ages of 28 and 36 in 1980 and again in 1990. They find increasing mobility in the PSID, but decreasing mobility in the NLS and GSS. Lee and Solon (2009) argue that past work has been plagued by non-classical measurement error. To correct this, they argue that rather than observing the son's outcomes once or twice and throwing away the rest of the data, researchers should make use of the full sample. Drawing on PSID data for cohorts of sons and daughters born between 1952 and 1975, Lee and Solon do just that. I focus on the Lee and Solon results for fathers and sons to keep in line with the analysis I am able to perform in my data. Controlling for a quartic in the ages of both parents and children, they only limit the sample to sons between 25 and 48. They find a simple average IGE of 0.44 over the period and no statistically significant trends in IGE between 1976 and 2000. Previously, I compared my estimated IGE parameter for income to the results in Lee and Solon's paper (see Table 2.1), showing that my estimates of mobility for the early twentieth century were higher than the contemporary estimates.

earnings.[22] Chetty, Hendren, Kline, and Saez (2014) present evidence from recent US tax data that suggests this assumption is false; Corak and Heisz (1999) show the same with Canadian tax data. At both tails of the income distribution, the linear relationship between father's log income and son's log income breaks down. Following Dahl and DeLeire (2008), Chetty, Hendren, Kline, and Saez (2014) and Chetty, Hendren, Kline, Saez, and Turner (2014) estimate a rank-rank parameter of intergenerational mobility, regressing the son's income percentile (within his cohort) against the father's income percentile (within his cohort). The graphical evidence presented in Chetty, Hendren, Kline, and Saez (2014) suggests that the implied linearity in the rank-rank specification is a more accurate fit of the data. For their sample of the US, Chetty, Hendren, Kline, and Saez (2014) estimate a rank-rank parameter of 0.341 overall and 0.336 for sons. However, similar to the IGE literature, there are few historical estimates of the rank-rank parameter: Chetty, Hendren, Kline, Saez, and Turner (2014) plot trends in mobility for sons born between 1971 and 1993, but they cannot extend their sample farther back in time.

I address three issues—measurement error, life-cycle bias, and sample homogeneity—to attempt to produce unbiased and consistent estimates of mobility with my historical sample that I can compare to contemporary estimates of mobility.[23] I will estimate β, the intergenerational elasticity of income:

$$y_i = \beta X_i + \epsilon_i$$

where y_i is the child's adult income and X_i is the parent's adult income.

First, measurement error may arise because, rather than permanent income, I observe only a single year of income for both generations.[24] If a single year of income is just a noisy proxy for permanent income, this biases coefficients towards zero. While I cannot add additional years of income observation to my historical data, I do limit the contemporary data to a single year of

[22] Specifically, a linear relationship between log father's income and log son's income.

[23] Though the literature has considered these measurement issues with respect to calculating the IGE parameter, in theory these same problems could bias the rank-rank parameter as well. However, Chetty, Hendren, Kline, and Saez (2014) document the general robustness of the rank-rank estimtes to single year samples, variation in father and son ages, and included control variables.

[24] Even standard permanent income may be the wrong measure. The parental income that matters for children's outcomes may be income while the child is in utero, young, or in school, depending on one's reading of the critical years literature.

observation as a parallel.

Second, Haider and Solon (2006) raise a more complex problem owing to trends in life cycle earnings. The literature on life cycle earnings suggests that people with higher permanent income experience more earnings growth earlier in their career. Depending on the relative life cycle bias for fathers or sons, this can amplify or attenuate the estimate of β (Mazumder, 2015). By drawing my sample from a set of fathers and sons with average ages of 40 and 35, I aim to minimize this problem. In addition, I shape the contemporary comparison samples in similar ways to maintain comparability.

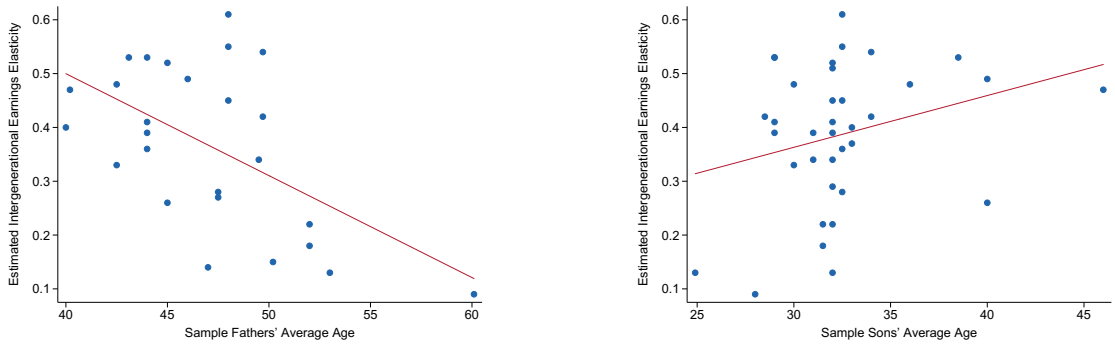
Third, Solon (1989) argues that more homogenous samples will tend to bias down IGE estimates. Though Iowa in 1915 was relatively representative of the nation, it is far more homogenous in terms of race than the country is today. Although it is impossible make my sample more racially heterogeneous, I can construct a racially homogenous sample using recent data and compare it to my earlier Iowa samples.

One particular consequence of the sample age bias is shown in Figure 2.2, based on the American IGE literature surveyed by Corak (2006). As the figure on the left shows, the older the average age of the fathers in the study samples, the lower the estimated intergenerational elasticity. The figure on the right shows a similar but weaker relationship holding in the opposite direction between estimated IGEs and the average age of sons in the sample.[25]

Parman (2011) also draws on data from the 1915 Iowa State Census. He matches adult men in the Iowa sample backwards in time to the 1900 Federal Census to construct childhood households. Though most are heads of household in 1915, these are Parman's "sons" in the analysis. The reconstructed households yield the name, state of birth, and other demographic characteristics of the "fathers" in 1900. Parman then matches these fathers forward in time to the 1915 sample, thus observing both fathers and sons in Iowa in 1915 and estimating IGEs based on income reported in the Iowa State Census. Parman finds an IGE of approximately 0.11 for all father-son pairs or 0.17 for non-farmer father-son pairs. These low estimates paint a picture of

[25] Both of these best fit lines are statistically significant in the univariate regression, but the relationship between father's age and estimated IGE is much stronger. The points graphed in Figure 2.2b suggest instead that with sons ages ranging from approximately 30 to 35, the estimated IGE should not be a function of the data sample.

Figure 2.2: Estimated IGE Affected By Sample Fathers' and Sons' Ages



(a) Correlation between estimated IGE and sample fathers' ages

(b) Correlation between estimated IGE and sample sons' ages

The average ages of both fathers and sons sampled may bias the estimated intergenerational mobility elasticity. Each point in the two scatterplots represent the estimated elasticity of income from studies of American intergenerational mobility reviewed by Corak (2006), plotted against the average age of either fathers or sons in the samples.

high levels of mobility in the early twentieth century.[26]

However, Parman's results are constrained by data limitations in two important ways.[27] First, all income data in Parman's study are drawn from a single state census in one year. Any fathers or sons leaving the state between 1900 and 1915 are omitted from the final dataset. The direction and magnitude of this bias is *a priori* unclear. If the moving sons are more likely to be unlike their fathers, father-son pairs with weaker relationships between their outcomes will be removed from the sample, thereby biasing up the IGE estimates. In this case, the early twentieth century may have been even more mobile than Parman suggests. Alternatively, the selection of which sons leave the state may make the final sample more homogenous; following the arguments in Solon (1989), this biases IGE estimates downward. Generally, the uncertain selection of out of state movers in Parman's sample complicates the interpretation of his results.

Figure 2.2 demonstrates the other major constraint on the interpretation of the IGE results in Parman (2011). The average age of fathers in Parman's sample is between 57 and 65, depending on the particular specification. This age range is on the far right tail of the IGE studies in the literature and very likely to present a very low IGE, due to life-cycle-induced measurement errors

[26] These estimates are similar but much lower than the income IGE estimates I will present later in this paper.

[27] Full access to the 1940 Federal Census, including all citizens names, was not available until April 2012, well after Parman completed his research.

(see Haider and Solon (2006) for the detailed econometric treatment of this issue). The average age of the sons in Parman's sample is between 25 and 30, and this may also, to an extent, bias his results towards a very low IGE. However, the strength of the bias based on the relationship presented in Figure 2.2b is less clear.

Thus, while the results in Parman (2011) suggest that income IGE was very low and that income mobility was very high in Iowa in 1915, data constraints complicate the comparison of the estimated IGE to other time periods and places.

2.3.2 *Intergenerational Education Mobility*

Due to data constraints, there has been little work on educational mobility in the US historically.[28] Hertz, Jayasundera, Piraino, Selcuk, Smith, and Verashchagina (2007) present the most comprehensive measures of intergenerational elasticity of education across many different countries and regions. They find an IGE of education for the US of 0.46, suggesting more mobility of education in the US than in South America (0.60) but less than in Western Europe (0.40).[29] Outside of the US, Checchi, Fiorio, and Leonardi (2008) study Italian cohorts born between 1910 and 1970 and find very high IGEs and low mobility of education in their early samples, relative to the recent period.

Though measurement error is a concern for education, as it was with income, the specific concerns are quite different. Education observed in a given year during adulthood is a much better proxy for permanent education.[30] However, other measurement error issues might include faulty reporting of years completed and artificial heaping at milestone numbers (four years of high school or college, for example). Kane, Rouse, and Staiger (1999) have shown how non-classical measurement error in years of education can lead to potential biases of Mincerian regressions on the return to education.[31]

[28] Parman (2011) measures the effects of public education on income mobility, but does not estimate father to son educational mobility directly.

[29] The use of the IGE term and elasticity more generally is a bit of an abuse of notation. The IGE literature on education estimates these parameters using levels on levels, rather than log on log.

[30] In my sample, it is very unlikely any of the fathers or sons continued education beyond when I observe them in their 30s or 40s.

[31] The inability to measure only quantity rather than quality of education is also a potential issue. It may be the case that years of education is simply a noisy signal of true education. Any noise will bias the estimated mobility

2.3.3 *Intergenerational Occupational Mobility*

Due to data limitations, the study of historical intergenerational mobility has focused on the study of occupational mobility. Early work on this topic is Thernstrom (1964, 1973), studying the occupations of successive generations in Boston and Newburyport, MA. Thernstrom tends to find quite high upward mobility, but a lot of white collar stability as well. Duncan (1965) finds more upward and less downward mobility in 1962 relative to the occupational transition matrices of 1952, 1942, or 1932, relying on data gathered from Occupational Changes in a Generation (OCG). However, neither Duncan nor any of the subsequent work based entirely on the OCG data is able to measure occupational mobility for earlier periods.

Guest, Landale, and McCann (1989) compare a nineteenth century sample, built by matching fathers and sons in the 1880 to 1900 censuses, to the OCG. They find less upward mobility and more occupational inheritance in the nineteenth century. However, for fathers and sons who are not farmers, the association is both economically smaller and statistically weaker. The results depend a great deal on where Guest et al. put farmers in the occupational distribution.

To avoid the fraught issue of how to rank occupations—especially without available average income, education, or wealth data by occupation—the economics literature has turned to occupational transition matrices, which are agnostic about movements up or down the occupational ladder and instead focus only on movements by the son out of the father's occupational category. In particular, Altham and Ferrie (2007) present the Altham statistic, which has become the standard measure of intergenerational occupational mobility in economics. To compute these measures of occupational mobility, fathers and sons are each grouped by occupation into one of four broad categories—farmer, white collar, skilled and semi-skilled labour, and unskilled labour—within an occupation transition matrix. The Altham statistic measures the strength of association between both the rows and columns of a transition matrix and between any two matrices. Altham statistics can be defined for any two matrices. Specifically, let both P and Q be $r \times s$ matrices with elements p_{ij} and q_{ij}. Then the Altham statistic is:

$$d(P,Q) = \left[\sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{l=1}^{r} \sum_{m=1}^{s} \left| \log \left(\frac{p_{ij} p_{lm} q_{im} q_{lj}}{p_{im} p_{lj} q_{ij} q_{lm}} \right) \right|^2 \right]^{1/2}$$

coefficient downwards, but to the extent that education quantity is always a noisy measure of education quality, this does not present a strong challenge to comparing results over different time periods.

Altham and Ferrie (2007) use the $d(P, Q)$ notation to convey the sense in which the Altham statistics are distance measures.[32] $d(P, I)$, where I is the occupation transition matrix of perfect mobility (that is, a matrix with ones in all rows and columns), can be used as a measure of distance from independence.

One of the strongest criticisms of using occupations to study long-term trends in intergenerational mobility is the difficulty in classifying farmers (Xie and Killewald, 2013).[33] Comparison of mobility measures across time is complicated—perhaps even driven—by the secular movement out of agriculture in the US. For example, Guest, Landale, and McCann (1989) conclude that there was more social mobility in the post-WWII period in the US than there had been in the nineteenth century, but they suggest that this reflects the high-heritability of farming and the declining shares of farmers since the late nineteenth century. In this paper, I attempt to control for that by comparing relatively homogeneous samples over time, particularly by constructing a sample in the PSID or other contemporary data that is as rural, white, and agricultural as my Iowa sample. I also focus more analysis on the urban Iowa sons, almost none of whom had farmers for fathers or became farmers themselves.

The second problem posed by farmers is their extreme distribution of earnings. In the standard census sources, including both the 1915 Iowa State Census and the 1940 Federal Census, an individual classified as a farmer may be a small scale tenant farmer, renting his land and equipment and working a small plot. However, owners of very large farms are also classified simply as farmers. It is quite possible for a father and son who are both farmers to have very different incomes. Similarly, a shift between father and son from farming to another occupational category may represent an increase or a decrease in income. Further, to what extent is the wide variation in income among farmers driven by measurement error or transitory income shocks (annual weather shocks, for example)?

When considering an intergenerational sample drawn from a population with a large share of farmers, measures of income mobility and occupational mobility may diverge. In this paper,

[32] As a distance measure, Altham statistics satisfy the triangle inequality. For any three $r \times s$ matrices A, B, C, it is true that $d(A, C) \leq d(A, B) + d(B, C)$.

[33] Income measures for farmers, when available, are not a panacea either. To the extent farmers are engaged in subsistence farming, income will be a poor measure between generations.

I measure both income and occupational mobility, as well as educational mobility, for such a population.

2.4 *Intergenerational Mobility Estimates for Iowa 1915 to 1940*

In Figure 2.3, I present the raw correlations between fathers' and sons' outcomes: log annual income (2.3a), years of education (2.3b), and occupation score based on 1950 scores (2.3c) and on 1915 scores (2.3d).[34] There is a strong positive relationship between outcomes for fathers in 1915 and sons in 1915, but the exact measure of the respective slopes of these lines—and how those slopes compare with the estimates of mobility from studies of the recent period—is the key question of this paper. In the sections that follow, I estimate the intergenerational mobility of income, education, and occupation, documenting that there was indeed more mobility historically than there is today.

2.4.1 *Intergenerational Mobility of Income*

I measure mobility in two primary ways. First, following the intergenerational mobility literature, I use intergenerational elasticities (IGE) (Corak, 2006; Solon, 1999; Black and Devereux, 2011). The canonical formulation regresses the son's adult outcome, in my case as measured in the 1940 Census, on the father's adult outcome, as measured in the 1915 Census. Let Y_i be the outcome of interest, either (log) income or education. The model I estimate can be summarised as:

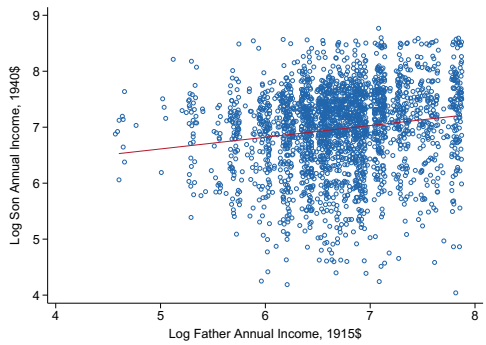
$$Y^s_{i,1940} = \alpha + \beta \cdot Y^f_{i,1915} + Q^s(age^s_{i,1915}) + Q^f(age^f_{i,1915}) + Q^s(age^s_{i,1915}) \times Y^f_{i,1915} + \epsilon_i$$

β can be thought of as a persistence parameter: larger estimates mean a tighter link between father and son and thus less mobility.

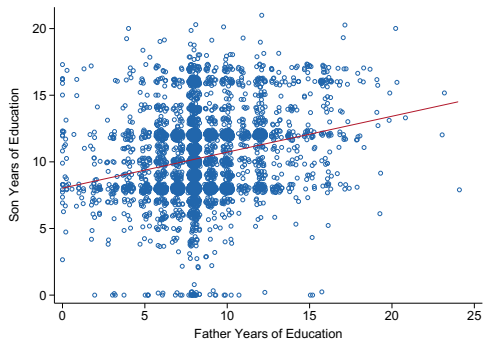
Second, following Dahl and DeLeire (2008) and Chetty, Hendren, Kline, Saez, and Turner (2014); Chetty, Hendren, Kline, and Saez (2014), I also use rank-rank estimates. Again, I regress the son's outcomes on the father's outcomes, but where outcomes are the relative positions or

[34] Naturally, there are many father-son pairs with the same outcome levels as other pairs. In an attempt to display this density at certain points on the graphs, I have used both hollow scatterplot markers and jittered the data. The best fit lines are, of course, drawn based on the full sample before jittering.

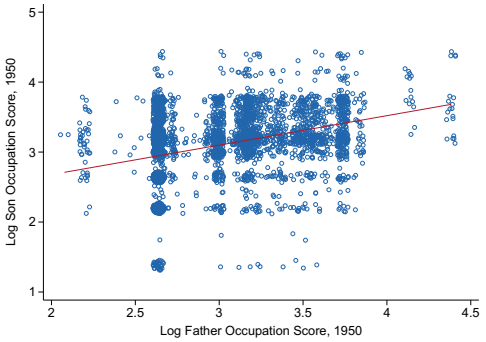
Figure 2.3: Estimated IGE Affected By Sample Fathers' and Sons' Ages



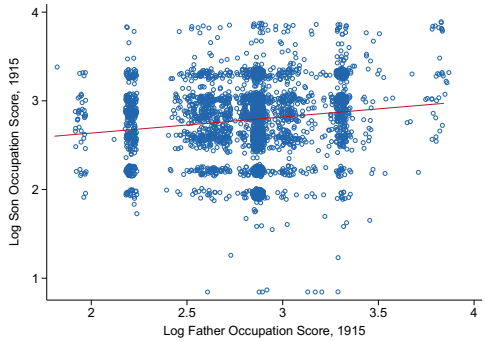
(a) Intergenerational Mobility of Income



(b) Intergenerational Mobility of Education



(c) Intergenerational Mobility of Occupation, 1950 Basis



(d) Intergenerational Mobility of Occupation, 1915 Basis

percentiles in the income distribution. For sons, observed in 1940, I use the full 1940 IPUMS census sample to calculate the full income distribution of white men, aged 28-42, matching the demographics of my sample. For fathers, income data are not available for a nationally representative sample. I instead calculate the full income distribution of white men in the Goldin-Katz Iowa 1915 census sample with the same age range as the fathers in my sample. Ranks are scaled as percentiles between 0 and 1; a rank of .5 indicates that the father or son is at the median for annual income.

To reduce any measurement error induced by life cycle income effects, I follow Lee and Solon (2009) and include quartic age controls for both the father and the son, defined as Q^s and Q^f above, as well as an interaction between the son's age and the father's outcome. In the interaction term, I normalise son's age in 1940 relative to age 40 (Haider and Solon, 2006).[35] The fact that I define my sample to observe sons between ages 28 and 42 in the 1940 Census also reduces life cycle driven measurement error. As some of my observed sons are brothers (and therefore have the same father data), I cluster standard errors at the family level. I also include county fixed effects (which subsumes a control for urban or rural sample). The results are robust to the inclusion of controls for family-size effects, county fixed effects, and the name string control variables described previously.[36]

Panel A of Table 2.7 presents my estimates of the IGE of income across a variety of samples. Both father's and son's incomes are measured as annual log earnings.[37] The first specification is a simple univariate regression of son's log earnings on father's log earnings. In specification

[35] With this normalization, the estimated β represents the relationship between son's and father's outcomes when the son is age 40. I follow Lee and Solon (2009) in normalizing to 40.

[36] The county fixed effects indicate the county of residence when the son is observed in Iowa in 1915. The name string controls include first and last name commonness, length, letter similarity, and Scrabble scores, all attempts to control for differential matching rates between the 1915 and 1940 censuses. While I include these various controls to reduce measurement error, both Chetty, Hendren, Kline, and Saez (2014) and Nybom and Stuhler (2014a) present extensive results that suggest the rank-rank measures of intergenerational mobility are much less susceptible to biases. Working with the universe of US tax records, Chetty, Hendren, Kline, and Saez (2014) show that estimates are stable even with just one year of income observed for both fathers and sons. Further, they document that the exact age when fathers or sons are observed has very little effect on the measurement of mobility, so long as the fathers are observed between the ages of 30 and 55 and the sons are observed after age 30. Nybom and Stuhler (2014a) replicate these lessons for the estimation of rank-rank mobility using Swedish data. The stability of my estimates of rank-rank mobility with and without various controls suggests that the rank-rank parameter is quite robust in my historical sample as well.

[37] To ensure comparability with contemporary estimates, I use annual earnings, not weekly earnings. Results using weekly earnings are similar and in fact lower than those presented in Table 2.7, suggesting even more mobility in the early twentieth century.

two, I include controls for name string properties that might affect matching, 1915 county of residence fixed effects, and quartic controls in father and son age. In the third specification, I also include an interaction between son normalised age and father log earnings to control for lifecycle measurement error (Haider and Solon, 2006).

My baseline estimates for the IGE parameter for the full sample of Iowa fathers and sons range from 0.199 to 0.258, as shown in the first row of Table 2.7. The literature suggests an IGE of 0.47 for income in the United States today (Corak, 2006). Lee and Solon (2009) argue that the IGE of income has been roughly stable for cohorts observed between the late 1970s and the early 2000s. My results suggest that this recent stability does not extend historically, and that there was much more intergenerational mobility of income in the early twentieth century US than there is today.[38]

Similar to my IGE results, I find much more income mobility historically than today. The rank-rank parameter ranges from 0.169 to 0.219 in the first row of Panel B of Table 2.7, the main sample with all linked father-son pairs. Chetty, Hendren, Kline, and Saez (2014) measure a rank-rank parameter of 0.341; among just male children, they find a rank-rank estimate between 0.307 and 0.317.

However, any measurement error will tend to bias down estimates of intergenerational mobility (Solon, 1999). Further, though Iowa is broadly representative of the US in 1915 (Goldin and Katz, 2000, 2008), the differences in my estimated IGE may reflect differences between Iowa and the rest of the country, not differences between time periods. In fact, according to contemporary data, children born in Iowa are among the most economically mobile in the entire country, across many measures (Chetty, Hendren, Kline, and Saez, 2013). To account for these threats to my analysis, I construct a sample of recent intergenerational data that is demographically comparable to my Iowa sample, drawing on data from the PSID. To do this, I limit the PSID to include only white father-sons pairs (99% of my linked Iowa sample is white). I also limit the PSID to

[38] One concern with the results presented thus far is the reliance on the log transformations of the income data. By logging income, the assumption made is that small changes in income for very poor fathers have much higher returns (to son's income) than smaller changes farther up the income distribution. In the appendix, I show that these results are robust to alternative transformations of the father's and son's income variables, including both levels and square roots.

Table 2.7: Intergenerational Mobility Estimates

	Specification				
	(1)	(2)	(3)	Observations	Clusters
A. Intergenerational Elasticity (IGE)					
Full Sample	0.209	0.199	0.258	2041	1669
	(0.032)	(0.031)	(0.081)		
PSID Iowa-Like Sample	0.330	0.350	0.502	3449	346
	(0.056)	(0.080)	(0.166)		
Urban Sample	0.287	0.275	0.310	1004	824
	(0.045)	(0.050)	(0.102)		
Rural Sample	0.156	0.168	0.233	1037	845
	(0.040)	(0.041)	(0.113)		
Excluding Sons of Farmers	0.302	0.259	0.391	1454	1201
	(0.037)	(0.038)	(0.095)		
Including Sons with Imputed Income	0.162	0.156	0.217	2866	2162
	(0.021)	(0.022)	(0.050)		
Sons Remaining in Iowa	0.148	0.148	0.173	1185	1002
	(0.043)	(0.043)	(0.121)		
B. Intergenerational Rank Rank Parameter					
Full Sample	0.173	0.167	0.220	2041	1669
	(0.022)	(0.022)	(0.046)		
PSID Iowa-Like Sample	0.258	0.240	0.323	3680	356
	(0.049)	(0.064)	(0.084)		
Urban Sample	0.217	0.206	0.212	1004	824
	(0.032)	(0.034)	(0.071)		
Rural Sample	0.142	0.152	0.230	1037	845
	(0.028)	(0.028)	(0.061)		
Excluding Sons of Farmers	0.232	0.201	0.258	1454	1201
	(0.026)	(0.027)	(0.059)		
Including Sons with Imputed Income	0.141	0.137	0.183	2866	2162
	(0.016)	(0.016)	(0.031)		
Sons Remaining in Iowa	0.125	0.130	0.188	1185	1002
	(0.028)	(0.027)	(0.059)		

Standard errors clustered by family in all regressions. In Panel A, son's annual log earnings in 1940 is the dependent variable. In Panel B, the son's rank in the income distribution in 1940 is the dependent variable. The income distribution in 1940 calculated using the 1940 IPUMS 1% sample. Specification 1 is a univariate regression of son's outcome on father's outcome (log earnings or income rank). Specification 2 adds name string controls, 1915 county fixed effects, and quartic controls in father and son age. Specification 3 adds an interaction between father's outcome and son's normalised age. Name string controls: first and last name commonness, length, letter similarity, and Scrabble scores. Son's ages are normalised relative to age 40 in 1940.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

sons who grew up in the Midwest.[39] The results are presented in the second rows of each panel in Table 2.7 (Panel A for IGE, Panel B for rank-rank).

To calculate a comparable contemporary IGE, rather than follow Lee and Solon (2009) and measure the father's income as the average of his income when the matched son is between 15 and 17 years old, I use the father's income when his son is 10.[40] In doing so, I attempt to replicate the noise in my historical data from only observing income once. The son's income is observed in each year that the son is in the PSID and is between the ages of 28 and 42, to match my 1940 census data. Both income variables are measured in 2000$ and logged, so as to interpret the estimated coefficients as intergenerational elasticities.[41] Limiting the PSID to sons born in the Midwest, I estimate an IGE between 0.33 and 0.50, depending on the use of state fixed effects and age controls.[42] Though the Iowa-like samples in the PSID are quite small, the results suggest that the lower mobility I find historically is driven neither by the demographic composition of my data nor by the single year measurements of income.

In the second row of Panel B of Table 2.7, I measure the rank-rank mobility using the PSID sample. For the Midwest sample, I measure a lower parameter than is found nationally; however, these estimates are far larger than what I find historically in the full sample. For an alternative construction of a comparable recent rank-rank parameter, I use the county level results reported by Chetty, Hendren, Kline, and Saez (2014) in Online Data Table 3. When I calculate the weighted average of rank-rank mobility, weighing by the shares of my sample living in each county in 1915, I find a contemporary parameter of 0.31, similar to the result from the contemporary PSID data and, more importantly, far larger than the rank-rank parameter of 0.169 to 0.219 that I find in my historical sample.[43]

---

[39] The Midwest region is defined in the PSID as Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, and South Dakota. I do not limit the PSID sample just to sons raised in Iowa as there are only 385 father-son pairs with the requisite data.

[40] I use 10 because this is the midpoint of my age range for sons in the 1915 sample. If I do not observe a father in the year when his son is 10, I use the year when the son is closest to 10 in the PSID sample.

[41] While I attempt to match my age and county fixed effects from my Iowa sample results with age quartics and "grew up" fixed effects, I do not observe either family size or name strings in the PSID.

[42] Given the litany of measurement concerns in the IGE literature, Specification 3, which includes the recommended controls, is likely the best measure of the Midwestern IGE.

[43] The exact weighted average of the contemporary data is 0.3097. I can also split the sample between the urban and rural counties in my analysis. The weighted average of rank-rank mobility is 0.3538 in the three urban counties and

In the appendix, I test the degree to which either false matches in my linking procedure between censuses or higher levels of measurement error in historical data could account for my estimates of lower IGE parameters (and thus higher mobility) in the 1915 to 1940 sample relative in the contemporary sample. I introduce both mismatches and measurement error into my Iowa-like PSID sample considered above. Simulation tests suggest that neither source of error is likely to account for the differences in estimated IGEs. The share of false matches would have to approach 50% for mismatching to account for the estimated differences in IGE parameters, which seems highly unlikely.[44] As detailed in the data section, the matches were carefully constructed based on first and last names, year of birth, state of birth, and gender. In addition, the measurement error simulations suggest that earnings measures from the 1915 and 1940 censuses would have to be considerably noisier than earnings measured in the recent period to generate the large difference in IGEs.

Are the higher rates of mobility that I find historically driven by the secular movement off the farm in the early twentieth century?[45] To answer this question, I compare differential mobility for both sons of rural and urban Iowa, splitting the sample according to where the sons were living when their fathers were first sampled in the 1915 Iowa State Census.[46] The results for these subsample analyses are presented in the third and fourth rows of Table 2.7. Only 15 of the urban sons has a father farmer in 1915 and only 65 are farmers in 1940. Sons observed in rural Iowa in 1915 are more mobile than their urban peers as measured both by IGE and rank-rank parameters, though the differences are not statistically significant for the rank-rank mobility estimate. All measures still show more mobility historically than is estimated in today. The much higher levels of mobility for rural sons may be driven by the large increases in access to public education even in remote, rural regions of Iowa (Parman, 2011). Alternatively, the high levels of mobility may be caused in part by movement off the farm; this finding is consistent

0.2714 in the 10 rural counties.

