Do Grandparents and Great-Grandparents Matter? Multigenerational Mobility in the US, 1910-2013

By JOSEPH FERRIE, CATHERINE MASSEY, AND JONATHAN ROTHBAUM*

Abstract

Studies of US intergenerational mobility focus almost exclusively on the transmission of (dis)advantage from parents to children. Until very recently, the influence of earlier generations could not be assessed even in long-running longitudinal studies such as the Panel Study of Income Dynamics (PSID). We directly link family lines across data spanning 1910 to 2013 and find a substantial "grandparent effect" for cohorts born since 1920. Although this may be due to measurement error, we conclude that estimates from only two generations of data understate persistence by about 20 percent.

* Ferrie: Northwestern University, 302 Donald P. Jacobs Center, 2001 Sheridan Road, Evanston, IL 60208 (email: ferrie@northwestern.edu); Massey: Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor MI 48104 (email: cgmassey@umich.edu); Rothbaum: US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 (email: jonathan.l.rothbaum@census.gov). This work was made possible by the Center for Administrative Records Research and Applications at the U.S. Census Bureau. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Do Grandparents and Great-Grandparents Matter? Multigenerational Mobility in the US, 1910-2013*

October 15, 2017

Abstract

Studies of US intergenerational mobility focus almost exclusively on the transmission of (dis)advantage from parents to children. Until very recently, the influence of earlier generations could not be assessed even in long-running longitudinal studies such as the Panel Study of Income Dynamics (PSID). We directly link family lines across data spanning 1910 to 2013 and find a substantial "grandparent effect" for cohorts born since 1920. Although we find evidence that the majority of the grandparent effect is due to measurement error, we conclude that estimates from only two generations of data understate persistence by about 25 percent.

Keywords: Intergenerational Mobility; Multigenerational Mobility
JEL Codes: J62, R00

*This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the US Census Bureau.

1 Introduction

Research since the mid-1970s on rising inequality in the US (e.g. Piketty and Saez 2014) has been accompanied by increased attention from economists on a related topic of long-standing interest among sociologists, equality of opportunity and intergenerational mobility. This is generally measured as the correlation across generations within family lines in economic and social status. Though research on intergenerational mobility has advanced dramatically over the past two decades, with a proliferation of methodological improvements (Solon 1992; Mazumder 2005), data sources (Chetty et al. 2014; Feigenbaum 2014; Feigenbaum 2015; Hilger 2015; Clark 2014; Grusky et al. 2015), and international comparisons (Björklund and Jäntti 1997; Solon 2002; Bourdieu et al. 2009; Long and Ferrie 2013), this work has focused almost exclusively on the transmission of (dis)advantage from parents to their children. The influence of grandparents and earlier generations, if addressed at all, is inferred through iteration of the parental influence, a procedure that is unsatisfying in a number of respects (Stuhler, 2012).

The focus on just parents and children has been primarily due to practical considerations: few longitudinal studies have been running long enough to capture the experiences of three or more generations through adulthood in the same family lines. Yet there are several reasons why we might expect that grandparents would matter. Mare (2011) describes several channels through which generations prior to parents could have an impact on children's outcomes, including inheritance of financial assets, inheritance of social networks, and the direct effects that personal contact between grandparents and their grandchildren could have in a society where increasing longevity means their lives will increasingly overlap.

The inability to assess a "grandparent effect" also shapes how we view, in a purely statistical sense, the parent-child transmission process. As Solon (1992) notes, an intergenerational correlation will be biased downward by measurement error in the parent generation. However, information on grandparents can reduce the impact of this downward bias, even if we are only interested in parent-child transmission (Solon, 2014). Finally, the presence of a grandparent effect has been detected in many of the places where it has been possible to look for it: rural China (Zeng and Xie, 2014), Sweden (Lindahl et al., 2015), Britain (Chan and Boliver, 2013), Germany (Hertel and Groh-Samberg, 2014), Chile (Celhay and Gallegos, 2015), Denmark (Boserup et al., 2013) and even the

pre-WWII US (Olivetti and Paserman 2015, Long and Ferrie 2012). In fact, one of the few places where such an effect has not been consistently observed is the modern US. Warren and Hauser (1997), using the Wisconsin Longitudinal Study, found no independent grandparent effect, though Mare (2011, p. 16) warns that "mid-twentieth century Wisconsin families may be a population in which multigenerational effects are unusually weak." One paper using the Panel Study of Income Dynamics (PSID) has found a grandparent effect (Hertel and Groh-Samberg, 2014), but neither Hodge (1966), Behrman and Taubman (1985), nor Peters (1992) found one.

We contribute to the research on multigenerational mobility in several ways. We link family lines from the 1910, 1920, and 1940 US Censuses of Population to the Current Population Survey (CPS) Annual Social and Economic Supplements (ASEC) of the 1970s and 1980s and to the 2000 US Census of Population Long Form (LF) and the 2001-2013 American Community Surveys (ACS). From these linkages, we construct one four-generation sample and two three-generation samples.[1] We measure mobility across generations using educational attainment (years of schooling), which suffers from less measurement error than single-year observations of income.[2] Our ability to observe many of the same individuals at multiple points in time (for example, many individuals can be observed as adults in both the 2000 Long Form Census and the 2001-2013 ACS) allows us to directly assess the degree of measurement error in reported education. Finally, we link both males and females throughout the 1910-2013 period we examine, making it possible to assess differences in the effects of male and female ancestors as well as differences in outcomes for males and females.[3]

Although we find a grandparent effect in our regressions, our results indicate that the majority of that effect is due to measurement error. However, we do find that intergenerational mobility of education follows an AR(2) process. We estimate that intergenerational mobility of education using two generations of data underestimate persistence by approximately 25 percent.

[1]The structure of our data also allows us to examine change over time in the two-generation parent-child transmission of (dis)advantage. In Ferrie et al. (2016), we show that mobility in educational attainment was high as high schools appeared in the early twentieth century, but fell for successive cohorts born 1895-1915 as access to education expanded across the U.S. through the early 1930s.

[2]There is now a substantial literature on the challenges in estimating intergenerational income elasticities with noisy data on parents' and/or children's incomes. See Mazumder (2016).

[3]The baseline estimates in this paper include all children, regardless of gender, race, or place of birth. In the appendix (Table A.3), we include a table showing separate results by the gender of children and their ancestors.

2 Data

2.1 Constructing Family Dynasties

To create a four-generation sample that spans the twentieth century, we use the complete count 1910, 1920, and 1940 Censuses, the 1973, 1979, and 1981-1990 CPS ASEC, the 2000 Long Form Census, and the 2001-2013 ACS. Throughout the analysis, we use males and females of all races and countries of origin. We refer to adults in the 1910 and 1920 censuses as the Great-grandparent generation. Those in the 1940 census are the Grandparent generation, and adults in the CPS ASEC are the Parent generation. Finally, adults in the 2000 Census and ACS, are the Child generation. These generations, which we capitalize as proper nouns, refer to the data source and general age cohorts of the adults in each sample.

We employ Person Identification Keys (PIKs), assigned by the Census Bureau, to link individuals across data sources from 1940 forward. PIKs are assigned by a probabilistic matching algorithm that compares characteristics of records in census and survey data to characteristics of records in a reference file constructed from the Social Security Administration (SSA) Numerical Identification System (or Numident) as well as other federal administrative data. These characteristics may include Social Security Number (SSN), full name, date of birth, address, place of birth, and parents' names depending on the information available in the census or survey.[4] The PIK uniquely identifies a particular person and is consistent for that person over time. The PIK thus allows us to link individuals across data sources. See the appendix for a full description of PVS and how the PIKs were assigned to records.

Figure 1 shows the process by which links are created; each box represents a single household. We begin with a sample of both male and female children assigned PIKs in the 1940 Census. We observe these children living with their parents in 1940, creating the parent-child link between the Grandparent and Parent generations (1). We then search for the Parent generation's PIKs in the 1973, 1979, and 1981-1990 CPS ASEC (2). From this linkage, we observe the outcomes of the Parent generation as adults and create another parent-child link from the Parent to the Child generation by observing the Parent generation residing with their own children in the CPS ASEC (3). We then locate the PIKs of the Child generation in the 2000 Long Form Census or the

[4]PIKs correspond one-to-one with a particular SSN.

2001-2013 ACS (4).

We use two approaches to link families from the 1910 and 1920 censuses to the 1940 Census. First, because the 1910 and 1920 censuses did not collect education attainment, we use probabilistic matching techniques to link individuals at least 25 years of age in 1910 and 1920 to 1940 to pick up years of schooling (5). We use first and last name, middle initial, age, and birthplace to conduct the linkage.[5] Next, we create family units in 1910 and 1920, identify children of parents linked to 1940 (6), and link those children forward to 1940 (7), again using probabilistic matching techniques. This allows us to observe education attainment for the Great-grandparent generation and link them to their children in the Grandparent generation, observed as adults in 1940.

[[Insert Figure 1 Here]]

We use an additional approach to match female Grandparents in 1940 back to their 1910 or 1920 childhood observation. First, we construct family units in the 1940 Census (1). Then we use the PIKs assigned to children to link to the Numident, which contains the mother's maiden name. We then append the mother's maiden name to the mother's 1940 record and conduct the record linkage back to 1910 or 1920 using maiden name instead of the reported surname (the reverse of link (7)). Consequently, the sample of female 1940 children for whom we observe a parent in the Great-grandparent generation depends on (i) successfully linking the Great-grandparent generation individual from 1910/1920 to 1940 to obtain education attainment (which necessitates the Great-grandparent surviving until 1940) and on (ii) whether the female Grandparent observation received a PIK in the 1940 Census.[6]

We rely on the reported relationship to household head to establish parent-child relationships and identify spouses. We omit subfamilies in the 1910, 1920, and 1940 censuses, but we include

[5] We describe our process to link 1910 and 1920 to 1940 in more detail in the appendix.

[6] Linkage of Great-grandparents implicitly requires survival from 1910/1920 to 1940. The life-expectancy of someone who survived to be 20 to 50 years old in 1909-1911 is 62.7 to 70.4 years, and the life-expectancy of someone who survived to be 20 to 50 years old in 1919-1921 is 65.6 to 72.2 (Arias, 2015). Consequently, most of the adults we observe in the Great-Grandparent generation from the 1920 Census were expected to live until 1940, but survival to 1940 would have been much lower for those 50 years and older in 1910. If higher-educated individuals tend to have higher levels of education attainment (Lleras-Muney, 2005), the necessity for the Great-Grandparents to survive from 1910 to 1940 may bias our results. In Figure 1, we plot reported education by year of birth and by decennial census. We also include the combined 1910-1940 and 1920-1940 linked samples (denoted as 1910-1940 Linked Sample). If higher-educated individuals live longer, we would expect to see higher average education lines for those born in earlier years and survived to the later censuses. This figure shows that 1960 and 1940 lie right on top of each other, as do the 1950 and 1970 censuses. This suggests the relationship between education and survival is small during this time period.

subfamilies observed in the CPS ASEC. We require that adults fall between the ages of 25-55 when we observe their education attainment. We also require that the age observed for a record across multiple data sources are within a five-year interval around what we would expect. For example, if a person is 3 years old in 1940, we expect that person to be roughly 53 years old in the 1990 CPS ASEC, and we drop this observation if they are younger than 48 or older than 58 years old.[7]

The Great-grandparent-Grandparent-Parent generation sample (data from 1910-1990) includes matched children from the Parent generation in the CPS ASEC and their parents from the 1940 Census. Their grandparents from the Great-grandparent generation in the 1910 and 1920 Censuses are added to the data set when available, but are not required for inclusion of the other two generations in the sample. Great-grandparent generation observations are available for about 10 percent of the sample (3,517 individuals from the Parent generation).[8] The Grandparent-Parent-Child sample (data from 1940-2013) includes matched children from the 2000 Long Form Census and 2001-2013 ACSs, their parents from the CPS ASEC, and their grandparents from the 1940 Census. For this sample, children are included only if we observe at least one parent and at least one grandparent.

2.2 Measuring Education Attainment

The 1940 Census was the first U.S. federal census to collect information on schooling. Respondents reported both highest grade of schooling completed and whether they were attending school on March 1, 1940. The 1940 census requested schooling information from all respondents, unlike subsequent decennial censuses that limited education questions to sample line individuals or the Long Form sample. The 1973-1990 CPS ASEC also collected highest grade completed. The 2000 Census and the 2001-2013 ACS, however, collected education attainment in categories. We construct a years of schooling variable for the 2000 Census and ACS that is consistent across all data sources from the education categories.

Measurement error in reported income is a well-documented problem in intergenerational analyses (Mazumder 2005 and Mazumder, Nybom and Stuhler 2016, Haider and Solon 2006, Böhlmark

[7]Note that the record linkage process did not allow a 5-year interval around age (see the data appendix for more detail). In practice, this is an independent check of the matching process because each data source is linked independently to the PVS reference file to receive a PIK, which we then use to merge the data by PIK.

[8]Results restricted to children in the Parent generation for whom at least one grandparent in the Great-grandparent generation is observed are shown in Table A.1.

and Lindquist 2006, Grawe 2006). In the absence of multiple observations of income or administrative earnings data, education attainment may provide mobility estimates less biased by measurement error. Unlike data sources used in other educational mobility studies, we can compare education of the same individual reported at two points in time to test whether this assumption holds.

