CARRA Working Paper Series
Working Paper #2018-09

Using Linked Data to Investigate True Intergenerational Change: Three Generations Over Seven Decades

Mark A. Leach
U.S. Census Bureau

Jennifer Van Hook
Pennsylvania State University

James D. Bachmeier
Temple University

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: August, 2018

Disclaimer: This paper is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Using Linked Data to Investigate True Intergenerational Change: Three Generations Over Seven Decades

Mark A. Leach, U.S. Census Bureau
Jennifer Van Hook, Penn State University
James D. Bachmeier, Temple University

Abstract

It is widely thought that immigrants and their families undergo profound cultural and socioeconomic changes as a consequence of coming into contact with U.S. society, but the way this occurs remains unclear and controversial due in large part to data limitations. In this paper, we provide proof of concept for analyses using linked data that allow us to compare outcomes across more "exact" family generations. Specifically, we are able to follow immigrant parents and their children and grandchildren across seven decades using census and survey data from 1940 to 2014. We describe the data and linkage methodology, evaluate the representativeness of the linked sample, test a method for adjusting for biases that arise from non-representative linkages, and describe the size, diversity, and socioeconomic characteristics of the linked sample. We demonstrate that large sample sizes of linked data will likely permit us to compare several national origin groups across multiple generations.

The population of the United States has been shaped by multiple waves of large-scale immigration (Alba and Nee 2003; Bean and Stevens 2003; Portes and Rumbaut 2006). But for the Latin American and Asian origins of today's immigrants, the early decades of the new millennium are a mirror image of the pre-WW1 years of the twentieth century that saw massive numbers of Catholic and Jewish immigrants from Southern and Eastern Europe migrate to what was then a protestant, black-white society. The story documented by sociologists of that era is one of rapid assimilation, in which persons of diverse European ancestry intermarried and became, simply, "white", while at the same time redefining the American mainstream (Gordon 1964; Warner and Srole 1945).

Fundamentally, scholarship on contemporary, post-1965 immigration probes the nation's absorptive capacity. Can the US assimilate Asians and Latin Americans in the same fashion that it absorbed diverse European ethnicities during the first half of the 20th century? Contemporary assimilation theories are largely reformulations of the classical assimilation model that emerged to describe patterns of inter-generational mobility observed among the children and grandchildren of large numbers of immigrants, mostly from Southern and Eastern European around the start of the 20th Century. As such, their objective is to provide a model to explain "post-1965" assimilation patterns during an era that is dramatically different with respect to institutional context, post-industrial labor market, and the national origins of the immigrants, who come overwhelmingly from Latin American and Asia, rather than from Europe.

The two leading contemporary assimilation theories (and their variants) – new assimilation theory (Alba and Nee 2003) and segmented assimilation theory (Portes and Zhou 1993) – disagree primarily with respect to the salience of "non-white" racial and ethnic status in today's, post-civil rights era institutional context and post-industrial labor market, largely bifurcated into "high" and "low" skill sectors (Piore 1979). As a result of this singular focus on the post-1965 immigration, both theories have implicitly assumed that studying immigrant inter-generational mobility consists, roughly, of comparing the attainment of an immigrant second-generation coming of age starting in the 1970s, with an eye toward accomplishments of their third-generation children who have not yet reached adulthood in large numbers (Myers 2007).

In short, contemporary assimilation theories are oriented toward describing trajectories of mobility from the first to the second generation among post-1965 immigrant groups, with an implicit (if vague) prediction that the pattern from 2nd to 3rd should be a continuation of that observed from the first to the second. From this perspective, it is problematic that the nation's largest immigrant group – the Mexican-origin population – consists of a generational composition that varies widely from the simple two-generation distribution assumed by contemporary theories. Unlike most other dominant contemporary immigrant groups, Mexican migration has persisted, largely uninterrupted, as a flow of low-skilled labor migrants across the U.S.-Mexico border for roughly a century (Massey et al. 1987), and thus precedes by decades the post-civil rights era context with which contemporary assimilation theory is primarily concerned.