[44] That is, 50% of sons that I find in the 1940 census and link back to 1915 on the basis of the son's first and last names, state of birth, and year of birth, would have to be the wrong person. For the rank-rank parameter, the mismatch error required to shrink the difference between the estimates is roughly 30%.

[45] Or are the results driven by the difficulty of accurately measuring income for farmers?

[46] As presented in Figure 2.1, the rural counties included in the Goldin-Katz sample are Adair, Buchanan, Carroll, Clay, Johnson, Lyon, Marshall, Mitchell, Montgomery, and Wayne and the urban cities are Davenport, Des Moines, and Dubuque.

with the model of human capital transmission presented by Nybom and Stuhler (2014b), which suggests periods of structural transformation in the economy weaken the links between parents' and children's outcomes.

Further isolating the effects on mobility of the shift away from agriculture, I limit the samples in the fifth rows of both panel A and B to only sons with fathers who were not farmers.[47] Again, mobility is lower than the contemporary estimates, though much closer to the urban sample than the rural sample. Overall, these urban and non-farmer-father subsamples suggest that the lower levels of mobility found historically are not artifacts of poor measurement of farmer income, whether that mismeasurement is driven by classical measurement error, by the difficulty of farmers to distinguish between net and gross income in census responses, or by transitory income shocks (such as adverse weather or crop-destroying pests).

As noted previously, the 1940 census collected data only for wage and salary workers and not capital income.[48] Thus, it is impossible to include sons in 1940 who were either farmers or business owners; this excludes data from 13.7% of the sons in 1940 who were farm owners or operators without income. These observations are not included in the measures of mobility previously discussed in Table 2.7. However, in row 6, I impute earnings for farmers using the 1950 census, which did collect data on capital income and non-wage and salary earnings. Earnings are imputed using years of education, age, state of residence, and state of birth. Using these imputed earnings, I estimate even higher levels of mobility than in my main results.[49]

Overall, my estimates suggest more mobility historically than today, measured both by IGE and with the rank-rank parameter. However, they are not as low as the results presented in Parman (2011). One key shortcoming of the data considered in that analysis—a sample with very old fathers and very young sons—was highlighted previously. Another data-driven limitation is

[47] This sample is made up of the urban sample and nearly half the rural sample.

[48] The data collected is the "total amount of money wages or salary" but enumerators were instructed: "Do not include the earning of businessmen, farmers, or professional persons derived from business profits, sale of corps, or fees."

[49] To impute total earnings, I regress the log of total income in 1950 on a full set of years of education indicators, a quartic in age, state of residence fixed effects, and state of birth fixed effects from data on farmers in the IPUMS 1% sample of the 1950 census. I use the results from that regression to predict income for the farmers in 1940, normalizing income from 1950$ to 1940$. For more details on the imputation of capital income in 1940 and graphs of the relationships between income and education and age for farmers in 1950, see Appendix B.3.

that the Parman (2011) sample can only match fathers to sons who still live in Iowa as adults.[50] How large is the bias of this restriction, and in what direction does it push the intergenerational income mobiility results? I use my sample to better understand the magnitude and direction of the problem by limiting my sample to only those sons who still live in Iowa in 1940. In the final rows of both panels A and B in Table 2.7, I calculate IGE and rank-rank parameters for this subset of sons in Iowa in 1940. The results suggest that the bias is both large and negative: selecting the sample using only sons remaining in Iowa in adulthood reduces both the measured IGE and rank-rank parameters. Because the state of residence in 1940 is, in part, jointly determined with the outcome of interest (income), controlling for it or splitting the sample based on it is problematic. In order to estimate an accurate IGE parameter for the early twentieth century, we need to be able to observe sons that remain in their father's state of residence and sons that move elsewhere.[51]

What is the probability that a son born in a given quintile will be in the same or another quintile in 1940? Table 2.8, an income quintile transition matrix, can be used to answer such questions. Each cell represents the probability that a son with a father in a given income quintile (identified by the column) will be in a given income quintile in 1940 (given by the row).[52] A son whose father is in the bottom quintile in 1915 has only a 14.6% chance of being in the top quintile in 1940, while the odds that a son born in the top quintile remains there in 1940 are 36.2%. What is the likelihood a son falls into the bottom quintile? Not surprisingly, those odds fall with father's rank: a son born in the top quintile is has only a 7% chance of being in the bottom quintile in 1940, while the odds are more than twice as high (nearly 15%) that a son born in the bottom quintile remains there. While the table clearly presents a degree of intergenerational immobility, when these percentages are compared to a transition matrix for the recent period (sons born in the 1980-82 cohorts) from Chetty, Hendren, Kline, and Saez (2014), there is in fact more mobility

[50] Again, this is due to the fact that the 1940 Federal Census was not available until 2012, 72 years after the survey was originally conducted and Parman (2011) instead matched fathers and sons within the 1915 Iowa State Census.

[51] My sample is technically restricted to those sons still living in the US and enumerated in the Federal Census. However, the number of sons moving abroad in this period is likely very low and thus the bias is likely to be insignificant.

[52] The columns sum to 100 because conditional on the father's quintile a son must be in one of the five groups. However, the rows do not sum to 100 because the sons' quintiles are based on the 1940 IPUMS 1% sample, not just the sons in my matched Iowa sample.

Table 2.8: Income Rank Quintile Transition Matrix

	Father's Quintile					
Son's Quintile	1	2	3	4	5	Total
1	14.5	7.7	9.2	9.5	7.2	9.4
2	23.4	22.6	15.3	18.3	15.1	18.9
3	27.2	24.8	25.7	18.8	18.1	22.8
4	20.7	26.1	24.7	23.4	23.3	23.8
5	14.2	18.9	25.2	29.9	36.2	25.1
Total	100.0	100.0	100.0	100.0	100.0	100.0

The cells in this table report the probability that a son with a father in a given income quintile in 1915 (column) will be a given income quintile in 1940 (row). The income distribution in 1940 calculated using the 1940 IPUMS 1% sample. The income distribution in 1915 is calculated using the Goldin-Katz 1915 Iowa State Census sample.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

between 1915 and 1940, at least at the lower end of the income distribution. Sons born in the bottom quintile are more than two times as likely to remain there as adults in the contemporary data than historically: 33.7% to 14.9%. Historically, sons born in the second quintile have a 69% chance of being in a higher quintile in adulthood; that probability is only 52% in contemporary data. At the top end, however, transition probabilities are similar between the two periods: sons born in the fourth or fifth quintiles have 23.7% and 36.2% probability of being in those quintiles as adults in the Iowa sample, compared with 24.4% and 36.5% probabilities in the Chetty, Hendren, Kline, and Saez (2014) sample.

2.4.2 *Alternative Measures of Intergenerational Mobility*

I also estimate the intergenerational mobility of education and present these results in Panel A of Table 2.9.[53] In addition to serving as a (potentially) more accurately measured check on my income results, the education IGE estimate is a valuable and important historical parameter. My fathers and sons are both observed at a pivotal moment of change in public education. The growth of mass public schooling in the United States, first in common schools during the later half of the nineteenth century and then through the high school movement in the early twentieth century, made education widely available and free (Goldin and Katz, 2008, 2011). Goldin and Katz (2008) also argue that this increase in human capital helped spur national growth and prosperity in the following century. Whether or not this massive public investment in education also

[53] Though the regressions are estimated in levels, I will follow the literature in describing these relationships as intergenerational elasticities.

reduced the strength of the relationship between a son's educational prospects and his father's educational outcomes can help scholars understand the role of public programs in shaping or changing inequality. The literature suggests an IGE of education of 0.46 (Hertz, Jayasundera, Piraino, Selcuk, Smith, and Verashchagina, 2007). As Table 2.9 shows, I find a much lower IGE parameter for schooling, between 0.206 and 0.264. This suggests that, like income, educational mobility in the US was higher in the early twentieth century than it is today.[54] Similar to the results presented on income mobility, there is more mobility of education among rural sons than urban sons as well, though these differences are not always statistically significant.

Though not my preferred measure of economic status or position, I can also estimate intergenerational mobility using occupation scores.[55] These scores measure the median earnings in a given occupation and may contain less measurement error than annual income observations. The occupation scores are not a panacea, even with income often unavailable in historical data. The occupation score commonly used is calculated by IPUMS from a 1950 census report. However, occupations in 1950 are difficult to link to occupations in earlier years, given the changing nature of tasks within an occupation and development or death of other occupations. Further, given the large changes in the returns to human capital and specific skills throughout the last two centuries, the median earnings for even the same exact occupation in two periods may be poorly correlated (Goldin and Katz, 2008). Nevertheless, given the widespread use of these measures in the historical literature on both intergenerational mobility and more broadly as a substitute or proxy for income, I can replicate my analysis with the occupation scores. I do so using both the standard 1950-based occupation scores from IPUMS as well as a 1915-based score that I construct from the full Iowa sample.[56]

The results, presented in Panels B and C of Table 2.9, suggest that, measured with occupation score, mobility was quite high in the early twentieth century, corroborating my findings with

[54] Concerns about noise driving down the estimated IGE parameters are less important for education, as any given annual measurement of years of schooling completed (for an adult) is a very accurate measure of lifetime years of education completed.

[55] For research on economic outcomes in periods without earnings data, many economists have turned to such occupation score measures, including Olivetti and Paserman (2015) and Abramitzky, Boustan, and Eriksson (2012, 2013b,a).

[56] In the appendix, I detail the construction of these measures and compare them. They are highly correlated, however some occupation groups are clear outliers suggesting that they moved up or down the income scale between generations.

Table 2.9: Alternative Intergenerational Mobility Estimates

	Sample	Specification (1)	(2)	(3)	Observations	Clusters
		A. Education Mobility				
Years of Education	Full Sample	0.264	0.241	0.206	3378	2505
		(0.023)	(0.023)	(0.040)		
	Urban Sample	0.297	0.269	0.301	1283	1028
		(0.035)	(0.035)	(0.058)		
	Rural Sample	0.235	0.223	0.142	2095	1477
		(0.029)	(0.030)	(0.056)		
	Excluding Farmer Sons	0.317	0.303	0.338	2006	1590
		(0.027)	(0.028)	(0.046)		
		B. 1915 Occupation Score Mobility				
Log Occupation Score 1915 Basis	Full Sample	0.167	0.162	0.265	3039	2280
		(0.030)	(0.030)	(0.052)		
	Urban Sample	0.229	0.226	0.308	1154	940
		(0.035)	(0.036)	(0.065)		
	Rural Sample	0.110	0.101	0.176	1885	1340
		(0.040)	(0.038)	(0.091)		
	Excluding Farmer Sons	0.180	0.182	0.289	1762	1415
		(0.030)	(0.030)	(0.053)		
		C. 1950 Occupation Score Mobility				
Log Occupation Score 1950 Basis	Full Sample	0.441	0.366	0.386	3204	2375
		(0.021)	(0.024)	(0.045)		
	Urban Sample	0.258	0.247	0.291	1220	980
		(0.036)	(0.036)	(0.068)		
	Rural Sample	0.424	0.419	0.401	1984	1395
		(0.030)	(0.030)	(0.070)		
	Excluding Farmer Sons	0.228	0.220	0.263	1869	1480
		(0.027)	(0.027)	(0.061)		

Son's completed years of education in 1940 is the dependent variable in panel A. The log of the son's occupation score, using either the 1915 or 1950 occupation score measures, is the dependent variable in panel B (1915) and C (1950). Standard errors clustered by family. Specification 1 is a univariate regression of son's outcome on father's outcome. Specification 2 adds name string controls, 1915 county fixed effects, and quartic controls in father and son age. Specification 3 adds an interaction between father's outcome and son's normalised age. Name string controls: first and last name commonness, length, letter similarity, and Scrabble scores. Son's ages are normalised relative to age 40 in 1940.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

income and education.[57] Specifically, I estimate an IGE parameter between 0.167 and 0.265for the occupational score measure based on income data from 1915 Iowa. This is higher than my IGE estimate for income, but lower than my IGE estimate for education. However, when I use the 1950-based occupation score measure, I find significantly larger IGE estimates, indicating less mobility. These measures are higher than any previous IGEs estimated in this paper, suggesting only slightly more mobility in the early twentieth century than today if any. These results are driven by farming fathers: when I subset the analysis to the urban sons or exclude sons of farmers, the results are more consistent across both measures of occupation score. There are a large number of farmers in my sample, and the relative positions of farmers and their median incomes changed quite a lot between 1915 and 1950. Based on the incomes reported in the 1915 Iowa State Census, farmers were at the median of the occupation distribution in 1915 (in Iowa) but in 1950, farmers ranked around the bottom 10th percentile of occupation groups by median income (nationally). This instability reflects the complication of using a measure like occupational score to determine intergenerational mobility, especially at a time of large structural change in the economy. It also reflects the difficulty of crudely classifying all famers with the same simple median income score, given the huge income differences in reality between wealthy owners of large farms and poorer tenant farmers.

2.4.3 *Occupation Results*

Similar to income, educational, and occupation score mobility, there appears to be more broad-category occupational mobility during my period of study than there is today. I measure occupational mobility using the Altham statistics discussed above (Altham and Ferrie, 2007). My occupational transition table from 1915 to 1940 is presented in Table 2.10. Occupations are categorised by linking occupational strings (exactly as entered by the census enumerators) to the 1940 occupational code charts for both the 1915 father and 1940 son samples.[58] Sample sizes are

[57] Occupation score is not available in the contemporary PSID sample and so I cannot compare these estimates to a parallel recent estimate. Further, given the availability of income data in the recent period, I am not aware of any studies that attempt to estimate intergenerational mobility of occupation score.

[58] In the margins of the original 1940 census manuscripts, exact occupation codes are included. Using both these occupation codes (when they are recorded) and the exact occupation strings, I have attempted to carefully match occupations from my data to the 1940 occupation master list. There may be measurement error in the exact matching. However, given that occupations are then collapsed to the four broad categories used in the Altham statistic, errors in

Table 2.10: Occupation Tranisitions Table for Fathers and Sons, 1915 to 1940

	Father's Occupation				
Son's Occupation	Farmer	Skilled	Unskilled	White Collar	Total
Farmer	599	56	44	31	730
Skilled	288	380	220	190	1,078
Unskilled	332	179	196	91	798
White Collar	244	313	159	337	1,053
Total	1,463	928	619	649	3,659

Father's occupation categories are determined from the 1915 Iowa State Census. Son's occupation categories are determined from the 1940 Federal Census. Total counts do not match previous totals for fathers and sons in other tables because some observations contain information on wages or education but with occupation descriptions that cannot be linked to broad categories.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

different from previous portions of the analysis because not all occupation strings are matchable to occupation codes or broad occupation groups.

As detailed earlier, the distance between the occupation transition table and the identity table, I, can be thought of as a measure of occupational immobility—the larger the distance, the more likely it is that sons enter the same occupational class as their fathers. Calculating the Altham statistic for Table 2.10 yields $d(IA1940, I) = 16.14$. Long and Ferrie (2007b) report $d(US1880, I) = 12.09$,$d(US1900, I) = 14.58$, and $d(US1973, I) = 20.76$. The Altham statistic generated by my linked sample of fathers and sons from 1915 to 1940 is larger—thus indicating less mobility—than the measures presented by Long and Ferrie (2007b) for the nineteenth century and statistically significantly different from these historical measures as well. These results are summarised in Figure 2.4 and in the first column of Table 2.11. In addition, echoing the results presented previously suggesting more mobility historically than today, there appears to be more mobility between 1915 and 1940 than between 1950 and 1976—as compared to the more contemporary estimates reported by Long and Ferrie.

The second and third columns of Table 2.11 present the Altham statistics for the urban and rural subsamples of my linked data. Relative to the full sample, I find more mobility among both the urban sample and the rural sample.[59] I also find more mobility in the urban sample

occupation matching will bias the final results only if occupations are coded into the wrong broad category.

[59] The fact that the Altham statistics of two disjoint sets can each be smaller than the Altham statistic of their union is algebraically allowed, but seems to a very undesirable property of Altham statistics. In theory, an alternative statistic might possess a form of continuity and the intermediate value theorem.

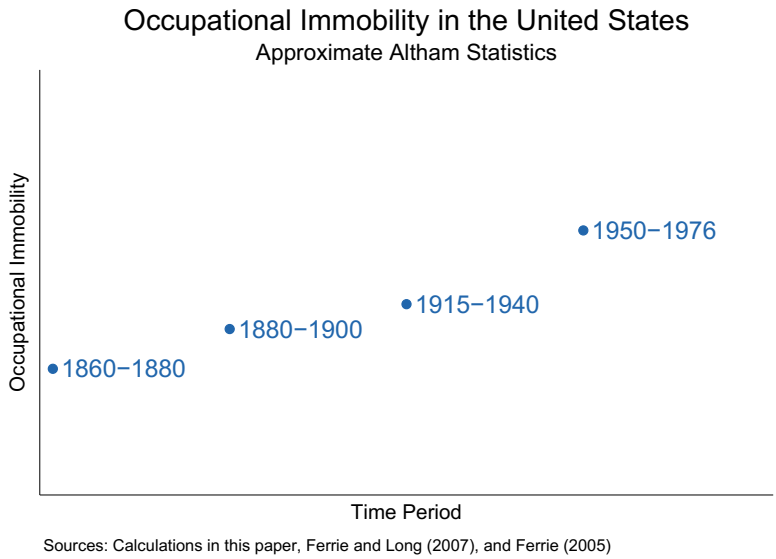


Figure 2.4: Occupational Immobility measured by Altham statistics from 1860 to 1976. This graph suggests that occupational immobility rose over time and that there is less mobility in the mid-twentieth century than there was between 1915 and 1940. The Altham statistics for each period presented in the plot are statistically different from one another. See Table 2.11 for exact values and distance tests.

Table 2.11: Altham Statistic Summary, Iowa 1915 to 1940

	Sample (P)		
	All Sons	Urban Sons	Rural Sons
Altham Statistic ($d(P, I)$)	16.14	10.74	14.34
$Pr(d(P, US1880) = 0)$	0.000	0.627	0.141
$Pr(d(P, US1900) = 0)$	0.000	0.017	0.511
$Pr(d(P, US1973) = 0)$	0.000	0.003	0.004
$Pr(d(P, IAsons) = 0)$	.	0.718	0.711
$Pr(d(P, IAurbansons) = 0)$	0.718	.	0.786
$Pr(d(P, IAruralsons) = 0)$	0.711	0.786	.

Father's occupation categories are determined from the 1915 Iowa State Census. Son's occupation categories are determined from the 1940 Federal Census. The Altham-Ferrie statistic is a distance metric; the distance from the identity matrix I can be interpreted as a measure of mobility with higher values implying less mobility. The distance metric can also compare two occupation transition matrices. Altham-Ferrie statistics for US1880 (a father-son linked sample between fathers in 1860 and sons in 1880) is 12.09, for US1900 (fathers in 1880 and sons in 1900) it is 14.58, and for US1973 (fathers in 1950 and sons in 1973) it is 20.76. The above results reject that occupation category transitions were the same between 1915 and 1940 and any of the other periods. In particular, there was more occupation transition mobility in the nineteenth century than the early twentieth century and more mobility in the early twentieth century than between 1950 and 1973.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census; Long and Ferrie (2007b)

than in the rural sample.[60] However, none of these differences is statistically significant, and I cannot reject that mobility was the same in Iowa overall as in the urban and rural subsamples. The subsamples also confirm the previous finding of more mobility historically than today.

2.4.4 *Geographic Mobility*

In addition to the standard measures of mobility considered thus far, my linked 1915 and 1940 samples also allow me to estimate the correlations of father's income or education with son's geographic mobility. Figure 2.5 presents a map of the residences, in 1940, of the sons included in my sample. The sons are located in almost every state in the US and most territories (territories not pictured on the map).[61] Table 2.12 gives the percentage living in each of the most common states and compares these results with the geographic locations among 28-42 year old white male Iowa natives in the 1940 IPUMS 1% sample (Ruggles, Alexander, Genadek, Goeken, Schroeder, and Sobek, 2010).[62] Nearly 64% remain in Iowa.[63] Neighboring states, especially Illinois and Minnesota, account for 17.3% of sons. Los Angeles county is the most prominent urban destination for the sons who left Iowa, with 3.9% of the sample population, followed by Cook county, Illinois (Chicago, 3.3%), Rock Island county, Illinois (1%), Douglas county, Nebraska (Omaha, 0.9%), and Hennepin county, Minnesota (Minneapolis, 0.9%); few travel farther east than Detroit.[64]

I also measure geographic mobility as the distance that the sons had moved between when they are first observed in 1915 and when they are observed again in 1940. Table 2.13 suggests that distance moved increases with either the father's income or father's education, though in

[60] This difference is the reverse of what I found with respect to both income and educational mobility in the previous section.

[61] Recall that I am matching from the 1915 Iowa State Census to the 1940 Federal Census. Thus, while I will be able to find sons in any of the 48 states or other territories included in the census, sons leaving the country will not be matched. There are no sons living in Delaware, New Hampshire, or Vermont. Hawaii and Alaska were not yet states and are not covered by the 1940 Federal Census sample used for son-matching.

[62] The conceptual construction of the 1% 1940 IPUMS sample does not match my sample exactly because not all sons in Iowa in 1915 (in my sample) were born in the state, but they are roughly similar.

[63] Long and Ferrie (2004) estimate geographic mobility in the US between 1850 and 1880 and find identical results for the earlier period: 64.7% of young men in their matched sample remain in the same state from 1850 to 1880.

[64] Rock Island, Illinois is across the Mississippi River from Davenport, Iowa; some sons remaining in Iowa travel fewer miles than those sons moving from Davenport to Rock Island, IL or Moline, IL.

Table 2.12: Sons of Iowa Residences in 1940

	Matched 1915-1940 Sample		1940 IPUMS Sample	
	Count	Share (%)	Count	Share (%)
Iowa	2859	63.8	1933	58.9
Illinois	301	6.7	163	5.0
California	292	6.5	221	6.7
Minnesota	177	4.0	159	4.8
Wisconsin	94	2.1	61	1.9
Nebraska	78	1.7	82	2.5
Missouri	66	1.5	71	2.2
South Dakota	61	1.4	70	2.1
New York	52	1.2	33	1.0
Other	498	11.1	487	14.8

This table compares the state residences of the sons matched between the 1915 Iowa State Census and the 1940 Federal Census with state residences of all men born in Iowa between 1898 and 1912 in the IPUMS 1% sample of the 1940 census.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

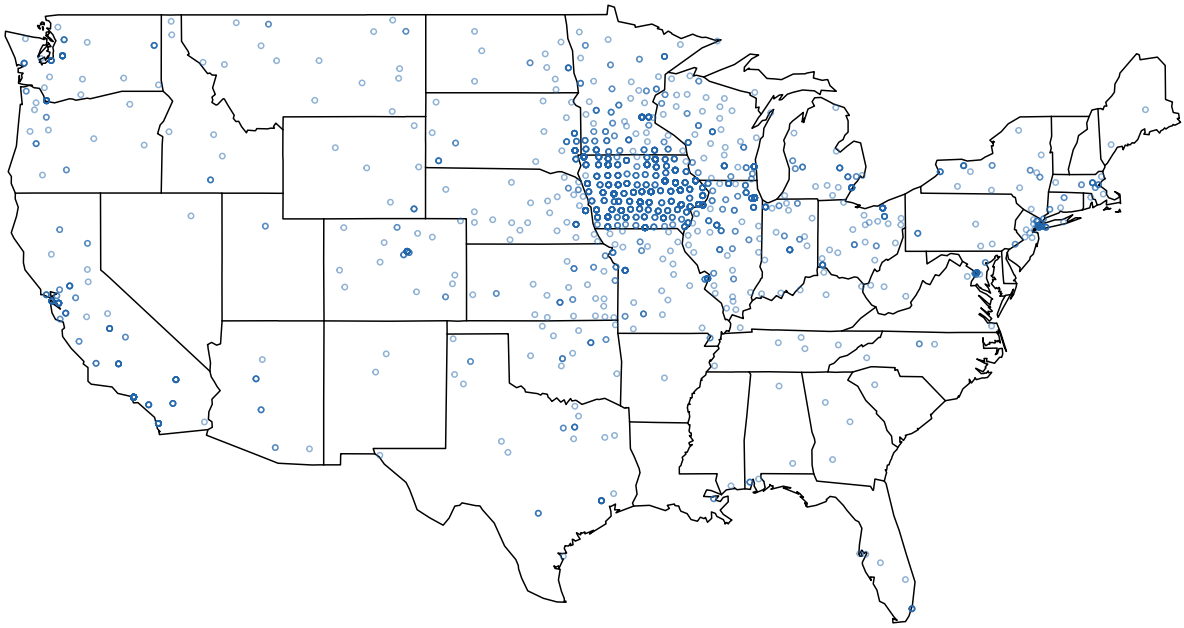


Figure 2.5: Sons County of Residence in 1940. The darker symbols implies greater density of points at a given latitude and longitude.

Table 2.13: Geographic Mobility: Miles Moved 1915 to 1940

	Full Sample			Urban		Rural	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Father Earnings	0.0474		-0.00973	0.252*		-0.102	
	(0.0616)		(0.0638)	(0.130)		(0.0738)	
Father Education		0.0578***	0.0478***	0.0419	0.0709***	0.0418**	0.0497***
		(0.0132)	(0.0145)	(0.0267)	(0.0227)	(0.0172)	(0.0161)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name String Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3657	3937	3651	1479	1563	2172	2374
Number of Clusters	2629	2836	2624	1141	1207	1483	1629
R-squared	0.0252	0.0291	0.0292	0.0467	0.0448	0.0180	0.0156

The log of 1 + miles moved by the son from 1915 to 1940 is the dependent variable. It is necessary to add one to the number of miles to avoid dropping sons who did not move counties between 1915 and 1940 from the analysis. Standard errors clustered by family. Name string controls: first and last name commonness, length, letter similarity, and Scrabble scores. Son's ages are normalised relative to age 40 in 1940.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

column 3, with both father-level variables included, only education has a significant effect. An additional year of education for the father increases the number of miles moved by the son by between 4.8% and 5.8% in the full sample. This relationship appears to be stronger for the urban sons (nearly 7% per year of education) relative to the rural sample (4% per year). While these results are more speculative, they suggest that enabling higher levels of geographic mobility may be one way in which better educated fathers (or richer fathers) improve potential outcomes for their sons. However, the importance of geographic mobility should not be overstated; earlier in this paper, I found that rural sons had more economic mobility even though they had less geographic mobility in Table 2.13.

In addition to providing the first national micro-records of income and years of education, the 1940 Federal Census was conducted on the heels of two major economic events of the twentieth century: the Great Depression and the Dust Bowl. The Dust Bowl may be a particularly important factor to consider when following the earnings and education trajectories of men growing up in Iowa during this time period. Huge dust storms during the 1930s blew topsoil off farms in the American plains, causing severe and lasting erosion to once highly productive and fertile farmland. Hornbeck (2012) uses variation in dust bowl severity to measure the effects of

environmental catastrophes on economic outcomes. He finds that the Dust Bowl led to immediate and persistent reductions in agricultural land vales and production and to large population out-migrations. Dust Bowl severity in Iowa was highly varied throughout the state (Hornbeck, 2012, Figure 2). While the central and northern portions of the state suffered little to no erosion, the southwestern and southern areas of the state endured high levels of topsoil erosion and loss. Of the rural counties included in the Goldin-Katz sample, Adair, Carroll, Johnson, Lyon, Montgomery, and Wayne were all in medium erosion zones; the other rural counties were in low erosion areas.

The regressions presented in Table 2.13 and used in all other analyses in this paper include county level fixed effects; these fixed effects control for differences in Dust Bowl exposure at the county level. However, it may be the case that fathers with higher income or higher levels of human capital (measured here by years of education) were better prepared to deal with the Dust Bowl (either through out-migration or changing their agricultural practices). As farms are often passed down from father to son, the probability of moving or the distance moved depends greatly for these sons on the interaction of the Dust Bowl and their father's attributes. Results on a sample restricted to fathers who were farmers in 1915 are limited.[65] It does not appear that the Dust Bowl interacted in any general way with either father's income or education to determine the son's geographic mobility.

2.5 *Conclusion*

This paper presents estimates of both income and educational mobility for men born in Iowa between 1900 and 1910, showing that their mobility rates were much higher than those of men born since 1960. I matched fathers from the Iowa State Census of 1915 to their sons in the 1940 Federal Census. In my sample of fathers and sons, I estimate a lower intergenerational elasticity of income than is found in contemporary studies of the United States, suggesting higher levels of income mobility. Estimates of an intergenerationl rank-rank parameter, relating the father's ranking in the income distribution in 1915 to the son's ranking in the income distribution in 1940, also show more mobility historically than in more recent settings. Intergenerational mobility of

[65] See Table B.1 in the appendix.

education is higher in my sample than in contemporary measures as well. I find sons in rural counties in 1915 to have more mobility of both income and education than urban sons. When I compute the standard measures of occupational mobility for my sample, I find generally higher levels of mobility between 1915 and 1940 than are found in recent estimates as well.

In addition to constructing estimates of intergenerational mobility of income, education, and occupation, my dataset also enables me to measure other forms of intergenerational mobility. Because I match into the Federal Census of 1940, I am not restricted to considering non-migrants. Thus, I am able to measure the relationship between parental outcomes and children's geographic mobility. Many of the sons in my sample remained in Iowa, but a large portion moved within the state to urban areas or to larger urban centers outside the state, such as Chicago and Los Angeles. I can also measure differential geographic mobility between rural and urban sons. My results suggest that the sons of more educated farmers are more mobile, especially those raised in urban Iowa, but that there is little to no effect of father's income once controlling for father's education.