We have data on education attainment from three time periods: 1940, 1973-1990, and 2000-2013. Through record linkage, we observe education attainment at two points at time for a large number of individuals from the Grandparent, Parent, and Child generations. This allows us to examine whether education reporting in 1940 was measured with significantly greater error relative to later years and to assess differences in education reporting over time more generally.[9] We report differences in observed education attainment from a sample of 25-55 year olds in 1940 linked to the CPS, a sample of 25-55 year olds in the CPS linked forward to 2000-2013, and a sample of 25-55 year in the 2000 Census or 2001-2007 ACS linked to the 2007-2013 ACS in Figure 2.[10]

The top row of Figure 2 shows the distribution of the difference between reported years of schooling at time $t + n$ and time t. For each linked sample, the difference in reported years of schooling appears tightly centered around zero. The lower row of figures shows the average difference between reported years of schooling (education at $t + n$ minus education at t) for each grade level at time t plotted against the distribution of years of schooling at time t. This line shows that differences in reported education are greatest for those with the least education. However, only a small portion of the education distribution falls within the lowest levels of education attainment. For the majority of grade levels, the difference in reported years of schooling between time t and $t + n$ is within one year. To more formally test whether the discrepancies in reported education across surveys behave non-classically, we regress education at time $t+n$ on education at time t and several interaction terms to account for age, sex, race, and socioeconomic status.

[[Insert Figure 2 Here]]

Table 1 reports the regression of education attainment at time $t + n$ on education attainment

[9] Goldin (1998) and Goldin and Katz (2000) argue that educational attainment in 1940 may be reported with error due to non-standardized education systems within and across states, the eventual standardization of middle school and high school that occurred from 1910 to 1940, and potential misunderstandings about answering the education attainment questions for the first time in a census.

[10] We require that 6 years pass between the first observation of education attainment and the second observation of education attainment for the 2000-2013 linked samples.

at time t. The bivariate regressions in Columns 4, 6, and 8 reveal high levels of correlation between reported education at time t and $t+n$. When we add covariates to the regression of highest grade completed from the CPS on highest grade in the 1940 Census in Column 5, sex appears to be the greatest predictor of differences in reported education. Men underreported their education by a third of a year in 1940 on average and there is a small, statistically significant coefficient on the interaction of the male dummy variable and education attainment, suggesting men with higher education levels in 1940 tend to report more similar levels of education attainment in the CPS. Otherwise, equivalent reporting of education attainment in the CPS and 1940 Census appears unrelated to age, race, and occupational prestige.

Columns 6 and 7 of Table 1 show the regression results from the CPS linked forward to the 2000 Census and 2001-2013 ACS. For these cases, men with more schooling in the CPS are more likely to report the same level of education attainment in the CPS and black and other-race respondents with more education are less likely to report the same level of education attainment in the CPS. These coefficients, although statistically significant, are small in magnitude. The regressions in Columns 8 and 9 exhibit similar patterns with the addition of income and age as statistically significant predictors of education attainment at time $t+n$. Although the interaction terms suggest some degree of non-classical error (at least in observables) in the reporting of education, the small magnitudes of these coefficients suggest small changes in the gap between education reported at time t and $t+n$, even for large differences in income.

[[Insert Table 1 Here]]

Differences between reported education across surveys may result from proxy response, errors in recall, differences in questionnaire design over time, respondents acquiring more education as adults, general confusion over how to respond, or from linkage error. It is impossible to disentangle the sources of error and each dataset may be affected by one source of error more than another. The 1940 Census, for instance, was the first census to collect education information and asked for the "highest grade of school completed."[11] The convention of "common" schools (e.g., grammar

[11]Respondents who never attended school or whose highest year of schooling was from 1-8 had that number reported, those who attended high school had their highest high school grade reported preceded by the letter "H" (so a respondent who attended high school only through sophomore year was recorded as "H2"), and those who attended post-secondary school had their highest post-secondary grade reported preceded by the letter "C" (so a respondent who attended college only through junior year was recorded as "C3"). The highest grade reported was "C5" for respondents who attained any years of schooling following four years in college.

school instead of graded schools) may result in misreported completion of secondary school as respondents may not have known how to correctly translate their own education experiences into grade numbers (Goldin and Katz, 2000).[12] The vast expansion and standardization of education in the US between 1910 and 1940 may improve our measure of education attainment for young adults observed in 1940 (Goldin, 1998), but our linked data show that, even within the Grandparent generation (born between 1885 and 1918), there is inaccurate reporting despite standardization of the education system.

Our results in Table 1 dispel some of the concerns surrounding accuracy of education reporting in the 1940 Census. Although there appear to be greater differences between reported education in the CPS and the 1940 Census relative to the CPS-LF/ACS linked sample and the LF/ACS-ACS linked sample, age in 1940 is not a significant predictor of education in the CPS. This suggests that Grandparents who were deciding whether to enter high school in 1899 were not any less consistent in their later-life education responses in the CPS as Grandparents entering high school in 1932 – in spite of the standardization of middle school and high school that began in 1910. We further examine the idea that adults in 1940 may better recollect their education attainment or that the translation between grade school and years of schooling became more concrete over time in Figure 3. This figure shows average education attainment by year of birth and Census source for all individuals (native and foreign-born) under the age of 85. Generally, average education attainment from the 1950-1970 censuses are within half a year of the average from 1940. Thus, recollection is not a great source of error in education reporting.

The discrepancies may arise, in part, from the transition from collection of highest grade completed towards the collection of education categories beginning in 2000. Categories that capture education milestones (e.g., graduation from college or high school) may be easier for survey respondents to recall. In the 2008 ACS, the number of categories changed, allowing us to see if reporting differences were smaller for years when the categories remained unchanged. If we limit the regressions in Columns 8 and 9 to the years 2000-2007, the regression results are nearly identical. This implies that even when the survey questions are identically measured and the years between recollection are fewer, we can still expect small differences in reported education attainment over

[12]The enumerator instructions provided some guidance: "For a person who completed his [sic] formal education in an ungraded school or a foreign country, enter the approximate equivalent grade in the American school system, or, if this cannot readily be determined, the number of years the persons attended school."

time for the 25-55 year old population. We explore how measurement error affects our estimates of mobility in the Results section.

[[Insert Figure 3 Here]]

2.3 Age, Education, and Sample Selection

Inclusion into the linked samples requires observing a parent and child living together and successfully linking that child to their adult observation in another data source. If we are more likely to link certain types of records, these requirements may bias the sample. We show the summary statistics for our two three-generation samples in Tables 2 and 3. In both cases, there are statistically significant differences between the full samples of adults in the surveys and those selected into our baseline regression samples that results from our ability to match across generations. For example, in Table 3, individuals in the regression sample in the Child generation are more likely to have a college degree and more likely to have completed high school. However, given the changes in the US population over the period, especially from immigration, it is not clear that large differences between the regression and full samples are a concern. What we would like is a counterfactual full sample of children whose grandparents resided in the US, which is not available.

Instead, we focus on the differences between the regression and full samples at the Grandparent generation, as in both cases inclusion in the regression is conditional on matching between at least the Parent and Grandparent generations. In the Great-grandparent-Grandparent-Parent generation sample in Table 2 we see that 1940 individuals in the regression sample are younger on average by 3.5 years than the full sample. This is likely because we are conditioning on matching with their children observed as late as in the 1990 CPS, 50 years later. However, the educational attainment numbers differ only slightly between the samples. Taking the difference between the regression and full samples, age is statistically significant different (-3.5 years), as are years of schooling (0.37 years), share with some college education (-0.3 percentage points), and share with a graduate degree (0.2 percentage points). There are also statistically significant differences between the full and regression samples of the Parent generation. However, as before we do not observe the counterfactual Parent generation whose parents (Grandparent generation) resided in the US in 1940.

[[Insert Table 2 Here]]

[[Insert Table 3 Here]]

We see the same pattern in the Grandparent generation in the Grandparent-Parent-Child generation sample in Table 3. Comparing the regression and full samples, we find statistically significant differences of comparable magnitudes as above: age (-3.7 years), years of schooling (0.44 years), and the share with less than high school education (-2.2 percentage points), high school education (1.1 percentage point), some college education (-0.3 percentage points), and graduate education (0.2 percentage points). Tables 2-3 suggest a small degree of sample selection. However, education attainment is similar across the full and regression samples for the oldest generation in each intergenerational and multigenerational sample, which is the only generation for which we observe the counterfactual population. We further examine sample selection in the Results section.

3 Empirical Specification

We regress the years of schooling of the child on the highest observed years of schooling in each ancestor generation, which can include parents, grandparents, and great-grandparents.[13] In each case, we include fixed effects for the survey year that each generation is observed, y_t (for $t \in 1973, ..., 2013$),[14] along with age and age-squared terms for the child and age and age-squared of the parent in each generation, X_i.[15] To measure mobility across three generations, we regress the childs years of schooling on the maximum observed years of schooling for both the child's parents, $g-1$, and grandparents, $g-2$:

$$Y_{ig} = \alpha + \beta_1 Y_{ig-1} + \beta_2 Y_{ig-2} + \delta X_i + \gamma_t + \tau_{g-1,g-2} + \epsilon_{ig} \tag{1}$$

As the number of observed grandparents varies across children, we also include fixed effects for whether we observe each of the four possible grandparents, $\tau_{g-1,g-2}$. This controls for differences in the grandparent-child relationship that may differ across grandparents. For example, the years

[13] In Table A.3, we also show results using different measures of education in the prior generation (e.g. average education, the education of each ancestor observed). Our choice of measure does not change the results we present.

[14] This is only relevant for the parent and child generations, which are observed across multiple surveys.

[15] The age terms are included to capture any education increases that may occur in adulthood as well as differences across cohorts in education.

of schooling of the maternal grandmother may be more or less correlated with child education than the years of schooling of the paternal grandfather. We do the same for our four-generation sample, with fixed effects for each of the eight possible great-grandparents:

$$Y_{ig} = \alpha + \beta_1 Y_{ig-1} + \beta_2 Y_{ig-2} + \beta_3 Y_{ig-3} + \delta X_i + \gamma_t + \tau_{g-1,g-2,g-3} + \epsilon_{ig} \tag{2}$$

In all the regressions, for each child generation g, the errors are clustered at the $g-1$ generation level. We require the child to be at least 25 years old when observed as an adult and the highest-educated spouse in each ancestor generation must be between 25 and 55 years old when they are observed as adults in the prior generations' survey.

We also report correlation coefficients for each regression. Figure 4 shows how the distribution of education attainment has changed from the Great-grandparent generation to the Child generation. Because the variance of education has decreased over time, our OLS regression results are scaled by the relative standard deviations of the different generations in each regression. For example, the coefficient on the parent's education is:

$$\beta_1 = \rho_{PC} \frac{\text{var}(y_C)^{1/2}}{\text{var}(y_P)^{1/2}} \tag{3}$$

[[Insert Figure 4 Here]]

To disentangle changes in the correlation from changes in the variance, we normalize years of schooling to have mean 0 and variance of 1 in each generation such that the correlation regression coefficient $\beta_1 = \rho_{g-1,g}$.

4 Results

4.1 Multigenerational Mobility of Education

Our data allows us to compare multigenerational mobility across two three-generation samples. With three generations, we can determine whether grandparent education predicts child education conditional on parent outcomes, which we call a grandparent effect. Before we move to the third generation and consideration of the extent to which ignoring it biases our perceptions of mobility,

we present baseline results on two-generation mobility in Table 4. We report two-generational results beginning with the grandparent generation because this is the first generation that we can observe reporting error in education. There are five other sources to which these two-generation results can be compared: (1) Couch and Dunn (1997) use the 1984 and 1988 PSID and report intergenerational correlations between 0.40 and 0.43, compared to our 0.42 to 0.44 in Panel B; (2) Hertz et al. (2008) who use data from the 2000 International Social Survey Programme (ISSP) and focus on adults age 20-69 and find an intergenerational correlation of 0.46 compared to our 0.42 to 0.44 in Panel B; (3) Hilger (2015) who uses a very different methodology and focuses on children when they are age 26-29, finding regression coefficients that range from 0.37 to 0.40 for whites and 0.24 to 0.40 for blacks for 1960-2000, compared to our 0.36 for both races combined in Column 2 of Panel A; (4) the Occupational Change in a Generation 1973 cohort (OCG73) for males born 1925-40 which yields a correlation of 0.47 compared to our 0.44 in Column 1 of Panel B; and (5) the General Social Survey (GSS) which yields a correlation of 0.43 for children born 1925-40 compared to our 0.44 in Column 1 of Panel B and yields a correlation of 0.50 for children born 1960-85 compared to our 0.42 in Column 2 of Panel B. With the exception of the GSS 1965-85 birth cohorts, then, our two-generation results are broadly similar to those obtained in other studies and with other samples and methodologies.[16]

[[Insert Table 4 Here]]

There are two shortcomings in all the work to which we have just compared our two-generation results: (1) they cannot account for measurement error; and (2) they cannot account for the effect of generations prior to the parents. We have already shown (Table 1) that educational attainment is reported with error. We now turn to three-generation mobility and report the grandparent-parent-child regression and correlation coefficients in Panels A and B of Table 5. The grandparent coefficient is statistically significant in both samples for both the regression and correlation coefficients. In other words, conditional on parent education, grandparents matter in predicting the educational outcomes of children.

[16] Hilger (2015) examines individuals age 26-29 and still co-resident with their parents (with adjustment for the experience of their non-co-resident siblings). The ISSP, OCG73, and GSS asked respondents at the time they were surveyed to report both their own education and the education of their parents. The comparison study closest in design to ours is the PSID which actually recorded parents' education when their children were young and then recorded the education of those children when they were themselves adults.

We add great-grandparents to our three-generation sample to create a four-generation sample spanning data 1910-2013 and individuals born between 1885 and 1988. In our sample of 10,890 children with matched parents and grandparents, we are able to match 1,444 to at least one of their eight possible great-grandparents. We test if great-grandparent education is associated with child education, conditional on parent and grandparent education, shown in Table 5, as well. We do not find a statistically significant relationship between education of the Great-grandparent and Child when controlling for both the Grandparent and Parent.

These results suggest grandparents directly influence their grandchildren beyond their indirect effect through the parents while great-grandparents have no discernable effect. Several studies also find independent grandparent effects in non-US countries using a variety of mobility measures (Lindahl et al. 2015; Zeng and Xie 2014; Hertel and Groh-Samberg 2014; Chan and Boliver 2013; Boserup et al. 2013), though the influence of grandparents is generally larger outside the US. These findings imply that an AR(1) process does not fully describe mobility in the twentieth century and undermine conclusions drawn about multigenerational mobility from two-generation samples.