Among contemporary immigrant groups, the Mexican-origin population thus is distinctive in that large numbers are of the "third-and-later" generation and trace their American origins to immigrant grandparents, great-grandparents, and beyond, whose inter-generational mobility trajectory is rooted in pre-civil rights era and even pre-WW II historical context. For example, assuming 25-year family generational intervals, a hypothetical 25-year old fourth-generation Mexican-origin woman in 2010 would be the daughter of a 3rd-generation mother born in 1960, the grand-daughter of a second-generation grandmother born in 1935, and the great-granddaughter of a Mexican immigrant woman born in 1910. In this example, half of the inter-generational lineage took place prior to the civil rights era, and in fact, largely before WWII, which, relative to the post-civil rights era, is a period in history marked by at least three contexts that were inimical to the upward mobility of Mexican-Americans: (1) high fertility and large sibships; (2) an underdeveloped public education bureaucracy and therefore considerable regional, class, and racial inequality in access to educational opportunity; and (3) overt racial discrimination, hostility, and even mob violence directed at persons of Mexican-origin.

The key implication of this "generational heterogeneity" (i.e., the fact that unlike other immigrant groups, Mexican-Americans extend well beyond the second and third immigrant generation), is that Mexican-American intergenerational mobility patterns have been differentially influenced by historical contextual factors that have varied significantly over time. Generational heterogeneity poses a problem for contemporary assimilation research that is compounded by the fact that nearly all available data sources used to study intergenerational mobility do not measure immigrant generations with the necessary precision to account for the distinctive historical contexts that have affected the mobility trajectories of later-generation Mexican Americans. At best, publicly available national surveys allow for the identification of the 1st, 2nd, and "3rd+" generation. Studies comparing, for example, age-adjusted educational attainment across these three generation groups typically find "stalled" or "declining" mobility between the 2nd and the 3rd+ generation. Implicit in this type of generational measurement and comparison is the flawed assumption that all members of the 3rd+ generation, like their 2nd generation contemporaries, fall on a mobility trajectory rooted entirely in a post-1965 immigration context. In fact, as noted above, 3rd-and-later generation Mexican-Americans inherit familial educational orientations and aspirations forged during pre-Civil Rights era and often pre-WWII contexts, and to date, we know very little, empirically, about the nature of the mobility patterns for Mexican immigrants and their descendants during this period in history.

The diversity of perspectives about assimilation may arise partially from data limitations. Immigrant assimilation is widely hypothesized to transpire over long periods of time across multiple generations. Yet no data sources follow families long enough to directly observe this kind of intergenerational change. As a consequence, intergenerational change is often deduced with cross-sectional data. That is, 1st generation immigrants are typically compared with the 2nd generation (U.S.-born children of immigrants) and the 3rd-or-higher generation (U.S.-born children of U.S.-born parents) at a single point in time, meaning that people in the 1st generation are not the actual parents and grandparents of the 2nd and 3rd generations. Yet research conducted by Julie Park and Dowell Myers (2010) provides indirect evidence that one would probably observe greater intergenerational upward mobility if one were to follow advancement within families across generation by linking children to their parents and grandparents. At this point in time, we know of only one study (Telles and Ortiz 2008) that actually follows families across generations over time, and it was conducted for a single national origin group, Mexicans, was limited in geographic scope (Los Angeles and San Antonio), and had a small sample size. As a consequence, our understanding of the assimilation process tends to rest on incomplete data with little to no direct observation of the phenomenon of intergenerational change.

In this paper, we introduce and provide proof of concept for a new data linking approach that promises to help fill some of these important data gaps. We build on a U.S. Census Bureau project which aims to link individuals across decennial censuses and survey data files from 1940 to the present (Alexander, Gardner, Massey, and O'Hara 2015). This data infrastructure is ideally suited to directly observe processes of immigrant intergeneration mobility. The data linkages are based on a process wherein individuals in each data source are given a unique, protected identifier that serves as a match key to link records for the same individuals across data sources. This enables us to follow families over seven decades and across three generations. We start with parents and their children in 1940. We follow the children from 1940 to a time when they are adults in the 1970s and 1980s. At that time, we are able to identify many of the grandchildren of the original 1940 parents that reside with the adult children. We then follow up on the grandchildren to a time when they are adults in the 2000s. The linked data thus enable us to investigate outcomes across generations from parents to children to grandchildren.

Additionally, the linked data will allow us to compare outcomes across more "exact" family generations than has been possible to observe ever before. Typically, survey data contains information on place of birth and (sometimes) parents' place of birth, which allows researchers to compare the first, second, and third-or-higher generations, but it is not possible to tell who is in the "exact" third generation versus the 4th-or-higher generations. However in the linked data, the parents in the 1940 Census may be of the 1st, 2nd, or 3rd-or-higher immigrant generations, so we will be able to identify their grandchildren as members of the 3rd, 4th, or 5th+ generations.