This paper demonstrates that there was more intergenerational mobility in the early twentieth century than there is today. An important question is why. The high school movement and the huge expansion of access to public education could have been one driver. Sons had on average two more years of schooling than their fathers. I find high levels of educational mobility: a son's completed years of schooling are only weakly related to his father's education. The wide availability of free schooling, it appears, could thus sever the link between father's and son's educational attainment. The general transition away from an agriculture-based economy may have also played a role. I estimate higher levels of mobility among rural sons, many of whom were the sons of farmers. However, the high levels of mobility persist in samples restricted to the sons of non-farmers in both urban and rural Iowa. Even without a change in the underlying mobility parameters, the changing composition of the country—fewer rural residents and farmers in each subsequent generation—would lower mobility over time.[66] There were high levels of geographic mobility throughout the country in the early twentieth century, both to the west and to urban areas. Geographic mobility and economic mobility are likely correlated, but this

[66] 42% of the fathers in my sample were farmers compared to only 21% of the sons.

pushes the question back a step. What determined geographic mobility? My results suggest that the odds of a son moving or his distance traveled may have been related to his father's outcomes, but that overall an individual's migration decision is difficult to predict. Finally, the Great Depression almost certainly altered the economic fortunes of most sons in my sample, perhaps by changing the relative returns to different professions or skills in unpredictable ways. However, the change induced by the dislocation of the Great Depression is unlikely to be the full explanation, given that I find high levels of educational mobility as well. Even the youngest sons in my sample were 19 in 1929, and the vast majority had already completed their schooling before the downturn.

Both Lee and Solon (2009) and Chetty, Hendren, Kline, Saez, and Turner (2014) find relative stability in intergenerational mobility over the past two to three decades. If mobility was higher among sons born between 1900 and 1910, then this stability could not be a permanent feature of intergenerational mobility in the United States. At what point in the twentieth century did economic mobility decline? Was there a sharp transition from one stable level of mobility to another, or was the shift a gradual decrease in mobility over several decades? Was there variation across the US in this change? And, finally, what caused this shift? Based on the results in this paper, it appears unlikely that the Great Depression drove this change, at least among the generation of sons born before but employed after it. Did the Great Compression, documented by Goldin and Margo (1992), induce a new era of lower mobility? Or did the change come later and affect sons born during mid-century and entering the labour force in the 1970s or 1980s? The data to answer this question exist but much of it is not yet accessible. With the 2022 release of the full non-anonymous 1950 census—and the 2032 release fo the 1960 census—it will be possible to track intergenerational mobility through the middle of the twentieth century.

3. A MACHINE LEARNING APPROACH TO CENSUS RECORD LINKING

3.1 *Introduction*

Thanks to the availability of new historical census sources and advances in record linking technology, economic historians are becoming big data genealogists. Linking individuals over time and between databases has opened up new avenues for research into intergenerational mobility, assimilation, discrimination, and the returns to education. To take advantage of these new research opportunities, scholars need to be able to accurately and efficiently match historical records and produce an unbiased dataset of links for downstream analysis.

The problems in record linkage facing economic historians are distinct from those faced by users of modern datasets. Where uniquely identifying variables such as social security numbers are available, it is mostly a question of getting access to restricted use files.[1] But such variables are rarely found in historical data. Instead, economic historians have access to other variables that can be combined to try and uniquely identify individuals, such as first and last names, year of birth, state of birth, and parents' place of birth. Unfortunately, historical data are also not as clean as modern data, and these variables may be mismeasured, including transcription errors, spelling mistakes, name changes, or name shortening.

In this paper, I detail a transparent census matching technique for constructing linked samples that can be replicated across a variety of cases. The procedure applies insights from machine learning classification and text comparison to the well known problem of record linkage, but with a focus on the sorts of costs and benefits of working with historical data. I use a supervised learning procedure, teaching an algorithm to discriminate between correct and incorrect matches based on training data generated by the researcher. The method begins by cross matching two census-like datasets and extracting a wide subset of possible matches for each record. I then

[1] For example, Chetty, Hendren, Kline, Saez, and Turner (2014) use tax payer IDs to link tax payer records over time and social security numbers to match dependents to heads of household across and within samples.

build a training dataset on a small share of these possible links and use the training data to tune a matching algorithm. The algorithm attempts to minimize both false positives and false negatives, taking into account the inherent noise in historical records. To make the procedure precise, I trace its application to an example from my own work, linking children from the 1915 Iowa State Census to their adult-selves in the 1940 Federal Census. Using my linking procedure, I am able to match nearly 60% of the sons in my data ahead to 1940.[2] The procedure follows many of the central ideas outlined by Goeken, Huynh, Lenius, and Vick (2011) regarding the Minnesota Population Center linkage project that ultimately built the IPUMS linked samples, but with an extension of the method to other linking procedures and more detail on the utilized record comparison characteristics included. The procedure I outline in this paper can be fully and transparently implemented with software commonly used by empirical social scientists—either Stata or R, for example.

In addition, I provide guidance on three practical questions for social scientists undertaking historical record linkage. First, I show how many records need to be manually coded as matches and non-matches by the researcher before tuning the match algorithm. Fortunately, the procedure is quite accurate even with a relatively small training data set. Second, I run a horse race between potential classification models—including probit and logit models, random forests, and support vector machines—and show that the probit model, familiar to all social scientists has the best cross-validation, test set performance. Third, I document that the matching algorithm avoids making one particularly bad type of false match—linking people who have died between census waves—using the records for Major League Baseball players active in the 1903 season.[3]

The linking of historical records by scholars is not new. Thernstrom (1964, 1973) matched generations in Boston and Newburyport, MA to study intergenerational mobility. The Minnesota Population Center, taking advantage of the 1880 complete count census, provides linked data from each census between 1850 and 1920 to 1880.[4] Joseph Ferrie linked records between the 1850

[2] I will also show that these matches are not just made but made accurately and that the algorithm is able to replicate the careful manual matching work done by a trained RA.

[3] Thanks to both the prominence of the people in the dataset and the interest among baseball fans, the Lahman database (Lahman, 2016) presents a unique source that includes both the biographical information necessary to link to the census, as well as detailed information on date of death.

[4] See `https://usa.ipums.org/usa/linked data samples.shtml`

and 1860 Federal Censuses, exploring various dimensions of economic mobility (Ferrie, 1996). However, no standard method for linking these records has emerged. Each scholar alters the process slightly based on the data at hand and the tools available.

In April 2012, the US Census Bureau released the full, un-anonymized 1940 Federal Census, opening many new possibilities for historical research. By law, the complete census, which includes the names of all respondents, must be sealed for 72 years after its completion.[5] The 1940 census was the first nationwide survey to include questions on educational attainment and annual income.[6] The data in the 1940 census has been used by researchers in the past to measure the returns to education, to quantify racial and gender discrimination, and to answer many other research questions. These analyses have all been possible through the use of an anonymized 1% 1940 Census sample collected by IPUMS (Ruggles, Alexander, Genadek, Goeken, Schroeder, and Sobek, 2010).

What specifically changed for researchers in April 2012? The full census allows for the matching of individuals from other datasets—other federal and state censuses, enlistment records, legal records, etc—by name to the 1940 Census. With such matched records, it is possible to conduct research that follows individuals over time or across generations— intergenerational mobility, stability of income over time, long run effects of exposure in childhood to pandemics, etc. But how can a researcher merge a list of names from one set of records into another, such as the 1940 census?

Any matching procedure should aspire to three important criteria: it should be efficient, accurate, and unbiased. I define these terms in the record linkage context:

- Efficient: A high share of the records to be searched for are found and matched. The match rate will naturally vary across applications and source or target databases, but generally, a procedure that requires thousands of records to match only a handful would be quite inefficient and not very useful for econometric analysis. An efficient match process will have a low share of type I errors. In the machine learning context, one measure of efficiency

[5] The 72 year seal is driven by privacy concerns. When the law was first passed, life expectancies were such that few census subjects would be alive 72 years later. That is less true today, but the privacy law remains in effect. The 1950 census will be unsealed in April 2022.

[6] In 1915, the Iowa State Census compiled similar data; see Goldin and Katz (2000).

is the true positive rate or TPR. This records the ratio of true positives with the total number of positive: $TPR = \frac{TP}{TP+FN}$.

- Accurate: A high share of the records matched are true matches and not false positives. Ideally, this rate would be close to 100%, but naturally the higher the bar for declaring two records matched, the less efficient it will be. An accurate match process will have a low share of type II errors. In machine learning, accuracy could be measured with the positive predictive value or PPV. This measures the ratio of the true positives to all of the records identified as matches by the algorithm: $PPV = \frac{TP}{TP+FP}$.

- Unbiased: A match procedure will generate a dataset for downstream analysis. To what extent is this final dataset representative of the records that the researcher attempted to link in the first place? Improvements in either efficiency and accuracy will necessarily decrease the bias in the resulting dataset. But non-random variation in either error rate will generate bias. One manifestation of bias would be an unrepresentative linked sample. Using spouse names to create links, for example, would increase the match rate among married people and over-represent them in final analysis; similarly matching on county or state of residence would bias against including interstate migrants in the sample (Goeken, Huynh, Lenius, and Vick, 2011).

Manual matching is one option for record linkage. Scholars can hire research assistants to search for each name in one dataset in the 1940 census, either via the index file or through a commercial provider like Ancestry.com. With dutiful RAs, this process could be quite efficient with assistants tracking down as many links as possible. And with skilled RAs, the process could be highly accurate. However, this method is costly and time-consuming.[7] Perhaps more importantly, it is inconsistent, certainly across projects but perhaps within a project as well. Different RAs will use different internal heuristics or decision rules in matching or not-matching close hits. While a clear set of rules can reduce such problems, a complete decision tree is impractical if not impossible.

[7] I have found RAs can search for approximately one record per minute on Ancestry.com. With a match rate of 50%, generating a database of 1000 matched records will cost $\$2000/60 \times .5 \times w$, where w is the RA's wage (or double that for double entry). Search time costs will certainly vary by researcher and project.

Researchers would be better off using a formalized matching algorithm that made consistent choices between potential matches in all scenarios. Goeken, Huynh, Lenius, and Vick (2011) describe the method used by the Minnesota Population Center for IPUMS.[8] The process makes use of the Freely Extensible Biomedical Record Linkage (FEBRL) software from Peter Christen and Tim Churches.[9] This method also relies on highly trained researchers to identify and approve of huge numbers of links between different censuses. The matches are made initially by FEBRL, comparing records by first name, last name, year of birth, state of birth, race,and gender. A subset of these possible matches are then completed by researchers, identifying true and false links on a training data set. Then, IPUMS uses a standard machine learning technique, Support Vector Machines (SVMs), trained on name and age similarity scores, to classify matched records as links or non-links.

The method I propose in this paper deviates from the IPUMS strategy in three key ways. First, the tools used in my method are all available in Stata and R and should be more familiar to most economists and other social scientists than either FEBRL or SVMs. This should make implementation of the procedure in different datasets easier. Second, unlike the FEBRL system used to collect possible matches, my method was designed for historical work, using datasets with transcription errors and other historical noise. Third, the method can be easily modified for use with different datasets that contain alternative potential matching variables.

3.2 *Procedure*

To fix ideas, I will describe the process used to locate matches between a smaller list of records and a full list of records in a 100% census. For example, in Feigenbaum (2014), I match a list of boys from the Iowa 1915 State Census into the 1940 Federal census. Other examples that could potentially use this method include Aizer, Eli, Ferrie, and Lleras-Muney (2014) which linked the male children of recipients of a Progressive era welfare program for poor families into the 1940

[8] Mill (2012) proposes another alternative record linking procedure, one that is a fully automated learning algorithm that does not require even a training dataset. Using an EM maximization process, the algorithm attempts to split the data into a set of matched and a set of unmatched records. One key strength of this method is that it does not require any training data and thus suffers from no human-induced bias in determining which records are or are not matches. However, the Bayesian learning process employed is constrained by independence between parameters.

[9] `http://sourceforge.net/projects/febrl/`

Census, WWII enlistment records, and death records or Collins and Wanamaker (2014) which linked black and white census respondents from the IPUMS sample of the 1910 census to the 1930 census.

The size of the first set of records can vary from project to project. The list of sons from Iowa 1915 was approximately 5500 observations. In other cases, it will be much larger: Collins and Wanamaker (2014) started with nearly thirty thousand observations of men from southern states in 1910.

3.2.1 *Extracting Possible Matches*

Call the first set of records $X1$. What variables does it need to contain to ensure a good (and unbiased match)? To begin with, first and last names for each record. Without names matching would be unlikely to work with any method. In addition, an age or year of birth,[10] gender, and a state of birth will be improve match quality. Mother and father state of birth have proven to be valuable matching variables in past work. However, in the 1940 complete count census, these questions were not asked of all respondents.[11]

I prefer not to use reported race in my matching algorithm. As documented by Mill and Stein (2012), individuals change their reported race in the census with some frequency. To the extent that changing a respondent's reported race is endogenous to whatever process the researcher ultimately hopes to study, then conditioning matches on race will likely bias the final sample. However, it may be useful to use race in both datasets as an outcome of interest in assessing match quality of the final data.[12]

Call the second set of records to be matched into $X2$. The records in $X2$ should include first name, last name, and any of the other variables to be used for matching that were also included in

[10] Of course, assuming that the year the records were compiled is known, year of birth and age are substitutes.

[11] Only those respondents entered on the 14th or 29th line of each enumeration page were asked additional questions. This is roughly 5% of the population. Mother and father place of birth were among these supplementary questions. See `http://www.archives.gov/research/census/1940/general-info.html#questions`. For this subsample, parent's place of birth, if it is available in the original list of records, can be used in testing the accuracy of the match algorithm and to compare between different possible match algorithms. But it would significantly reduce the sample size to require these variables in the record linking.

[12] Mill (2012) uses race as well as county of residence between multiple census waves in such a way.

$X1$, often gender, state of birth and year of birth (or age).[13] 1940 complete count census available from Minnesota Population Center and IPUMS (via NBER) has these variables and more, but other censuses do as well (complete counts of 1880 and 1850 are both available via IPUMS, for example). To this point, I have listed gender as one of the variables used in the census record linking. However, given the frequency with which women change their surnames in marriage, the task of linking women across censuses is much more difficult than linking men. While this will mean that there are important questions that I cannot answer, I focus here on linking only men. Thus, as a preliminary step, imagine limiting both $X1$ and $X2$ to only male observations.

The first step is to extract records in $X2$ that could be matched with records in $X1$. What defines which records could be matched? In an extreme case, the entire Cartesian product of $X1$ and $X2$ has $X1 \times X2$ observations. The matching algorithm could compare each of these potential matches, but even with a smaller original list (for example, the sons in Iowa in 1915), matching it with any complete census index would result in a huge database, full of potential matches that are clearly not matches.[14]

To limit the comparison to links that have some non-trivial likelihood of being matches, I first extract the records in $X2$ with the attributes sufficiently similar to the attributes of the record in $X1$. The attributes I focus on are birth year, state of birth, and names. Similarity of birth year is measured by the absolute difference in years. An indicator variable for matching or non-matching state of birth is clear to define as well. How does one measure string similarity for first and last names? I follow IPUMS, Goeken, Huynh, Lenius, and Vick (2011), and Mill (2012) in relying on the Jaro-Winkler string distance as a measure of string dissimilarity. Though other string distance measures, notably Editex and syllable-comparison, show better results in identifying common English homonyms, the differences between these measures on American names are less striking. Moreover, algorithms to calculate Jaro-Winkler distances are available for most statistical software packages;[15] an Editex module only exists for Python. For more details on the Jaro-Winkler distance and its properties, see Winkler (1994). Traditionally, strings that

[13] And a unique identifier that will be used to merge our final data back into $X2$ to collect the variables of interest in any analysis.

[14] More the 700 billion for the 1915 sons and the full 1940 census.

[15] For stata, see my own package: `https://github.com/jamesfeigenbaum/jarowinkler-ado`. For R, see the `stringdist` package.

match will have a Jaro-Winkler distance of 1 and strings that are not similar at all will have a distance of 0. However, I use $1 - JW$ here, so that increases in distance correspond with words that are less similar, as one would expect for a distance measure.[16]

Returning to $X1$ and $X2$, I limit the set of possible links between the two datasets to be those records with the same reported state of birth, a year of birth distance less than 3 years, a first name Jaro-Winkler distance of less than .2, and a last name Jaro-Winkler distance of less than .2.

Why are these good starting point conditions? The state blocking condition is the same used by Mill (2012). IPUMS does not use a state blocking condition explicitly, however, of the 98,317 links between the 1880 census other decennial censuses, only 21 list different states of birth between matches, 20 of which were in the link between 1880 and 1850. The state blocking requirement reduces the complexity of the matching problem and the computing power required to solve it: string distance measures and matrix Cartesian products are both computationally expensive. For a given individual, state of birth should only change between censuses due to random enumerator error.

The birth year limits have been used in past matching work: The IPUMS project merging various historical census samples into the 1880 complete count census used a year of birth distance of 7 (Goeken, Huynh, Lenius, and Vick, 2011). However, in the final matched samples produced by IPUMS, no matches were made between records with birth years differing by more than 3 years. I follow these results and use 3 years as the distance bound here. As I show in Figure 3.1, based on data from the IPUMS matching procedure between 1880 and other censuses, the vast majority of matches are either 0 or 1 years apart, in either direction.[17]

The name distances may be more controversial. The stricter the bounds, the fewer potential matches will emerge and the quicker the entire matching process will be. The looser the bounds, however, the less likely it is that good matches will be lost in this early stage. Figure 3.2, presents a scatter plot of first and last name Jaro-Winkler distances for all the matched records between

[16] Mill (2012) makes the same translation of Jaro-Winkler string distance.

[17] Why consider any potential matches that do not share the exact birth year? Aside from simple errors of mistranscription, there may also be "translation" errors. In many waves, the census asked respondents not their year of birth but their age. With census enumeration taking place over time (and the final census day varying slightly throughout the 19th and 20th century), the exact same respondent enumerated ten years apart may have a different constructed year of birth.

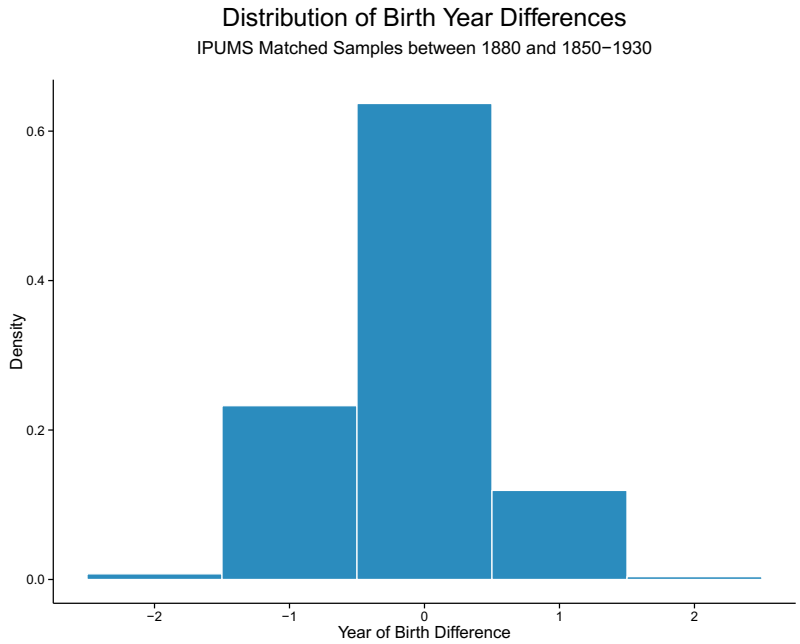


Figure 3.1: Distribution of Birth Year Differences. Differences are defined as the reported year of birth in the less recent census less the reported year of birth in the more recent census. For a record linked between 1880 and 1900 with a birth year of 1850 in 1880 and 1851 in 1900, the difference would be -1. All year differences are between -2 and 2.

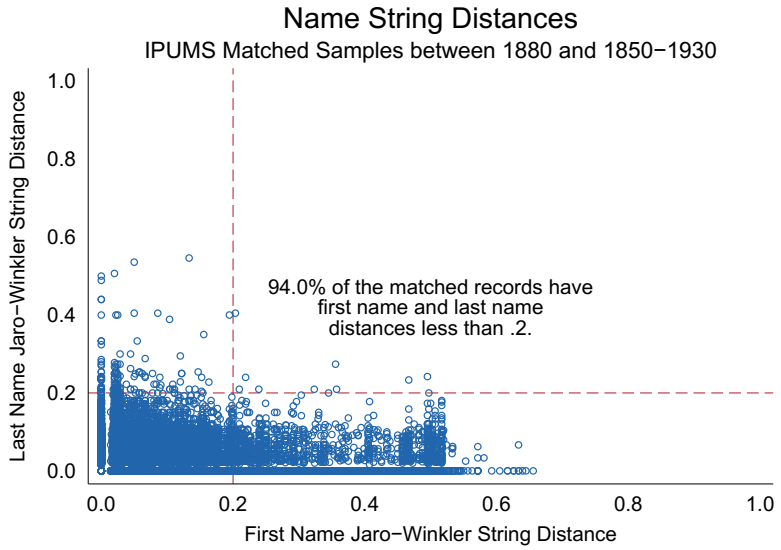


Figure 3.2: Distribution of First and Last Name String Distances, calculated with the Jaro-Winkler string distance metric. 94% of the records linked between the IPUMS samples have first and last name distances of less than 0.2.

Table 3.1: IPUMS Linked Samples: Matched Record Name Similarity

	Share of Links within Bounds		
Linked Years	First Name Distance $\leq .2$	Last Name Distance $\leq .2$	Both Conditions
1850-1880	94.07	99.92	93.99
1860-1880	92.86	99.81	92.68
1870-1880	94.44	99.84	94.30
1880-1900	94.50	99.79	94.31
1880-1910	95.01	99.69	94.72
1880-1920	95.47	99.94	95.41
1880-1930	96.23	99.96	96.19

Share of records matched between the 1880 complete count census and various IPUMS census samples by the Minnesota Population Center. These samples were built not using Jaro-Winkler distances but by trained research assistants comparing names (and other census data) manually. However, no more than 8% of the matched sample in any given year falls outside the .2 distance bounds for both first and last name Jaro-Winkler distances. Most linked records outside the bounds are farther apart in Jaro-Winkler distance in first names. Manual inspection reveals that most of these links match full names to initials or vice versa.

Sources: IPUMS Linked Representative Samples, 1850-1930

the 1880 census and other IPUMS samples. Only 6% of the matched sample falls outside the .2 distance bounds for both first and last name distances. Most of these records outside the bounds appear to be farther apart in Jaro-Winkler distance in first names. Manual inspection reveals that most of these links match full names to initials or vice versa.[18] In Table 3.1, I show that for each IPUMS linked sample between 1880 and another census year, between 92.16% and 95.65% of the records are within these bounds.

3.2.2 *Two-Step Match Scores*

Call the set of all possible matched data XX. Again, these are the records from our original data, $X1$, matched with records from the index file, $X2$, with matching state of birth, year of birth difference less than 3, and first and last name Jaro-Winkler distances less than .2. With the set of possible matches between each record of $X1$ and $X2$, we now how to decide which of these links (if any) are "good" links, in the sense that it is very likely that the record in $X1$ and $X2$ refer to the same individual. To do so, we need to compile training data: a set of records for which the researcher (or research team) have determined that a record is either a match or not a match. I

[18] For data with a large number of first names recorded as initials, a looser distance bound may be useful.

discuss in more detail in a later section how to generate the training data and how large it needs to be. For now, consider it simply as a subset of XX called XX_T , where each record is either described as a match or a non-match.[19] One key feature is that for every unique record in $X1$, there can be at most one match in the XX_T data.[20] However, this one match is not a requirement. If there are no good matches for a given unique record in $X1$, then all the links will be marked as non-match. However, it need not be the case that each record in $X2$ have a match. Obviously in the case where $X1$ is a short list of people—like sons between 3 and 17 in Iowa in 1915—not every record in the entire 1940 census will be available.

I use the term training data because we will train our match algorithm using it as an input. However, there is one key way in which the training data in this context differs from training data more generally. In traditional machine learning applications, training data is a set of true or correctly classified observations. In the case of historical census matching, such "truth" does not exist. Instead these are the matches the researcher would be confident labeling as links or not. These decisions are clearly not infallible, but any algorithmic approach can (and should) approximate the best efforts of a trained researcher assessing matches.

Each record, $xx \in XX$ matches a record from $X1$ to a record in $X2$. In addition to first and last names and year of birth from both original datasets, I create a number of variables to describe the features of the potential match. In principle, this set of variables could differ across matching procedures with each new dataset. However, I found a relatively consistent set of variables to be necessary and sufficient to produce accurate matches. I list these variables, along with a brief description in Table 3.2. Many of these variables are link specific, such as the Jaro-Winkler distance in first name ($fdist$) or the absolute birth year distance in years ($ydist$). Others are constant for all records containing a given original record, $x1 \in X1$. For example, *hits* is the number of potential matches for a given record in $X1$ that we find in the entire $X2$ file. In addition to Jaro-Winkler string distances, which I use to compute both $fdist$ and $ldist$, I also use the Soundex system to generate another pair of important matching variables. These variables indicate whether or not the soundex scores for the first (or last) names match between the two

[19] XX_T should be stratified on unique values in $X1$ such that if one possible match from $X1$ to $X2$ is in XX_T then all possible matches are in the training data.

[20] Christen (2012) defines this as one-to-one matching.

possible records. Soundex groups (and other phonetic groupings such as NYSIIS and phonex) have been used in the past by economic historians and others performing record linkage. The scores attempt to encode the sound of a word, relying on the first letter and sound grouping for further letters.[21]

How were the variables in Table 3.2 chosen? For model selection, one could imagine using k-fold cross validation or another formal model selection process to pick variables from the space of all variables that might describe the similarity of two given records. These techniques are common to machine learning procedures, but less frequently used in the social sciences. While a more formal method will identify the variables that are most important to the classification procedure, such efforts may be unnecessary. Instead, I rely on my experience of constructing matches between several difference census datasets to specify the variables listed in Table 3.2. In the next section, when I walk through the example of matching sons from Iowa 1915 into the 1940 census, I show that the set of variables in Table 3.2 produces a classification method that is both accurate and efficient. More generally, there may be other variables that could improve the matching fit for other datasets and the procedure could be easily modified to include them.[22]

Once the variables describing each potential link have been constructed, we are ready to estimate the relative importance of these variables in determining links in the training data, XX_T , and to assign links outside of our training data. The generation of links is a two step procedure. First, I note one special variable in Table 3.2, *exact.all.mult*, which indicates whether there are multiple hits in $X2$ which are exact matches on first name, last name, state of birth, and year of birth for a given record $x1 \in X1$. While somewhat unlikely, there are of course certain common names for whom this will occur. In most populous states in most year, for example, there are multiple men named John Smith. When *exact.all.mult* is 1, it should be clear that we cannot decide which of these exact matches is correct and we mark all of these potential matches are

[21] See the National Archives description of Soundex for a more detailed description of the Soundex method. I use Soundex because it is commonly used in census linking, but also because there are prebuilt soundex functions in Stata and in the RecordLinkage package in R.

[22] Such variables may include alternative versions of string distance or phonetic classification of names. If linking records based on non-American data, language specific phonetic codes may be valuable. If the matching is into the 1930 Census, for example, mother and father state of birth might be a useful variable to include to improve match accuracy. In my experience, syllable-count comparisons, string length comparisons, and higher orders of the string distances above are not useful features to include but this may vary between datasets.

Table 3.2: Census Matching Variables

Variable Names	Variable Description
	Record Variables
fname1	First name of record in $X1$, including middle initial if available
fname2	First name of record in $X2$, including middle initial if available
lname1	Last name of record in $X1$
lname2	Last name of record in $X2$
yob1	Year of birth in $X1$
yob2	Year of birth in $X2$
	Constructed Variables
fdist	Jaro-Winkler string distance between first names
ldist	Jaro-Winkler string distance between last names
ydist	Absolute value difference between year of birth in $X1$ and $X2$
hits	Number of records in $X2$ matched for given $X1$ observation
hits2	$hits^2$
exact	Indicator if both $lname1 \equiv lname2$ and $fname1 \equiv fname2$
exact.all	Indicator if $lname1 \equiv lname2$, $fname1 \equiv fname2$, and $yob1 \equiv yob2$
f.start	Indicator if the first letter of the first names match
l.start	Indicator if the first letter of the last names match
f.end	Indicator if the last letter of the first names match
l.end	Indicator if the last letter of the last names match
mimatch	Indicator if the middle initials match
exact.mult	Indicator if more than one hit in $X2$ for a given record in $X1$ matches first and last names exactly
exact.all.mult	Indicator if more than one hit in $X2$ for a given record in $X1$ matches first name, last name, and year of birth exactly
fsoundex	Indicator if the soundex codes of the first names match
lsoundex	Indicator if the soundex codes of the last names match

If there are any records in $X1$ with multiple exact matches in $X2$—that is exactly the same first name string, last name string, and year of birth—then we will be unable to pick between these possible matches. All possible matches are equally as likely to be the true match. Instead, we score any record links in XX with failure if $exactmult1$ is not 0. Thus, the variable $exactmult1$ is not used directly in the prediction algorithm.

non-matches.

After removing the set of records with multiple exact matches, the second step is to run a probit model on the training data, XX_T. Using the probit, I calculate the probability a given record is a match.[23] Unlike other—more common—classification problems, there is a special feature of the census linking problem. Namely, if one record from $X1$ is coded as matched to a record in $X2$, then we do not want that record from $X1$ to be coded as matched to any other records in $X2$.[24] There are some variables generated and used in the matching procedure that account for this feature, including the total number of potential hits for a given $X1$ observation. However, the probit regression does not directly account for this fact. Instead, we use a second step to ensure records are not double-matched.

In the second step, we take the generated probit scores and define matches only as those records $xx \in XX$ that meet the three following criteria. First, the score for xx is the highest score for that record in $X1$.[25] Second, the score of the match is sufficiently large (more on the parameter in the next section). Third, we require that the score of the second-best link is sufficiently small, relative to the top score, that we are confident that the best link is a match. Choosing these hyper-parameters or meta-parameters has an important role in the matching procedure and is discussed in the next section.