[[Insert Table 5 Here]]

As the regression coefficients suggest that an AR(2) model better describes the data, we can calculate how assuming an AR(1) model understates the persistence of advantage or disadvantage over the long term. To do that we calculate the AR(1) coefficient that would result in the same mobility over 10 generations. With a parent-child correlation of 0.416 and an additional grandparent effect of 0.06, advantage persists as if the parent-child correlation were 0.529.[17] In other words, the true underlying persistence implied by the multigenerational mobility regression is 26 percent higher than would be estimated with only two generations of data. Suppose a child's grandparents and parents have years of schooling one standard deviation above the mean. Under the AR(2) process, an average of 23 percent more of that advantage persists for the child, 14 percent more persists for grandchildren, 9 percent more for great-grandchildren, and 5 percent more for great-great-grandchildren.

However, it is important to emphasize that the grandparent relationship may not be causal and could spuriously arise as a result of measurement error, omitted group effects, cultural inheritance,

[17] We chose to calculate the AR(1) equivalent at 10 generations to abstract away from the importance of the education level of the initial child's grandparents.

or other unobserved factors (Solon, 2015). Using linked data, we can explore how measurement error affects our estimates.

4.2 Measurement Error

In Table 1, we find evidence of plausibly classical measurement error. Under classical measurement error, the observed regression coefficient is attenuated by the magnitude of the variance of the measurement error (v) relative to the variance of the true underlying independent variable (y^*), or:

$$\hat{\beta} = \beta \left(\frac{1}{1 + \sigma_v^2 / \sigma_{y^*}^2} \right). \tag{4}$$

If we assume education for the same individual in two surveys are both observed with error, the regression of one observation on the other is:

$$\left(y_{it+1}^* + v_{it+1}\right) = \beta \left(y_{it}^* + v_{it}\right) + \eta_{it+1}. \tag{5}$$

If $E(v_{it}, v_{it+1}) = 0$, then $\hat{\beta}$ from this regression identifies $\sigma_v^2 / \sigma_{y^*}^2$ because the true β from a regression of a variable on itself is 1.[18]

However, because we are linking individuals across cross-sectional surveys, another potential source of measurement error is mis-linking. If we have incorrectly assigned a PIK to a subset of individuals in either survey, then the regression coefficients will be attenuated. With the assumption that incorrect PIK assignment is independent across surveys,[19] the attenuation from the regression of education observed in period 1, y_1, on education observed in period 2, y_2 is:

$$\hat{\beta} = (1 - m_1)(1 - m_2)\beta, \tag{6}$$

given m_1 and m_2 as the share of individuals with incorrect PIK assignments in periods 1 and

[18] Or approximately 1 in this case, as individuals could gain education as they age. We assume that this change is small, as evidenced by the differences in reported education across observations for the same individuals shown in Table 1. The assumption that $E(v_{it}, v_{it+1}) = 0$ is a strong one. Alternatively, we could model misreporting as $v_{it} = c_i + u_{it}$ where c_i is the persistent misreporting of education of individual i and u_{it} is the transitory measurement error in survey year t. In this case, we would be estimating $\sigma_u^2 / \sigma_{y^*+c}^2$ from a regression of $y_{it+1}^* + c_i$ on $y_{it}^* + c_i + u_{it}$. The possibility of persistent education misreporting is an interesting research question that unfortunately cannot be answered in the absence of a benchmark data source on educational attainment.

[19] This is a strong assumption. However, if incorrect assignment is correlated across surveys, then the attenuation from mis-linking is smaller as there is no additional attenuation from multiple incorrect assignments.

2 respectively. Given the assumption that $\beta \approx 1$, with both misreporting and mis-linking, the observed coefficient is:

$$\hat{\beta} = (1 - m_1)(1 - m_2) \left(\frac{1}{1 + \sigma_v^2 / \sigma_{y^*}^2} \right). \tag{7}$$

The same attenuation is present for regression estimates of intergenerational mobility coefficients. From the parent-child regression with true parameter β, with measurement error from misreporting and incorrect PIK assignment rates of m_P and m_c in the parent and child data sets,[20] the estimated coefficient is the same as in 7, but with m_P and m_C in place of m_1 and m_2, respectively and $\sigma_v^2 / \sigma_{y_P^*}^2$ in the denominator of the fraction.

However, for the grandparent coefficient in multigenerational mobility regressions, mis-assignment of PIKs and classical measurement error result in biases with opposite signs. Under our linking strategy, a three-generation link of the Grandparent, Parent, and Child generations requires four links to PIKs to successfully identify the correct individuals in each generation: 1) the Parent generation linked as a child in the 1940 census, 2) the Parent generation linked as an adult in the CPS, 3) the Child generation linked as a child in the CPS and 4) the Child generation linked as an adult in the 2000 census and 2001-2013 ACS.

Assuming independent link probabilities within families, this implies there are three relevant mis-linking probabilities, m_G, m_P, and m_C, for the Grandparent, Parent, and Child generation data sets respectively. For simplicity, let the variance be constant and equal to 1 across all generations. Further, let the true intergenerational correlations across pairs of generations be ρ_{GP} and ρ_{PC} for the Grandparent-Parent and Parent-Child generations respectively. Finally, let $\rho_{GC} = \rho_{GP}\rho_{PC} + \rho_{GC}^I$.

In the absence of measurement error, the two-generation regression coefficients would equal the corresponding $\rho_{g-1,g}$ parameter and the multigenerational coefficient for grandparents would be $\beta_G = \rho_{GC}^I$.[21] However, with both classical measurement error and mis-linking, the observed $\hat{\beta}_G$ is:

$$\hat{\beta}_G = \frac{(1 - m_G)(1 - m_P)^2(1 - m_C)\left[(1 + \sigma_{vP}^2)\rho_{GC}^I + \sigma_{vP}^2 \rho_{GP}\rho_{PC}\right]}{(1 + \sigma_{vP}^2)(1 + \sigma_{vG}^2) - (1 - m_G)^2(1 - m_P)^2 \rho_{GP}^2}. \tag{8}$$

[20] In each case, the relevant mistaken PIK is for the child. Observing the correct parent education level is conditional on assigning the correct PIK to the child in the parent data set. Observing the correct child education level as an adult is conditional on correctly assigning the PIK to the child in the child data set.

[21] As before let g index the generation in the family line, so that $g = G, P, C$ and $g - 1$ represents the prior generation. If $g = C$, then $g - 1 = P$.

If we further assume that there is not independent grandparent effect on children, or that $\rho_{GC}^{I} = 0$ and $\rho_{GC} = \rho_{GP}\rho_{PC}$, then:

$$\hat{\beta}_{G}^{NoG} = \frac{(1-m_G)(1-m_P)^2(1-m_C)\sigma_{vP}^2\rho_{GP}\rho_{PC}}{(1+\sigma_{vP}^2)(1+\sigma_{vG}^2) - (1-m_G)^2(1-m_P)^2\rho_{GP}^2}. \tag{9}$$

In this case, there is a spurious grandparent effect in the regression coefficient if there is measurement error in the parent generation ($\sigma_{vP}^2 > 0$). This is attenuated by linkage error because for children with incorrect links to their parents and grandparents, it does not matter if there is measurement error in each ancestor generation's reported education as there is no correlation regardless.

4.3 Mobility Estimates with Measurement Error Correction

From the measurement error estimates in Table 1 and Equation 7, we can construct bounds on the amount of measurement error, whether through misreporting or linkage error, in the data. For example, for individuals linked across surveys between the 2000 Long Form Census and ACS to a later ACS, the coefficient for years of schooling in the later survey regressed on years of schooling in the earlier survey is 0.860. If all of the attenuation were due to measurement error, then $\sigma_v^2/\sigma_{y^*}^2 = 0.163$. If, on the other hand, all of the attenuation were due to linkage error, then the implied linkage error rate in the 2000 Long Form and ACS would be 7.3 percent. We can do the same calculations for the measurement in the 1940 Census and the CPS ASECs used in this paper. In the data set for each generation, we can define α_g as the share of the measurement error due to linkage error and then derive the implied misreporting error. In the above examples for the Child Generation (2000 Long Form and ACS), we assumed $\alpha_C = 0$ and $\alpha_C = 1$.

We define a series of measurement error scenarios (values for each α_g) and derive the implied misreporting and linkage error parameters. With a given set of parameters for σ_{vg}^2 and m_g, we can calculate the "true" intergenerational correlations ρ_{GP} and ρ_{PC} that are consistent with the observed two-generation mobility regression results. In the case of all misreporting ($\alpha_g = 0, g = G, P, C$), with Parent-Child years-of-schooling correlation coefficient from Table 4 of 0.420, the implied true ρ_{PC} is 0.499. The implied true Grandparent-Parent years-of-schooling correlation is

0.542.[22]

We can then calculate the multigenerational mobility grandparent regression coefficient we would observe with no true grandparent effect ($\rho^I_{GC} = 0$) as in Equation 9. In the misreporting only case, $\hat{\beta}^{NoG}_G = 0.043$, which is within the 95 percent confidence interval of the observed grandparent coefficient of 0.06.

We can solve for the implied ρ^I_{GC} given the observed grandparent coefficient in the multigenerational mobility regression using Equation 8. In the misreporting only case, $\rho^I_{GC} = 0.012$. Finally, we can calculate the persistence given an AR(2) model and the derived ρ_{PC} and ρ^I_{GC}. With misreporting only, the equivalent AR(1) model to the case with $\rho_{PC} = 0.499$ and $\rho^I_{GC} = 0.012$ is 0.526, 25 percent higher than the parent-child regression coefficient of 0.420.

In Table 6, we show three measurement error scenarios. Scenario A assumes all measurement error is due to misreporting, with no linkage error as describe above. Scenario B uses estimated upper bounds on the linkage error in each data set from validation studies (see the Appendix). In Scenario C, we hold misreporting fixed across all surveys given the linkage error rate in the 2000 Long Form and ACS.[23] In each scenario, the measurement error parameters are shown at the top with the results below.

As misreporting and linkage error have the same effect on two-generation regression coefficients, the true two-generation correlations are the same in all three scenarios, with $\rho_{GP} = 0.542$ and $\rho_{PC} = 0.499$. This implies that our intergenerational mobility regressions understate the true two-generational persistence by 24 percent between the Grandparent-Parent sample and 19 percent for the Parent-Child sample.

Additionally, in all three cases, the implied true ρ^I_{GC} is considerably smaller than the value observed in the data, at 0.012, 0.034, and 0.027 in scenarios A, B, and C. Putting the error-corrected estimates for ρ_{PC} and ρ^I_{GC} into an AR(2) model, we find that the equivalent AR(1) persistence varies from 0.526 in scenario A to 0.567 in scenario B. This is an increase in persistence of 25 to 35 percent over the two-generational regression estimate of 0.420.

Whether we assume that grandparents matter from the multigenerational mobility regression or that measurement error biases the grandparent coefficient up by as much as 500 percent, we

[22] $0.499 = 0.420(1 + \sigma^2_{vP})$ where $\sigma^2_{vP} = 0.189$, and $0.547 = 0.437(1 + \sigma^2_{vG})$ where $\sigma^2_{vG} = 0.244$.

[23] We use the linkage error from the 2000 Long Form and ACS as the linkage process and error rate has been more thoroughly studied in more recent data, as in Layne et al. (2014).

estimate that the true persistence of advantage is at least 25 percent higher than when estimated with two generations of data. However, in each of the measurement-error scenarios, more than half of that additional persistence can be accounted for by the underestimate of the parent-child correlation as opposed to the additional persistence associated with grandparents.

4.4 Selection into the Multigenerational Samples

We also indirectly test for non-random selection into the regression samples by comparing the two-generation regression coefficients from the three-generation samples to those estimated from the two-generation samples in Ferrie et al. (2016). The assumption is that if selection into the matched regression sample is biasing our results, that bias should be more severe when conditioning on selection into a sample of three generations rather than two.

Figure 5 plots the average child years of schooling for each parent year of schooling across each set of two generations in each of our two- and three-, and four-generation samples. Panels A, B, and C compare the years of schooling gradients between the Great-grandparent and Grandparent generations estimated from three samples, in Panel A: the four-generation Great-grandparent to Child sample, Panel B: the three-generation Great-grandparent to Parent sample, and Panel C: the two-generation Great-grandparent to Grandparent sample. Panels D and E compare the years of schooling gradients between the Grandparent and Parent generations estimated from two samples, in Panel D: the four-generation Great-grandparent to Child sample and Panel E: the three-generation Great-grandparent to Parent sample. Panels F and G compare the years of schooling gradients between the Parent and Child generations estimated from two samples, in Panel F: the four-generation Great-grandparent to Child sample and Panel G: the two-generation Parent to Child sample.[24] We do not see evidence of selection in the figure.

[[Insert Figure 5 Here]]

We formally test the different regression and correlation coefficients in Table 7. Models (1) and (2) match Panels A-C above comparing mobility from the Great-grandparent to Grandparent

[24] A three-generation Grandparent to Child sample is equivalent to the four-generation Great-grandparent to Child sample as inclusion in the four-generation sample does not require a successful match to the Great-grandparent generation. For the same reason, the three-generation Great-grandparent to Parent sample is equivalent to a two-generation Grandparent to Parent sample.

generations, and neither the regression nor the correlation coefficients are statistically significantly different from each other. Models (3) and (4) make the comparison of Panels D and E above from the Grandparent to Parent generations, again with no statistically significant differences. Models (5) and (6) make the comparison of Panels F and G above from the Parent to Child generations. In this case, the regression coefficients, but not the correlation coefficients, are statistically different.

[[Insert Table 7 Here]]

We interpret the difference in summary stats between our regression and full samples in Tables 2-3 as evidence of some selection into our regression samples. However, the bias seems relatively small when comparing the regression results in Table 7, which should be differentially affected by selection into the two- vs. three-generation samples.