In the paper we first describe the data and linkage methodology and demonstrate the feasibility of the approach. We specifically evaluate the representativeness of the linked sample, test a method for adjusting for biases that arise from non-representative data linkages, and describe the size, diversity, and basic socioeconomic characteristics of the linked sample. We next demonstrate the value of the linked data by using it to explore how early disparities in grandparents' and parents' educational attainment help explain contemporary disparities across racial-ethnic groups.

Data

Data Sources and Linkage

The CLIPP data infrastructure enable us to link individuals across multiple data sources: (1) the 1940 Census (100% file), (2) the 1973, 1979, and 1981-1990 Current Population Surveys (CPS), and (3) Census 2000 (long form data) and the 2001-2015 American Community Surveys (ACS). The 1940 census data contains a record for everyone in the population but the other data are large sample surveys, so only a small subset of those in the 1940 census may be observed in the CPS data. Likewise, only a sample of those in the CPS will link to records in the Census 2000 long form data or ACS data.

To link data records from different sources for the same individuals, the Census Bureau attempts to assign a unique identifier to each individual in each data source based on Social Security Number (SSN) when it is available (in essence, the unique identifier is a "scrambled" SSN). For records without an SSN, personally identifiable information such as name, date of birth, and address are used for probabilistic matching to assign a unique identifier. The fields used for matching are compared against the same fields in a master reference file that contains a unique identifier for individuals whose identity has previously been validated. Depending on availability of SSNs and the quality of the name, address, and date of birth data, the percentage of records assigned a unique person identifier varies greatly from data source to data source. Personal information is removed from each data source before a researcher may link the data sets together and use them for research purposes. For more information on the linking process, see Wagner and Layne (2014).

To construct our analytical database, we first identify all parents and their co-residential children, regardless of age, in the 1940 Census (for illustration, we show in Figure 1 the population ages 16+, a proportion of which are parents and reside with at least one child). We refer to the co-residential parents as the 1st family generation and to their children as the 2nd family generation (not to be confused with immigrant generation, which we define below). We then link the records of the 2nd family generation children in the 1940 Census to the records for the same individuals as adults observed in the 1973, 1979, or 1981-1990 CPS data. Among the linked individuals, we then identify the co-residential of the linked adults in the CPS. The co-residential children of the linked 2nd family generation comprise the 3rd family generation sample. We then link the 3rd family generation children in the CPS to the records for the same individuals in the Census 2000 long form data and in the 2001-2015 ACS data to observe those individuals as adults.

Measures

Our primary independent variables of interest are national or regional origin and immigrant generation. By including in our analysis only co-residential parents and their children in the 1940 Census, we can determine both national or regional origin and immigrant generation from the place of birth of the 1st family generation parents that we identify in the 1940 Census data. National or regional origin is determined by the place of birth of the 1st family generation parents. We categorize foreign-born parents into nine regions or countries of birth and native-born parents into three race categories, creating a total of 12 mutually exclusive origin groups (we later aggregate these into five categories for statistical analysis). We assign the same origin to the 2nd family generation children and 3rd family generation grandchildren. If one parent is native born and the other is foreign born, we assign the origin of the foreign-born parent to the child(ren). If two foreign-born parents are of different origins, we generally assign the origin with a smaller U.S. immigrant population in 1940, with some exceptions. Given our interest in Mexican origin progress, for example, we categorize a child as Mexican origin if they have a Mexican born parent and another parent born in some other Latin American country. And while Southern and Eastern European immigrants had relatively larger U.S. populations in 1940 than immigrants from the United Kingdom and Canada, we assign children to the origin of the parent from the former given their disadvantage relative to immigrants from the U.K. and Canada.

Immigrant generation of the 1940 2nd family generation children are determined by their birthplace, if foreign born, or the birthplaces of their parents. Foreign-born children are in the 1st immigrant generation. Children with two foreign-born parents are in the 2nd immigrant generation. Children with one foreign-born parent and another native-born parent are in the 2.5 generation. Children with two native-born parents are in the 3rd or higher (3rd+) immigrant generation, and we cannot determine how many family generations have been in the United States. An exception are children who reside with a parent and grandparent, a very small percentage of our sample. If the grandparent is foreign-born and the parents are native born, we know the grandchild is 3rd generation. We determine the immigrant generation of the 3rd family generation children that we observe in the CPS by advancing by one the immigrant generation of the 2nd family generation parents.