This second stage is important in another way. For many common names in $X1$, there might be a variety of possible links in $X2$ that are all slightly different. For example, we might match Jonny to John and Jon, with exactly matched last names and years of birth. Alternatively, there may be many John Smiths matched to different years of birth. While the algorithm might prefer the match to one, if this preference is not sufficiently strong either absolutely or relatively, we should not be comfortable declaring a match. Rejecting some of these close matches will necessarily increase the rate of false negatives, but it will also decrease the false positive rate and replicate the manual matching procedure which should attempt to limit the close judgment calls.

[23] In section 3.5, I justify the use of a probit model rather than another machine learning classifier. The probits are both familiar to social scientists performing record linkage and do extremely well in out of sample prediction, minimizing both false positives and false negatives.

[24] The same is true in the reverse; two different records from $X1$ should not link to the same record in $X2$.

[25] That is, if "James Feigenbaum" is a record in $X1$, then the match in XX could only be a match according to the final algorithm if it is the best link for $x1$.

While false negatives will lower our final samples, the biases driven by false positives are likely to be quite problematic.

After running the second stage, we can limit XX to the records that are coded as matches and proceed with analysis.[26] The match rate—the share of records in $X1$ which we are able to link to a record in $X2$—will vary across datasets. It also depends on the hyper-parameters, as shown in a different matching context by Mill (2012). In practice, I have found that the match rates from the procedure outlined here will be as good if not better than other methods traditionally used in economic history.

3.3 *An Example of The Record Linking Procedure: Iowa 1915 to 1940*

In Feigenbaum (2014), I measure intergenerational mobility of income, linking fathers from the 1915 Iowa State Census to their sons in the 1940 Federal Census. I utilize the Iowa State Census sample digitized by Claudia Goldin and Lawrence Katz for their work on the historical returns to education (Goldin and Katz, 2000, 2008). To construct my sample for census matching, I limit the Goldin-Katz sample to families with boys aged between 3 and 17 in 1915. These sons will be between 28 and 42 when I observe them again in 1940. Of course, I restrict my analysis to sons in 1915, because name changes make it impossible to locate most daughters in the 1940 Census. This leaves me with a sample of 7,580 boys, each of whom is a son in 1915 Iowa. For each son, I know his first and last name, age in 1915, and state of birth—exactly the variables I need to match into the 1940 census.

As described above, I start by extracting, for each son, the set of matches in the 1940 census with a year of birth distance of less than 3, a first name string distance of less than .2, a last name string distance of less than .2, and a matching state of birth. This returns a dataset, XX, of 79,047 records. Only 6,889 of the 7,580 boys are in this sample, suggesting that there are nearly 700 sons for whom no possible matches can be found in the 1940 census.[27] On average, each of these 6,889

[26] In the interest of speeding up the matching process and conserving memory, XX likely does not include any variables that we might want to use in analysis, like income, occupation, education, etc. The true final stage is simply locating the chosen records in $X1$ and $X2$ and bringing together all the variables of interest.

[27] These sons could no longer be living by 1940 or they may have moved out of the country or, perhaps more likely, the measurement error in recording their names, ages, or state of birth in either 1915 or 1940 was sufficiently severe that they cannot be matched at all.

sons is matched to 11.4 records in 1940. However, the distribution of potential matches is quite skewed. 1,005 of the sons return only one possible match and the median is 6 matches; 37 sons return more than 100 matches.[28]

With this dataset, I then construct an indicator variable for matched and non-matched records. For the purposes of demonstrating the procedure in this paper, I do this for the entire sample. For every son in the Iowa 1915 sample, I indicate which link, if any, in 1940 is the correct match. The rest of the links are indicated as incorrect non-matches. Based on both my own speed and the speed several research assistants with experience assessing linked records, in an hour, 500 records in $X1$ could be assessed. Of course, constructing all of these links for every observation in the full sample obviates one reason for an algorithmic approach; the algorithm allows us to match large datasets without the time or monetary costs of manually matching. But it is necessary to have "correct" classifications for all observations to properly build the matching model in this paper and assess its accuracy.

I then randomly partition the data into two equal parts, sampling at the $x1$ record level so as to keep all possible matches for a given son in Iowa in the same data partition. One partition will be the training data, where I will fit my probit model. The other partition will be held back to test the accuracy of my algorithm out of sample.[29]

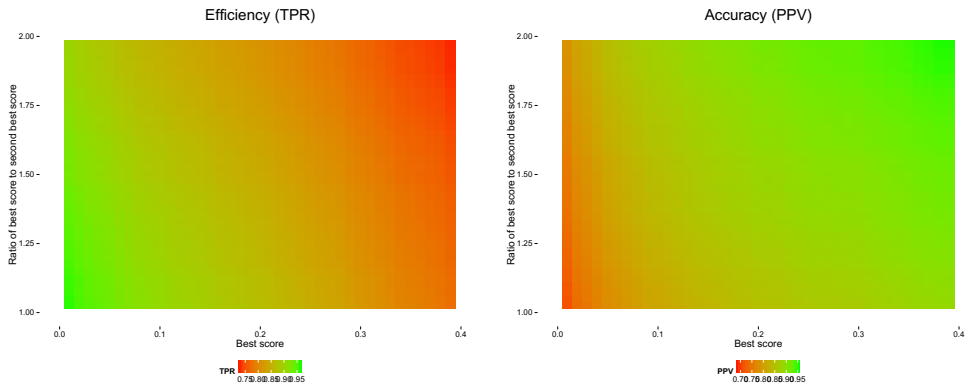
There are many possible variables that could describe how likely a given matched pair of records are to be a true match. I focus first on the set of variables listed in Table 3.2. As described previously, I run a probit regression on the dummy variable indicating whether or not the record was declared a match or not-match in the human review process on these variables. I rely on the `caret` package in R, a standard package used in machine learning to train my probit model. The model is trained with bootstrap sampling from the training data, taking 25 samples to estimate the final probit coefficients.[30]

The results of this regression are presented in Table 3.3. I show both the probit that I use

[28] John Barile, born in Pennsylvania in 1909 returns 364 possible matches. John is a very common name, Barile is very close in string distance to a large number of names (Bradley, Bailey, Barnes, Barrett, and Riley), and Pennsylvania had a large state population in 1909.

[29] In section 3.4, I show that in fact, training on 50% of the data is overkill and that the method yields accurate predicted matches using a much smaller share of data for calibration.

[30] The model parameters are largely the same when using k-fold cross validation to choose parameters or simply running one probit regression on the full sample of training data.



(a) TPR Values (Share of Type I Errors) **(b)** PPV Values (Share of Type II Errors)

Figure 3.3: Match Algorithm Quality and Hyper-Parameter Values

and an alternative logit model. The parameters presented are the direct model coefficients, not marginal effects as is common in a probit model.

Using the coefficients estimated in the probit in Table 3.3, I calculate the predicted score. These scores run from 0 to 1; higher values suggest a stronger likelihood of the records being true matches.[31] Following the method described earlier, I define matches only as those records that meet the three following criteria. First, for a given son in the Iowa 1915 sample, the score is the highest score of all matches. Second, the score of the match is sufficiently large. Third, that the score of the second-best link for the Iowa 1915 son is sufficiently small, relative to the top score, that we are confident that the best link is a match.

Clearly, at this point we cannot proceed unless we define the two meta-parameters referenced above: the absolute threshold for declaring a match and the relative threshold for declaring a match better than the next best alternative. How to pick these meta-parameters? I attempt to minimize false positives and false negatives. To make this concrete, I use two standard machine learning assessment measures, the true positive rate (TPR) and the positive prediction rate (PPR). The TPR is the ratio of true positive to total positive matches in our training data. The PPR is the ratio of the true positives to the total matches made by the algorithm.

Call the first meta-parameter b_1 and the second b_2. b_1 will take a value between 0 and 1, matching the range of our scores. In theory, b_2 should range from 1 to ∞ (it is the ratio of two scores between 0 and 1). I search over the grid of possible values of b_1 and b_2 for the maximums of

[31] As these scores are based on predictions from a probit model, the scores are simply $\Phi(X'\hat{\beta})$.

Table 3.3: Iowa 1915 to 1940 Census Linking Model

	Probit	Logit
First and Last name match	0.632^{***}	1.129^{***}
	(0.086)	(0.168)
First name distance, Jaro-Winkler	-6.071^{***}	-11.543^{***}
	(0.525)	(0.994)
Last name distance, Jaro-Winkler	-10.285^{***}	-19.145^{***}
	(0.487)	(0.954)
Absolute Value Difference in Year of Birth is 1	-0.708^{***}	-1.308^{***}
	(0.044)	(0.083)
Absolute Value Difference in Year of Birth is 2	-1.562^{***}	-2.893^{***}
	(0.065)	(0.126)
Absolute Value Difference in Year of Birth is 3	-2.316^{***}	-4.370^{***}
	(0.102)	(0.208)
First name Soundex match	0.153^{***}	0.294^{***}
	(0.054)	(0.100)
Last name Soundex match	0.698^{***}	1.341^{***}
	(0.069)	(0.135)
Hits	-0.064^{***}	-0.123^{***}
	(0.002)	(0.005)
Hits-squared	0.0003^{***}	0.001^{***}
	(0.00002)	(0.00004)
More than one match for first and last name	-1.690^{***}	-3.217^{***}
	(0.093)	(0.183)
First letter of first name matches	0.871^{***}	1.593^{***}
	(0.130)	(0.245)
First letter of last name matches	0.886^{***}	2.003^{***}
	(0.148)	(0.356)
Last letter of first name matches	0.147^{***}	0.312^{***}
	(0.053)	(0.101)
Last letter of last name matches	0.649^{***}	1.239^{***}
	(0.070)	(0.139)
Middle Initial matches, if there is a middle initial	0.537^{***}	0.908^{***}
	(0.097)	(0.186)
Constant	-1.479^{***}	-3.087^{***}
	(0.225)	(0.480)
Observations	38,091	38,091
Log Likelihood	−2,440.877	−2,444.649
Akaike Inf. Crit.	4,915.753	4,923.298

Table 3.4: Table Summary

	Hyper-Parameters		Algorithm Quality	
Relative Weight on PPV	b1	b2	PPV	TPR
0.500	0.050	1.125	0.784	0.931
0.750	0.090	1.300	0.831	0.897
1	0.140	1.375	0.858	0.875
1.250	0.200	1.975	0.912	0.815
1.500	0.200	1.975	0.912	0.815

The utility function in PPV and TPR is defined as TPR + γPPV where γ is the relative weight on PPV in column 1. The algorithm quality metrics, PPV and TPR, are calculated with respect to the training partition of the data.

TPR and PPV, calculated from the same training data we used to generate the probit coefficients previously. Figure 3.3 presents a graphical representation of these grid searches. As we change parameters to increase PPV, TPR tends to fall. This makes sense: the more restrictive we are on matching matches, the fewer false matches we will find, but fewer true matches will be found as well. One solution is to maximize some utility function with two arguments, PPV and TPR. The simplest function would be a weighted sum of the two values. Choosing weights is subjective and may vary between projects, depending on the costs of including false positives relative to the benefits of finding more true positives. In Table 3.4, I show the optimal hyper-parameters under various weighting schemes. In the analysis that follows, I use a weight of 1 and thus select b_1 = 0.14 and b_2 = 1.375.

With these hyper-parameters selected, I can return to my matched dataset, identify the links (and the non-links) and proceed to my other analysis. We can see from Table 3.4 that the PPV is 85.8% and the TPR is 87.5%. But these are the in-sample results and reflect only how well my model is tuned to the data I built it with. How does the algorithm do out of sample, predicting matches and non-matches on the data that I held back from the training? To test this, I create predictions on the test set, using the probit coefficients listed in Table 3.3 and the hyperparameters b_1 = 0.14 and b_2 = 1.375. Table 3.5 presents a confusion matrix, listing the counts of true and false positives and the true and false negatives, as well as the PPV and TPRs from applying the algorithm to the test data. This out-of-sample prediction is a much more difficult test for the record linking algorithm than the in-sample predictions made in Table 3.4. However, the

Table 3.5: Confusion Matrix: Out of Sample Predictions

	True Status	
Algorithm Prediction	Not Matched	Matched
Not Matched	35,608	295
Matched	272	1,869

The True Positive Rate is the number of true positives over the total number of matches, which is the sum of the true positives and the false negatives. This rate indicates the efficiency of the algorithm: how many of the matches in the full data were identified by the algorithm. The out-of-sample TPR is 86.4%.

The Positive Predictive Value is the number of true positives over the total number of positives, which is the sum of the true positives and the false positives. This rate indicates the accuracy of the algorithm: how many of the matches made by the algorithm are in fact matches in the true data. The out-of-sample PPV is 87.3%.

algorithm does extremely well. The TPR is 86.4%, less than one percentage point worse on the test data relative to the training data. The PPV is 87.3% which is, in fact, slightly higher than in the training data. The interpretation is that of the matches identified by the algorithm, 87.3% were coded by RAs as a match and that of the matches coded by RAs, 86.4% were identified by the matching algorithm.

Rather than using the TPR and PPV, economic historians undertaking matching procedures typically judge a link between two censuses on its matching rate. Matching rates vary in the literature between different samples and procedures, but are often between 20% and 50%. Of the original 7,580 boys in my Iowa sample, I match 4,349 uniquely and accurately with my matching algorithm, for a match rate of 57.37%. While the algorithm is far from perfect—neither the PPV nor TPR hits 100%, nor does the match rate—the improvements are large. For samples that begin relatively small, an increase in the match rate would be even more important to the potential analysis.

3.4 How Much Training Data is Enough?

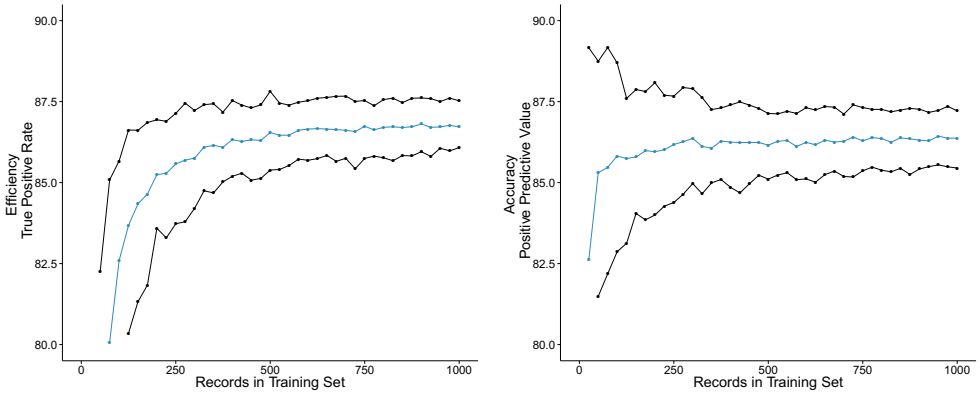
In Table 3.3, I estimated a set of coefficient values that do well in locating matches between the Iowa 1915 State Census and the 1940 Federal Census. The coefficients were estimated using only half of the full dataset, the training data. In Table 3.5, I evaluated the predictions on the

other half of the data—data that had not been used to tune the parameter. This suggested that the record linking procedure and coefficients were very able to identify matches out of sample. In future matching projects, for example between the 1910 and 1930 censuses, one could simply calculate the variables described in Table 3.2 and apply the coefficients to an entirely new sample. Given the good out-of-sample performance of the algorithm, this is likely to produce reasonably good matches. However, one might also create training data based on a subsample of the *new* data in question and follow the steps outlined above to generate a new project specific matching algorithm and hyper-parameters. Naturally, there are benefits and costs to either choice. Using the matching algorithm presented here will save time and research funds because no manual training data will be necessary. The major cost is common to any use of out-of-sample predictions: less accuracy. This cost is likely to be especially high when matching between datasets that are very different than the two I built the algorithm with in the first place. In this case, out of sample has two very different meanings. The algorithm was tested on data held back from its training, but it was not tested on data that was of the same type as the training data but built from a wholly different set of censuses.

If a researcher is going to use training data specific to the datasets to be used, one important practical question is how big should the training data be? How many records need to be manually adjudicated? The traditional empirical adage that more data is usually better applies here, but I suggest a more exact answer. To do this, I randomly subset the data used in the previous section, train a match algorithm, and test how much the algorithm changes and, more importantly, how much the measures of accuracy and efficiency change as the subset of data used for training grows.

The procedure is a straightforward cross-validation exercise. I start by drawing a π random sample of my data, blocking at the unique $X1$ observation level.[32] This draw will be the training data for a given iteration. I then train the match algorithm on my training data and apply the algorithm to the test data—here the test data is the $N - \pi$ of the data not sampled. Normally, it would be impossible to measure the error rates on the non-training sample. However, because in reality I have matches for the entire sample, I can compare the predictions based on the π

[32] That is, I include all possible links for a given record in the $X1$ data to the $X2$ data.



(a) True Positive Rate **(b)** Positive Prediction Value

Figure 3.4: Algorithm Efficiency and Accuracy increases as the size of the training data set increases, but most gains are realized with no more than 500 records of the data used in training.

sample with the test data and record a measure of efficiency (TPR) and accuracy (PPV). I draw 100 random samples for each π-sized block of my data from $\pi = 25$ to $\pi = 1000$, by 25. I graph the resulting TPR and PPV in Figure 3.4, along with 95% confidence intervals.

The speed of convergence to the full-sample results is quite rapid. With around 500 records trained—on average, less than 10% of the original $X1$ data—the algorithm identifies links with a TPR of nearly 87% and a PPV of 86%. These TPR and PPV results are approximately as large as the corresponding rates when the algorithm had 50% or more of full data to train with. This implies that, in fact, the answer to "how much training data is required?" is quite small relative to the overall size of the project.

How long does it take to create training data based on 1400 or so observations? Based on records from my Iowa matching project, RAs can assess the links for about 500 $X1$ records per hour. Thus, in less than 3 hours of work (6 with double entry), training data can be constructed which—once used in the matching procedure outlined above—generate a final matched sample with a nearly 90% efficiency and match accuracy.

3.5 *Machine Learning Method Selection*

I use a probit model to assign scores to each potential match. However, there are many different classification techniques in the machine learning toolbox. Why probit? For one, probit regression is a well-understood concept among economic historians and other social scientists

who might want to create census links. Beyond familiarity, I have found that the probit model, while perhaps simple when compared to newer methods in machine learning, performs extremely well out of sample. In this section, I compare the cross validation matching results from the probit model to two other common machine learning classifiers—random forests and support vector machines—as well as methods commonly used by economic historians. As I will show, my probit-based method makes an ideal choice to use in record linking procedures.

I use cross validation to assess the various record linkage procedures on my Iowa 1915 sons data. Because I have fully hand-linked this sample ahead to the 1940 census, I am able to apply the predictions from each model to the held-out testing data and assess the accuracy of each, based on both the PPV and the TPR. I begin by splitting the set of potential matches in the 1940 census for the sons of the Iowa 1915 census into a training and testing set. Each set contains 50% of the sons from 1915. I sample at the son level, rather than the potential match level, because matches are only identified relative to other potential matches for a given son. I use the same set of independent variables or features in each model, those listed in Table 3.2.[33]

I find the probit classifier to be the best regression method for record linkage. In addition to probit, I also assess a logit and OLS classifier. All three of these models are built in the same way. In the first stage, I regress an indicator variable for match or not match on the linkage features. In the second stage, I declare matches based on the absolute and relative scores from the models. As I document in the first three rows of Table 3.6, both the probit and logit models have strong accuracy and efficiency on the testing data.[34] The OLS model, not surprisingly given the dichotomous nature of the outcome, is slightly worse on both efficiency and accuracy.

Many machine learning classification or discrimination tasks are solved not with regression models but with more flexible, non-parametric solutions. I assess both random forests and support vector machines (SVMs) in the third and fourth rows of Table 3.6. While both do quite well on the training data, the accuracy and efficiency of both are no better than my main probit method on the testing data which the models were not trained on.

[33] While feature discovery is relatively standard in the machine learning literature, it is much less frequently used in the social sciences. I follow this convention and specify the features in advance.

[34] The training results for the probit model replicate the results from Table 3.4. For each model in Table 3.6, I have chosen the hyper-parameters to maximize the sum of the TPR and the PPV.

Table 3.6: Comparing Matching Algorithms

	Training Data		Testing Data	
Model	Efficiency (TPR)	Accuracy (PPV)	Efficiency (TPR)	Accuracy (PPV)
My Method	0.881	0.854	0.879	0.858
Logit	0.874	0.861	0.873	0.862
OLS	0.884	0.792	0.856	0.792
Random Forest	0.973	0.947	0.756	0.868
SVM	0.851	0.904	0.827	0.891
Exact	0.192	0.830	0.195	0.858
Exact (yob $\pm$ 3)	0.292	0.861	0.295	0.883
Soundex	0.631	0.782	0.601	0.766

All models fit on the same training data with the same set of features described in Table 3.2. Training data is a 50% random sample of the full Iowa sons linked data from 1915 to 1940. Hyper-parameters are optimized using an equal weighting scheme on both the TPR and PPV in the training data.

Random forest classifiers may be best thought of as an advancement of simple decision trees. In the record linkage case, a decision tree could be used to split records into matches and non-matches. At the first node, the tree might split at some critical value of first name string distance and at the second on a critical value of last name distance. Later nodes could split based on matching first or last name soundex results, whether or not middle initials match, etc. In Figure 3.5, I show one possible decision tree with 2 levels. At the first node, the records are split on the Jaro-Winkler string distance in last name. Records with a last name distance of more than 0.047 are sent in one direction (and ultimately classified as non-matches), while records with relatively close last names (≤ 0.047) go to a second node. At this second node, records are classified based on the Jaro-Winkler string distance in first name. Records with small first name distances are coded as matches and the rest are coded as non-matches. For such a painfully simple model, the tree does reasonably well. Of the 1844 records in the matched node, 47% are matches; the non-match nodes are reasonably accurate as well.[35] However, this was only one of the many trees that could be grown to classify the census data. Further, this tree has only two levels of decisions—it might be possible to extract more precision by using additional variables.

[35] Obviously, I have yet to utilize the special one-to-one matching feature of my data and these results are not directly comparable to those presented previously.

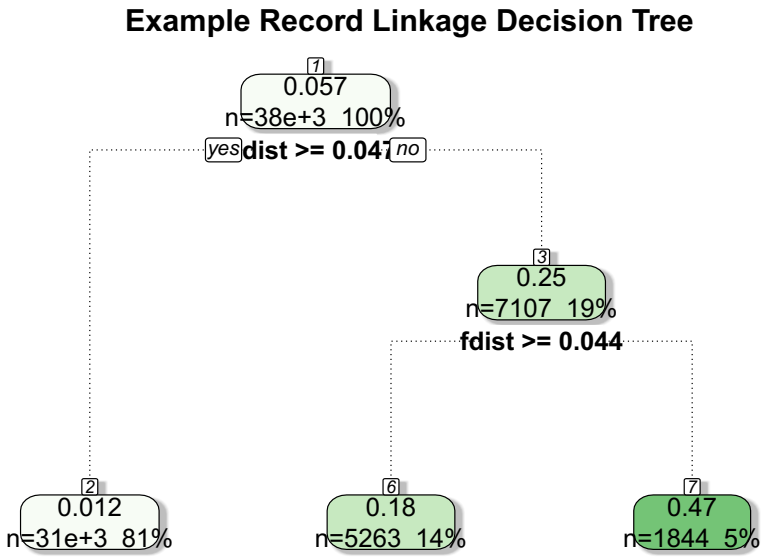


Figure 3.5: Example Decision Tree based on first and last name Jaro-Winkler String Distance

Random forests consist of the averaging of many different decision trees grown to much greater depth than the tree in Figure 3.5. The randomness is introduced in two ways. First, the trees are trained on bootstrapped samples of the training data. This induces natural variation in both the optimal splitting variables and in the critical values to split at. Second, the features available to split at each level are a random subset of all features in the data.[36] Within this sample of the training data and with these feature constraints at each node, the optimal decision tree is tuned. By averaging over each tree's prediction, the random forest algorithm can generate a classification or a score. In my case, this is a score from 0 to 1 and can be interpreted as the likelihood that a given record is in fact a match. As with the probit model, I then compare the scores for a record in $X1$ across all possible matches and identify as matched only those records that are the highest score at the $X1$ level, sufficiently large, and sufficiently larger than the next best record.[37]

Support vector machines (SVMs) search for the optimal separating hyperplane between two

[36] Where p is the number of features or independent variables, usually only $\sqrt{p}$ or $p/3$ features are considered at each node. This avoids high correlation across trees, which might arise if certain features are very powerful predictors.

[37] As before, the metaparameters used to distinguish how large of a score is needed, both absolutely and relatively, to be a match are generated by maximizing the sum of the TPR and the PPV in the training data. The metaparameters for the random forest are different from those of the probit or logit models.

groups. Each feature in my dataset—the first and last name string distances, the year of birth difference, the soundex score indicator, etc—is a dimension. Again, the SVM returns a score for each possible match of a given record in $X1$ and I use the same second step to identify the matches that are sufficiently strong and relatively strong.

In the fourth and fifth rows of Table 3.6, I give the PPV and TPR in the training data and the testing data for the Random Forest classifier and the SVM. The probit beats both the random forest classifier and the SVM on the test data. First, it is quite clear that the random forest outperforms the probit on the training data. The random forest scores a TPR of 97%, suggesting that the model locates nearly all of the matches identified by a research assistant. The PPV is similarly high at nearly 95%—very few of the records suggested by the algorithm were incorrect. The SVM is comparable to the probit model on the training data, with less efficiency but more accuracy.

However, the results on the testing data are much less promising for the random forest method. Again, the testing data is the half of the full sample that was held back during training of the models and is an accurate assessment of the various model's out of sample properties. The random forest model has a TPR on the test data of only 75.6%. Thus, many of the actual matches coded manually are not identified by the algorithm. The PPV for the random forest model is similar the the PPV for the probit and logit models. Thus, the random forest is no more or less accurate than the probit model, but it is much less efficient, tagging many true matches as non-matched.

Finally, I compare my machine learning approach to a few standards from the economic history literature.[38] In all cases, I require state of birth to match exactly. The exact model requires matches to agree exactly on first name, last name, and year of birth. Not surprisingly, these method is much less efficient than the other models presented. Good matches that may be off because of a small transcription error will not be matched by the exact method. In addition, I find the accuracy of the exact procedure to be no better than my main method. I realize the exact match criteria slightly in the seventh row of Table 3.6: records can still only be declared matches if the first and last names match exactly, but I now allow a three year window for year of birth,

[38] Because none of these models are trained, the efficiency and accuracy metrics are roughly the same on the training data and the testing data.

accounting for small errors in age in the census. While this flexibility improves the efficiency somewhat, it is still far below the machine learning methods, suggesting that many matches that would be made by hand are not included.

How does my machine learning approach compare to a linking procedure based on phonetic codes, as is used by Long and Ferrie (2007a), Collins and Wanamaker (2015), and others? In the final row of Table 3.6, I apply a soundex model to my Iowa data and report the results. The method requires that candidate links agree on both soundex of first names and soundex of last names. Within that set, if one record matches on year of birth exactly, it will be declared a match. If no such record exists, the year of birth error grows to ± 1, then ± 2, and then ± 3, as is standard. While the soundex model outperforms exact matching, it is significantly worse on both efficiency and accuracy than my machine learning method. The efficiency loss is straightforward: any names with transcription errors in the first few characters will have different soundex codes.[39] The accuracy loss is likely driven by the same sorts of problems: if the correct match drops out because of early string transcription errors, another match may be just good enough to take its place.

The machine learning approach outperforms the standard record linkage in the economic history literature. Ultimately, this is not surprising: the machine learning approach embeds the other methods, but also increases the degrees of freedom. That is, if the name strings and years of birth match exactly or if the soundex scores match, that will increase the match score. But, rather than putting all the decision weight on just a few features, the machine learning approach allows the training data (and implicitly the researcher) to determine the weights over a large feature set.

Based on the results presented in Table 3.6 and the simplicity of implementing a probit classifier[40]—I suggest using the probit model when implementing the record linkage procedure outlined in this paper.

[39] For example, I have seen "Feigenbaum" mistranscribed in the 1910 census as "Teigenbaum" likely because the cursive F and T are so similar looking.

[40] Easily done in Stata, R, etc. Random forest packages exist for these statistical programs and for others but are not often used by economists.

3.6 *Does the Algorithm See Dead People?*

Does my record linking procedure making incorrect links? One particularly troublesome error would be linking the record of someone who has died in between the collection of the two datasets. However, it is not usually possible to test for these sorts of bad matches; of course, the 1900 census does not include information on when (in the relative future) respondents will die. Instead, I turn to another source of records, the Lahman baseball encyclopedia, that includes both name, state of birth, and year of birth information, as well as externally validated date of death data. When linking the universe of Major League Baseball players ahead from 1903 to 1940, I find a "ghost matching" rate more than 45 percentage points lower than the live matching rate in the same sample, suggesting that my algorithm is usually able to avoid erroneous matches.

Constructed by journalist and sabermetrician Sean Lahman, the Lahman (2016) database records statistics for baseball players from 1871 to 2015. In addition to batting average, on-base percentage, and ERA, the database also includes biographical information for the more than 18 thousand players in Major League Baseball history. In particular, I draw on information about players' names, dates of birth, place of birth, and date of death that enable me to link players to the census.

I link the 381 active, American-born players in the 1903 season to the 1940 census. Of the 381 players, only 220 were still alive in 1940, according to the Lahman database. I am able to find 52% of these players in the 1940 census, using my algorithm.[41] The match rate is far smaller for the 161 players no longer alive in 1940: the algorithm has a ghost matching rate of only 6%. While this small handfull of matches are incorrect, the low number—both absolutely and relative to the live players—underscores the quality of the algorithm to make links between datasets, even when evaluated on a feature or metric not available during training.