5 Conclusion

As inequality in the US continues to increase, so does the importance of pinpointing its roots. An important factor underlying inequality is its persistence across generations. Much work by economists and sociologists examines the persistence of social status across two generations from parent to child. We extended this two-generational focus to include an analysis of educational mobility across three and four generations of families in the US.

We examined multigenerational mobility in educational attainment by linking families across multiple data sources. These sources include the 1910 Census, the 1920 Census, the 1940 Census, the CPS-ASEC spanning 1973-1990, the 2000 Census and the 2001-2013 ACS. From these linkages, we constructed two grandparent-parent-child samples and one four-generation sample including great-grandparents observed in the 1910 and 1920 censuses.

Our multigenerational analysis found a small grandparent effect and no evidence of a great-grandparent effect. Even a small independent grandparent-grandchild relationship can result in considerably slower convergence to the mean over the long-term for individuals from advantaged or disadvantaged educational backgrounds than is indicated by intergenerational education regressions with only two generations. This finding was robust to inclusion of multiple grandparents and both parents, as well as for matriarchal and patriarchal lines followed and analyzed separately.

Unlike other educational mobility research, we were able to evaluate measurement error in reported educational attainment in each generation. We found that many individuals do not consistently report education across surveys and that inconsistent reporting was greatest for our earliest source of education data, the 1940 Census. We further showed that misreported educational attainment – in the magnitudes we found – could explain half or more of the grandparent effect. Although this finding could be interpreted as support for the literature's focus on parent-child transmission when studying mobility of education in the US, the same measurement error that spuriously caused a positive grandparent effect also results in two-generation estimates that underestimate persistence by 25 percent. Together with the small grandparent effect, we estimate that persistence is understated by 25 or more with only two generations of data. This underestimate is of a similar magnitude whether the grandparent effect we find in our three generation is primarily due to measurement error or we assume it to be a true causal effect.

References

Alexander, J. T., T. Gardner, C. G. Massey, and A. O'Hara (2014): "Creating a Longitudinal Data Infrastructure at the Census Bureau," `http://paa2015.princeton.edu/uploads/152688`.

Arias, E. (2015): "United States Life Tables, 2011," *National Vital Statistics Report*, 64, 1–62.

Behrman, J. and P. Taubman (1985): "Intergenerational earnings mobility in the United States: some estimates and a test of Becker's intergenerational endowments model," *The Review of Economics and Statistics*, 144–151.

Björklund, A. and M. Jäntti (1997): "Intergenerational income mobility in Sweden compared to the United States," *The American Economic Review*, 87, 1009–1018.

Böhlmark, A. and M. J. Lindquist (2006): "Life-cycle variations in the association between current and lifetime income: Replication and extension for Sweden," *Journal of Labor Economics*, 24, 879–896.

Boserup, S. H., W. Kopczuk, and C. T. Kreiner (2013): "Intergenerational wealth mobility: Evidence from danish wealth records of three generations," *Univ. if Copenhagen mimeo*.

Bourdieu, J., J. P. Ferrie, and L. Kesztenbaum (2009): "Vive la différence? Intergenerational mobility in France and the United States during the nineteenth and twentieth centuries," *Journal of Interdisciplinary History*, 39, 523–557.

Celhay, P. and S. Gallegos (2015): "Persistence in the transmission of education: evidence across three generations for Chile," *Journal of Human Development and Capabilities*, 16, 420–451.

Chan, T. W. and V. Boliver (2013): "Social mobility over three generations in Britain," *American Sociological Review*, 78, 662–678.

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," *The American Economic Review*, 104, 141–147.

Clark, G. (2014): *The son also rises: surnames and the history of social mobility*, Princeton University Press.

Couch, K. A. and T. A. Dunn (1997): "Intergenerational correlations in labor market status: A comparison of the United States and Germany," *Journal of Human Resources*, 210–232.

Feigenbaum, J. J. (2014): "A new old measure of intergenerational mobility: Iowa 1915 to 1940," *Unpublished Manuscript*.

——— (2015): "Intergenerational Mobility during the Great Depression," *Unpublished working paper*.

Ferrie, J., C. Massey, and J. Rothbaum (2016): "Changes in U.S. Intergenerational Mobility in Educational Attainment, 1895-2013," *Unpublished working paper*.

Goeken, R., L. Huynh, T. Lynch, and R. Vick (2011): "New methods of census record linking," *Historical methods*, 44, 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, 345–374.

Goldin, C. and L. F. Katz (2000): "Education and income in the early twentieth century: Evidence from the prairies," *The Journal of Economic History*, 60, 782–818.

Grawe, N. D. (2006): "Lifecycle bias in estimates of intergenerational earnings persistence," *Labour economics*, 13, 551–570.

Grusky, D. B., T. M. Smeeding, C. M. Snipp, D. B. Grusky, T. M. Smeeding, and C. M. Snipp (2015): "A new infrastructure for monitoring social mobility in the United States," *The ANNALS of the American Academy of Political and Social Science*, 657, 63–82.

Haider, S. and G. Solon (2006): "Life-cycle variation in the association between current and lifetime earnings," *The American Economic Review*, 96, 1308–1320.

Hertel, F. R. and O. Groh-Samberg (2014): "Class mobility across three generations in the US and Germany," *Research in Social Stratification and Mobility*, 35, 35–52.

Hertz, T., T. Jayasundera, P. Piraino, S. Selcuk, N. Smith, and A. Verashchagina (2008): "The inheritance of educational inequality: International comparisons and fifty-year trends," *The BE Journal of Economic Analysis & Policy*, 7, 10.

Hilger, N. G. (2015): "The Great Escape: Intergenerational Mobility in the United States Since 1940," *National Bureau of Economic Research*.

Hodge, R. W. (1966): "Occupational mobility as a probability process," *Demography*, 3, 19–34.

Layne, M., D. Wagner, and C. Rothhaas (2014): "Estimating record linkage false match rate for the Person Identification Validation System," *Center for Administrative Records Research and Applications Working Paper*, 2.

Lindahl, M., M. Palme, S. S. Massih, and A. Sjögren (2015): "Long-Term Intergenerational Persistence of Human Capital An Empirical Analysis of Four Generations," *Journal of Human Resources*, 50, 1–33.

Lleras-Muney, A. (2005): "The relationship between education and adult mortality in the United States," *The Review of Economic Studies*, 72, 189–221.

Long, J. and J. Ferrie (2012): "Grandfathers matter (ed): Occupational mobility across three generations in the US and Britain, 1850-1910," *Unpublished manuscript*.

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

Mare, R. D. (2011): "A multigenerational view of inequality," *Demography*, 48, 1–23.

Massey, C. G. (2017): "Playing with matches: An assessment of accuracy in linked historical data," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 1–15.

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, 235–255.

——— (2016): "Estimating the Intergenerational Elasticity and Rank Association in the United States: Overcoming the Current Limitations of Tax Data," in *Inequality: Causes and Consequences*, Emerald Group Publishing Limited, 83–129.

Michelson, M. and C. A. Knoblock (2006): "Learning blocking schemes for record linkage," in *AAAI*, 440–445.

Nybom, M. and J. Stuhler (2016): "Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation," *Journal of Human Resources*, 51, 239–268.

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

Peters, H. E. (1992): "Patterns of intergenerational mobility in income and earnings," *The Review of Economics and Statistics*, 456–466.

Piketty, T. and E. Saez (2014): "Inequality in the long run," *Science*, 344, 838–843.

Solon, G. (1992): "Intergenerational income mobility in the United States," *The American Economic Review*, 393–408.

——— (2002): "Cross-country differences in intergenerational earnings mobility," *The Journal of Economic Perspectives*, 16, 59–66.

——— (2014): "Theoretical models of inequality transmission across multiple generations," *Research in Social Stratification and Mobility*, 35, 13–18.

——— (2015): "What do we know so far about multigenerational mobility?" *National Bureau of Economic Research*.

Stuhler, J. (2012): "Mobility across multiple generations: The iterated regression fallacy," *IZA Discussion Paper No. 7072*.

Sweeney, L. (2000): "Simple demographics often identify people uniquely," *Health (San Francisco)*, 671, 1–34.

Wagner, D. and M. Layne (2014): "The Person Identification Validation System: Applying the Center for Administrative Records and Research and Applications' Record Linkage Software." .

Warren, J. R. and R. M. Hauser (1997): "Social stratification across three generations: New evidence from the Wisconsin Longitudinal Study," *American Sociological Review*, 561–572.

Winkler, W. E. (1995): "Matching and record linkage," *Business survey methods*, 1, 355–384.

Zeng, Z. and Y. Xie (2014): "The effects of grandparents on children's schooling: Evidence from rural China," *Demography*, 51, 599–617.

Table 1: Differences in Reported Education in Linked Data

	Summary Statistics			Regressions					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	1940-CPS	CPS-LF/ACS	LF/ACS-LF/ACS	1940-CPS	1940-CPS	CPS-LF/ACS	CPS-LF/ACS	LF/ACS-ACS	LF/ACS-ACS
1940 Census Education	10.08			0.804***	0.983***				
	(3.40)			(0.006)	(0.176)				
CPS Highest Grade	9.90	12.97				0.841***	0.735***		
	(3.12)	(2.63)				(0.003)	(0.068)		
LF/ACS Highest Grade		12.68	13.43					0.860***	0.689***
		(2.64)	(2.32)					(0.0002)	(0.0057)
ACS Highest Grade (t+1)			13.31						
			(2.36)						
Age	31.75	37.51	41.01		0.0265		-0.067		-0.0467***
	(5.87)	(8.49)	-8.46		(0.106)		(0.047)		(0.0039)
Age Squared	1042.72	1479.04	1753.04		-0.000583		0.00016		0.00014***
	(408.74)	(666.52)	-686.63		(0.00152)		(0.0006)		(4.73e-05)
Male	0.42	0.46	0.48		-0.818***		-0.522***		0.0865***
	(0.49)	(0.50)	-0.5		(0.178)		(0.093)		(0.0081)
White	0.95	0.91	0.87						
	(0.22)	(0.28)	(0.34)						
Black	0.05	0.06	0.06						
	(0.21)	(0.24)	(0.24)						
Other Race	0.00	0.03	0.07						
	(0.06)	(0.16)	(0.26)						
Occupation Score	13.18				0.0168***				
	(13.61)				(0.006)				
Total Personal Income		16056	35,746				0.0313*		-0.152***
(Logged in Regression)		(16,831)	(49,900)				-0.0176		(0.001)
Age*Edu					-0.011		0.0023		0.00056*
					(0.010)		(0.0035)		(0.00028)
Age Squared*Edu					0.00014		3.88e-06		1.72e-05***
					(0.00015)		(4.43e-05)		(3.51e-06)
Male*Edu					0.0480***		0.0278***		-0.0121***
					(0.0163)		(0.0069)		(0.0005)
Black*Edu					-0.0138		-0.0124***		-0.0022***
					(0.0132)		(0.0018)		-0.00018
Other Race*Edu					0.0387		-0.0092***		-0.0066***
					(0.0282)		(0.0027)		(0.00017)
Occupation Score*Edu					-0.00034				
					(0.00055)				
Log Income x Edu							-0.00036		0.0129***
							(0.0014)		(7.93e-05)
Year FE						No	Yes	No	Yes
Constant				2.120***	2.227	2.300***	5.220***	1.975***	4.944***
				(0.0675)	(1.804)	(0.0417)	(0.914)	(0.00326)	(0.0781)
Observations	23,454	76,633	3,841,233	23,454	23,454	76,633	76,633	3,841,233	3,841,233
R-squared				0.546	0.551	0.711	0.733	0.768	0.772

Notes: The summary statistics columns report the mean and the standard deviation in parentheses. The regression columns report robust standard errors in parentheses. We used log income in the regressions. The CPS, 2000 Census, and ACS samples do not include imputed education.
Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 2: Age and Education in the Great-Grandparent, Grandparent, and Parent Three-Generation Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Great-Grandparent		Grandparent			Parent	
Individual	Full	Regression	Full	Full Born US	Regression	Full	Regression
Age	38.03	39.27	39.30	38.60	35.79	38.68	46.74
	(8.60)	(7.27)	(8.37)	(8.29)	(7.81)	(8.64)	(5.51)
Education							
Years of Schooling		7.40	8.29	8.61	8.66	12.68	12.79
		(3.08)	(3.66)	(3.51)	(3.24)	(2.91)	(2.61)
< High School		0.86	0.77	0.75	0.76	0.19	0.19
		(0.34)	(0.42)	(0.43)	(0.42)	(0.39)	(0.39)
High School		0.07	0.14	0.15	0.14	0.39	0.32
		(0.26)	(0.35)	(0.36)	(0.35)	(0.49)	(0.47)
Some College		0.02	0.05	0.06	0.05	0.21	0.28
		(0.16)	(0.22)	(0.23)	(0.22)	(0.41)	(0.45)
College		0.01	0.03	0.03	0.03	0.12	0.08
		(0.10)	(0.17)	(0.17)	(0.16)	(0.32)	(0.28)
Graduate		0.03	0.01	0.01	0.01	0.09	0.12
		(0.17)	(0.09)	(0.09)	(0.12)	(0.29)	(0.32)
Race/Ethnicity							
White	0.90	0.97	0.92	0.91	0.96	0.88	0.96
	(0.30)	(0.16)	(0.27)	(0.28)	(0.20)	(0.32)	(0.20)
Black	0.09	0.02	0.07	0.08	0.04	0.08	0.04
	(0.28)	(0.15)	(0.26)	(0.28)	(0.19)	(0.28)	(0.19)
Hispanic			0.02	0.01	0.02	0.10	0.02
			(0.12)	(0.10)	(0.15)	(0.30)	(0.15)
Born in the US	0.74	0.83	0.86	1.00	0.89		1.00
	(0.44)	(0.38)	(0.35)	0.00	(0.31)		(0.05)
Max of Spouses							
Age	38.41	39.56	40.30	39.67	38.04		
	(8.64)	(7.21)	(8.35)	(8.28)	(7.64)		
Education							
Years of Schooling		7.25	9.12	9.41	9.53		
		(3.29)	(3.59)	(3.44)	(3.16)		
< High School		0.88	0.70	0.68	0.68		
		(0.32)	(0.46)	(0.47)	(0.47)		
High School		0.07	0.18	0.19	0.19		
		(0.26)	(0.38)	(0.39)	(0.39)		
Some College		0.03	0.07	0.08	0.07		
		(0.16)	(0.26)	(0.27)	(0.26)		
College		0.01	0.04	0.05	0.04		
		(0.11)	(0.20)	(0.21)	(0.20)		
Graduate		0.01	0.01	0.01	0.01		
		(0.07)	(0.11)	(0.12)	(0.11)		
Individual Observations	99,577,077	3,928	33,039,289	28,442,897	70,038	623,822	35,820

Notes: Standard deviations shown in parentheses. Summary statistics are shown for the full survey sample and, where available, the sample of those born in the US. The regression sample was constructed by conditioning on the match in the Grandparent and Parent generations, in effect creating a two-generation sample. Afterwards, any available data on Great-grandparents was added, but the existence of the Great-grandparent link was not required for inclusion in the data. The full sample includes all adults between the ages of 25 and 55 when observed in the survey.

Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 3: Age and Education in the Great-Grandparent, Grandparent, Parent, and Child Four-Generation Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Great-Grandparent		Grandparent			Parent		Child		
Individual	Full	Regression	Full	Full Born US	Regression	Full	Regression	Full	Full Born US	Regression
Age	38.03	39.39	39.30	38.60	35.61	38.68	44.32	40.07	40.23	39.31
	(8.60)	(7.08)	(8.37)	(8.29)	(7.72)	(8.64)	(5.99)	(8.47)	(8.48)	(6.73)
Education										
Years of Schooling		7.53	8.29	8.61	8.73	12.68	12.86	13.14	13.34	13.82
		(3.13)	(3.66)	(3.51)	(3.25)	(2.91)	(2.59)	(2.89)	(2.48)	(2.27)
¡ High School		0.85	0.77	0.75	0.75	0.19	0.19	0.13	0.10	0.06
		(0.36)	(0.42)	(0.43)	(0.43)	(0.39)	(0.39)	(0.34)	(0.30)	(0.23)
High School		0.08	0.14	0.15	0.15	0.39	0.30	0.36	0.37	0.32
		(0.26)	(0.35)	(0.36)	(0.36)	(0.49)	(0.46)	(0.48)	(0.48)	(0.47)
Some College		0.03	0.05	0.06	0.05	0.21	0.31	0.23	0.25	0.25
		(0.18)	(0.22)	(0.23)	(0.22)	(0.41)	(0.46)	(0.42)	(0.43)	(0.43)
College		0.01	0.03	0.03	0.03	0.12	0.08	0.18	0.19	0.25
		(0.11)	(0.17)	(0.17)	(0.17)	(0.32)	(0.28)	(0.39)	(0.39)	(0.43)
Graduate		0.03	0.01	0.01	0.01	0.09	0.12	0.09	0.09	0.12
		(0.18)	(0.09)	(0.09)	(0.12)	(0.29)	(0.33)	(0.29)	(0.29)	(0.33)
Race/Ethnicity										
White	0.90	0.98	0.92	0.91	0.96	0.88	0.95	0.80	0.85	0.93
	(0.30)	(0.13)	(0.27)	(0.28)	(0.20)	(0.32)	(0.21)	(0.40)	(0.35)	(0.25)
Black	0.09	0.02	0.07	0.08	0.04	0.08	0.04	0.10	0.11	0.05
	(0.28)	(0.13)	(0.26)	(0.28)	(0.19)	(0.28)	(0.20)	(0.31)	(0.31)	(0.21)
Hispanic			0.02	0.01	0.03	0.10	0.03	0.12	0.07	0.04
			(0.12)	(0.10)	(0.16)	(0.30)	(0.18)	(0.33)	(0.25)	(0.19)
Born in the US	0.74	0.83	0.86	1.00	0.89		1.00	0.87	1.00	0.99
	(0.44)	(0.38)	(0.35)	0.00	(0.32)		(0.06)	(0.34)	0.00	(0.08)
Max of Spouses										
Age	38.41	39.60	40.30	39.67	37.84	38.95	46.04			
	(8.64)	(7.04)	(8.35)	(8.28)	(7.55)	(8.53)	(5.52)			
Education										
Years of Schooling		7.39	9.12	9.41	9.60	13.27	13.58			
		(3.32)	(3.59)	(3.44)	(3.17)	(2.82)	(2.40)			
¡ High School		0.87	0.70	0.68	0.67	0.13	0.11			
		(0.33)	(0.46)	(0.47)	(0.47)	(0.34)	(0.31)			
High School		0.07	0.18	0.19	0.20	0.36	0.28			
		(0.26)	(0.38)	(0.39)	(0.40)	(0.48)	(0.45)			
Some College		0.03	0.07	0.08	0.08	0.24	0.33			
		(0.18)	(0.26)	(0.27)	(0.27)	(0.43)	(0.47)			
College		0.01	0.04	0.05	0.04	0.14	0.10			
		(0.11)	(0.20)	(0.21)	(0.21)	(0.34)	(0.30)			
Graduate		0.01	0.01	0.01	0.01	0.13	0.19			
		(0.08)	(0.11)	(0.12)	(0.11)	(0.34)	(0.39)			
Individual Observations	99,577,077	1,662	33,039,289	28,442,897	28,490	623,822	20,680	28,363,195	24,592,127	10,890

Notes: Standard deviations shown in parentheses. The regression sample was constructed by conditioning on the match in the Grandparent, Parent, and Child generations, in effect creating a three-generation sample. Afterwards, any available data on Great-grandparents was added, but the existence of the Great-grandparent link was not required for inclusion in the data. The full sample includes all adults between the ages of 25 and 55 when observed in the survey. Summary statistics are shown for the full survey sample and, where available, the sample of those born in the US.

Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 4: Educational Mobility over Time

A. Regression Coefficients

	(1) Parent 1973,79,81-90		(2) Child 2000-2013
Grandparent 1940	0.361***	Parent 1973,79,81-90	0.363***
	(0.005)		(0.006)
R^2	0.20	R^2	0.19
Observations	35,820	Observations	39,998

B. Correlation Coefficients

	(1) Parent 1973,79,81-90		(2) Child 2000-2013
Grandparent 1940	0.437***	Parent 1973,79,81-90	0.420***
	(0.006)		(0.007)
R^2	0.20	R^2	0.19
Observations	35,820	Observations	39,998

Notes: Each regression reports the coefficient of years of schooling of the child regressed on years of schooling for the most educated parent regressed. Errors are clustered at the parent family level. For the parent and child generations, each regression includes dummies for the year of the survey. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This allows us to account for the increase in the variance of education over time.
Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 5: Educational Mobility over Time across Three and Four Generations

A. Regression Coefficients

	Great-grandparent, Grandparent, and Parent Sample			Grandparent, Parent, and Child Sample			Full Four Generation Sample			
	Both Together	Each Separately		Both Together	Each Separately		All Together	Each Separately		
	(1) Parent 1973,79, 81-90	(2) Parent 1973,79, 81-90	(3) Parent 1973,79, 81-90	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013	(9) Child 2001-2013	(10) Child 2001-2013
Parent				0.394***	0.418***		0.394***	0.418***		
1973, 79, 81-90				(0.010)	(0.009)		(0.010)	(0.009)		
Grandparent	0.360***	0.361***		0.042***		0.180***	0.042***		0.180***	
1940	(0.005)	(0.005)		(0.008)		(0.008)	(0.008)		(0.008)	
Great-Grandparent	0.033***		0.116***				-0.001			0.059***
1910, 1920	(0.012)		(0.014)				(0.017)			(0.018)
R^2	0.20	0.20	0.05	0.21	0.21	0.08	0.21	0.21	0.08	0.04
Observations	35,820	35,820	3,517	10,890	10,890	10,890	10,890	10,890	10,890	1,444

B. Correlation Coefficients

	Great-grandparent, Grandparent, and Parent Sample			Grandparent, Parent, and Child Sample			Full Four Generation Sample			
	Both Together	Each Separately		Both Together	Each Separately		All Together	Each Separately		
	(1) Parent 1973,79, 81-90	(2) Parent 1973,79, 81-90	(3) Parent 1973,79, 81-90	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013	(9) Child 2001-2013	(10) Child 2001-2013
Parent				0.416***	0.442***		0.416***	0.442***		
1973, 79, 81-90				(0.011)	(0.010)		(0.011)	(0.010)		
Grandparent	0.436***	0.437***		0.060***		0.255***	0.060***		0.255***	
1940	(0.006)	(0.006)		(0.011)		(0.011)	(0.011)		(0.011)	
Great-Grandparent	0.031***		0.110***				-0.001			0.074***
1910, 1920	(0.011)		(0.013)				(0.021)			(0.023)
R^2	0.20	0.20	0.05	0.21	0.21	0.08	0.21	0.21	0.08	0.04
Observations	35,820	35,820	3,517	10,890	10,890	10,890	10,890	10,890	10,890	1,444

Notes: Each regression reports the coefficient of years of schooling of the child regressed on years of schooling for the most educated parent regressed. Errors are clustered at the parent family level. For the parent and child generations, each regression includes dummies for the year of the survey. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This allows us to account for the increase in the variance of education over time.

Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 6: Mobility Estimates with Measurement Error Correction

A. All Misreporting

Parameters	Grandparent	Parent	Child
Linkage Error (m_g)	0	0	0
Misreporting Rate (σ^2_{vg})	0.244	0.189	0.163

"True" Two-Generation Results		Counterfactual Multigenerational Regression Results	
Grandparent-Parent Correlation (β_{GP})	0.542	Grandparent Coefficient if $\rho^I_{GC} = 0$	0.043
Parent-Child Correlation (β_{PC})	0.499	Implied True ρ^I_{GC}	0.012
		Equivalent AR(1)	0.526

B. Upper Bound of Linkage Error

Parameters	Grandparent	Parent	Child
Linkage Error (m_g)	0.040	0.051	0.010
Misreporting Rate (σ^2_{vg})	0.133	0.117	0.140

"True" Two-Generation Results		Counterfactual Multigenerational Regression Results	
Grandparent-Parent Correlation (β_{GP})	0.542	Grandparent Coefficient if $\rho^I_{GC} = 0$	0.026
Parent-Child Correlation (β_{PC})	0.499	Implied True ρ^I_{GC}	0.034
		Equivalent AR(1)	0.567

C. Constant Misreporting with Linkage Error Fixed for Child Generation

Parameters	Grandparent	Parent	Child
Linkage Error (m_g)	0.054	0.032	0.010
Misreporting Rate (σ^2_{vg})	0.140	0.140	0.140

"True" Two-Generation Results		Counterfactual Multigenerational Regression Results	
Grandparent-Parent Correlation (β_{GP})	0.542	Grandparent Coefficient if $\rho^I_{GC} = 0$	0.032
Parent-Child Correlation (β_{PC})	0.499	Implied True ρ^I_{GC}	0.027
		Equivalent AR(1)	0.554

*Notes:*This table reports the result of a counterfactual estimate of intergenerational and multigenerational mobility under measurement error. First, using the measurement error estimates from Table 1, we can determine the total measurement error present in each data set. Under scenarios A, B, and C in this table, we divide the measurement error into linkage error and misreporting error parameters. From that, we can determine the "true" underlying intergenerational correlation coefficients that are consistent with the observed regression results from Tables 4 and 5 and the assumed error parameters, shown under "True" Two-Generation Results in each scenario. Next, we can calculate the multigenerational coefficient for grandparents (β_G) that would be observed given the error structure and two-generation correlations if there were no true underlying grandparent effect ($\rho^I_{GC} = 0$ and $\rho_{GC} = \rho_{GP}\rho_{PC}$). Given our observed grandparent coefficient from the multigenerational mobility regression, we can solve for the implied underlying grandparent effect, ρ^I_{GC}. Finally, we calculate the AR(1) that would result in the same long-term persistence as an AR(2) model with the derived ρ_{PC} and ρ^I_{GC}.
Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table 7: Comparison of Two-Generation Mobility in the Three-Generation Samples

	A. Regression Coefficients					
	3-Gen	2-Gen	3-Gen	2-Gen	3-Gen	2-Gen
	(1)	(2)	(3)	(4)	(5)	(6)
	Grandparent	Grandparent	Parent	Parent	Child	Child
	1940	1940	1973, 79, 81-90	1973, 79, 81-90	2001-2013	2001-2013
Parent					0.418***	0.363***
1973, 79, 81-90					(0.009)	(0.006)
Grandparent			0.341***	0.361***		
1940			(0.009)	(0.005)		
Great-Grandparent	0.237***	0.240***				
1910, 1920	(0.016)	(0.001)				
R^2	0.08	0.09	0.19	0.20	0.21	0.19
Observations	3,627	1,188,042	14,646	35,820	10,890	39,998

	B. Correlation Coefficients					
	3-Gen	2-Gen	3-Gen	2-Gen	3-Gen	2-Gen
	(1)	(2)	(3)	(4)	(5)	(6)
	Grandparent	Grandparent	Parent	Parent	Child	Child
	1940	1940	1973, 79, 81-90	1973, 79, 81-90	2001-2013	2001-2013
Parent					0.442***	0.420***
1973, 79, 81-90					(0.010)	(0.007)
Grandparent			0.426***	0.437***		
1940			(0.011)	(0.006)		
Great-Grandparent	0.262***	0.262***				
1910, 1920	(0.018)	(0.001)				
R^2	0.08	0.09	0.19	0.20	0.21	0.19
Observations	3,627	1,188,042	14,646	35,820	10,890	39,998