To illustrate the above definitions, if two parents were born in Mexico and their co-residential child was born in the United States, all three individuals would be Mexican origin, the parents would be in the 1st immigrant generation, and the child would be 2nd immigrant generation. Alternatively, if the parents and children were all U.S. born, origin is unknown, so we use racial identification to assign the parents to a 2+ generation race category (non-Hispanic White, Black, or Other), and categorize the child as a 3+ generation race.

In addition, we use age, sex, race and ethnicity, and completed years of schooling in each family generation.

Methods

Microsimulation of the Universe of the 3rd Family Generation

Our linking procedures first identify the 3rd family generation in the 1973, 1979, and 1981-1990 CPS data. However, in order to identify a child in the 3rd family generation in a particular CPS, several conditions must be satisfied: (1) the child's 2nd family generation parent must have lived with a 1st family generation parent in 1940, (2) the child must have been born by one of the years of available CPS data, (3) the child must not have died prior to the year of CPS data, (4) the child's parent must not have died prior to the year of CPS data, and (5) the child must still be living with his/her parent in the year of CPS data. To illustrate the years during which these conditions are most likely to be satisfied, we used micro-simulation techniques to project the fertility histories for women born between 1921 and 1940 based on observed parity- and age-specific cohort fertility rates (National Center for Health Statistics 1947). These are the mothers that we would expect to observe as 2nd family generation children in the 1940 Census. We also account for mortality of the women and their children based on age-specific estimates from 1940 to the present (University of California, Berkeley and Max Planck Institute for Demographic Research 2016), and whether children live with a parent based on age-specific estimates of living arrangements from the 1950 through 2000 decennial census data (Ruggles, Genadek, Goeken, Grover, and Sobek 2015). The simulation allows us to compare the years in which the future mothers observed in 1940 would likely co-reside with their own children relative to the years of CPS data we have available to observe co-residential children.

Adjusting for Selectivity Bias in the Linked Sample

An inherent challenge in analyzing linked records across data sources and over time is to understand and account for selection bias among the families who are successfully linked and included in our final analytical sample. There are several factors that may cause a family's exclusion from our linked sample including an inability to assign unique person identifiers to individuals in each family generation, data availability and aging, mortality, or migration.

A primary source of bias in our sample comes from the process of assigning unique identifiers to individuals. The Census Bureau is not able to assign identifiers to every record and such assignments are not random. Whether and how individuals obtain a SSN has changed over time, and the availability of SSNs and other personally identifiable information across data sources varies greatly. Previous research shows that probabilistic matching techniques systematically assign a unique identifier to certain types of records and not others (Bond, Brown, Luque, and O'Hara 2014; Wagner and Layne 2014). About seventy percent of the 2nd family generation children observed in the 1940 Census were assigned an identification key, and the characteristics of the children assigned a key differ from those not assigned a key. The assignment rate in the years of CPS data is much lower, about 36 percent of all records, so linking to these data potentially increases selectivity bias to a great degree. By contrast, the assignment rate in more recent data from Census 2000 and the ACS is about 90 percent.

The particular years of cross-sectional data available to us may also introduce selectivity bias. The 2nd family generation children who co-reside with a parent in 1940 likely differ from children who did not reside with a parent, whether because the parent had not immigrated to the United States or had died. The 1973 data aside, most of the CPS data we have were collected more than 40 years after the 1940 Census. As such, a 2nd family generation parent may have died before they could be observed in the CPS and linked, a 3rd generation child may have died or moved out of the family home before observing them in the CPS. We thus are likely only to observe and link 3rd family generation children of relatively older parents in the 1980s.

Following the methods developed by Hong and Raudenbush (2008) and Sampson, Sharkey, and Raudenbush (2008), we attempt to account for selectivity bias in our linked sample by calculating Inverse Probability Treatment Weights (IPTWs) that adjust for observed differences between those who are included in our sample and those who are excluded. We model such differences in two stages. We first model the probability that the 2nd family generation children who we observe in the 1940 Census are linked to a CPS record *and* has a co-residential child in the CPS (the 2nd generation sample). We use age, race, immigrant generation, and origin of the 2nd family generation children, and years of education completed by their 1st family generation parents as independent variables.

We then model separately the probability that the 3rd family generation children that we identify in the CPS are linked to a record in the Census 2000 or ACS (the 3rd generation sample). We use similar independent variables as the first model.