3.7 *Conclusion*

In this paper, I detail a transparent census matching technique for constructing linked samples that can be replicated across a variety of cases. The procedure applies insights from machine

[41] Because the number of players overall is so small, I do not train a new record linking algorithm. Instead, I use the algorithm trained with the Iowa 1915 sons, described above, and the weights shown in Table 3.3.

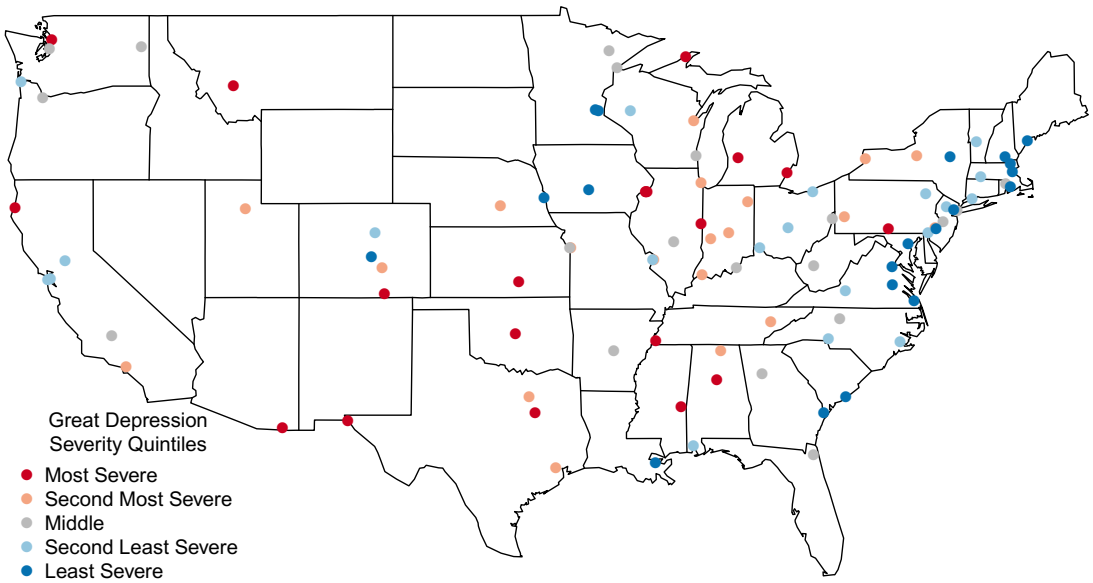
learning classification and text comparison to the well known problem of record linkage, but with a focus on the sorts of costs and benefits of working with historical data. Using tools readily available to economists and other social scientists in Stata or R, the method can automate the linking of records with minimal manual matching work and high levels of efficiency and match accuracy.

APPENDIX

A. APPENDIX TO CHAPTER 1

A.1 Additional Figures and Tables

Figure A.1: City-level Great Depression Severity, measured using per capita growth in retail sales from 1929 to 1933 (Fishback, Kantor, and Wallis, 2003).

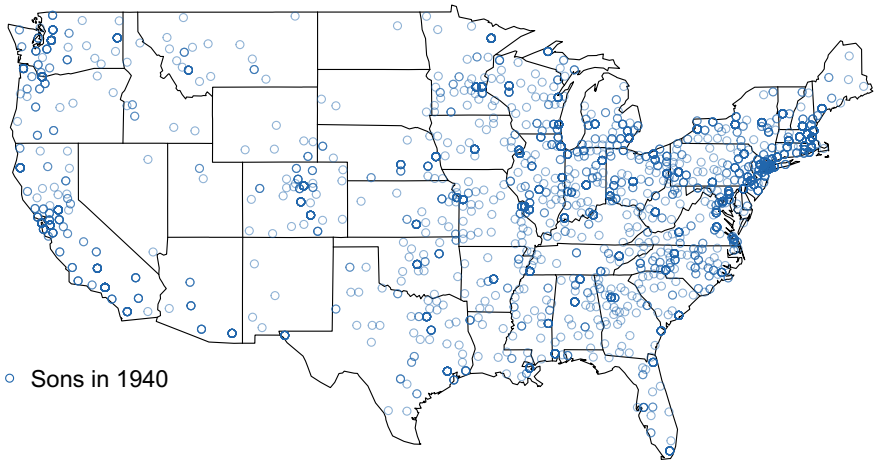


This map illustrates the variation in Great Depression downturn across the 99 American cities from the BLS Cost of Living Survey. Downturn severity is split into quintiles. Cities with the most severe downturns are in red, cities with the most mild downturns are in blue. Great Depression severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933.

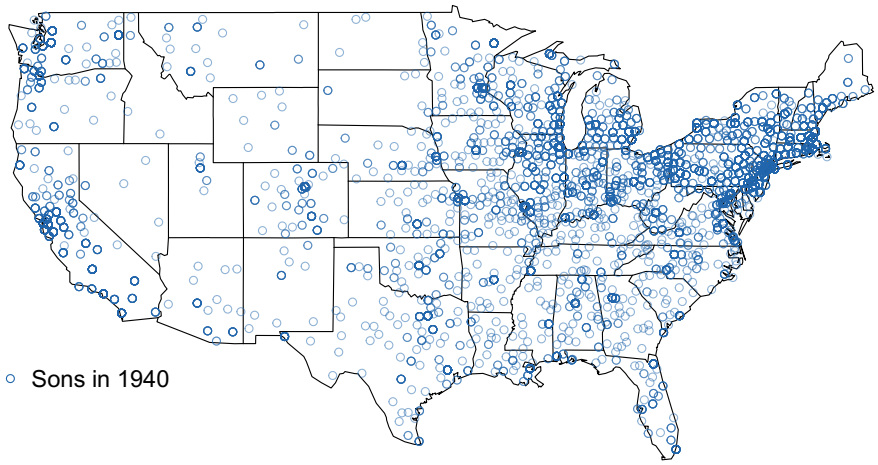
A.2 What Predicts Local Great Depression Severity?

In this section, I review the literature on the Great Depression and its local variation across the United States and empirically test theories of local variation on the cities in my sample.

Figure A.2: The 1940 location of sons in my sample. Sons are plotted in the county in which they reside.



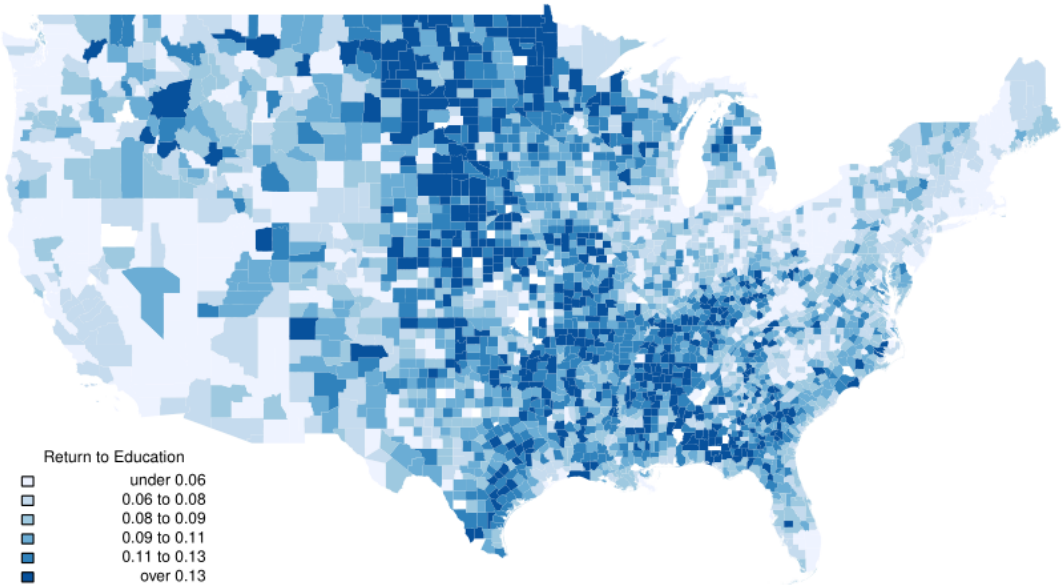
(a) Sons in 1940 from the BLS Sample



(b) Sons in 1940 from the full IPUMS Sample

These maps illustrate the locations in the 1940 census of the sons in my two intergenerational samples.

Figure A.3: The Estimated Return to Education from a Mincerian Earnings Regression by County in 1940



I estimate the return to a year of education in the complete 1940 census. I observe annual earnings in 1939, completed years of education, age, and place of residence in 1940. Separately for each of the 3,071 counties in the country, I run a simple Mincerian returns to education regression, regressing log earnings on years of education and age dummies for all full-time employed white men between the ages of 16 and 65 in the county. The coefficient on education in each regression is the observed return to education, mapped above. These estimated returns to education are descriptive and not causal. The returns to education correlate negatively with county population: an increase in county population by 1% correlates with a 1% smaller estimated return. It is also apparent from the map that the human capital returns were higher in the plains states—the Dakotas, Nebraska, and Kansas—as well as the in mid-South. The returns to education were extremely low in New England and New York, as well as in the industrial Midwest—Ohio, Illinois, Michigan, and Indiana—and on the West Coast.

Table A.1: Great Depression Severity Decreases Intergenerational Mobility: Local Cost of Living Adjusted Earnings

	IGE				Rank-Rank			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Father Earnings 1920	0.256***	0.281***	0.173***	0.182***				
	(0.042)	(0.044)	(0.055)	(0.058)				
Log Father Earnings 1920 X GD Normalized Severity	0.074*	0.109**						
	(0.043)	(0.046)						
Log Father Earnings 1920 X GD Above Median Severity			0.165**	0.196**				
			(0.082)	(0.085)				
Father Earnings Rank 1920					0.178***	0.186***	0.102**	0.101**
					(0.029)	(0.031)	(0.044)	(0.047)
Father Earnings Rank 1920 X GD Normalized Severity					0.074**	0.096***		
					(0.029)	(0.033)		
Father Earnings Rank 1920 X GD Above Median Severity							0.142**	0.162**
							(0.058)	(0.062)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4730	4730	4730	4730	4952	4952	4952	4952
Clusters	99	99	99	99	99	99	99	99
Adjusted R^2	0.165	0.175	0.165	0.175	0.137	0.142	0.137	0.141

Earnings of both fathers and sons are adjusted for variations in local prices. For fathers, I use median rents by city in 1920, drawing on data from the BLS Cost of Living Survey. For sons, I use median rents by county in 1940 from the 1940 Federal Census. Earnings are divided by the rental price index in the city of residence (for fathers based on 1920, for sons based on 1940). These yield estimates of real annual earnings. This table replicates, but with the local price adjustment, the results in Table 1.4. Estimates of intergenerational mobility based on a linked sample from the BLS survey of urban families in 1918-1919 to the 1940 Federal census. Each column is a regression of the son's outcome in 1940 on the father's corresponding outcome in 1918-1919, a measure of Great Depression severity in the city of residence in 1918-1919, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. In the odd columns, I include state fixed effects and direct controls for Great Depression severity (normalized in columns 1 and 5, above or below median in columns 3 and 7) but omit the point estimates from the table. In the even columns, these controls are absorbed by the city fixed effects. All fixed effects are based on the city of residence in 1918-1919. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. IGE is the intergenerational elasticity of income, and the dependent variable is log son earnings in 1940. Rank-rank mobility compares the son's position in the earnings distribution in 1940 to the father's position in 1918-1919 and the dependent variable is son earnings rank in 1940.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample; 1940 Complete Count Census.

Table A.2: Great Depression Severity Decreases Intergenerational Mobility in All Cities

	Son OccScore (IPUMS, All Cities)					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Father Occupation Score 1920	0.261***	0.264***	0.266***	0.268***	0.244***	0.248***
	(0.007)	(0.008)	(0.007)	(0.007)	(0.011)	(0.010)
Log Father Occupation Score 1920 X GD Normalized Severity			0.020**	0.019*		
			(0.010)	(0.010)		
Log Father Occupation Score 1920 X GD Above Median Severity					0.043***	0.037**
					(0.015)	(0.015)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	No	Yes	No	Yes	No
City Fixed Effects	No	Yes	No	Yes	No	Yes
Observations	35035	35035	33067	33067	33067	33067
Clusters	161	161	143	143	143	143
Adjusted R^2	0.120	0.116	0.122	0.118	0.122	0.118

Estimates of intergenerational mobility based on occupation scores for all cities. Mobility estimates in all columns are based on a linked sample from the IPUMS 1% sample of the 1920 census to the full 1940 census. Rather than restrict the sample to sons living in one of the 99 cities surveyed by the BLS in 1918-1919, I include all sons living in cities in 1920. Each column is a regression of the son's log occupation score in 1940 on the father's log occupation score in 1920. In the first two columns, I present an overall measure of mobility, comparable to the occupation score elasticities estimated in Table 1.3. In the four final columns, I include both a measure of Great Depression severity in the city of residence in 1920 and an interaction of severity and the father's occupation score. These columns are comparable to Table 1.5, showing the effect of Great Depression on occupation score based mobility. Controls include quartics in the son's and father's ages. In the odd columns, I include state fixed effects and direct controls for Great Depression severity (normalized in column 3, above or below median in column 5) but omit the point estimates from the table. In the even columns, these controls are absorbed by the city fixed effects. All fixed effects are based on the city of residence in 1920. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. Occupation scores are calculated as the national median income for men in the occupation.

Source: IPUMS 1920 1% Census Sample; 1940 Complete Count Census.

Table A.3: Matching Procedure Does Not Drive Results

	IGE			Rank-Rank		
	(1) Original	(2) Remove Best	(3) Remove Worst	(4) Original	(5) Remove Best	(6) Remove Worst
Log Father Earnings 1920	0.280***	0.305***	0.279***			
	(0.043)	(0.049)	(0.050)			
Log Father Earnings 1920 X GD Normalized Severity	0.108**	0.111**	0.121**			
	(0.042)	(0.047)	(0.050)			
Father Earnings Rank 1920				0.213***	0.216***	0.210***
				(0.033)	(0.037)	(0.037)
Father Earnings Rank 1920 X GD Normalized Severity				0.100***	0.089**	0.097**
				(0.036)	(0.041)	(0.037)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4730	3629	4131	4952	3788	4324
Clusters	99	99	99	99	99	99
Adjusted R^2	0.188	0.178	0.189	0.153	0.147	0.154

Estimates of intergenerational mobility based on a linked sample from the BLS survey of urban families from 1918-1919 to the 1940 Federal census. Each column is a regression of the son's outcome in 1940 on the father's corresponding outcome in 1918-1919, a measure of Great Depression severity in the city of residence in 1918-1919, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. In the second and fifth columns, I exclude observations enumerated by agents in the top quarter of match rates ($\geq$ 63%). In the third and sixth columns, I exclude observations enumerated by agents in the bottom quarter of match rates ($\leq$ 48%). All fixed effects are based on the city of residence in 1918-1919. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. IGE is the intergenerational elasticity of income. Rank-rank mobility compares the son's position in the earnings distribution in 1940 to the father's position in 1918-1919.

Source: BLS Cost of Living Survey 1918-1918; 1940 Complete Count Census; Census of Retail Sales

Table A.4: Depression Effects on Intergenerational Mobility Only for Migrants

	IGE			Rank-Rank		
	Full Sample (1)	Migrants (2)	Non-Migrants (3)	Full Sample (4)	Migrants (5)	Non-Migrants (6)
Log Father Earnings 1920	0.280***	0.352***	0.236***			
	(0.043)	(0.069)	(0.051)			
Log Father Earnings 1920 X GD Normalized Severity	0.108**	0.154**	0.063			
	(0.042)	(0.063)	(0.047)			
Father Earnings Rank 1920				0.213***	0.252***	0.183***
				(0.033)	(0.057)	(0.045)
Father Earnings Rank 1920 X GD Normalized Severity				0.100***	0.102**	0.074
				(0.036)	(0.050)	(0.048)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4730	2001	2729	4952	2090	2862
Clusters	99	99	99	99	99	99
Adjusted R^2	0.188	0.173	0.243	0.153	0.135	0.206

Estimates of intergenerational mobility based on a linked sample from the BLS survey of urban families from 1918-1919 to the 1940 Federal census. Each column is a regression of the son's outcome in 1940 on the father's corresponding outcome in 1918-1919, a measure of Great Depression severity in the city of residence in 1918-1919, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. The first and fourth columns replicate my main results from Table 1.4, columns (2) and (6). In the second and fifth columns, I restrict the sample to those sons migrating from their 1920 city of residence in 1940. In the third and sixth columns, I restrict the sample to those sons remaining in their 1920 city of residence. The results suggest that the Depression effected the mobility parameter through the migrating sons, not the sons remaining in their 1920 cities. All fixed effects are based on the city of residence in 1918-1919. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. IGE is the intergenerational elasticity of income. Rank-rank mobility compares the son's position in the earnings distribution in 1940 to the father's position in 1918-1919.

Source: BLS Cost of Living Survey 1918-1918; 1940 Complete Count Census; Census of Retail Sales

Table A.5: New Deal Spending Did Not Affect Intergenerational Mobility

	Panel A. Intergenerational Elasticity						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Father Earnings 1920	0.266***	0.264***	0.259***	0.267***	0.265***	0.265***	0.268***
	(0.040)	(0.040)	(0.040)	(0.039)	(0.040)	(0.039)	(0.038)
Log Father Earnings 1920 X GD Normalized Severity	0.105***	0.105***		0.105***	0.103***	0.104***	0.102***
	(0.040)	(0.038)		(0.040)	(0.038)	(0.037)	(0.035)
Log Father Earnings 1920 X Total New Deal Spending		-0.043	-0.044				
		(0.037)	(0.038)				
Log Father Earnings 1920 X New Deal Relief Spending				0.011			0.025
				(0.041)			(0.039)
Log Father Earnings 1920 X New Deal Public Works					-0.040		-0.040
					(0.047)		(0.048)
Log Father Earnings 1920 X New Deal Real Estate and Insurance						-0.075*	-0.076*
						(0.040)	(0.040)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4730	4730	4730	4730	4730	4730	4730
Clusters	99	99	99	99	99	99	99
Adjusted R^2	0.199	0.199	0.197	0.198	0.199	0.199	0.199

	Panel B. Rank-Rank Mobility						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Father Earnings Rank 1920	0.203***	0.203***	0.205***	0.205***	0.202***	0.204***	0.206***
	(0.033)	(0.033)	(0.034)	(0.033)	(0.033)	(0.032)	(0.032)
Father Earnings Rank 1920 X GD Normalized Severity	0.077**	0.077**		0.076**	0.077**	0.076**	0.074**
	(0.034)	(0.034)		(0.034)	(0.033)	(0.033)	(0.033)
Father Earnings Rank 1920 X Total New Deal Spending		-0.007	-0.005				
		(0.037)	(0.039)				
Father Earnings Rank 1920 X New Deal Relief Spending				0.022			0.026
				(0.035)			(0.035)
Father Earnings Rank 1920 X New Deal Public Works					0.015		0.013
					(0.040)		(0.040)
Father Earnings Rank 1920 X New Deal Real Estate and Insurance						-0.040	-0.043
						(0.030)	(0.029)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4952	4952	4952	4952	4952	4952	4952
Clusters	99	99	99	99	99	99	99
Adjusted R^2	0.151	0.151	0.150	0.151	0.151	0.151	0.151

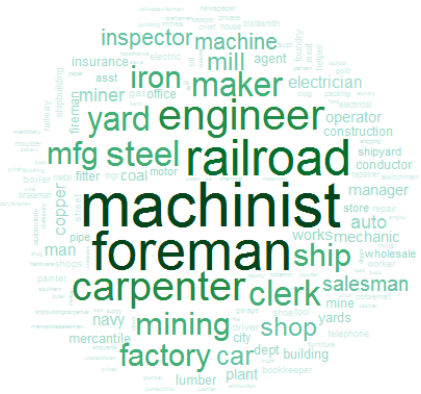
Great Depression severity is the normalized growth (decline) in retail sales per capita from 1929 to 1933. New Deal spending is drawn from Fishback, Kantor, and Wallis (2003). All spending measures are in per capita terms and normalized. Total New Deal spending includes spending on public works, relief, and loans, as well as other smaller programs.

Source: BLS Cost of Living Survey 1918-1918; IPUMS 1920 1% Census Sample; 1940 Complete Count Census; Census of Retail Sales

Figure A.4: Occupation titles in the BLS survey among fathers

(a) All BLS Fathers

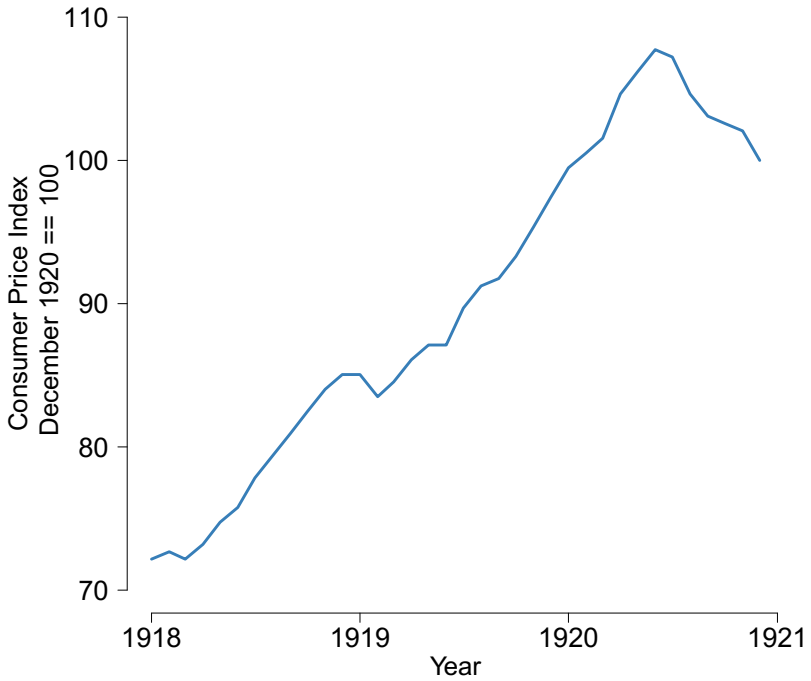
(b) Bottom Earnings Quartile



(c) Top Earnings Quartile

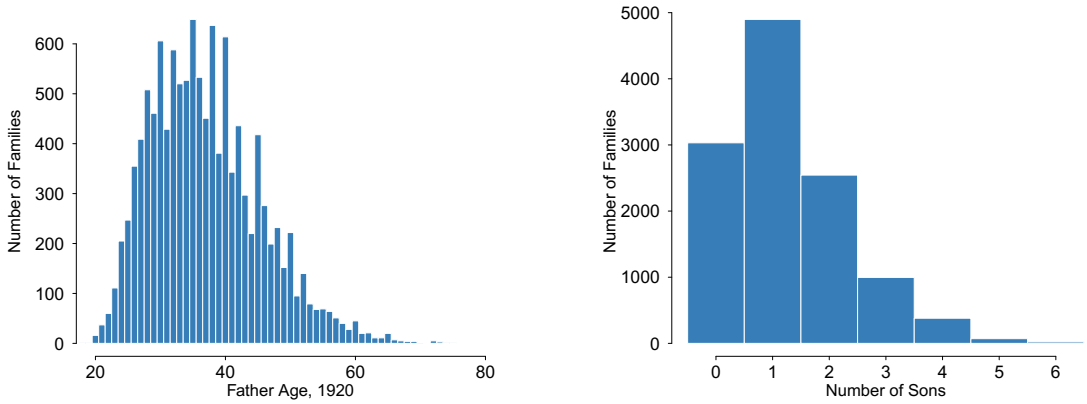
The word clouds represent the most frequent single word strings in occupation titles, scaled and colored by frequency. Among the fathers with the least earnings in 1918 or 1919, the most common occupation descriptions included the words laborer and factory. Among the fathers with the most earnings, however, descriptions like machinist, foreman, railroad, carpenter, and engineer were much more common.

Figure A.5: Consumer Price Index, 1918 to 1920



The final year of the First World War and the two following years of peace were a period of rapid price changes. The families surveyed by the BLS were asked about earnings in the previous 12 months, relative to survey dates from July 1918 to February 1919. I use the monthly CPI to standardize and normalize the earnings responses. CPI data for all urban consumers are drawn from a BLS series, stored at the Federal Reserve Bank of St. Louis archive, FRED.

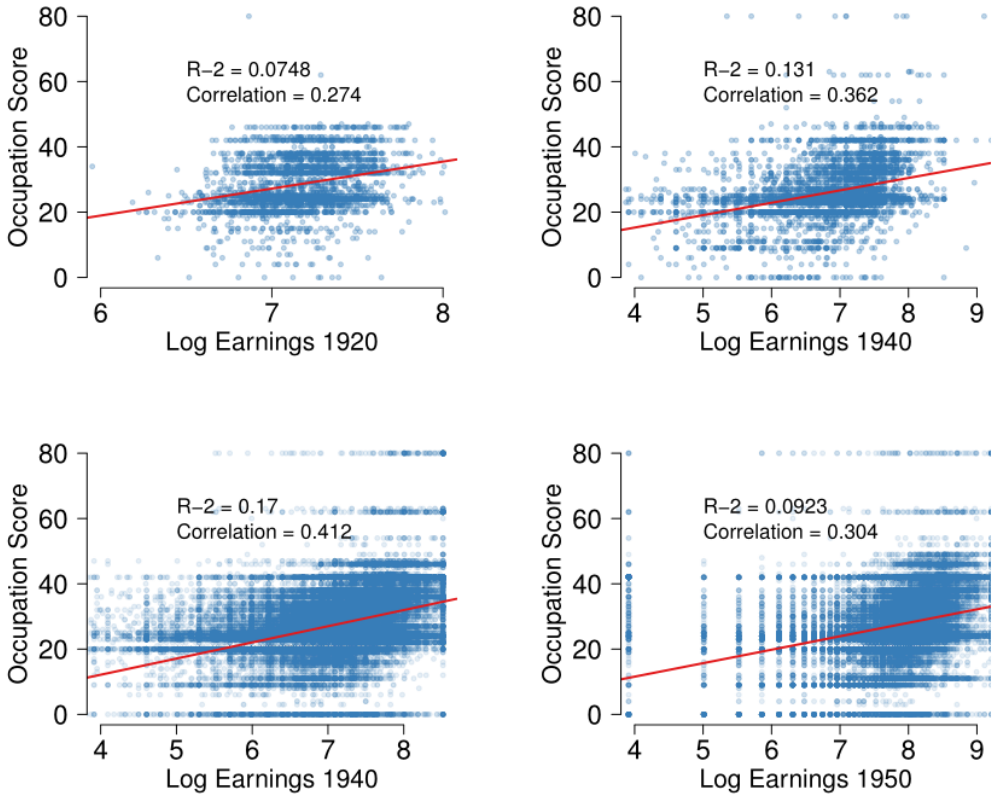
Figure A.6: Demographics of the BLS survey



(a) Father's Age

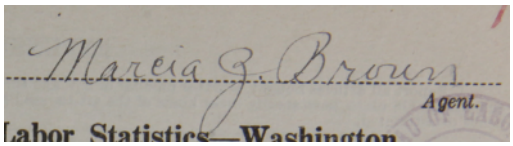
(b) Number of Sons

Figure A.7: The correlation of occupation score and log earnings

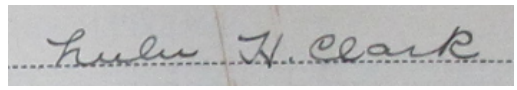


Occupation scores are calculated by IPUMS based on median earnings by occupation in 1950. Linking occupations across time is imprecise: some occupations may rise in earnings and prestige, while others fall, and other occupations may not still exist in the same form in 1950 as in 1920. However, generally, occupation score is a reasonable measure of earnings throughout my sample period. In the first plot, I graph log earnings in 1920 in my sample of fathers from the BLS survey against their occupation score. In the second plot, I graph log earnings in 1940 in my sample of sons from the BLS survey, matched into 1940, against their occupation score. In the bottom row, I show the same correlations for the 1940 IPUMS 1% and the 1950 IPUMS 1% sample.

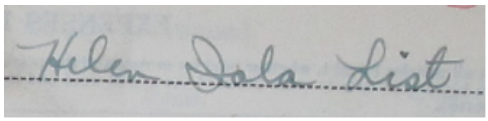
Figure A.8: Examples of Enumerator Signatures from the Original BLS Surveys



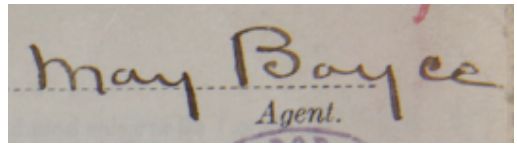
(a) Marcia G. Brown



(b) Lulu H. Clark



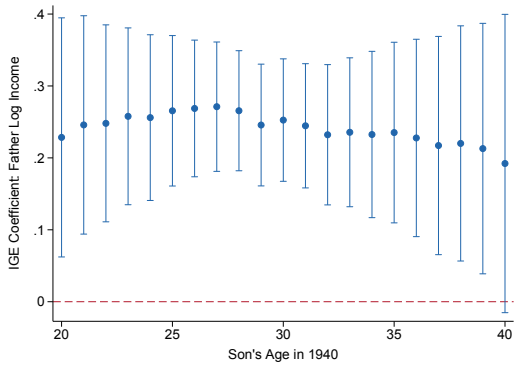
(c) Helen Iola List



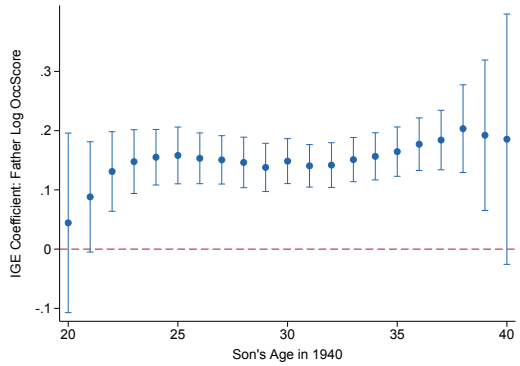
(d) May Boyce

Records written up by the enumerators in the top row have very low match rates from the BLS sample to the 1940 census, while entries recorded by the enumerators in the bottom row have very high match rates. The quality and clarity of the enumeration of the original records is a strong determinant of the accuracy with which names and other biographical information can be transcribed.