Notes: In this table, we compare the two-generation regression coefficients for our three-generation samples compared to the larger two-generation samples. Each regression reports the coefficient of years of schooling for the most educated parent regressed on the years of schooling of their child. Errors are clustered at the parent family level. For the parent and child generations, each regression includes dummies for the year of the survey. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This allows us to account for the increase in the variance of education over time.
Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Figure 1: Steps in the Linkage Process

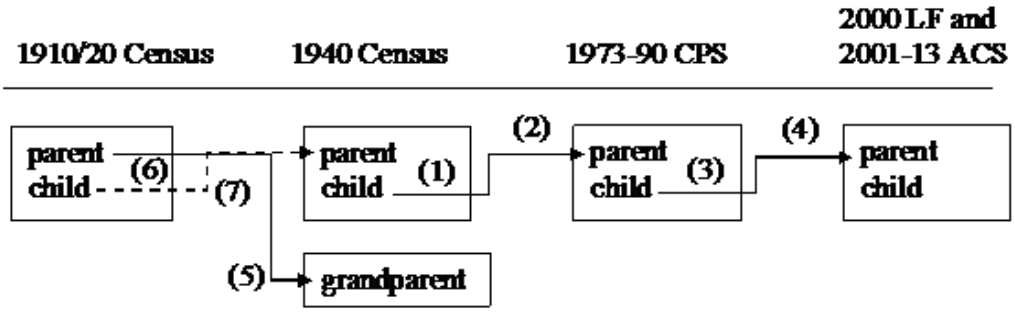
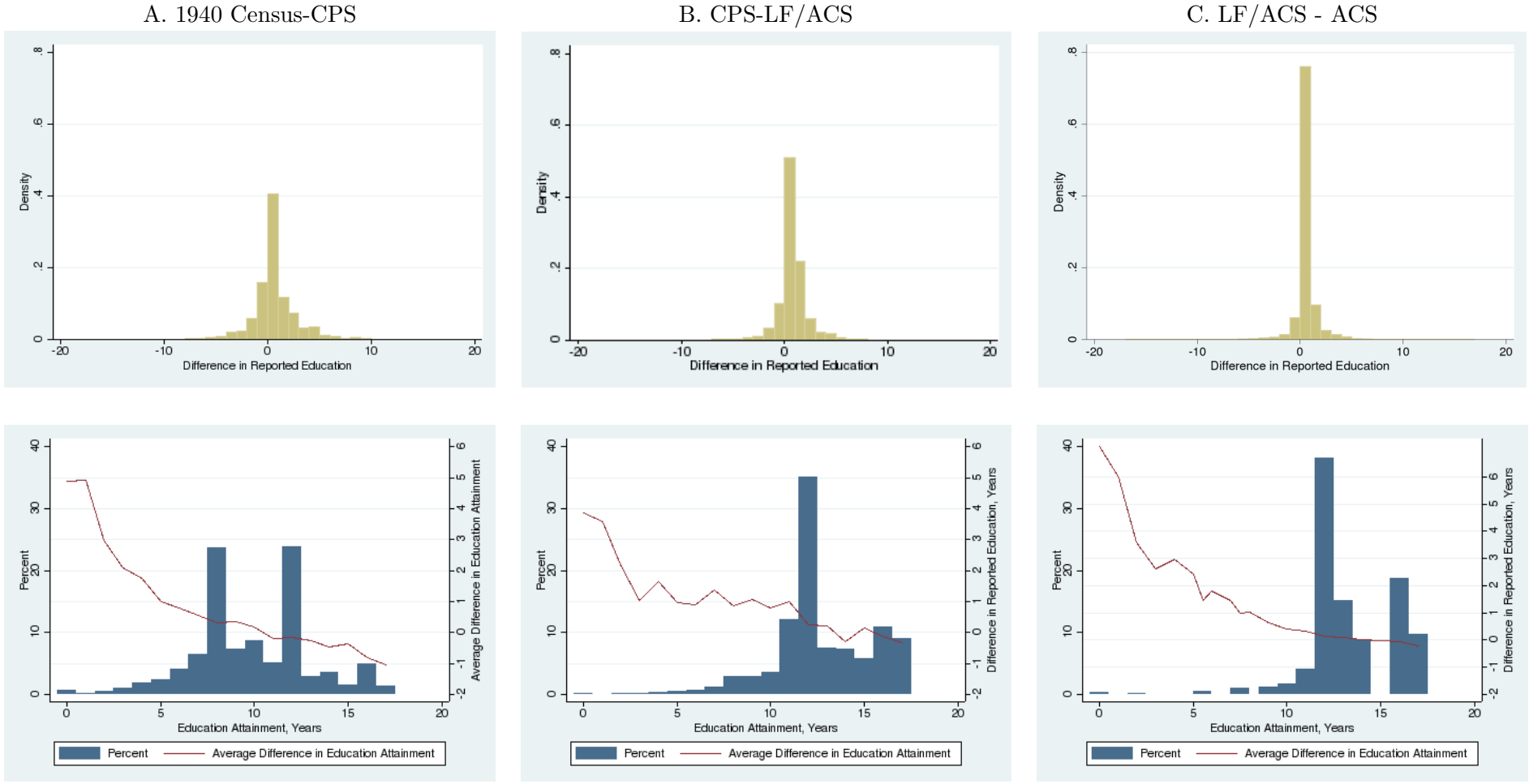
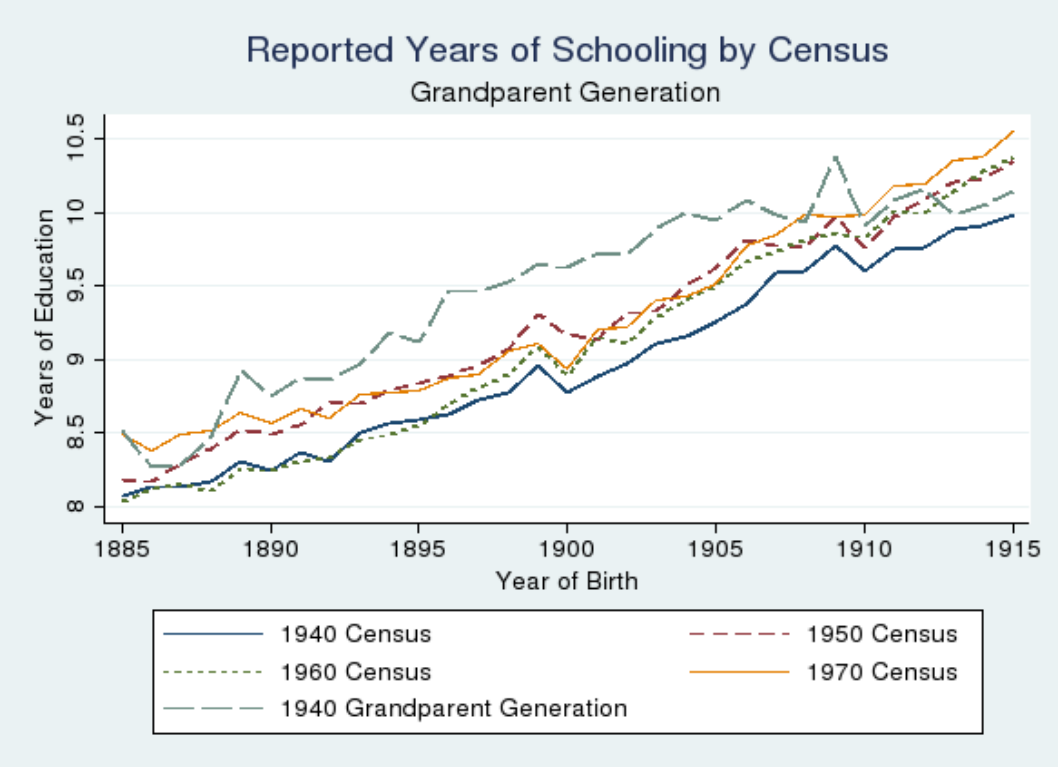


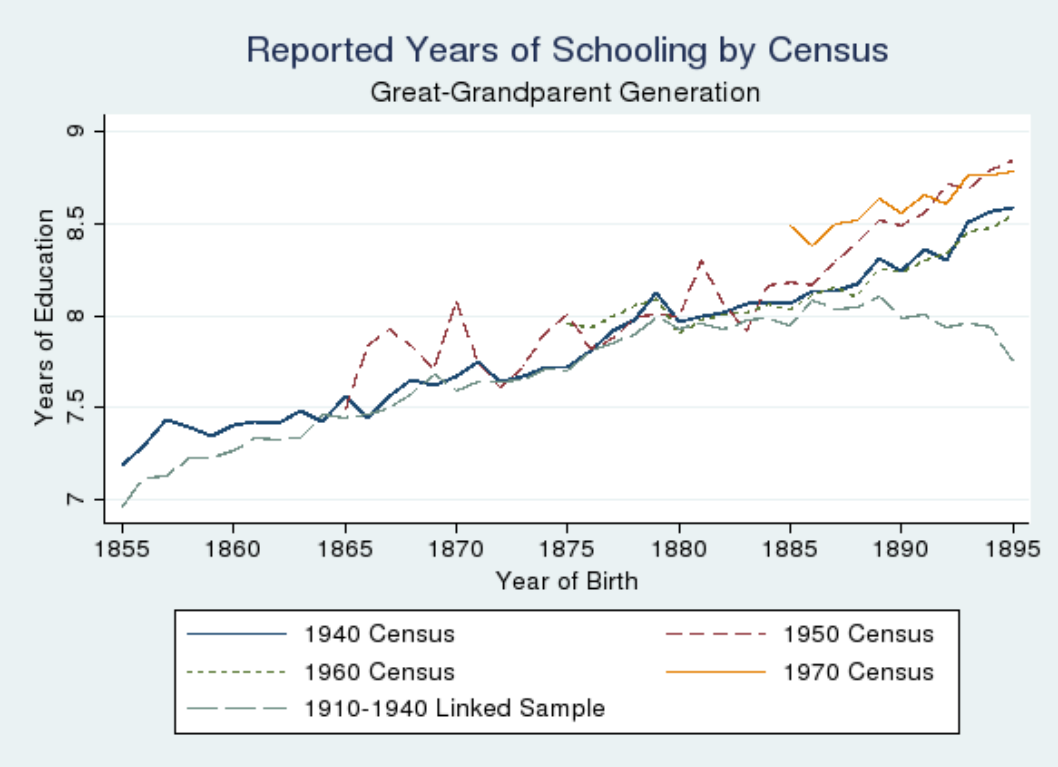
Figure 2: Within-Person Discrepancies in Education Attainment across Surveys



Notes: The reported education difference is equal to education attainment at time $t + n$ minus education attainment at time t.
Source: Linked 1910/1920 Census, 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplements, 2000 Long Form Census, and 2001-2013 American Community Survey data.

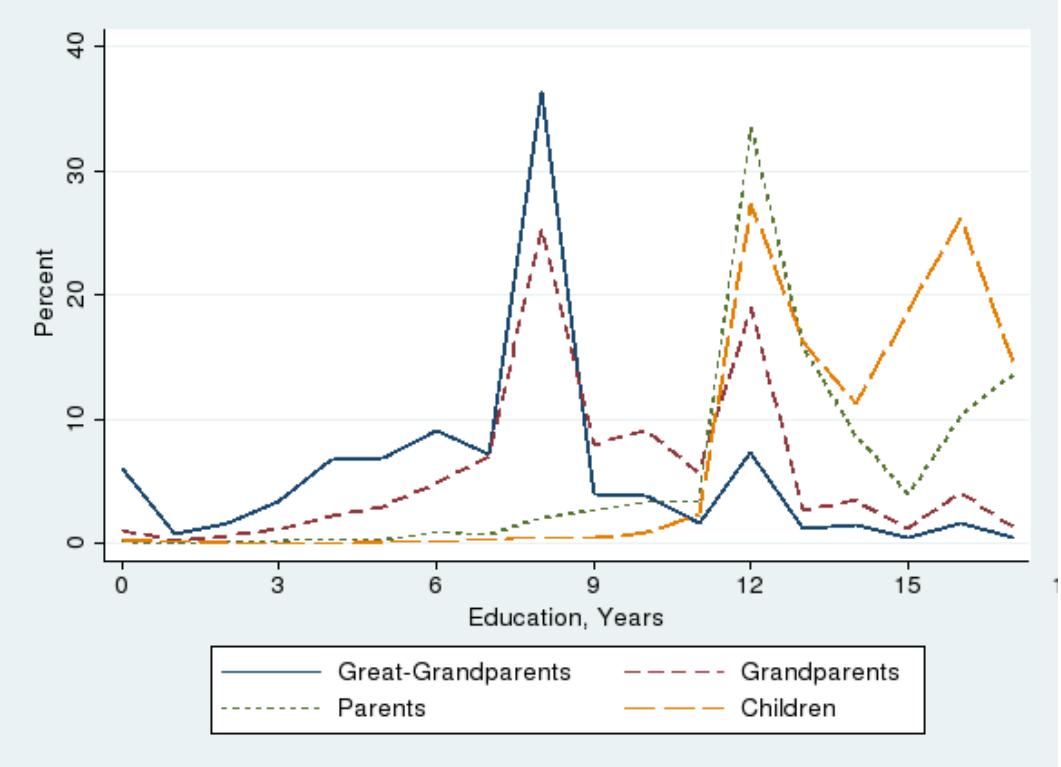
Figure 3: Reported Education Attainment by Year of Birth and Census