To calculate an IPTW for each 3rd generation individual included in our linked sample, we multiply the estimated probability of selection into the 2nd generation sample with the estimated probability of selection into the 3rd generation sample, and take the reciprocal of the product to obtain each IPTW. We then normalized the IPTWs so that the final person weights sum to the number of cases in our analytic sample. The calculation for our adjusted person weights is as follows:

$$Weight_{adj} = \frac{\left(\frac{1}{P_{Gen2} * P_{Gen3}}\right)}{\frac{\Sigma\left(\frac{1}{P_{Gen2} * P_{Gen3}}\right)}{n}}$$

Where P_{Gen2} and P_{Gen3} are the probabilities of selection into the 2nd generation and 3rd generational samples, respectively, and n is the number of cases in our final linked sample of 3rd family generation adults observed in Census 2000 or the ACS. The adjusted person weights essentially give individuals who are least likely to be included in our sample, given observed characteristics, relatively greater influence in our analyses.

Modeling Intergenerational Change in Years of Education

Finally, to explore the extent to which current inequalities between ethnic groups are related to differences in earlier immigrant generations, we estimate a series of models predicting the number of years of education as a function of racial/ethnic origin among members of the 3rd family generation, all of whom were linked as adults age 25+ to Census 2000 and the ACS in the 2000s.

Results

Number of Successfully-Linked Records

We first consider the number of records that we are able to link for three family generations in order to assess the feasibility of studying intergenerational change and mobility using the linked data. We are very encouraged by the number of records in our linked sample with which to observe outcomes in each successive generation. Figure 1 depicts our data sources and numbers of records that are assigned unique identifiers and for which we were able to link for each family generation. We start with more than 44 million records of 1st family generation parents who co-reside with at least one child in the 1940 Census. We find 56 million 2nd family generation children who reside with the 1st generation parents. Of the 56 million children, the Census Bureau was able to assign a unique identifier to about 65 percent (36 million).

[Figure 1 here]

We successfully link 109,000 of the 56 million 1940 child records to a record in the CPS data. Of the 109,000 linked 2nd generation records in the CPS, we observe that 52,000 resided with a 3rd generation child. The 52,000 2nd family generation linked individuals with a co-residential child (of the 56 million child records that we observe in the 1940 Census data) comprise our "2nd generation sample" for the purposes of modeling selectivity bias as described above.

We identify almost 80,000 3rd family generation children who co-reside with a parent in the 2nd generation sample. Our sample of 3rd generation children is comprised of almost 80,000 CPS records. About 81 percent of the child records, or 65,000, were assigned individual identifiers, and 18,900 were linked to a record in Census 2000 or the 2001-2015 ACS data. These linked individuals comprise our "3rd generation sample" that we use to assess and model selectivity and analyze differences in educational attainment.

Demographic Characteristics of the Linked Sample

Table 1 shows distributions of the observed and linked individuals by various characteristics. In addition to the total sample size, we also are encouraged that similar numbers of men and women are in our final linked sample. This potentially will enable us to conduct future analyses by gender, an important component of immigrant incorporation. The age distributions for the successive family generations are according to our expectations. The vast majority of 1st family generation parents in 1940 are between ages 25 and 50, and their 2nd generation children are younger than 25. In accordance with the years of CPS data available, the ages of 2nd family generation children that we are able to link to the CPS skew older, mostly in their 40s and 50s, and most of their co-residential 3rd generation children are between ages 10 and 24. Our analytical sample of 3rd family generation adults who linked to Census 2000 or the ACS are between 30 and 54 years old.

We find that 15.6 percent of the population of adults with co-residential children in 1940, our 1st family generation, were foreign born, or 1st generation immigrants. In accordance with the construction of our linked database, there are no foreign-born 3rd family generation adults in our final linked sample. The vast majority of individuals are in the 4th+ immigrant generation. In other words, our linked sample of 3rd family generation individuals includes about 4,000 records in the 3rd immigrant generation who had a foreign-born grandparent. These individuals comprise the core sample with which we will analyze inter-generational immigrant incorporation in future analyses.

Consistent with immigration flows in the first half of the 20th Century, most 1st family generation parents in 1940 were of European origins. Of particular interest to the authors, we observed almost 500,000 Mexican origin individuals, most of whom are Mexican born by definition. In our final linked sample includes about 500 Mexican origin records. We are encouraged that this should be sufficient to make group comparisons, but we hope to increase this when additional years of CPS data become available.