Figure A.9: Estimates of intergenerational mobility are relatively stable across sons by age



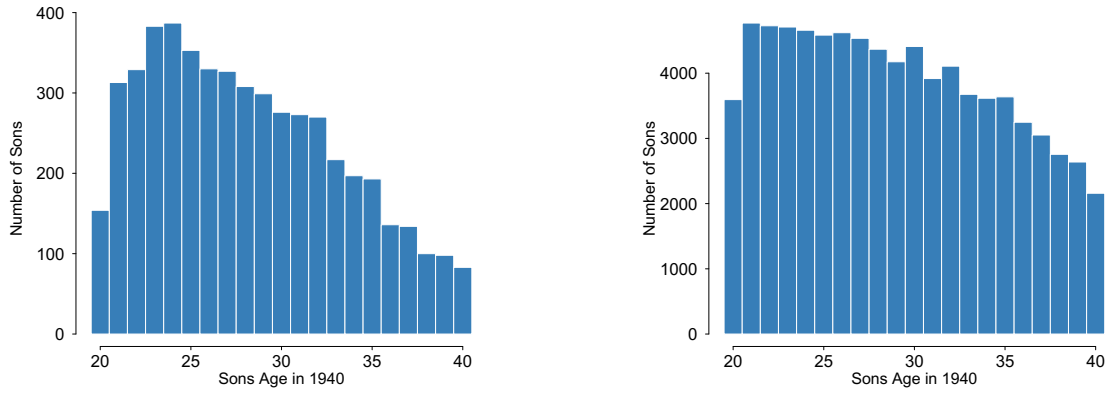
(a) BLS Sample



(b) IPUMS 1920-1940 Sample

Higher IGE coefficients implies a stronger link between fathers and sons and less mobility. The father and son ages at which income or occupations are observed often complicate the intergenerational mobility estimates. For example, many early IGE studies have been criticized for observing sons who were too young and fathers who were too old, which biases parameter estimates down and implies spuriously large levels of mobility (Corak, 2006; Grawe, 2004; Mazumder, 2015). In this figure, I explore this potential issue in my sample. In the BLS sample, I regress son's log earnings in 1940 on father's log earnings in 1920 interacted with a set of son's age dummy variables, controlling for a quartic in father's age and city fixed effects. In the IPUMS sample, I calculate mobility rates using occupation scores, running a similar regression of son's log occupation score in 1940 on father's log occupation score in 1920 interacted with a set of son's age dummy variables, controlling for a quartic in father's age and city fixed effects. I find similarly stable estimates of the occupation score mobility after age 23. The IPUMS-based estimates are more precise than those based on my BLS sample, as the sample is nearly four times larger. In Figure A.10, I plot the sample sizes underlying these figures. Using contemporary administrative data, Chetty, Hendren, Kline, and Saez (2014) do not find stabilization mobility estimates until sons in their late 20s. However, given the differences in mean education between 1940 and recent decades, parameter stabilization does occur at similar points of labor market experience.

Figure A.10: Distribution of son's adult ages in 1940.



(a) BLS Sample

(b) IPUMS 1920-1940 Sample

The primary explanation of the Depression's regional variation is pre-1929 industry mix: cities and states specializing in durable manufacturing or extraction of raw materials used in durable manufacturing saw the largest downturns.

The consensus in the macroeconomic literature is that monetary factors caused the Great Depression (Friedman and Schwartz, 1963; Bernanke, 2000), particularly reliance on the gold standard (Bernanke, 1995; Eichengreen, 1992; Temin, 1989). But while such factors may explain international variation, monetary policy did not vary within the US.[1]

There was large regional variation in the severity of the Depression. As John Wallis writes, "neither the Great Crash nor the economic recovery that followed was evenly felt across the different regions of the country" (Wallis, 1989). The downturn was worst in the mountain states but more mild in the upper south (Rosenbloom and Sundstrom, 1999). I explore three possible explanations for this variation: industry mix, 1920 growth and subsequent bank failures, and state capital status.

First, American cities all have different mixes of industries. At the onset of the Depression, for example, Detroit specialized in auto manufacturing, Pittsburgh in steel, and Lawrence, MA in textiles. Industries like mining, metal work, and timber declined much more during the Depres-

[1] Financial regulatory policy did vary across Federal Reserve districts. As Richardson (2009) show, rates of bank failure did change dramatically at the border between activist and conservative Fed districts. Ziebarth (2013) extends these results to the real economy using a sample of plants from the Census of Manufactures. However, variation across Fed districts cannot explain why the downturn varied within districts: Pittsburgh endured a huge decline while Cleveland had a relatively mild recession.

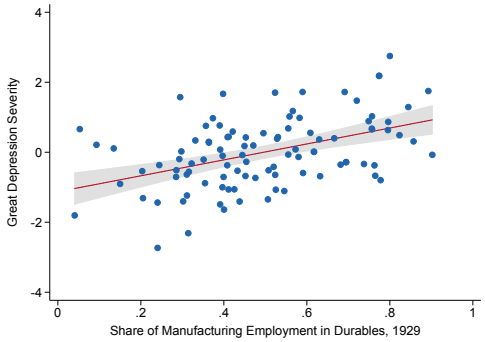
sion than others, so the cities with a higher share of employment or income from these industries would also decline more. At the state level, this also explains why the heavy-industry-focused midwest (particularly census region East North Central) and the extractive-natural-resources-focused mountain states (specializing in lumber and mining) suffered the worst Depression downturns (Wallis, 1989). Inelastic demand for cigarettes and other tobacco products may explain the milder declines in the upper south (Heim, 1998). I use two metrics to test these industry effects. First, drawing on Romer (1990), I measure the share of manufacturing employment in durables: durable consumption suffered a more immediate decline than nondurables in the first few years of the Depression. Following Boone and Wilse-Samson (2014), I calculate the share of manufacturing employment in durables in the IPUMS 1930 1% sample using the standardized IPUMS industry codes. Second, I construct a Bartik shock, compiling industry mix before the Depression and assuming that city-level employment in a given industry would grow (or shrink) at the same rates as the industry grew nationally.[2] In the top panel of Figure A.11, I show that the severity of the Depression is strongly correlated with both the share of manufacturing employment in durables (a) as well as the predicted decline based on industry mix (b).

Second, the booms of the 1910s and 1920s were also regionally varied. I measure the 1920s booms in each city with the growth rate of manufacturing employment from 1920 to 1929, as measured in the Census of Manufactures. I normalize these growth rates to be mean zero and standard deviation of one. In addition, I draw on bank failure data at the state level from FDIC Data Bank, ICPSR dataset 7. The bubble of the preceding decades made banks more susceptible to failure during the Depression, as they had invested in inflated assets like farm real estate (Calomiris and Mason, 2003). Because interstate banking laws remained restrictive during this time period, the banks in a given city would be especially sensitive to local or state-wide bubbles. I count the number of banks failing between 1929 and 1933 and divide by the number of banks in each state in 1929.[3] In the bottom panel of Figure A.11, I show that growth rates during the 1920s as well as state-level bank failure rates are positively correlated with Depression severity, though the boom-bust relationship is weaker.

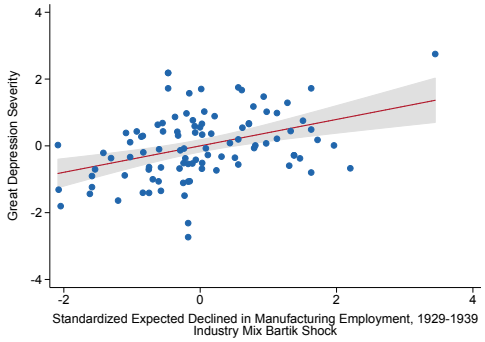
[2] I draw on the Census of Manufactures in 1929 to calculate industry mix before the Depression. For a handful of cities, industry-level data is not available, and I use state industry mix to proxy.

[3] Failure rates cannot be weighted by bank size, as assets data is only available in aggregate.

Figure A.11: Local correlates of Great Depression severity.

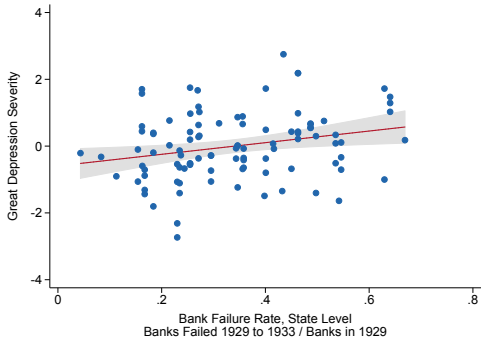


(a) Durable Share



(b) Industry Mix Bartik Shock

(c) 1920s Growth in Manufacturing Employment



(d) Bank Failures

Third, state capitals across the country appear to have suffered less severe downturns, perhaps because the response of government employment in the capital to the downturn was less elastic than private sector employment. Across all counties in the United States, I find that the growth rate of retail sales per capita from 1929 to 1933 in counties including a capital city was about 8 percentage points larger than the growth rate in all other counties: capitals declined by 0.39 log points, while non-capitals declined by 0.47 log points. 15 of the 99 cities in my sample are state capitals. When I regress Depression severity on capital status in my sample of BLS cities, I find that capitals declined 0.06 fewer log points than non-capitals.

While the bivariate correlations presented in Figure A.11 are suggestive, I turn to regressions to gauge the relative importance of local factors in determining city-level Depression severity. To facilitate the comparisons in Table A.6, I standardize every variable (except for the capital status indicator) to have mean zero and standard deviation one. I find that the industry mix variables, either the share of manufacturing workers in durables or a standardized Bartik shock,

Table A.6: Correlates of Local Great Depression Severity

	Normalized Great Depression Severity			
	(1)	(2)	(3)	(4)
Share of Manufacturing Employment in Durables	0.268**	0.377***		0.283***
	(0.103)	(0.095)		(0.098)
Standardized Bartik Shock	0.227**		0.347***	0.133
	(0.107)		(0.096)	(0.100)
Employment Growth in 1920s	0.197*	0.186*	0.232**	0.165
	(0.103)	(0.104)	(0.103)	(0.108)
Bank Failures 1929 to 1933 Share of Banks in 1929	0.116	0.135	0.156*	0.115
	(0.082)	(0.082)	(0.089)	(0.086)
Capital City of State	-0.257	-0.255	-0.351	-0.265
	(0.225)	(0.220)	(0.248)	(0.221)
Latitude				-0.025
				(0.022)
Longitude				-0.014**
				(0.006)
Log Population				-0.003
				(0.075)
Observations	99	99	99	99
Adjusted R^2	0.266	0.233	0.221	0.293

Source: Census of Retail Sales; Census of Manufactures; IPUMS 1930 1% sample.

have the strongest effects on severity. The two measures are highly collinear, but, when included separately, a standard deviation increase in either increases Depression severity by more than a third of a standard deviation. The effects of a standard deviation increase in either bank failures or growth during the 1920s, meanwhile, increases Depression severity by between 0.15 and 0.25 of a standard deviation. Finally, the capital effect is the expected sign, as capital cities had less severe Depressions; however, this effect is statistically insignificant.[4]

Garrett and Wheelock (2006) conduct a similar study at the state level, exploring the decline in per capita state income from 1929 to 1933, and find that only industry structure and per capita income before the Depression correlate significantly with Depression severity: states with a higher concentration of industries declining during the Depression declined more, as did states with higher incomes in 1929. However, the authors conclude that spatial spillovers and industrial

[4] Recall that 15 of the 99 cities in my sample are capitals, meaning the indicator variable has a standard deviation of 0.36. Thus, multiplying the coefficient on capital by 0.36 allows me to interpret it in the same way as the other coefficients and yields a standardized effect of about 0.13.

composition fully explain the effect of initial income in 1929 on variation in Depression downturn.

A.3 Ranking Income in 1920

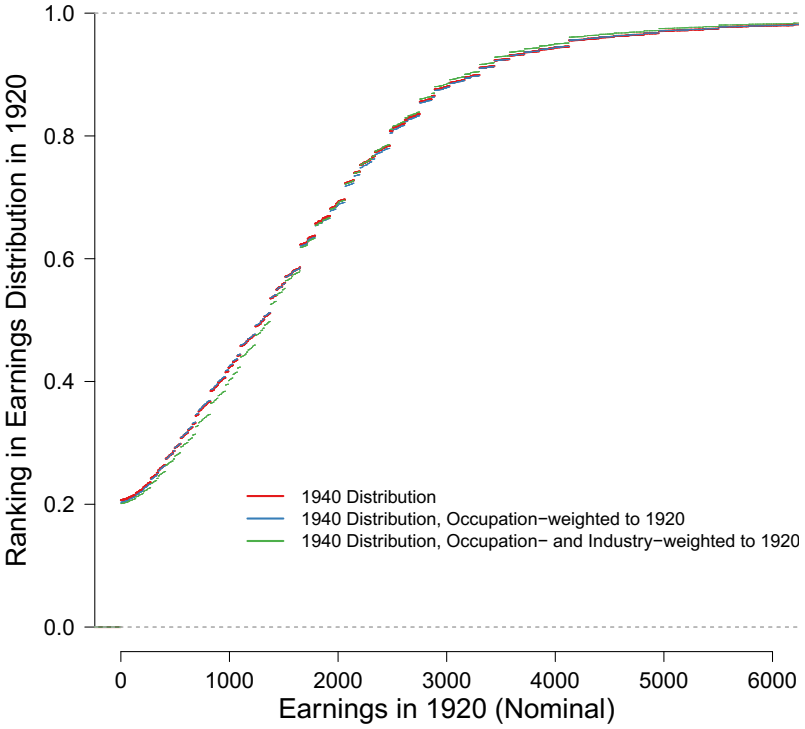
One measure of mobility I use in this paper is the rank-rank parameter. In this section, I describe the construction of the earnings ranking for my sample of fathers from 1918 to 1919, where the full earnings distribution is unknown. I construct three different ranking distributions, and my results are robust to choosing any of the three.

The rank-rank parameter is the coefficient from a regression of the son's ranking in the earnings distribution on his father's earnings distribution rank, as well as controls for the son's and father's age. Four measures are required to calculate the rank-rank parameter: (1) earnings of the father, (2) earnings of the son, (3) the relevant earnings distributions for fathers, and (4) the relevant earnings distribution for sons. I have the complete 1940 census, which includes labor earnings information for every person in the country. This data enables me to observe a son's earnings in 1940 as well as his ranking in the distribution. For the fathers in 1918-1919, I observe earnings from the BLS sample. However, this sample is one of the only sources for earnings information for this period. While it would be simple to calculate each father's position within the BLS sample, this position is a very misleading representation of the father's position overall. The BLS targeted middle income respondents, as the survey takers were hoping to study the costs of living faced by urban consumers who were wage workers or had salaries under $2000 per year. Instead, I construct an earnings distribution in 1920, leveraging the complete and known earnings distribution for 1940.

I reconstruct the earnings distribution in 1920 in three different ways in order to rank fathers in the earnings distribution. I base each ranking on the known 1940 earnings distribution. In the first, I use the 1940 distribution and simply convert earnings from 1940 dollars to 1920 nominal dollars using the CPI. In the second, I use the 1940 distribution, but reweight observations by the relative prevalence of occupations in the 1920 census and convert 1940 dollars to 1920 dollars. In the third, I use the 1940 distribution, but reweight observations by the relative prevalence of occupations and industries in the 1920 census and convert 1940 dollars to 1920 dollars.

In all three ranking constructions, I use the CPI to deflate the distribution from 1940 dollars

Figure A.12: Estimated 1920 Earnings Cumulative Distribution



I reconstruct the earnings distribution in 1920 in three different ways in order to rank fathers in the earnings distribution. In the first, I use the 1940 income distribution, only converting earnings from 1940 to 1920 nominal levels (plotted in red). In the second, I use the 1940 income distribution, but reweight observations by the relative prevalence of occupations in the 1920 census and convert 1940 dollars to 1920 dollars (plotted in blue). In the third, I use the 1940 income distribution, but reweight observations by the relative prevalence of occupations and industries in the 1920 census and convert 1940 dollars to 1920 dollars (plotted in green).

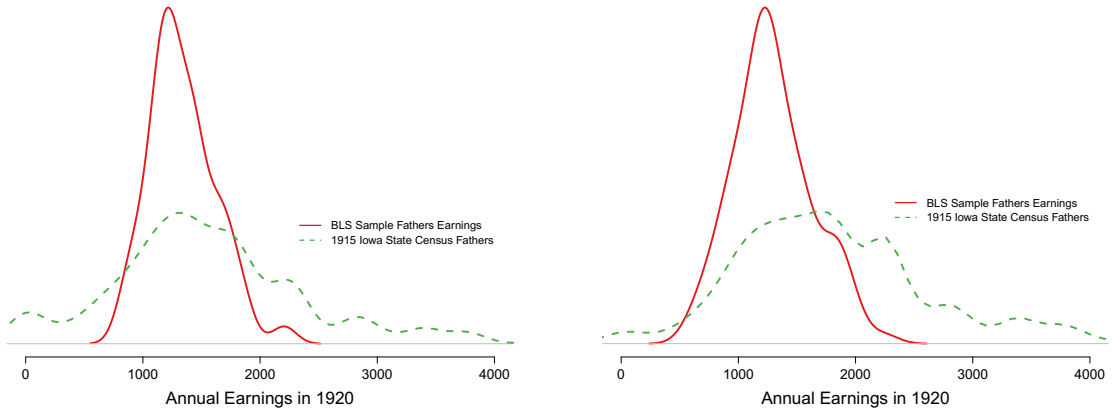
to 1920 dollars. I choose the CPI rather than a measure of mean nominal earnings because I need a monthly price deflator to convert the BLS earnings themselves. The BLS sample was collected over the course of eight months, from July 1918 to February 1919, during a period of significant price changes, as I document in Figure A.5. Earnings were reported relative to the 12 months prior to the survey data, and thus they vary across respondents.

To reweight by occupation and occupation-by-industry, I collect the counts of men in each occupation or occupation-by-industry cell in the 1920 IPUMS 1% sample. Occupations and industry are collected by census enumerators as free-form strings but standardized by IPUMS. Specifically, I rely on the occ1950 and ind1950 variables, which are occupation and industry codes that have been standardized both within and across census waves to 1950 definitions. There are 258 unique occupation codes in the 1920 IPUMS sample and 3662 unique occupation-by-industry codes. For the occupation reweighting, the weight for each observation in the 1940 earnings distribution is simply the share of men with that occupation in 1920 divided by the share of men with that occupation in 1940. For example, there were 3.5 times as many blacksmiths in 1920 as there were in 1940, yielding a higher weight on blacksmiths in the estimated 1920 distribution than in the actual 1940 distribution. Conversely, there were 30 times as many "Airplane pilots and navigators" in 1940 than in 1920, leading to a very small weight. The reweighting process is parallel for the occupation and industry version, but with smaller cells defined by both occupations and industries. I plot the CDFs of these three ranking functions in Figure A.12.

The correlation in earnings rankings of the fathers in my sample across the three possible methods is extremely high. The unweighted method correlates with the occupation weighted ranking at 0.99996 and with the occupation and industry weighted ranking at 0.99974. The occupation weighted and occupation and industry weighted rankings correlate at 0.99960. Given these high correlations, it is not surprising that my results are robust to using any of the rankings. Throughout the paper, I rely on the first, unweighted earnings distribution.

The 1940 census is not the only possible source with which to build an earnings distribution for 1920. The 1915 Iowa State Census also collected earnings data from respondents. Two Iowa cities, Davenport and Des Moines, are among the 99 cities surveyed by the BLS sample. I compare the distribution of earnings in those two Iowa cities from the BLS sample to the distribution of earnings in 1915, drawn from a sample of the state census, digitized originally by Goldin and

Figure A.13: Distribution of Earnings in the BLS Sample in Iowa cities compared to the 1915 Iowa Stata Census



(a) Davenport, IA **(b)** Des Moines, IA

The differences in the distributions of earnings in the BLS Sample in Davenport and Des Moines and the distributions of earnings in those two cities based on the 1915 Iowa State Census are comparable to the distributional differences shown in Figure 1.2, which compared the full BLS sample to the national 1920 earnings distribution. All earnings are measured in 1920 dollars. The complete earnings distribution for 1915 for the two Iowa cities is calculated using the 1915 Iowa state census.

Katz (2000), in Figure A.13.

A.4 Imputed Business and Self-Employment Earnings in 1940

I use earnings data from the 1940 Federal Census to measure the adult outcomes of the sons in my intergenerational sample. However, the 1940 census includes detailed records only for earnings from wages and salaries; the exact amounts of other forms of income, including self-employment earnings, were not recorded—only an indicator for more than $50 in such earnings. In this section, I show that this data challenge cannot explain my results by imputing the income of sons using the 1950 Federal Census. My estimates of the effect of variation in Great Depression severity on intergenerational mobility are robust to supplementing my original data with imputed income for sons with non-wage or salary income.

Unlike the 1940 census, the 1950 census recorded earnings data for respondents from both wages and salary, as well as from business income. I draw on the IPUMS 1% sample of the 1950 census, which I limit to white men, aged 20 to 40, replicating my sample. Among the men in the 1950 sample with business earnings, I predict log earnings from business on age, years of

Table A.7: Great Depression Severity Decreases Intergenerational Mobility, Including Imputed Business Earnings

	IGE				Rank-Rank			
	(1) Main	(2) Imputed	(3) Main	(4) Imputed	(5) Main	(6) Imputed	(7) Main	(8) Imputed
Log Father Earnings 1920	0.280***	0.252***	0.181***	0.162***				
	(0.043)	(0.042)	(0.056)	(0.053)				
Log Father Earnings 1920 X GD Normalized Severity	0.108**	0.098**						
	(0.042)	(0.039)						
Log Father Earnings 1920 X GD Above Median Severity			0.197**	0.175**				
			(0.082)	(0.082)				
Father Earnings Rank 1920					0.213***	0.225***	0.124**	0.168***
					(0.033)	(0.028)	(0.051)	(0.038)
Father Earnings Rank 1920 X GD Normalized Severity					0.100***	0.070**		
					(0.036)	(0.030)		
Father Earnings Rank 1920 X GD Above Median Severity							0.175***	0.112**
							(0.066)	(0.056)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4730	5468	4730	5468	4952	5690	4952	5690
Clusters	99	99	99	99	99	99	99	99
Adjusted R^2	0.188	0.196	0.188	0.196	0.153	0.244	0.153	0.244

Estimates of intergenerational mobility based on a linked sample from the BLS survey of urban families from 1918-1919 to the 1940 Federal census. Each column is a regression of the son's outcome in 1940 on the father's corresponding outcome in 1918-1919, a measure of Great Depression severity in the city of residence in 1918-1919, and an interaction of severity and the father's outcome. Controls include quartics in the son's and father's ages. All fixed effects are based on the city of residence in 1918-1919. Great Depression Severity is measured using the decline in per capita retail sales at the county level from 1929 to 1933. The odd columns, titled "Main," are the standard results from Table 1.4 with city fixed effects. In the even columns, titled "Imputed," I impute capital earnings for the 12% of the sample reporting more than $50 in non-wage or salary income in 1940. The imputations use the IPUMS 1% sample of the 1950 census, which includes data on business income. Log business income is predicted using a multi-level model including age, years of education, state of residence, occupation code, and industry code. IGE is the intergenerational elasticity of income. Rank-rank mobility compares the son's position in the earnings distribution in 1940 to the father's position in 1918-1919.

Source: BLS Cost of Living Survey 1918-1918; 1940 Complete Count Census; Census of Retail Sales

education, state of residence, occupation code, and industry code.[5] Returning to my sample of sons in 1940, I calculate total earnings for each son, either by adding predicted business earnings to labor earnings for sons earning both labor and business income in 1940, or by using predicted business earnings directly for sons without any labor earnings.

My main results do not change significantly when supplementing my sample with imputed capital earnings for the sons reporting such earnings in Table A.7. Only 12% of the sons in my sample of 1940 report earning more than $50 in non-wage or salary income, and less than half of that group earned business income exclusively.

[5] I convert the earnings in 1950 to 1940 dollars using the CPI. To improve the predictive fit, I use a multi-level model that shrinks estimated fixed effects towards the mean when the cell sample sizes are smaller (Gelman and Hill, 2007). Both age and years of education enter the prediction non-parametrically. Cross-validation rejected any gains from including interactions of the predictors in the model.

A.5 Intergenerational Mobility Model and Measurement Error

I only observe father's income in one year—1918 or 1919—which may be a noisy proxy for permanent income. For the fathers living in cities with larger Great Depression downturns, this one year of income, before the Depression, may be a particularly noisy measure. In this section, I show that noise driven by the Depression cannot explain my results; in fact, if the Great Depression increased measurement error, then it would tend to lower the estimated persistence parameter and thus increase the estimated mobility.

The true model is

$$\log(Y_{it}) = \mu^* + (p\theta + \omega)\log(Y_{i,t-1}) + pe_{it}$$

But suppose only noisy measures of log income are observed: $\log(\tilde{Y_{i,t-1}}) = \log(Y_{i,t-1}) + u_{i,t-1}$ and $\log(\tilde{Y_{it}}) = \log(Y_{it}) + u_{it}$. The probability limit of the estimated persistence parameter, β, will be

$$\begin{aligned} p\lim\beta =& \frac{Cov\left(\log(\tilde{Y_{it}}), \log(\tilde{Y_{i,t-1}})\right)}{Var\left(\log(\tilde{Y_{i,t-1}})\right)} \\ & \frac{Cov\left(\mu^* + (p\theta+\omega)\log(Y_{i,t-1}) + pe_{it} + u_{it}, \log(Y_{i,t-1}) + u_{i,t-1}\right)}{Var\left(\log(Y_{i,t-1}) + u_{i,t-1}\right)} \end{aligned}$$

But μ^* is a constant, and $u_{i,t-1}$ and u_{it} are random noise and do not covary with the other terms.

$$p\lim\beta = \frac{Cov\left((p\theta+\omega)\log(Y_{i,t-1}) + pe_{it}, \log(Y_{i,t-1})\right)}{Var\left(\log(Y_{i,t-1}) + u_{i,t-1}\right)}$$

Inspecting the denominator, I have $Var(\log(Y_{i,t-1}) + u_{i,t-1}) = Var(\log(Y_{i,t-1})) + Var(u_{i,t-1}) + 2Cov(\log(Y_{i,t-1}, u_{i,t-1})) = Var(\log(Y_{i,t-1})) + Var(u_{i,t-1})$.

$$p\lim\beta = \frac{(p\theta+\omega)Var(\log(Y_{i,t-1})) + pCov(e_{it}, \log(Y_{i,t-1}))}{Var(\log(Y_{i,t-1})) + Var(u_{i,t-1})}$$

Let $Var(u_{i,t-1}) = Var(u_{it}) = \sigma_u^2$. $Var(e_{i,t-1}) = Var(e_{it}) = Var(\delta + \lambda e_{i,t-1} + v) = \lambda^2 Var(e_{i,t-1}) + \sigma_v^2$. That implies that $Var(e_{i,t-1}) = \frac{\sigma_v^2}{1-\lambda^2}$.

$$\begin{aligned} Var(\log(Y_{i,t-1})) &= Var(\log(Y_{it})) \\ &= Var(\mu^* + (p\theta+\omega)\log(Y_{i,t-1}) + pe_{it}) \\ &= (p\theta+\omega)^2 Var(\log(Y_{i,t-1})) + p^2 Var(e_{it}) + 2(p\theta+\omega)pCov(\log(Y_{i,t-1}), e_{it}). \end{aligned}$$

That implies that $Var(\log(Y_{i,t-1})) = \frac{p^2 Var(e_{it}) + 2(p\theta+\omega)pCov(\log(Y_{i,t-1}),e_{it})}{1-(p\theta+\omega)^2}$.

$$\begin{aligned} Cov(\log(Y_{i,t-1}), e_{it}) &= Cov(\log(Y_{i,t-1}), \delta + \lambda e_{i,t-1} + v) \\ &= \lambda(p\theta+\omega)Cov(\log(Y_{i,t-1}), e_{it}) + \lambda p Var(e_{it}) \end{aligned}$$

That implies that $Cov(\log(Y_{i,t-1}), e_{it}) = \frac{\lambda p Var(e_{it})}{(1-\lambda(p\theta+\omega))}$. Putting this together we have

$$\begin{aligned} p\lim\beta = &\frac{(p\theta+\omega)Var(\log(Y_{i,t-1})) + pCov(e_{it}, \log(Y_{i,t-1}))}{Var(\log(Y_{i,t-1})) + Var(u_{i,t-1})} \\ &\frac{p^2\sigma_v^2(p\theta+\omega+\lambda)}{\sigma_v^2 p^2((p\theta+\omega)\lambda+1) + \sigma_u^2(1-\lambda^2)(1-(p\theta+\omega)\lambda)(1-(p\theta+\omega)^2)} \end{aligned}$$

As σ_u^2 increases, the denominator increases, so the fraction decreases. Thus, the estimated persistence parameter is decreasing in measurement error, and mobility $(1-\beta)$ is increasing. If the Great Depression caused an increase in measurement error, that would bias the results towards finding more mobility in more severely shocked cities, not less.

A.6 The Great Depression Does Not Predict Current Mobility

If local fixed factors influenced both Depression severity and economic intergenerational mobility, a relationship between the Depression and mobility might be evident in the contemporary mobility data. I find that this is not the case: mobility today is no different in the cities in my sample with the worst Depression downturns and in the cities with the mildest downturns.