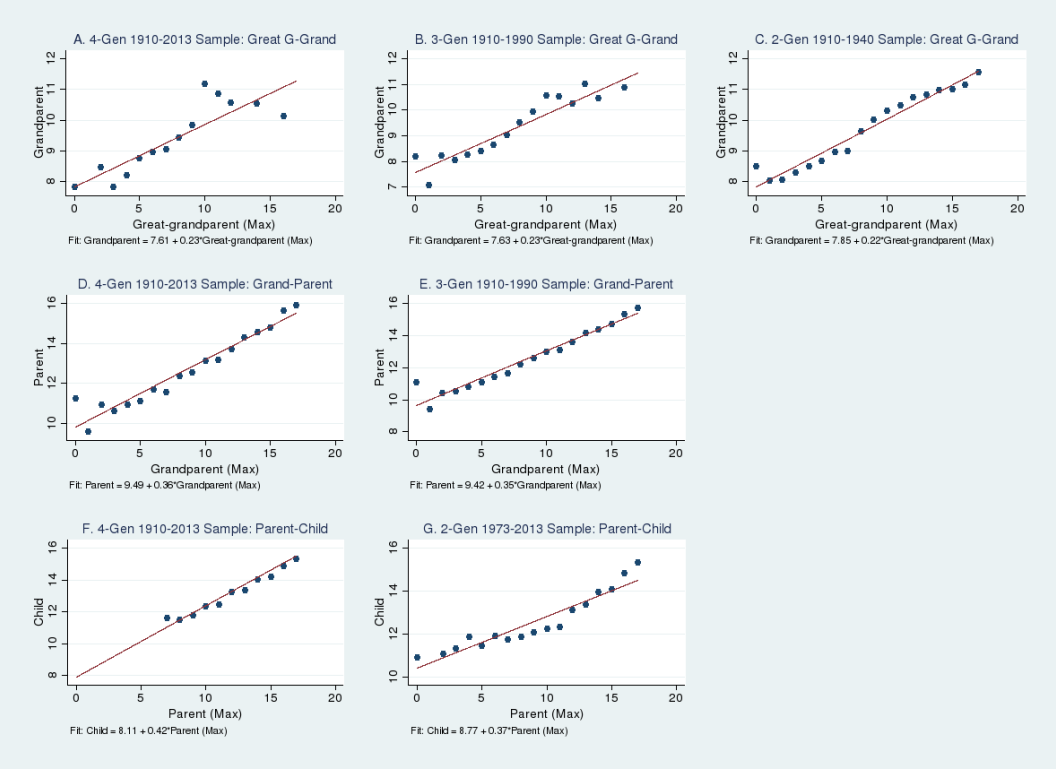
Notes: This figure plots the average years of schooling by age cohort and census sample. The 1940 Grandparent generation includes all adults 25-55 in the 1940 Census. The 1910-1040 linked sample includes all individuals in the 1910 and 1920 Censuses linked to themselves in the 1940 Census.
Source: Linked 1910/1920 Census, 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplements, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Figure 4: Distribution of Education Attainment by Generation



Notes: This figure plots the distribution of education in each generation. In each generation, only individuals aged 25-55 are included. The Great-grandparent generation comes from the 1910 and 1920 Censuses (linked to the 1940 Census for education). The Grandparent generation is from the 1940 Census. The Parent generation is from the 1973, 79, 81-90 CPS ASEC. The Child generation is from the 2000 Long Form Census and 2001-2013 ACS.
Source: Linked 1910/1920 Census, 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplements, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Figure 5: Intergenerational Education Gradients By Two-, Three, and Four-Generation Samples



Notes: This figure shows the intergenerational education gradient across each set of generations for each sample (two-, three-, and four-generations) used in the paper. The four-generation sample in Panels A, D, and F include all families with matched individuals in the grandparent generation from the 1940 Census, parent generation from the 1973-1990 CPS ASECs, and child generation from the 2000 Long Form Census and 2001-2013 ACSs. The great-grandparents from the 1910 and 1920 Census are added to the data set when available, but are not required for inclusion of the other three generations in the sample. The three-generation sample in Panels B and E includes matched parents in the grandparent generation from the 1940 Census and children in the parent generation from the 1973-1990 CPS ASECs. The great-grandparents from the 1910/1920 Census are added to the data set when available, but are not required for inclusion of the other two generations in the sample. The two-generation sample in Panel C includes matched parents in the great-grandparent generation from the 1910/1920 Census and their children in the grandparent generation from the 1940 Census. The two-generation sample in Panel G includes matched parents from the 1973-1990 CPS ASECs and children from the 2000 Long Form Census and 2001-2013 ACSs. For each pair of parent-child generations, the dots show the average child education at each parent education level (with at least 25 parent observations). The fitted lines show the observation-weighted regression coefficient of the average child education regressed on the parent education levels shown.
Source: Linked 1910/1920 Census, 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplements, 2000 Long Form Census, and 2001-2013 American Community Survey data.

For Online Publication: Appendix

Record Linkage Techniques

Our ability to accurately link individuals across data sources relies on the assignment of the Protected Identification Keys (PIK). The Center for Administrative Records and Research (CARRA) uses the Person Identification Validation System (PVS) to assign unique PIKs to person records in census and survey data to facilitate record linkage (Wagner and Layne, 2014). PIKs correspond one-to-one with a particular person and, once assigned, allow researchers to link individuals across files that have been "PIKed." To assign the PIKs, the PVS uses a probabilistic matching algorithm to compare personally identifiable information (PII) – such as Social Security Numbers (SSNs), full name, full date of birth, and address – of census and survey records to person records in a reference file.

In probabilistic matching, each available piece of PII, or matching variable, is compared and assigned a score depending on how well they match. The score is a function of how likely two matching variables will agree. For example, it is more likely that the age or state of birthplace of two random records agree than the last name because there are many more last names than possible ages or birthplaces. Therefore, higher weights are applied to name agreement than age and birthplace. PVS employs a probabilistic matching technique similar to that defined by Fellegi and Sunter (1969), which scores each matching variable and assigns agreement and disagreement weights to each variable depending on their similarity. Once each matching variable is scored, the scores are summed over all matching variable fields to produce a total score for a potential match. Chosen cut off values distinguish between matches and non-matches.

Overview of the Person Identification Validation System

The PVS assigns a PIK to a census or survey observation whose characteristics sufficiently match the characteristics of the reference record associated with that PIK. The reference file consists of Social Security Administration (SSA) Numident data as well as other federal administrative data. Each SSN in the reference file corresponds uniquely with a PIK. The reference file includes nearly 500 million SSNs and is formatted to include additional records for each name and date of birth change made for a particular SSN through the SSA. Consequently, the reference file is large and accounts

for name changes associated with marriage as well as nicknames. The reference file is enhanced using other federal administrative data to obtain additional variables not in the Numident, such as place of residence. The PVS follows the typical steps in record linkage: preprocessing, sorting into blocks, identifying potential matches, scoring matches, and resolving multiple matches. Person records are preprocessed to standardize the blocking and matching fields to ensure comparability across the census file and the reference file. Next, because the reference file is large, PVS sorts the input and reference files into blocks to create reasonably sized search spaces (Michelson and Knoblock, 2006).

Blocking refers to the creation of chunks of data where one more matching fields are exactly the same across two records. For example, Ferrie (1996) and Abramitzky et al. (2013) require that phonetically-coded last name and state or country of birth agree exactly across matched records, creating last name-birthplace blocks in which to compare age. PVS employs many blocking strategies, referred to as modules. These include blocking on the first three digits of SSNs, various definitions of geography (zip code, three-digit zip code), first and last initial, and birthdate information The PVS creates the Cartesian product of the census and reference records falling within the same block, comparing every census record to every reference record falling within the same block. The PVS then scores the similarity between the census and reference records in this comparison space.

The PVS assigns potential matches a total score depending on the similarity of the characteristics of the input records and reference file records by summing over the similarity scores assigned to each matching variable. PVS employs a string comparator program to measure Jaro-Winkler distances between first and last names (Winkler, 1995). These distances serve as a metric of how closely two names match, while allowing for some degree of misspelling and are typically indexed as a value between 0 and 1, where 1 indicates an exact match and 0 indicates two strings are not similar at all. If two strings perfectly agree, PVS assigns that matching field the full agreement weight in the score. If the strings do not agree, but fall above a user-specified cutoff, the agreement score is weighted by the Jaro-Winkler score. If the strings disagree and the Jaro-Winkler score is less than the cutoff, the matching field is assigned the disagreement weight. For numeric variables, such as year of birth, a maximum acceptable difference between the variable value in the input and reference record is dictated by the researchers, but is typically no more than 2 years. This also

allows for creation of an interval, or band, around year of birth to permit inexact matches. When numerical variables do not agree but fall within the user-specified band, PVS prorates the score of numerical variables between the specified agreement and disagreement weights. If numerical variables do not agree and fall outside the user-specified band, the disagreement weight is assigned to that matching variable. A potential match's total score is calculated as the sum of the agreement and disagreement weights attributed to each matching variable (Fellegi and Sunter, 1969).

The PVS identifies potential matches within each blocking strategy, or module, retaining only those scoring greater than a user-specified cutoff score as potential matches. Input records that do not receive a match in one module move to the next specified module. Once the input data has been processed through all passes of a module, with each pass having more refined blocking schemes, potential matches are grouped into one file and sorted by person and by score. The final step of a module evaluates the potential matches. The matches with the highest scores are processed using a decision rule to determine if the PVS will assign the PIK. If one potential match has a higher score than all the other potential matches for a particular input record, then the PIK associated with that reference record is assigned to that input observation. If there are multiple potential matches for a particular input observation with the same high score, then no PIK is assigned in that module. Records that fail to find a match in a module are passed along to the next module.

The PVS includes several modules that employ various blocking schemes. These include the Verification module, which blocks on SSN, the Geosearch module, which blocks by the first three digits of a zip code, the Namesearch module, which blocks on the first letter of first and last name, the Date of Birth module, which blocks on month and day of birth, the ZIP3 Adjacency module, which creates clusters of zip-3 areas that border each other, and the Household Composition module, which uses family structure information to assign PIKs to household members in households where at least one person was assigned a PIK in one of the other modules. The choice of module depends on the information available in each dataset. For instance, only datasets with SSNs are processed by the Verification module. The Census Applications Branch (CAB) within CARRA processed the 2000 Census and the 2001-2013 American Community Surveys (ACS) through the PVS. CAB assigned PIKs to these files using full name (first, last, and middle), full date of birth (month, day, and year), and street address. To assign PIKs to the 1940 Census and the 1973-1990 Current Population Surveys (CPS) Annual Social and Economic Supplements (ASEC), we modified the

PVS to assign PIKs tailored to the information available on each file. We report PIK rates by data source in Table A.4.

2000 Census and 2001-2013 ACS

To assign PIKs to the 2000 Census and 2001-2013 ACS, the Census Bureau used a highly-vetted configuration of PVS, which produces high-quality matches to the Numident. To assign these PIKs, PVS employs full name, full date of birth (month, day, and year of birth), sex, and full street address. Both full date of birth and street address provide a great deal of PII for constructing high-quality matches. In fact, (Sweeney, 2000) shows that 87 percent of the population is uniquely identifiable by zip, sex, and date of birth alone. The addition of full name as a matching variable further increases the Census Bureau's ability to disambiguate between multiple potential matches and increase the match rate beyond 87 percent while maintaining accuracy. Research conducted on the quality of this approach shows that PVS achieves type I error rates of less than 1 percent while achieving match rates as high as 94 percent for the ACS (Layne et al., 2014).

1940 Census

The 1940 Census contains PII that is not traditionally used to assign PIKs. We adapted the Census Bureau's matching software to use state or country of birth, location in 1940 and 1935, and parents' names in addition to more traditional PII such as first name, middle initial, last name, and age (see Alexander et al. (2014) for a detailed description of the process used to PIK the 1940 Census). To incorporate place of birth, we coded birthplace in the Numident to match the five-digit IPUMS birthplace (BPL) codes in the 1940 Census, accounting for both territories and changes in country names over time. To match on age, we calculated age on April 1, 1940 using full date of birth in the reference file and compared this to reported age in the 1940 Census.

We used six customized modules to PIK the 1940 Census. The first module blocked on the first three digits of the IPUMS BPL code. The second module blocked on age on April 1, 1940. The third module blocked on the first letter of first and last name. The fourth and fifth modules compared location in 1940 and 1935 observed from the 1940 Census to states where records received their SSN in the reference file. For these modules, we used the first three digits of the SSN (called the area number) to determine location of SSN issue in the reference file and allowed no more

than 2 years difference between age in 1940 or 1935 and the age a person acquired their SSN. The final module blocks on county observed in 1940 and 1935 to state and county of birth observed in the reference file. We only processed person-records less than 2 years old in 1940 or 1935 through this module to ensure we did not introduce significant bias from migration. Within each module, potential matches were scored based on the similarity of first name, middle initial, last name, age, five-digit state or country of birth BPL code, and parents' first name. Following Goeken et al. (2011), we dictated a cutoff value of 0.9 for the Jaro-Winkler string distance used to score the similarity of first, middle, and last name.

To produce cutoff values and scoring weights for each variable, we minimized type I error in simulated matches using the ground truth data used in Massey (2017), which was the 2005 CPS linked to the Numident by SSN with the same name and full date of birth, in addition to ground truth data we created linking the 2000 Long Form to the Numident by the Census Bureau's PIK, which was assigned using name, street address, and full date of birth. We calculated and tested weights and cutoff values that minimized erroneous links to less than 4 percent (within the truth data) using name, age, state or country of birth, and parents' name as our linkage variables.

Linking 1940 to 1910 and 1920

We used probabilistic matching techniques to link the 1940 Census back to the 1910 and 1920 censuses to obtain education for the Great-Grandparent generation. We used the 1920 Census in addition to the 1910 Census to reduce bias introduced by requiring survival of the Great-Grandparent generation to 1940. To link 1910 and 1920 to 1940, we used first name, middle initial (if available), last name, age, sex, and state or country of birth. Once again, we used our ground truth data to determine our cutoff values and minimize the instance of type I error. We were successful in achieving a type I error rate of less than 6 percent when linking on first name, last name, age, sex, and state or country of birth. We employed two blocking strategies, first blocking on place of birth then blocking on the first letter of the first and last name. We allowed a tolerance of three years in age between a 1910 or 1920 observation and a potential match in the 1940 Census. Once the probabilistic matching algorithm identified and scored all potential matches, we used only the highest-scoring, unique match.