[Table 1 here]

Representativeness of the 3rd Family Generation

We assess representativeness of our linked 3rd family generation sample in several ways. First, we show the results of simulating the universe of 3rd family generation children who co-reside with a parent in Figure 2. The y-axis of the chart represents the proportion of the universe of children born to mothers who were born between 1921 and 1940. During the 1950s, only a small share (19 percent, blue area) of the universe of 3rd family generation would have been born and co-residing with a 2nd family generation mother who we observed in 1940. The key reason was that most 3rd family generation children had not been born yet (shown in green). The largest proportion of 3rd family generation children had been born and co-resided with a parent by the mid- to late-1960s (blue area). We can see that beginning in the early 1970s, the proportion of children who had moved out of their parents' household began to grow (shown in orange). The first year of CPS data available to us is 1973. By 1979, the next year of available data, a majority of 3rd generation children no longer resided with a parent. It is also during the 1970s and 1980s that the children's parents start to die (light blue), which contributes to our inability to make parent-child linkages.

[Figure 2 here]

These results suggest that we would have linked an even larger share of the 3rd generation had we looked for them in data collected during the mid- to late-1960s rather than the 1970s and 1980s. At this time, however, we are unable to link to data from the 1960s because the respondents in these data have not yet been assigned unique identifiers. These results (not shown) also suggest that the 3rd generation children in our linked sample are more likely be later-born children in their families (i.e., higher parity children of older parents, born in later years) than their siblings who are not represented in our linked sample.

As described above, we created a sampling weight that, when applied, attempts to make the 2nd and 3rd family generations in the linked data more representative of the children and grandchildren of the original parents in the 1940 Census. In Table 2, we illustrate the impact of applying this weight by comparing the weighted and unweighted characteristics of those age 25+ in the 3rd family generation who were successfully linked to Census 2000 or the American Community Survey in the 2000s (N = 17,881). We specifically examine average years of education for the 3rd family generation and the educational attainment of their parents and grandparents, by ethnic origin. In future work, we will evaluate other characteristics besides educational attainment.

[Table 2 here]

In general, the weighted estimates of educational attainment are lower than the unweighted estimates. For example, among non-Hispanic whites in the 3rd family generation, the average years of education was 0.2 years lower when weighted (13.9) than unweighted (14.1). The differences are even greater for the educational attainment of their parents (0.4 years lower) and grandparents (.5 years lower). Similar patterns can be seen among African Americans and Mexicans. This suggests that more highly educated families were more likely to be linked across the census data sources. Other Hispanics were the only exception to this pattern. For this group, the weighted estimates tended to be higher than the unweighted estimates. Regardless of the direction of the bias, these results suggest that it is important to develop and use weights to adjust for sample selection bias.

Illustration of the Substantive Value of the Linked Data

Finally, we demonstrated the value of the data by using it to explore a substantive question that would have been difficult to answer without linked data. There are several topics we could have selected, such as explorations of how intermarriage, ethnic identity, socioeconomic status, place of residence, and fertility change across generations. Here, we focus on race/ethnic differences in educational attainment.

It is well known that racial and ethnic groups in the United States vary considerably with respect to their level of educational attainment. However, data limitations make it difficult to assess how much these differences are related to current limitations in children's opportunities, or whether they are also tied to a much longer process of intergenerational transmission of disadvantage. It also is widely known that racial and ethnic minorities have not had equal access to education, especially prior to the Civil Rights era of the late 1950s and 1960s, a time period during which many of the parents and grandparents of the 3rd family generation were obtaining their educations. Given how children's educational attainment is strongly linked to parental educational attainment for all racial and ethnic groups, is it possible that the current inequalities we see today are the remaining (and unaddressed) vestiges of this earlier era?

To explore this question, we estimate a series of models predicting the number of years of education as a function of racial/ethnic origin among members of the 3rd family generation, all of whom were linked as adults age 25+ to the ACS in the 2000s (N=17,881). The results are shown in Table 3. To establish a baseline for the current observed level of racial/ethnic inequality in education, Model 1 includes race/ethnicity and minimal controls (grandparents' immigration status and region[1]). To assess whether parents' and/or grandparents' levels of education help explain the baseline race/ethnic differences, the remaining models add parents' (Model 2) and grandparents' (Model 3) years of education (averaged across all observed parents or grandparents). All models are weighted to adjust for selection into the sample.

[Table 3 here]

[1] Models that include demographic characteristics such as age and sex did not show different results.

As shown in Model 1, all racial/ethnic groups have significantly fewer years of education than non-Hispanic whites. The most disadvantaged groups were Mexicans (1.5 fewer years of education than whites) and African Americans (.72 fewer years of education). The "Other" race/ethnic group also exhibited low levels of education (b = -1.0), but it is unclear to us at this time who this "Other" group represents. It probably does not include many Asian-Americans because most of them trace their ancestry to immigrants who arrived after 1965. More research is required to better understand the finding for this "Other" group.