Chetty, Hendren, Kline, Saez, and Turner (2014); Chetty, Hendren, Kline, and Saez (2014) measure local economic mobility for the recent period, using administrative tax records to generate economic mobility parameters. Their primary estimates of mobility are derived from rank-rank regressions, where the children's rank in the income distribution as an adult is regressed on the parent's rank in the income distribution. They focus on children born in the 1980 to 1982 cohorts and report results at the county level, assigning linked pairs of parents and children to the county of residence during childhood. Arguing that the rank-rank relationship is approximately linear throughout all samples, Chetty, Hendren, Kline, and Saez (2014) note that mobility can be fully described by two parameters: the slope of the rank-rank coefficient and the expected rank for children born at the 25th percentile.[6] Within the sample of cities in the BLS survey, Morris

[6] Given the linearity of the results, the slope and the intercept would also fully characterize mobility, but the 25th percentile measure has more rhetorical interest.

County, New Jersey (Newark) has the highest expected income rank for children born at the 25th percentile (49.5), and St. Louis has the lowest (32.2). Meanwhile, San Francisco has the weakest relationship between parents' and children's ranks (slope of 0.18), and Richmond, VA has the strongest (0.50).

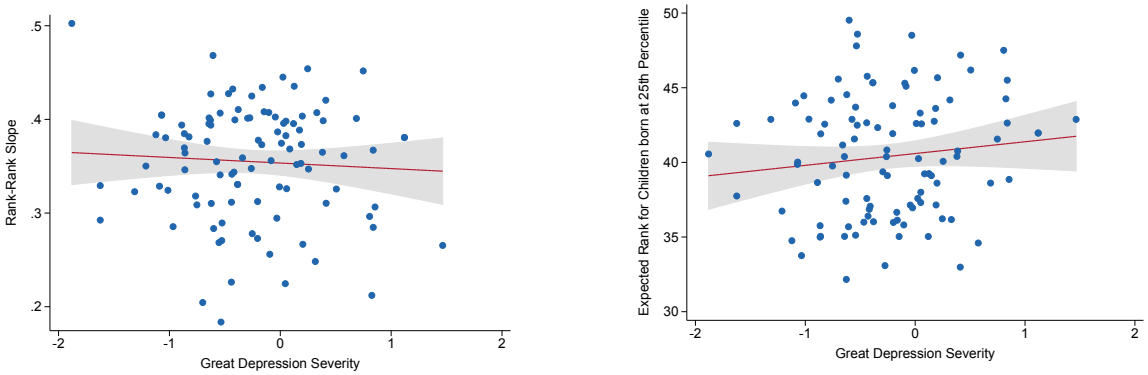
Neither of these measures of mobility today correlates with historical mobility or the Great Depression. If this were the case, again, that might prompt concern that the observed severity effect on mobility in the previous section was driven by other fixed local factors that determined mobility and severity. As I documented in Table 1.7 in the main text and present as well in Figure A.14 here, I show that there is no relationship between Great Depression severity and rates of economic mobility today, measured either by the slope of the rank-rank regressions or the 25th percentile expectation. I merge my county level measures of Great Depression severity to the county level measures of recent mobility.[7] These results further strengthen my claim that the cities shocked by the Great Depression do not have fundamentally different rates of mobility from cities with milder downturns.[8]

When the outcome is the rank-rank slope, the sign changes depending on the specification, and the effects are generally quite small. A positive relationship between Depression severity and the slope would imply that regions with more severe downturns have less mobility today, echoing my result that cities with more severe downturns had less mobility between 1920 and 1940. When the outcome is the expected adult rank of a child born into the 25th percentile, a positive coefficient implies a more severe Great Depression increases mobility as measured by the expected outcomes for poor children.

[7] Cities comprised of multiple counties like New York and St. Louis, enter the sample multiple times.

[8] The results also rule out any very long run persistence of the effect of the Great Depression on mobility many generations later. While economic outcomes are likely the result of many generations worth of family inputs (Long and Ferrie, 2015), given the high levels of geographic mobility in the US during the last 80 years, these null results are not surprising and are not the ideal test of long run economic shocks on multi-generation mobility. A better test would link people today to their grandfathers or great-grandfathers during the Depression and use the location of the grandfather rather than the location of the child to measure Depression severity. However, due to privacy restrictions on census data, such multigenerational matches ending with final generations in the recent period are not yet possible.

Figure A.14: Great Depression Severity Does Not Correlate with Contemporary Measures of Mobility



(a) Rank-Rank Coefficient

(b) Expected Rank for Children born at 25th Percentile

Drawing on data from Chetty, Hendren, Kline, and Saez (2014), I show that there are no lasting effects today of the Great Depression on local intergenerational mobility. Chetty, Hendren, Kline, and Saez (2014) use administrative tax data to estimate both relative and absolute mobility between the generation born between 1980 and 1982 and their parents by county. I plot these estimates of mobility for each city included in my BLS sample against Great Depression severity, measured by the decline in per capita retail sales between 1929 and 1933. There is no clear relationship between the two measures, suggesting that places with high or low mobility today are no more or less likely to have suffered large downturns during the Great Depression.

B. APPENDIX TO CHAPTER 2

B.1 Additional Figures and Tables

Table B.1: Geographic Mobility and the Dust Bowl, 1915 to 1940

	Log Miles Moved		In Iowa in 1940		In Same County in 1940	
	(1)	(2)	(3)	(4)	(5)	(6)
Log Father Earnings	-0.194		-0.00134		0.0586	
	(0.147)		(0.0355)		(0.0366)	
Father Education		-0.00533		0.00340		0.00677
		(0.0298)		(0.00753)		(0.00799)
Dust Bowl Severity	-1.720	-0.0596	0.119	-0.00497	0.502	0.138
	(1.222)	(0.356)	(0.299)	(0.0879)	(0.318)	(0.0933)
Dust Bowl Severity × Earnings	0.248		-0.0171		-0.0675	
	(0.178)		(0.0433)		(0.0462)	
Dust Bowl Severity × Education		0.000415		0.00188		-0.0117
		(0.0415)		(0.0103)		(0.0110)
Son Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Father Age Quartic	Yes	Yes	Yes	Yes	Yes	Yes
Name String Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	1421	1513	1421	1513	1421	1513
Number of Clusters	930	987	930	987	930	987
R-squared	0.00805	0.00312	0.00412	0.000813	0.00107	-0.0000557

In the first two columns, the dependent variable is the log of 1 + the number of miles between the son's county in 1940 and in 1915. In the second two columns, the dependent variable is an indicator variable for whether or not the son lives in Iowa in 1940. In the final two columns, the dependent variable is an indicator variable for whether or not the son lives in the same county in 1940 as he lived in 1915. Standard errors clustered by family. Dust Bowl severity is coded as 0 (low erosion), 1 (medium erosion), or 2 (high erosion), following Hornbeck (2012), measured at the county level. Name string controls: first and last name commonness, length, letter similarity, and Scrabble scores. Son's ages are normalised relative to age 40 in 1940.

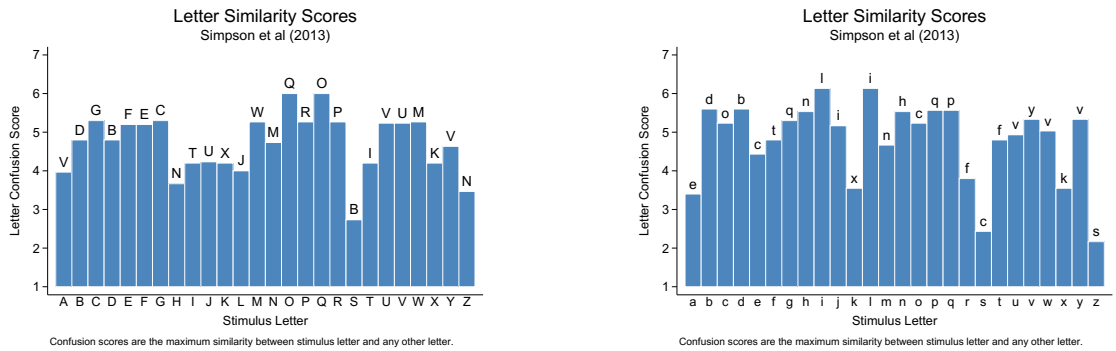
Sources: 1915 Iowa State Census Sample; 1940 Federal Census

Table B.2: Census Matching Variables

	Probit	Logit
First and Last name match	0.632^{***}	1.129^{***}
	(0.086)	(0.168)
First name distance, Jaro-Winkler	-6.071^{***}	-11.543^{***}
	(0.525)	(0.994)
Last name distance, Jaro-Winkler	-10.285^{***}	-19.145^{***}
	(0.487)	(0.954)
Absolute Value Difference in Year of Birth is 1	-0.708^{***}	-1.308^{***}
	(0.044)	(0.083)
Absolute Value Difference in Year of Birth is 2	-1.562^{***}	-2.893^{***}
	(0.065)	(0.126)
Absolute Value Difference in Year of Birth is 3	-2.316^{***}	-4.370^{***}
	(0.102)	(0.208)
First name Soundex match	0.153^{***}	0.294^{***}
	(0.054)	(0.100)
Last name Soundex match	0.698^{***}	1.341^{***}
	(0.069)	(0.135)
Hits	-0.064^{***}	-0.123^{***}
	(0.002)	(0.005)
Hits-squared	0.0003^{***}	0.001^{***}
	(0.00002)	(0.00004)
More than one match for first and last name	-1.690^{***}	-3.217^{***}
	(0.093)	(0.183)
First letter of first name matches	0.871^{***}	1.593^{***}
	(0.130)	(0.245)
First letter of last name matches	0.886^{***}	2.003^{***}
	(0.148)	(0.356)
Last letter of first name matches	0.147^{***}	0.312^{***}
	(0.053)	(0.101)
Last letter of last name matches	0.649^{***}	1.239^{***}
	(0.070)	(0.139)
Middle Initial matches, if there is a middle initial	0.537^{***}	0.908^{***}
	(0.097)	(0.186)
Constant	-1.479^{***}	-3.087^{***}
	(0.225)	(0.480)
Observations	38,091	38,091
Log Likelihood	−2,440.877	−2,444.649
Akaike Inf. Crit.	4,915.753	4,923.298

The results of two different matching alogrithms—a probit model and a logit model—trained on the 30% sample of Iowa sons link to the 1940 census, constructed manually by trained research assistants. Each observation is a possible link between a son in 1915 and records in 1940. I use the probit model to generate a score for each possible link. Possible links are coded as actual links if (1) the score is the top score for the given son in 1915, (2) the score is larger than 0.14, and (3) the ratio of the top score to the second best score is larger than 1.2. These parameters were chosen to maximize the accuracy and efficiency of the model through cross-validation. If there are any sons in 1915 with multiple exact matches in 1940—that is exactly the same first name string, last name string, and year of birth—then I am unable to pick between these possible matches. All possible matches are equally as likely to be the true match. Instead, I score any record links of this type with failure it is not used directly in the prediction algorithm.

Figure B.1: Letter Similarity Scores Used to Calculate Typographical Errors



(a) Upper Case Letters **(b)** Lower Case Letters

Letter Similarity Scores used to calculate typographical errors. The letters listed on the x-axis are most similar to the letters printed on the column chart. For example, O and Q are most similar upper case letter pair, with a score of 6. S is the upper case letter least likely to be confused as its most similar match is B with only a score of 2.73. Among lower case letters, l and i are most similar (score of 6.13); z is the most distinct.

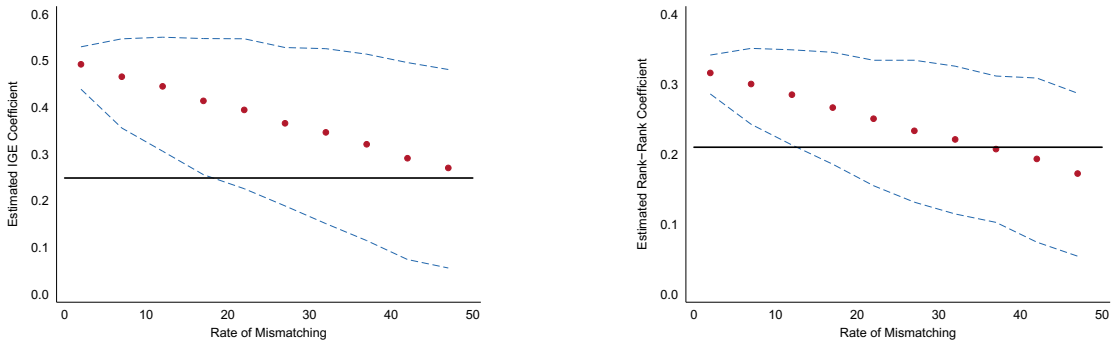
B.2 Matching Bias and Measurement Error

B.2.1 Matching Bias

The analysis that I conduct in this paper requires the construction of a dataset that links fathers and sons over time, between two censuses. The linking procedure, though carefully conducted, likely introduces a type of measurement error and bias to the estimation of IGE parameters for historical periods that might not be present in contemporary data. As these errors (or the mismatching rate) grow, the likely estimate of the IGE or of the rank-rank parameter will fall. To some extent, this could explain why an estimate of the historical IGE is smaller than a contemporary estimate of the same parameter, even if the true values are the same. I follow a mismatching simulation procedure based on the one used by Parman (2011) to gauge the magnitude of these biases.[1] While it is impossible to know exactly the rate of mismatches in the my linked sample between Iowa 1915 and the Federal 1940 census, I can introduce different levels of mismatch into the PSID data and measure the effect on estimated IGE parameters for matching error.

[1] Though Parman also draws on the 1915 Iowa Census, the construction of my dataset of linked fathers and sons varies somewhat from that used by Parman (2011). Thus, my simulation method differs from his so as to properly replicate the possible points of measurement error in the matching.

Figure B.2: Simulated Intergenerational Mobility of Income in the PSID Iowa-like sample as the rate of mismatch between fathers and sons varies



(a) Simulated Intergenerational Elasticity of Income

(b) Simulated Intergenerational Rank-Rank Correlation of Income

To determine the appropriate mismatching simulation for my data, I begin by reexamining the actual matching procedure used between the 1915 and 1940 censuses. I observe families with fathers and sons in 1915 where sons are between 3 and 17 years of age. There is no restriction on father's ages in these sample families. I then search for the sons in the 1940 Federal Census, using uniquely identifying information such as first and last name, state of birth, and year of birth. However, despite my best efforts at ensuring a unique and correct match, I may identify the "wrong" son in 1940.[2]

Suppose the match error rate is π. That is, if I make 100 matches, then $\pi \times 100$ matches will be erroneous. To replicate a π share of matching errors in the PSID, I drop the son's income and education data for π of the father-son pairs. Then, I randomly draw new son outcome data (income or education, independently), conditional on the true son's age.[3] Using this new data I estimate an IGE parameter, following the regression specified in the main empirical section of this paper. I simulate 1000 draws for each π and, following Parman (2011), I determine the IGE for π ranging from 2 to 50 percent. I repeat the same procedure (with a new simulation of mismatches) to test the stability of the rank-rank parameter as well.

[2] Wrong sons, in this case, would be a man with the same name, state of birth, and year of birth (within a 1 or 2 year bandwidth). This is not a common name or "John Smith" problem, as there are likely too many John Smiths born in a given state and year. Rather this might be a "John Smitherson" problem if there are two John Smithersons but one is not found in 1940 (possibly because he is dead, out of the country, or had his name transcribed incorrectly into the census as, for example, John Smithson).

[3] I do this all conditional on the son's age because I observe the true son's age in the original 1915 sample.

Figure B.2 presents the estimated β parameter from these mismatching tests, with the rate of mismatching, π, on the x-axis. In Figure B.2a, the solid horizontal line at $\beta = 0.258$ represents my largest estimate of the IGE between 1915 and 1940 from Panel A of Table 2.7. These tests suggest that a mismatching rate of more than 50% would be required to generate an IGE as low as I find in the historical period, if the true IGE were the same as in contemporary samples.[4] Given that matches are made on first and last names, states of birth, and years of birth, such a high rate of mismatch seems extremely unlikely. Similarly, Figure B.2b presents the same tests but for the rank-rank parameter, following Dahl and DeLeire (2008) and Chetty, Hendren, Kline, Saez, and Turner (2014). In this case, the solid horizontal line is drawn at $\beta = .220$, the largest estimate of the rank-rank parameter for the full sample in Panel B of Table 2.7. In this case, the mismatch rate would have to be at least 30% to induce such a low rank-rank estimate of intergenerational mobility in the recent data as I find historically.

B.2.2 Measurement Error

As Haider and Solon (2006); Solon (1989) show, measurement error will cause serious problems for estimates of IGE parameters. I have attempted to minimise these issues with age quartic controls, age quartic interaction controls, and by sampling fathers and sons at the middle of both lifecycles. In addition, I compared my historical estimates to contemporary estimates generated with just a single year of income data observed for fathers and sons, and the difference in the results remained. Finally, my results are quite consistent between several measures of father and son outcomes—income, education, and occupational standing. These measures all suffer from their own measurement problems, but taken together the consistent results are reassuring that intergenerational mobility was in fact lower in the early twentieth century than it is in the recent period.

As a further test of the measurement error effects, I introduce measurement error into the presumably well-measured PSID data. Let ζ be a $N(0, \sigma^2)$ shock. I add this random noise to either the father's income, the son's income, or both (in this case, the shocks are uncorrelated). I then reestimate the IGE parameter. I simulate 1000 draws for each σ and let σ vary from 0 to 1.

[4] Doing a similar test, Parman similarly finds a mismatching rate of 50% would be required to overturn his IGE findings.

Figure B.3: Simulated Intergenerational Elasticity of Income in the PSID as the noise in earnings varies

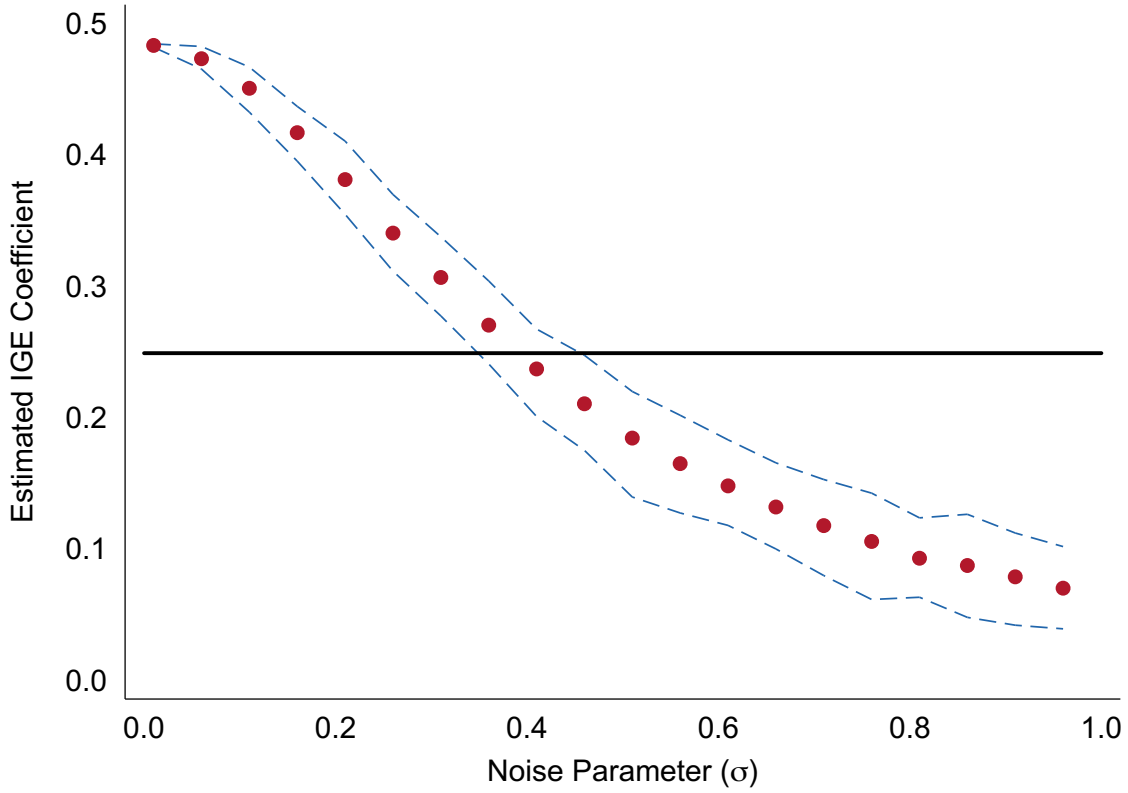


Figure B.3 presents the estimated betas for measurement error in both measures of the father's and son's earnings.

B.3 Farmer Income in 1940

For each individual listed in the 1940 Federal Census, annual wage and salary earnings and weeks of work are reported.[5] However, the census does not include information on either business or farm income as in later censuses. In practice, farm owners and other business proprietors reported working a full year (52 weeks) and having zero income in 1940. Thus, for any observed sons in 1940 who are either farm owners or business proprietors, I do not observe any measure of earnings in 1940. This restriction does not apply to farm labourers: farm labour income is

[5] IPUMS reports the specific enumeration instructions. The entry should be the "total amount of money wages or salary" but "Do not include the earning of businessmen, farmers, or professional persons derived from business profits, sale of corps, or fees."

reported in the same way as any other form of wage or salary income. Of the 4,478 matched sons in my sample, 1,177 report zero earnings in 1940. Of these, more than half (610) are farmers or farm owners or farm operators. The other 567 are a variety of occupations, including proprietors (36), operators (31), labourers (29), owners (23). and various forms of doctors and lawyers.[6]

In the main results presented in Panels A and B of Table 2.7, I drop all of these observations with no earnings in 1940. However, to the extent that sons with either very high or very low intergenerational mobility select into farming in 1940, this restriction could bias my estimates. It may be the case that the sons are farmers in 1940 because they have inherited the family farm from their fathers and thus their incomes, driven perhaps in large part by the productivity in the same plot of land, are highly correlated. Given the large changes in agriculture during this period, owing both to mechanization, the discovery of new irrigation sources in the Ogallala Aquifer, and especially the Dust Bowl (Hornbeck and Keskin, 2014; Hornbeck, 2012), this correlation may not be as strong in the early twentieth century as during other eras.

My results are consistent across other measures of mobility, particularly educational mobility, which do not suffer from this same missing data problem in 1940. As a further robustness check, I impute farm income in 1940 and re-compute the main results on intergenerational mobility below.[7] The estimated IGE and rank-rank parameters including sons with imputed capital income are presented in the sixth rows of Table 2.7 and suggest even more mobility historically, relative to contemporary estimates, than do my baseline results. Thus, the fact that I have to exclude the sons who are farmers in 1940 in the main results because I do not observe their incomes is not driving the measured result that income mobility was higher in the early twentieth century than it is today.

Here, I detail the imputation process. First, I collect data from the 1950 IPUMS 1% sample, which includes measures of both wage and salary income (which I observe in 1940), as well as total income and business and farm income. In a sample of only farmers in 1950, I regress

[6] The most common occupation among the non-farmers without earnings is actually to have no occupation listed (83 of the 567). These men are likely unemployed and not working for the WPA and thus the zero reported earnings are a correct measure not unreported data.

[7] I do this only the 610 farmers in my sample in 1940 without reported earnings. For the other proprietors and occupations, imputation would be much less accurate given the smaller sample sizes of these various occupations in 1950 and the (perhaps) more idiosyncratic nature of earnings in these professions.

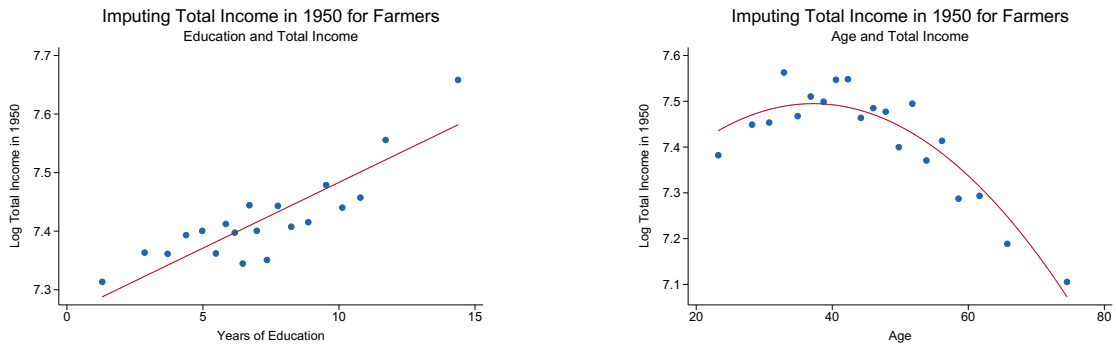


Figure B.4: Binscatter graphs presenting the correlation between years of education and age with log total income in 1950 for farmers from the IPUMS 1% sample. Both figures include controls for weeks worked, state of birth, state of residence, and education or age (when the variable is not on the x-axis). The slope in the figure on the left is approximately 0.0225.

total income on years of education,[8] a quartic in age, number of weeks worked, and indicator variables for state of birth and state of residence in 1950. Using the results of this regression, I impute the total income of farmers in my 1940 sample, assuming that the earnings function in 1940 resembles that in 1950 with respect to the effects of education, age (experience), weeks of work, and location fixed effects. An additional week of work in 1950 increases total income by 0.22%. As Figure B.4 shows, the relationship between years of education and the log of total income is nearly linear, with a slope of 2%,[9] while the relationship between age and total income is non-linear. Both the state of residence and state of birth fixed effects are quite strong as well. Overall, the R^2 is 15.52. I convert the imputed 1950 earnings to 1940 earnings with the price deflator.

B.4 Imputing Relationships

The 1915 Iowa State Census lacks household relationships. The raw data is stored not in tabular form, as is the case for federal censuses, but rather, in card form, with one card for each individual in the state. Goldin and Katz (2008) create household families in the data based on card position and last name and address matching. However, to link fathers and sons between

[8] With a large sample in 1950, I measure the effect of education nonparametrically with indicator variables for each year of education.

[9] The returns to education are quite low in this imputation because I am also controlling for state of residence. Among farmers, differential mobility and location choice is likely one of the channels through which education determines earnings.

1915 and 1940, the existing family identification is insufficient. Instead, I need to assign roles within each family, to identify the father, mother, and children, as well as any other non-nuclear family members in the household. I create an algorithm to assign probable family roles to each member. 85% of families observed in the Iowa 1915 data have two married people and the rest single, with the married people of the opposite sex and the other family members younger than the married couple. In these cases, assigning roles is trivial.[10] For the rest, I use simple rules based on marital status of family members, sex, and age.

I test my algorithm on the IPUMS 1% samples for 1910 and 1920 for Iowa that does include household roles. The results are presented below. I report the true census relationship from IPUMS across the table and my imputed family relationship down the table.[11] Generally, my family relationship imputation does quite well in replicating the family positions for citizens of Iowa in 1910 and 1920. Among fathers in the 1910 and 1920 IPUMS samples, my imputation algorithm identifies 98.7% as fathers and only 1.2% of the identified fathers are false positives. Among children, 98.8% are identified properly and only 3.1% of the identified children are false positives.

	True Census Relation				
	child	father	mother	other	Total
child	18265	17	2	569	18853
father	38	6391	0	38	6467
mother	40	0	6521	84	6645
other	151	69	47	804	1071
Total	18494	6477	6570	1495	33036

[10] Of course, some of those children could be step-siblings or half-siblings or live-in cousins. Unfortunately, there is no way for me to know this with any certainty. In terms of comparability with recent data however, studies of intergenerational mobility using the PSID, for example, rely not on measures of biological fathers', but on income of the male head of household in the child's house during childhood. Thus, misassignment of step-fathers as fathers is not a major problem.

[11] I should note that the so-called true census relationship are in fact imputed by IPUMS as well, based on family roles relative to the head of household reported on the census, as well as age, sex, and name.

B.5 Intergenerational Mobility under Alternative Function Forms

Measuring intergenerational mobility of income using the log-log specification, as is done in the main section of this paper and in the intergenerational mobility literature, constrains the relationship between the incomes of successive generations to have a very particular function form. The log function assigns the same weights to percentage changes in income, rather than to absolute changes in income. Thus very small changes in income at the bottom of the distribution are given the same weight as much larger change in income elsewhere in the distribution. To the extent that farmers were growing their own food and not selling production on the market, their reported or cash incomes would understate their true incomes or status.

However, I show here that my standard results are robust to alternative transformations of annual income. First, in Panel A of Table B.3, I show that normalizing earnings to be weekly, rather than annual, increases historical levels of mobility.[12]

In Panel B of Table B.3, I present intergenerational "elasticity" estimations[13] with the income variables in levels. Panel C of Table B.3 uses a square root transformation.

I also recompute the intergenerational elasticity of education using a log-log specification. In this case, the weighting implied by a log transformation is somewhat unnatural. Lemieux (2006) argues that in contemporary data the returns to education in a traditional Mincerian framework are convex, suggesting that each additional year of schooling is actually more valuable than the previous year.[14] By logging both the father's and the son's years of education, this specification implies that the return to each year of schooling for the father is decreasing (where the "returns" are measured as the years of completed for the son, rather than in wages, as is usual). The estimates presented in Panel E of Table B.3 suggest that the IGE parameter is smaller than I found earlier in this paper. Using the more traditional levels version, I found an IGE for education of between 0.187 and 0.275. Here, in logs, the IGE is between 0.10 and 0.19 and the confidence

[12] In 1915, the Iowa State Census measured the number of months unemployed for respondents. In 1940, the Federal Census measured the number of weeks employed. Using these variables, I can easily construct earnings per week employed.

[13] This is a slight abuse of notation common to the IGE literature. When the father's and son's incomes are no longer logged, the parameters are not truly elasticities.

[14] There is no work, that I am aware of, in the spirit of Lemieux that reconsiders the exact polynomial function of education that best fits the data in a Mincerian wage regression. Mincer (1974) uses untransformed years of schooling in his canonical study.

intervals for these sets of estimates do overlap.