When the initial match between 1910/1920 and 1940 was complete, we appended education data

from the 1940 census to adults in the 1910 and 1920 censuses. We then used the relationship-to-household-head variable to construct family units and identify children. These children were linked forward to the 1940 Census. To account for name-changes of women, we appended mother's maiden names from the Numident to children in the 1940 Census using the PIKs discussed previously. Once we knew a child's mother's maiden name, we appended the maiden names to mothers observed in 1940 to link backwards to 1910 and 1920.

Current Population Surveys

To fill in the gap between 1940 and 2000, we produced PIKed versions of the 1973, 1979, and the 1981-1990 CPS. We employed multiple techniques to PIK this data. For 1973, 1979, and 1981-1985 data, we used probabilistic matching techniques to assign PIKs using SSN, first name, middle initial, last name, age (or full date of birth if available), and sex. We observed SSN for a large number of observations over the age of 15. For children, we also used parents' first names in the PIKing algorithm. We used two blocking procedures: one blocking on age and one blocking on first and last name initials.

The 1986-1990 CPS did not contain first or last name. To PIK this data, we merged the CPS to the Numident using SSN. If sex and age agreed, we assigned the PIK associated with that SSN. We were able to PIK 99.5 percent of the 248,670 respondents who provided a SSN (out of 386,630 total respondents). Because we do not observe SSN for children under the age of 15, we took additional steps to increase the number of parent-child associations possible from the CPS. After PIKing adults who provided an SSN, we used the PIKs to append first and last name from the Numident to the 1986-1990 CPS. We then took these appended these names to observations of their children, allowing us to PIK children using age, sex, and parents' first and last names. Any linkage using SSN is essentially a merge of the two datasets by SSN with additional comparisons of the other matching variables to confirm or reject the linkage. As a result, we assume type I error rates in these linkages are small if present at all. For small subset of pre-1986 linkages without SSN, we have full name, age, sex, and parents' names to ensure high quality linkages and Massey (2017) finds the probabilistic matching used here produced error rates as low as 5.1 percent when using first name, last name, age, and place of birth. Although the CPS linkages do not employ place of birth, the availability of parents' name and more detailed date of birth information imply

that 5.1 percent may be an upper bound for these linkages. Only 4 percent of the 1986-1990 CPS data was PIKed using age, sex, and parents' names. To ensure these linkages were accurate, we used stricter cutoff values than our previous linkages calibrated using clerical review.

Table A.1: Educational Mobility over Time across Three and Four Generations
(All Family Lines Observed in All Generations)

A. Regression Coefficients

	Great-grandparent, Grandparent, and Parent Sample			Grandparent, Parent, and Child Sample			Full Four Generation Sample			
	Both Together	Each Separately		Both Together	Each Separately		All Together	Each Separately		
	(1) Parent 1973,79, 81-90	(2) Parent 1973,79, 81-90	(3) Parent 1973,79, 81-90	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013	(9) Child 2001-2013	(10) Child 2001-2013
Parent				0.394***	0.418***		0.322***	0.391***		
1973, 79, 81-90				(0.010)	(0.009)		(0.029)	(0.025)		
Grandparent	0.389***	0.397***		0.042***		0.180***	0.100***		0.242***	
1940	(0.016)	(0.015)		(0.008)		(0.008)	(0.025)		(0.022)	
Great-Grandparent	0.029**		0.116***				0.001			0.059***
1910, 1920	(0.013)		(0.014)				(0.018)			(0.018)
R2	0.23	0.23	0.05	0.21	0.21	0.08	0.23	0.20	0.12	0.04
Observations	3,517	3,517	3,517	10,890	10,890	10,890	1,444	1,444	1,444	1,444

B. Correlation Coefficients

	Great-grandparent, Grandparent, and Parent Sample			Grandparent, Parent, and Child Sample			Full Four Generation Sample			
	Both Together	Each Separately		Both Together	Each Separately		All Together	Each Separately		
	(1) Parent 1973,79, 81-90	(2) Parent 1973,79, 81-90	(3) Parent 1973,79, 81-90	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013	(9) Child 2001-2013	(10) Child 2001-2013
Parent				0.416***	0.442***		0.347***	0.423***		
1973, 79, 81-90				(0.011)	(0.010)		(0.031)	(0.026)		
Grandparent	0.452***	0.462***		0.060***		0.255***	0.133***		0.321***	
1940	(0.018)	(0.017)		(0.011)		(0.011)	(0.033)		(0.029)	
Great-Grandparent	0.051**		0.202***				0.001			0.092***
1910, 1920	(0.023)		(0.024)				(0.028)			(0.028)
R2	0.23	0.23	0.05	0.21	0.21	0.08	0.23	0.20	0.12	0.04
Observations	3,517	3,517	3,517	10,890	10,890	10,890	1,444	1,444	1,444	1,444

Notes: Each regression reports the coefficient of years of schooling of the child regressed on the years of schooling for the most educated observed ancestors (parent and grandparent and great-grandparent, when relevant). Errors are clustered at the parent family level. The great-grandparent generation sample is from the 1910 and 1920 Censuses linked to the 1940 Census with age data from 1910/1920 and education data from 1940. For the parent and child generations, each regression includes dummies for the year of the survey. We also include dummies whether a match exists for each specific ancestor to account for the presence of more ancestors for some individuals compared to others. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent and grandparent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Because the match rate is lowest in the 1910/1920 Census, we include all parents and grandparents from our two-generation sample and add any information on the great-grandparents with dummies for the presence of each of the four possible individuals. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This is allows us to account for the increase in the variance of education over time. This differs from Table 5 in that all regressions that include the Great-Grandparent generation include only those in the subsequent generation with an observed ancestor in the Great-Grandparent generation.

Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table A.2: Comparison of Two-Generation Mobility in the Three- and Two-Generation Samples (All Family Lines Observed in All Generations)

A. Regression Coefficients

	3-Gen Great G-Grand-Parent (1)	2-Gen (2)	3-Gen Great G-Grand-Parent (3)	3-Gen Grand-Parent-Child (4)	2-Gen (5)	3-Gen Grand-Parent-Child (6)	2-Gen (7)
	Grandparent 1940	Grandparent 1940	Parent 1973, 79, 81-90	Parent 1973, 79, 81-90	Parent 1973, 79, 81-90	Child 2001-2013	Child 2001-2013
Parent 1973, 79, 81-90						0.418***	0.363***
						(0.009)	(0.006)
Grandparent 1940			0.397***	0.363***	0.361***		
			(0.015)	(0.009)	(0.005)		
Great-Grandparent 1910	0.237***	0.240***					
	(0.016)	(0.001)					
R2	0.08	0.09	0.23	0.21	0.20	0.21	0.19
Observations	3,627	1,188,042	3,517	14,543	35,820	10,890	39,998

B. Correlation Coefficients

	3-Gen Great G-Grand-Parent (1)	2-Gen (2)	3-Gen Great G-Grand-Parent (3)	3-Gen Grand-Parent-Child (4)	2-Gen (5)	3-Gen Grand-Parent-Child (6)	2-Gen (7)
	Grandparent 1940	Grandparent 1940	Parent 1973, 79, 81-90	Parent 1973, 79, 81-90	Parent 1973, 79, 81-90	Child 2001-2013	Child 2001-2013
Parent 1973, 79, 81-90						0.442***	0.420***
						(0.010)	(0.007)
Grandparent 1940			0.462***	0.441***	0.437***		
			(0.017)	(0.011)	(0.006)		
Great-Grandparent 1910	0.262***	0.262***					
	(0.018)	(0.001)					
R2	0.08	0.09	0.23	0.21	0.20	0.21	0.19
Observations	3,627	1,188,042	3,517	14,543	35,820	10,890	39,998

Notes: In this table, we compare the two-generation regression coefficients for our three-generation samples compared to the larger two-generation samples. Each regression reports the coefficient of years of schooling for years of schooling of the child on the years of schooling of the most educated parent regressed. Errors are clustered at the parent family level. For the parent and child generations, each regression includes dummies for the year of the survey. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This is allows us to account for the increase in the variance of education over time. This differs from Table 7 in that all regressions that include the Great-Grandparent generation include only those in the subsequent generation with an observed ancestor in the Great-Grandparent generation.
Source: Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table A.3: Educational Mobility over Time across Three Generations
Robustness for Gender and Measure of Ancestor Education

A. Regression Coefficients
Grandparent, Parent, and Child Sample

	All Children All Ancestors	Sons All Ancestors	Daughters All Ancestors	Sons, Fathers, and Grandfathers	Daughters, Mothers, and Grandmothers	All Children, All Ancestors Mean Ancestor Education	All Children, All Ancestors Max and Individual Ancestors	All Children, All Ancestors Max and Mean of Ancestors (Max reported)
	(1) Child 2001-2013	(2) Child 2001-2013	(3) Child 2001-2013	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013
Parent	0.394***	0.399***	0.388***	0.330***	0.386***	0.435***	0.297***	0.105***
1973, 79, 81-90	(0.010)	(0.014)	(0.014)	(0.014)	(0.018)	(0.011)	(0.016)	(0.025)
Grandparent	0.042***	0.049***	0.037***	0.035***	0.027*	0.017**	0.031**	0.031*
1940	(0.008)	(0.011)	(0.011)	(0.012)	(0.014)	(0.009)	(0.014)	(0.018)
R2	0.21	0.22	0.21	0.21	0.19	0.23	0.22	0.22
Observations	10,890	5,605	5,285	3,912	3,184	10,890	10,890	10,890

B. Correlation Coefficients
Grandparent, Parent, and Child Sample

	All Children All Ancestors	Sons All Ancestors	Daughters All Ancestors	Sons, Fathers, and Grandfathers	Daughters, Mothers, and Grandmothers	All Children, All Ancestors Mean Ancestor Education	All Children, All Ancestors Max and Individual Ancestors	All Children, All Ancestors Max and Mean of Ancestors (Max reported)
	(1) Child 2001-2013	(2) Child 2001-2013	(3) Child 2001-2013	(4) Child 2001-2013	(5) Child 2001-2013	(6) Child 2001-2013	(7) Child 2001-2013	(8) Child 2001-2013
Parent	0.416***	0.413***	0.421***	0.415***	0.396***	0.447***	0.314***	0.111***
1973, 79, 81-90	(0.011)	(0.015)	(0.015)	(0.018)	(0.019)	(0.011)	(0.017)	(0.027)
Grandparent	0.060***	0.068***	0.054***	0.051***	0.038*	0.022**	0.044**	0.044**
1940	(0.011)	(0.015)	(0.015)	(0.017)	(0.020)	(0.011)	(0.020)	(0.025)
R2	0.21	0.22	0.21	0.21	0.19	0.23	0.22	0.22
Observations	10,890	5,605	5,285	3,912	3,184	10,890	10,890	10,890

Notes: Each regression reports the coefficient of years of schooling for the most educated observed ancestors (parent and grandparent) regressed on the years of schooling of their child. Column (1) includes the full sample of children regressed on the most educated ancestor in each generation. Columns (2) and (3) include sons and daughters respectively regressed against the most educated ancestors in each generation. Column (4) shows a regression of sons' years of schooling on the years of schooling of their fathers and paternal grandfathers. Column (5) shows a regression of daughters' years of schooling on the years of schooling of their mothers and maternal grandmothers. Column (6) uses mean ancestor education in each generation instead of maximum in the regression. Column (7) reports the coefficients on the most educated ancestor when all individual ancestor years of schooling is also included. Column (8) includes both mean and max education in each ancestor generation, with the coefficient on the most educated ancestor reported. In column (8), The corresponding coefficients in B. for mean education are 0.15*** for parents and -0.80 for grandparents. Errors are clustered at the parent family level. For the parent and child generations, each regression includes dummies for the year of the survey. We also include dummies whether a match exists for each specific ancestor to account for the presence of more ancestors for some individuals compared to others. To be included in the regression, the child must be between 25 and 55 years old and the oldest parent and grandparent must be between 25 and 55 years old. Each regression also includes age and age-squared terms for all generations to account the differences in ages for the observed parents and children and any increases in education achieved as adults. Panel B shows the results for the same regressions in Panel A where each education variable has been normalized to have a mean of 0 and a standard deviation of 1. This allows us to account for the increase in the variance of education over time. *Source:* Linked 1940 Census, 1973, 1979, 1981-1990 Current Population Survey Annual Social and Economic Supplement, 2000 Long Form Census, and 2001-2013 American Community Survey data.

Table A.4: PIK Rate by Data Source

Data Source	N	% PIK	Data Source	N	% PIK
1940 Census	53,783,480	41	2001 American Community Survey	1,301,715	93
1973 Current Population Survey	114,936	88	2002 American Community Survey	1,086,240	94
1979 Current Population Survey	36,433	85	2003 American Community Survey	1,197,788	93
1981 Current Population Survey	53,854	87	2004 American Community Survey	1,198,111	93
1982 Current Population Survey	70,729	87	2005 American Community Survey	3,995,392	92
1983 Current Population Survey	68,743	85	2006 American Community Survey	4,170,573	92
1984 Current Population Survey	65,975	82	2007 American Community Survey	4,069,624	91
1985 Current Population Survey	64,560	81	2008 American Community Survey	3,947,955	91
1986 Current Population Survey	49,466	63	2009 American Community Survey	3,930,050	90
1987 Current Population Survey	48,521	62	2010 American Community Survey	4,023,576	94
1988 Current Population Survey	48,690	62	2011 American Community Survey	4,382,393	91
1989 Current Population Survey	44,877	62	2012 American Community Survey	4,865,153	91
1990 Current Population Survey	29,831	38	2013 American Community Survey	4,474,328	92
2000 Long Form Census	39,497,455	93			

Notes: This table reports the share of the sample each survey and year that was successfully linked to the Social Security Administration's Numident file and assigned a Protected Identification Key (PIK). The N column reports the number of linked individuals.