Adjusting for parents' education in Model 2 explains most of the race/ethnic disadvantages seen in Model 1. The disadvantages observed among Mexicans drop from 1.5 to .41 years, a decline of 72%, and African-Americans' disadvantages drop to zero. After adjusting for grandparents' education in Model 3, Mexican's disadvantages drop further to .29 years. Thus accounting for the fact that Mexican-American's parents and grandparents had very low levels of education accounts for 80 percent of their current educational disadvantage relative to non-Hispanic whites. Among African-Americans, accounting for grandparents' education reveals a non-significant *advantage* relative to non-Hispanic whites.

Clearly more work is required to fully understand these findings. Nevertheless, these analyses are intriguing because they suggest that the inequalities in educational attainment we see today have deep roots that extend back several generations. It would not be possible to show this without the linked data.

Conclusion

The purpose of this paper is to assess the feasibility of using linked data from several censuses and national surveys to directly observe how immigrants and their descendants change over time and across generations. Our preliminary results are encouraging, suggesting that it is possible to observe nearly 20 thousand families across three generations and over seven decades (and over 50 thousand families across two generations and over four decades) using data linkage techniques. Such large sample sizes will permit detailed explorations of how the assimilation process varies across groups and communities. We further expect that as more data are assigned unique identifiers, the number of families that can be linked across generations will increase.

In the present study, we explored the characteristics of the linked sample of 1st family generation parents, their 2nd family generation children, and the 3rd family generation grandchildren to assess the suitability of these linked data for research on immigrant assimilation. We also illustrate and evaluate the magnitude and sources of sample selection bias conducting a microsimulation of when we would expect the 3rd generation children to reside with a parent relative to the years of data currently available for data linkage. We adjust for selection bias into our linked sample by modeling the likelihood of linking 2nd generation children from 1940 to the CPS and the likelihood of linking their 3rd generation children observed in the CPS to Census 2000 and ACS, and calculating weights that adjust for such bias.

Finally, we demonstrate the value of the linked data by using it to explore an important substantive question about immigrant assimilation, whether contemporary educational inequalities appear to be rooted in differences that existed upon earlier immigrant generations' arrival in the United States.

References

Alba, R. D. and V. Nee. 2003. *Remaking the American Mainstream*: *Assimilation and Contemporary Immigration.* Cambridge, MA: Harvard University Press.

Bean, F. D. and G. Stevens. 2003. *America's Newcomers and the Dynamics of Diversity*. New York, NY: Russell Sage Foundation.

Bond, B., J. Brown, A. Luque & A. O'Hara. 2014. The Nature of Bias When Studying Only Linked Person Records: Evidence from the American Community Survey. *CARRA Working Paper #2013-08.*

Gordon, M.M. 1964. *Assimilation in American Life.*: Oxford University Press.

Hondagneu-Sotelo, P.and E. Avila. 1997. "I'm here, but I'm there: The Meanings of Latina Transnational Motherhood." *Gender and Society* 11:548-571.

Massey, D. S., R. Alarcon, J. Durand, and H. Gonzalez. 1987. *Return to Aztlan: The Social Process of International Migration from Western Mexico.* Berkeley, CA: University of California Press.

Myers, D. 2007. *Immigrants and Boomers*. New York, NY: Russell Sage Foundation.

National Center for Health Statistics. 1947. *Vital Statistics Rates in the United States: 1900 – 1940*. United States Government Printing Office: Washington, D.C. Available for download at https://www.cdc.gov/nchs/data/vsus/vsrates1900_40.pdf.

Park, J. and D. Myers. 2010. "Intergenerational mobility in the post-1965 immigration era: Estimates by an immigrant generation cohort method." *Demography* 47(2):369-392.

Piore, M. J. 1979. *Birds of Passage: Migrant Labor and Industrial Societies*. Cambridge, UK: Cambridge University Press.

Portes, A. and R. G. Rumbaut. 2006. *Immigrant America*. Berkeley, CA: University of California Press.

Portes, A. and M. Zhou. 1993. "The new second generation: Segmented assimilation and its variants." *The Annals of the American Academy of Political and Social Science* 530:74-96.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek. 2015. *Integrated Public Use Microdata Series: Version 6.0* [1950-2000]. Minneapolis: University of Minnesota.

Telles, E.E .and V. Ortiz. 2008. *Generations of Exclusion*. New York: Russell Sage Foundation.