B.5.1 Intergenerational Mobility using Family Income

To generate the contemporary estimates of the intergenerational elasticity of income, researchers typically measure income at the family level rather than at the individual level, on both the right hand side (fathers) and on the left hand side (sons) (for example Lee and Solon, 2009). This is a necessary definition of earnings when the goal is to measure the relation, broadly, between outcomes from one generation to the next. However, given historical patterns of female labour force participation (Goldin, 2006), I have chosen in this paper to measure income at the individual level. First, this more accurately replicates the occupational mobility literature, which measures the occupational categories of fathers and sons, ignoring mothers and spouses. Second, the collection of wives' income from the 1940 census would have added an additional round of costly data collection and little usable data, given how few married women worked in this period. In the Goldin-Katz 1915 Iowa sample, the overall correlation between family income and the income of the head of household is 0.9951; among my sample of fathers (limited to those with matched sons in 1940), the correlation is 0.9976. Thus, while for some families, possibly those with disabled or sick fathers, the mother's income could be a valuable resource, in practice the father's income and the family's income are nearly identical. The results presented in Panel D of Table B.3 underscore that expectation.

B.6 Construction of Occupational Score from 1915

Prior to 1940, the United States Federal Census did not ask respondents to report annual income. Economic historians and others interested in income and occupational standing have instead used reported occupations to measure social status, linking the occupations to median income by occupation from 1950. These so-called occscores are provided by IPUMS in all census data extracts before 1950. However, while such occscores likely provide some information on the expected income of a given census respondent, the signal to noise ratio falls as the analysis shifts to earlier census data. This occurs for two main reasons. First, the measurement error in matching occupations across time increases with time. While the tasks performed by an

Table B.3: Intergenerational Mobility Estimates: Alternative Function Forms

	Specification				
	(1)	(2)	(3)	Observations	Clusters
	A. Intergenerational Elasticity (IGE), Log Weekly Earnings				
Full Sample	0.207	0.194	0.266	2039	1667
	(0.032)	(0.032)	(0.084)		
Urban Sample	0.286	0.280	0.311	1004	824
	(0.046)	(0.052)	(0.105)		
Rural Sample	0.147	0.161	0.244	1035	843
	(0.040)	(0.041)	(0.116)		
	B. Annual Income in Levels				
Full Sample	0.319	0.319	0.438	2116	1731
	(0.044)	(0.042)	(0.104)		
Urban Sample	0.541	0.548	0.558	1025	842
	(0.082)	(0.084)	(0.208)		
Rural Sample	0.193	0.203	0.352	1091	889
	(0.045)	(0.045)	(0.107)		
	C. Square Root of Annual Income				
Full Sample	0.189	0.184	0.251	2116	1731
	(0.031)	(0.030)	(0.069)		
Urban Sample	0.298	0.298	0.286	1025	842
	(0.057)	(0.058)	(0.141)		
Rural Sample	0.120	0.132	0.217	1091	889
	(0.035)	(0.035)	(0.076)		
	D. Log Family Annual Income				
Full Sample	0.208	0.195	0.260	1955	1595
	(0.032)	(0.032)	(0.085)		
Urban Sample	0.280	0.265	0.292	946	774
	(0.046)	(0.052)	(0.109)		
Rural Sample	0.157	0.168	0.246	1009	821
	(0.041)	(0.041)	(0.118)		
	E. Log Years of Education				
Full Sample	0.186	0.170	0.100	2437	1959
	(0.020)	(0.020)	(0.049)		
Urban Sample	0.200	0.179	0.235	1107	900
	(0.034)	(0.034)	(0.060)		
Rural Sample	0.175	0.166	0.027	1330	1059
	(0.025)	(0.026)	(0.058)		

Standard errors clustered by family in all regressions. In Panel A, son's weekly log earnings in 1940 is the dependent variable. In Panel B, son's annual income (in levels not logs) is the dependent variable. In Panel C, the square root of the son's annual income is the depdendent variable. In Panel D, the depdendent variable is the son's individual earnings in 1940 in logs; the key independent variable is the log of 1915 family income, rather than father's income. In Panel E, educations is measured as the log of years of completed schooling. Specification 1 is a univariate regression of son's outcome on father's outcome. Specification 2 adds name string controls, 1915 county fixed effects, and quartic controls in father and son age. Specification 3 adds an interaction between father's outcome and son's normalised age. Name string controls: first and last name commonness, length, letter similarity, and Scrabble scores. Son's ages are normalised relative to age 40 in 1940.

Sources: 1915 Iowa State Census Sample; 1940 Federal Census

accountant or bookkeeper were very similar between 1950 and 1940, they are far different from the tasks performed by accountants at the turn of the twentieth century.[15] Second, in the response to both uneven technological change over time as well as shifting supply and demand for various types of labour, the returns to some occupations will fall and the returns to other will rise. The magnitudes of these changes to technology, supply, and demand are likely to grow over time.

Taking advantage of the 1915 Iowa State Census, which was the earliest census in the US to record respondents' incomes, occupation, and education level, I construct two variants on the traditional measures of occupation score, measured not in 1950 but in 1915. While these measure are highly correlated with the occupation score provided by IPUMS based on the 1950 census, they vary in important ways and likely allow for a more accurate assessment of the income in a given occupation in the United States in the early twentieth century.

IPUMS defines the "OCCSCORE" on a 1950 basis as:

> The occupational income score indicates the median total income – in hundreds of dollars – of the personas [sic] in each occupation in 1950. It is calculated using data from a published 1950 census report. For the post-1950 period, the score reflects the weighted average income of the 1950 occupational components of each contemporary occupation. In practice, this has only a small effect, but it means that the measure can vary slightly across census years for a given occupation.[16]

The 1950 census source used by IPUMS includes a median income for men and for women. IPUMS then weighs these medians by the sex share in each occupation to get one score for a given occupation. Using the Goldin and Katz (2008) sample of the 1915 Iowa State Census, I can create a similar median wage for each occupation group.

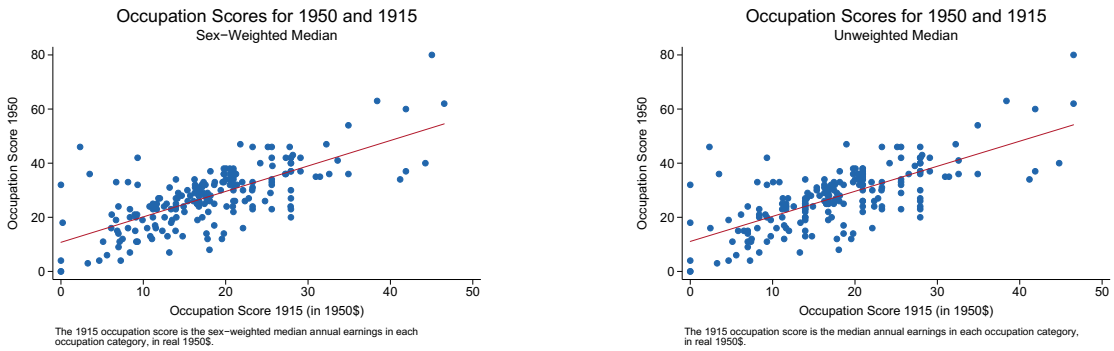
The Goldin and Katz (2008) Iowa sample includes a variable linking each observation to the 1940 occupation codes used in IPUMS. To generate a crosswalk between the 1940 and 1950 occupation codes, I collect the IPUMS 1940 1% sample of the census and contract the data by 1940-occupation and 1950-occupation.[17] Merging this crosswalk onto the 1915 Iowa data allows

[15] On the historical occupation tasks of accounting and bookkeeping specifically, see Rosenthal (2013).

[16] `https://usa.IPUMS.org/usa/chapter4/chapter4.shtml`

[17] The exact variables in the IPUMS sample are `occ` and `occ1915`.

Figure B.5: Comparing Occupation Score Measures between 1915 and 1950



me to link observations of income in the Iowa data to occupation categories in the 1950 data. I then calculate both the simple median and sex-weighted median income within each occupation group.[18]

Figure B.5 presents scatter plots of the occupation scores for 1950 and 1915. Both measures of occupational score in 1915 are highly correlated with the 1950 measure: the sex-weighted median is correlated at 0.7091 and the simple median at 0.7059. Given the high correlation with the traditional occupation score measure and the high correlation between my two constructed measures, I will focus on the sex-weighted median, particularly because the construction of that variable follows the IPUMS construction of the 1950 occupation score variable.

The points farthest from the best fit line may be of some interest. These are the occupations for which the returns changed the most between 1915 and 1950. Potentially consistent with increasing returns to human capital or education, the two of the occupation categories with the largest difference between the occupation score in 1950 and 1915 are "Physicians and surgeons" and "Optometrists".[19] "Mechanical engineers" and "Power station operators" are both relatively low-paid positions in 1915, but by 1950 they are in the upper quartile of incomes. "Attendants, recreation and amusement" is an occupation that was relatively middle-ranked in 1915, but by 1950 is towards the bottom of the occupational ladder.

[18] To find the sex-weighted median, I first calculate the median income for each occupation category by sex. Then I calculated the weighted average of these two medians, where the weights are the shares of men or women in each occupation. For the occupations where all observations are the same sex, the simple median and the sex-weighted median are the same. For example, in Iowa 1915, of the 100 civil engineers observed, none are women. Conversely, of the 27 private family laundresses in the sample, all are women.

[19] An alternative story for these divergences would be the difference between the rural, agrarian economy of Iowa in 1915 versus the whole US economy in 1950.

BIBLIOGRAPHY

Aaronson, D., and B. Mazumder (2008): "Intergenerational Economic Mobility in the United States, 1940 to 2000," *Journal of Human Resources*, 43(1), 139–172.

Abramitzky, R., L. P. Boustan, and K. Eriksson (2012): "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102(5), 1832–1856.

——— (2013a): "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 122(3), 467–717.

——— (2013b): "Have the poor always been less likely to migrate? Evidence from inheritance practices during the age of mass migration," *Journal of Development Economics*, 102, 2–14.

Aizer, A., S. Eli, J. P. Ferrie, and A. Lleras-Muney (2014): "The Long Term Impact of Cash Transfers to Poor Families," http://www.nber.org/papers/w20103.

Altham, P., and J. P. Ferrie (2007): "Comparing Contingency Tables Tools for Analyzing Data from Two Groups Cross-Classified by Two Characteristics," *Historical Methods*, 40(1), 3–17.

Athey, S. (2015): "Machine Learning and Causal Inference for Policy Evaluation," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Autor, D. H., L. F. Katz, and M. S. Kearney (2008): "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics*, 90(2), 300–323.

Becker, G. S., and N. Tomes (1979): "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *Journal of Political Economy*, 87(6), 1153–1189.

Bernanke, B. S. (1995): "The Macroeconomics of the Great Depression: A Comparative Approach," *Journal of Money, Credit and Banking*, 27(1), 1–28.

——— (2000): *Essays on the Great Depression*. Princeton University Press, Princeton, NJ.

Biavaschi, C., C. Giulietti, and Z. Siddique (2013): "The Economic Payoff of Name Americanization," IZA, https://www.econstor.eu/dspace/bitstream/10419/89864/1/dp7725.pdf.

Black, S. E., and P. J. Devereux (2011): "Recent developments in intergenerational mobility," *Handbook of Labor Economics*, 4(11), 1487–1541.

Bloome, D. (2015): "Income Inequality and Intergenerational Income Mobility in the United States," *Social Forces*, 93(3), 1047–1080.

Boone, C. D. A., and L. Wilse-Samson (2014): "Modernization, Rural Migration, and Market Withdrawal: Evidence from the Great Depression," http://www.columbia.edu/ cdb2129/Boone_jmp.pdf.

Bordo, M., C. Goldin, and E. N. White (1998): "The Defining Moment Hypothesis: The Editors' Introduction," in *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*, pp. 1–20. University of Chicago Press, Chicago, IL.

Boustan, L. P., M. E. Kahn, and P. W. Rhode (2012): "Moving to Higher Ground: Migration Response to Natural Disasters in the Early Twentieth Century," *American Economic Review*, 102(3), 238–244.

Calomiris, C. W., and J. R. Mason (2003): "Consequences of Bank Distress During the Great Depression," *American Economic Review*, 93(3), 937–947.

Checchi, D., C. V. Fiorio, and M. Leonardi (2008): "Intergenerational Persistence in Educational Attainment in Italy," IZA Discussion Papers, http://www.econstor.eu/dspace/bitstream/10419/34814/1/575205105.pdf.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2013): "The Equality of Opportunity Project," http://www.equality-of-opportunity.org/.

——— (2014): "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 129(4), 1553–1623.

Chetty, R., N. Hendren, P. Kline, E. Saez, and N. Turner (2014): "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility," *American Economic Review: Papers & Proceedings*, 104(5), 141–147.

Christen, P. (2012): *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*.

Clark, G. (2014): *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton University Press, Princeton, NJ.

Cohen, L. (1991): *Making a New Deal: Industrial workers in Chicago, 1919-1939*. Cambridge University Press, Cambridge, England.

Collins, W. J., and M. 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 A), 220–252.

Collins, W. J., and M. H. Wanamaker (2015): "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants," *The Journal of Economic History*, 75(04), 947–992.

Corak, M. (2004): "Generational income mobility in North America and Europe: an introduction," in *Generational income mobility in North America and Europe*, ed. by M. Corak, chap. 1, pp. 1–37. Cambridge University Press, Cambridge, England.

——— (2006): "Do poor children become poor adults? Lessons from a cross country comparison of generational earnings mobility," in *Dynamics of Inequality and Poverty (Research on Economic Inequality, Volume 13)*, ed. by J. Creedy, and G. Kalb, pp. 143–188. Elsevier Press, The Netherlands.

——— (2013): "Income Inequality, Equality of Opportunity, and Intergenerational Mobility," *Journal of Economic Perspectives*, 27(3), 79–102.

Corak, M., and A. Heisz (1999): "The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data," *Journal of Human Resources*, 34(3), 504–533.

Costa, D. (1997): "Less of a Luxury: The Rise of Recreation since 1888," Discussion paper, National Bureau of Economic Research, Cambridge, MA.

——— (1999): "American Living Standards: Evidence from Recreational Expenditures," Discussion paper, National Bureau of Economic Research, Cambridge, MA.

——— (2000): "American Living Standards, 1888-1994: Evidence From Consumer Expenditures," Discussion paper, National Bureau of Economic Research, Cambridge, MA.

Cullen, J., and P. V. Fishback (2006): "Did Big Government's Largesse Help the Locals? The Implications of WWII Spending for Local Economic Activity, 1939-1958," Discussion paper, National Bureau of Economic Research, Cambridge, MA.

Dahl, M., and T. DeLeire (2008): "The association between children's earnings and fathers' lifetime earnings: estimates using administrative data," Institute for Research on Poverty, http://irp.wisc.edu/publications/dps/pdfs/dp134208.pdf.

de Tocqueville, A. (1839): *Democracy in America*.

Duncan, O. D. (1965): "The Trend of Occupational Mobility in the United States," *American Sociological Review*, 30(4), 491–498.

Eichengreen, B. (1992): *Golden fetters: The Gold Standard and the Great Depression, 1919-1939*. Oxford University Press, New York, NY.

Elder, G. H. (1999): *Children of the Great Depression: Social Change in Life Experience*. Westview Press, Boulder, CO.

Feigenbaum, J. J. (2014): "A New Old Measure of Intergenerational Mobility: Iowa 1915 to 1940," http://scholar.harvard.edu/jfeigenbaum/publications/new-old-measure-intergenerational-mobility-iowa-1915-1940.

——— (2015): "Automated Census Record Linking: A Machine Learning Approach," http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf.

Ferrie, J. P. (1996): "A New Sample of Americans Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript," *Historical Methods*, 29, 141–156.

——— (2005): "History Lessons: The End of American Exceptionalism? Mobility in the United States Since 1850," *Journal of Economic Perspectives*, 19(3), 199–215.

Fishback, P. V., M. Haines, and S. Kantor (2007): "Births, deaths, and New Deal relief during the Great Depression," *Review of Economics and Statistics*, 89(1), 1–14.

Fishback, P. V., W. C. Horrace, and S. Kantor (2006): "The impact of New Deal expenditures on mobility during the Great Depression," *Explorations in Economic History*, 43(2), 179–222.

Fishback, P. V., W. C. Horrace, and S. E. Kantor (2005): "Did New Deal grant programs stimulate local economies? A study of Federal grants and retail sales during the Great Depression," *Journal of Economic History*, 65(1), 36–71.

Fishback, P. V., R. S. Johnson, and S. Kantor (2010): "Striking at the Roots of Crime: The Impact of Welfare Spending on Crime during the Great Depression," *Journal of Law and Economics*, 53(4), 715–740.

Fishback, P. V., S. Kantor, and J. J. Wallis (2003): "Can the New Deal's three Rs be rehabilitated? A program-by-program, county-by-county analysis," *Explorations in Economic History*, 40(3), 278–307.

Friedman, M., and A. J. Schwartz (1963): *A Monetary History of the United States, 1867-1960*. Princeton University Press, Princeton, NJ.

Fryer, R. G., and S. D. Levitt (2004): "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 119(3), 767–805.

Garrett, T. A., and D. C. Wheelock (2006): "Why Did Income Growth Vary Across States During the Great Depression?," *Journal of Economic History*, 66(02), 456–466.

Gelman, A., and J. Hill (2007): "Data analysis using regression and multilevel/hierarchical models," *Policy Analysis*, pp. 1–651.

Goeken, R., L. Huynh, T. Lenius, and R. Vick (2011): "New Methods of Census Record Linking," *Historical methods*, 44(1), 7–14.

Goldin, C. (1998): "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century," *The Journal of Economic History*, 58(2), 345–374.

——— (2006): "The Quiet Revolution That Transformed Women's Employment, Education, and Family," *American Economic Review*, 96(2), 1–21.

Goldin, C., and L. Katz (1997): "Why the United States Led in Education: Lessons from Secondary School Expansion, 1910 to 1940," Discussion Paper 6144, National Bureau of Economic Research, Cambridge, MA.

Goldin, C., and L. F. Katz (2000): "Education and Income in the Early Twentieth Century: Evidence from the Prairies," *Journal of Economic History*, 60(3), 782–818.

——— (2008): *The Race Between Education and Technology*. The Belknap Press of Harvard University Press, Cambridge, MA.

——— (2011): "Mass Secondary Schooling and the State: The Role of State Compulsion in the High School Movement," in *Understanding Long Run Economic Growth*, ed. by D. L. Costa, and N. R. Lamoreaux. Chicago, IL.

Goldin, C., and R. A. Margo (1992): "The Great Compression: The Wage Structure in the United States at Mid-Century," *Quarterly Journal of Economics*, 107(1), 1–34.

Grawe, N. D. (2004): "Intergenerational mobility for whom? The experience of high- and low-earning sons in international perspective," in *Generational income mobility in North America and Europe*, ed. by M. Corak, chap. 4, pp. 58–89. Cambridge University Press, Cambridge, England.

Grayson, K. Y. (2012): "Essays on Income Inequality and Health During the Great Depression," Ph.D. thesis.

Guest, A. M., N. S. Landale, and J. C. McCann (1989): "Intergenerational Occupational Mobility in the Late 19th Century United States," *Social Forces*, 68(2), 351–378.

Haider, S., and G. Solon (2006): "Life-Cycle Variance in the Association between Current and Lifetime Earnings," *American Economic Review*, 96(4), 1308–1320.

Heim, C. E. (1998): "Uneven Impacts of the Great Depression: Industries, Regions, and Nations," in *The Economics of the Great Depression1*, ed. by M. Wheeler, chap. 2, pp. 29–62. W.E. Upjohn Institute, Kalamazoo, MI.

Hertz, T., T. Jayasundera, P. Piraino, S. Selcuk, N. Smith, and A. Verashchagina (2007): "The Inheritance of Educational Inequality: International Comparisons and Fifty-Year Trends," *The B.E. Journal of Economic Analysis & Policy*, 7(2).

Hilger, N. (2015): "The Great Escape: Intergenerational Mobility Since 1940," Discussion paper, National Bureau of Economic Research, Cambridge, MA.

Hornbeck, R. (2012): "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe," *American Economic Review*, 102(4), 1477–1507.

Hornbeck, R., and P. Keskin (2014): "The Historically Evolving Impact of the Ogallala Aquifer: Agricultural Adaptation to Groundwater and Drought," *American Economic Journal: Applied Economics*, 6(1), 190–219.

Jencks, C., and L. Tach (2006): "Would Equal Opportunity Mean More Mobility?," in *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, ed. by S. Morgan, D. Grusky, and G. Fields, pp. 22–58. Stanford University Press.

Kane, T. J., C. Rouse, and D. Staiger (1999): "Estimating Returns to Schooling When Schooling is Misreported," NBER, http://www.nber.org/papers/w7235.

Kantor, S. E., and P. V. Fishback (1996): "Precautionary Saving, Insurance, and the Origins of Workers' Compensation," *Journal of Political Economy*, 104(2), 419–442.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015): "Prediction Policy Problems," *American Economic Review*, 105(5), 491–495.

Krueger, A. B. (2012): "The Rise and Consequences of Inequality," .

Krug, J. (1945): "Wartime Production Achievements and the Reconversion Outlook," Discussion paper, U.S. War Production Board, Washington, DC.

Kuhn, M., and K. Johnson (2013): *Applied Predictive Modeling*. Springer, New York.

Lahman, S. (2016): "The Lahman Baseball Database," .

Lebergott, S. (1964): *Manpower in Economic Growth: The American Record Since 1800*. McGraw-Hill Book Company, New York, NY.

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

Lemieux, T. (2006): "The Mincer Equation Thirty Years After Schooling, Experience and Earnings," in *Jacob Mincer A Pioneer of Modern Labor Economics*, ed. by S. Grossbard, vol. 79, pp. 127–145. Springer US.

Levine, D. I., and B. Mazumder (2002): "Choosing the Right Parents: Changes in the Intergenerational Transmission of Inequality Between 1980 and the Early 1990s," Federal Reserve Bank of Chicago, http://www.ssrn.com/abstract=323881.

Linder, F. E., and R. D. Grove (1947): *Vital Statistics Rates in the United State, 1900-1940*. United States Public Health Service, Washington, DC.

Long, J., and J. P. Ferrie (2004): "Geographic and Occupational Mobility in Britain and the U.S., 1850-1881," Working Paper, Northwestern University, http://www.colby.edu/economics/faculty/jmlong/research/usbritainmobility.pdf.

——— (2007a): "'Everything in Common ... [But] Language'?: Intergenerational Occupational Mobility in Britain and the U.S. Since 1850?," .

——— (2007b): "The Path to Convergence: Intergenerational Occupational Mobility in Britain and the US in Three Eras," .

——— (2013): "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850," *The American Economic Review*, 103(4), 1109–1137.

——— (2015): "Grandfathers Matter(ed): Occupational Mobility Across Three Generations in the U.S. and Britain, 1850-1910," https://sites.google.com/site/jasonmlongecon/papers/GrandfathersMattered.pdf.

Malamud, O., and A. Wozniak (2014): "The Impact of College on Migration: Evidence from the Vietnam Generation," *Journal of Human Resources*, 47(4), 913–950.

Marcin, D. (2014): "The Revenue Act of 1924: Publicity, Tax Cuts, Response," http://www-personal.umich.edu/ dmarcin/JMP.pdf.

Mayer, S. E., and L. M. Lopoo (2005): "Has the Intergenerational Transmission of Economic Status Changed?," *Journal of Human Resources*, 40(1), 169–185.

Mazumder, B. (2005): "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data," *Review of Economics and Statistics*, 87(2), 235–255.

——— (2015): "Estimating the Intergenerational Elasticity and Rank Association in the US: Overcoming the Current Limitations of Tax Data," Federal Reserve Bank of Chicago, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2647277.

McGraw, G., J. Rehling, and R. Goldstone (1994): "Letter Perception: Toward a conceptual approach," in *Sixteenth Annual Conference of the Cognitive Science Society*, pp. 613–618, Atlanta, GA.

Mendershausen, H. (1946): *Changes in Income Distribution During the Great Depression*, vol. I. NBER.

Mill, R. (2012): "Assessing Individual-Level Record Linkage between Historical Datasets," Stanford University.

Mill, R., and L. C. D. Stein (2012): "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America," http://econweb.umd.edu/ davis/eventpapers/MillRace.pdf.

Mincer, J. A. (1974): "Individual Acquisition of Earning Power," in *Schooling, Experience, and Earnings*, ed. by J. A. Mincer, vol. I, pp. 7–22. Columbia University Press, New York, NY.

Moehling, C. M. (2001): "Women's Work and Men's Unemployment," *Journal of Economic History*, 61(4), 926–949.

——— (2005): ""She Has Suddenly Become Powerful": Youth Employment and Household Decision Making in the Early Twentieth Century," *Journal of Economic History*, 65(2), 414–438.

Mulligan, C. B. (1997): *Parental Priorities and Economic Inequality*. University of Chicago Press, Chicago and London.

Nix, E., and N. Qian (2015): "The Fluidity of Race : "Passing" in the United States, 1880-1940," National Bureau of Economic Research, http://www.nber.org/papers/w20828.pdf.

Nybom, M., and J. Stuhler (2014a): "Biases in Standard Measures of Intergenerational Income Dependence," https://janstuhler.files.wordpress.com/2013/06/lifecycle_bias_pt2_november_2014.pdf.

——— (2014b): "Interpreting Trends in Intergenerational Mobility," http://www.sofi.su.se/polopoly_fs/1.170138.1394463038!/menu/standard/file/WP14n03.pdf http://www.homepages.ucl.ac.uk/ uctpjst/stuhler_interpreting_trends_(JMP).pdf.

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

Olney, M. L. (1998): "When your word is not enough: race, collateral, and household credit," *The Journal of Economic History*, 58(2), 408–431.

Parman, J. M. (2011): "American Mobility and the Expansion of Public Education," *The Journal of Economic History*, 71(01), 105–132.

Pfeffer, F. T., and A. Killewald (2015): "How Rigid is the Wealth Structure and Why? Inter- and Multigenerational Associations in Family Wealth," Population Studies Center Research Report, http://www.psc.isr.umich.edu/pubs/abs/9754.

Piketty, T., and E. Saez (2003): "Income Inequality in the United States," *Quarterly Journal of Economics*, 118(1), 1–39.

Richardson, G. (2009): "during the Great Depression : Quasi- Experimental Evidence from a Federal Reserve District Border , 1929 1933 William Troost," 117(6), 1929–1933.

Roberts, E. (2003): "Labor Force Participation by Married Women in the United States Results from the 1917/19 Cost-of-Living Survey and the 1920 PUMS," *Science History Association Conference*, (November).

Romer, C. D. (1986): "Is the Stabilization of the Postwar Economy a Figment of the Data?," *The American Economic Review*, 76(3), 314–334.

Romer, C. D. (1990): "The Great Crash and the Onset of the Great Depression," *Quarterly Journal of Economics*, 105(3), 597–624.

Rosenbloom, J. L. (2002): *Looking for Work, Searching for Workers: American Labor During Industrialization*. Cambridge University Press, Cambridge, England.

Rosenbloom, J. L., and W. A. Sundstrom (1999): "The Sources of Regional Variation in the Severity of the Great Depression: Evidence from U.S. Manufacturing, 1919-1937," *Journal of Economic History*, 59(3), 714–747.

Rosenthal, C. C. (2013): "From Memory to Mastery: Accounting for Control in America, 1750-1880," Ph.D. thesis, Harvard University.

Ruggles, S., J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek (2010): *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]*. Minnesota Population Center [producer and distributor], Minneapolis, MN.

Shanahan, M. J., G. H. J. Elder, and R. A. Miech (1997): "History and Agency in Men's Lives: Pathways to Achievement in Cohort Perspective," *Sociology of Education*, 70(1), 54–67.

Simpson, I. C., P. Mousikou, J. M. Montoya, and S. Defior (2013): "A letter visual-similarity matrix for Latin-based alphabets.," *Behavior research methods*, 45(2), 431–9.

Solon, G. (1989): "Biases in the Estimation of Intergenerational Earnings Correlations," *Review of Economics and Statistics*, 71(1), 172–174.

——— (1999): "Intergenerational Mobility in the Labor Market," *Handbook of Labor Economics*, 3.

——— (2004): "A model of intergenerational mobility variation over time and place," in *Generational Income Mobility in North America and Europe*, ed. by M. Corak, chap. 2, pp. 38–47. Cambridge University Press, Cambridge, England.

Temin, P. (1989): *Lessons from the Great Depression*. MIT Press, Cambridge, MA.

Terkel, S. (2013): *Hard Times: An Oral History of the Great Depression*. The New Press, New York.

Thernstrom, S. (1964): *Poverty and progress; social mobility in a nineteenth century city*. Harvard University Press, Cambridge, MA.

——— (1973): *The other Bostonians : poverty and progress in the American metropolis, 1880-1970*. Harvard University Press, Cambridge, MA.

Tyack, D., R. Lowe, and E. Hansot (1984): *Public Schools in Hard Times: The Great Depression and Recent Years*. Harvard University Press, Cambridge, MA.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2), 3–28.

Wallis, J. J. (1989): "Employment in the Great Depression: New data and hypotheses," *Explorations in Economic History*, 26(1), 45–72.

Watkins, L. R. (ed.) (2000): *A Generation Speaks: Voices of the Great Depression*. The Chapel Hill Press, Inc, Chapel Hill, NC.

Winkler, W. (1994): "Advanced methods for record linkage," (1991).

——— (2006): "Overview of record linkage and current research directions," *Bureau of the Census*.

Wozniak, A. (2010): "Are college graduates more responsive to distant labor market opportunities?," *Journal of Human Resources*, 45(4), 944–970.

Xie, Y., and A. Killewald (2013): "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850: Comment," *American Economic Review*, 103(5), 2003–2020.

Yagan, D. (2014): "Moving to Opportunity? Migratory Insurance Over the Great Recession," http://eml.berkeley.edu/ yagan/MigratoryInsurance.pdf.

Ziebarth, N. L. (2013): "Identifying the effects of bank failures from a natural experiment in Mississippi during the great depression," *American Economic Journal: Macroeconomics*, 5(1), 81–101.