University of California, Berkeley and Max Planck Institute for Demographic Research. 2016. *Human Mortality Database.* Available at www.mortality.org or www.humanmortality.de. Data downloaded on September 15, 2016.

Wagner, D. and M. Layne. 2014. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. *CARRA Working Paper #2014-01*.

Warner, W. L., and L. Srole. 1945. *The Social Systems of American Ethnic Groups.* New Haven, CT: Yale University Press.

Tables and Figures

Table 1. Demographic Characteristics of each Family Generation Observed and Linked Across Data Sources, 1940-2015.

Data Source:	1940 Census				1973, 1979, 1981-1990 CPS				2000 Census, 2001-2015 ACS	
Family Generation:	1st		2nd		2nd		3rd		3rd	
	Obs	Pct	Obs	Pct	Obs	Pct	Obs	Pct	Obs	Pct
Total	44,700,458	100	56,108,820	100	109,842	100	79,897	100	18,869	100
Sex										
Male	20,649,934	46.2	29,116,646	51.9	55,501	50.5	42,303	53.0	9,593	50.8
Female	24,050,524	53.8	26,992,174	48.1	54,341	49.5	37,594	47.1	9,276	49.2
Age										
0-4	272	0.0	10,285,421	18.3	-	-	3,355	4.2	-	-
5-9	147	0.0	10,358,978	18.5	-	-	8,333	10.4	-	-
10-14	17,380	0.0	11,182,545	19.9	5	0.0	17,825	22.3	7	0.0
15-19	382,966	0.9	10,573,388	18.8	12	0.0	26,065	32.6	39	0.2
20-24	2,548,747	5.7	5,912,291	10.5	20	0.0	16,011	20.0	167	0.9
25-29	4,964,530	11.1	2,890,124	5.2	48	0.0	5,106	6.4	825	4.4
30-34	5,991,849	13.4	1,722,647	3.1	1,810	1.7	1,875	2.4	2,263	12.0
35-39	6,159,934	13.8	1,172,669	2.1	4,400	4.0	823	1.0	3,694	19.6
40-44	5,822,453	13.0	812,645	1.5	10,060	9.2	306	0.4	3,969	21.0
45-49	5,333,622	11.9	560,821	1.0	19,613	17.9	104	0.1	3,634	19.3
50-54	4,278,459	9.6	345,229	0.6	22,869	20.8	41	0.1	2,532	13.4
55-59	3,094,007	6.9	178,690	0.3	20,658	18.8	27	0.0	1,204	6.4
60+	6,106,092	13.7	113,372	0.2	30,342	27.6	26	0.0	530	2.8
Race / Ethnicity										
Non-Hispanic White	40,419,771	90.4	49,264,828	87.8	100,488	91.5	69,582	87.1	16,751	88.8
Non-Hispanic Black	3,446,347	7.7	5,417,046	9.7	6,308	5.7	6,464	8.1	1,024	5.4
Other Non-Hispanic	155,980	0.4	262,798	0.5	567	0.5	717	0.9	369	2.0
Hispanic	678,360	1.5	1,164,148	2.1	2,479	2.3	3,134	3.9	725	3.8
Immigrant Generation										
1st	6,970,987	15.6	868,110	1.6	758	0.7	N/A	N/A	N/A	N/A
2nd, 2.5, or 2+	35,746,573	80.0	10,554,099	18.8	20,879	19.0	457	0.6	105	0.6
3rd or 3+	1,982,898	4.4	44,666,678	79.6	88,151	80.3	16,953	21.2	4,083	21.6
4th or 4+	N/A	N/A	19,933	0.0	54	77.4	62,441	78.2	14,670	77.8
5+	N/A	N/A	N/A	N/A	N/A	N/A	46	0.1	11	0.1
Origin										
Africa / Other	14,475	0.0	29,307	0.1	66	0.1	71	0.1	12	0.1
Mexico	498,652	1.1	921,618	1.6	2,061	1.9	2,589	3.2	546	2.9
Other Americas	179,965	0.4	257,810	0.5	386	0.4	427	0.5	99	0.5
USSR States & Asia	1,077,104	2.4	1,611,814	2.9	2,865	2.6	2,027	2.5	507	2.7
Central / Eastern Europe	2,537,764	5.7	3,955,158	7.1	7,139	6.5	5,635	7.1	1,432	7.6
Southern Europe	1,449,187	3.2	2,583,723	4.6	4,928	4.5	4,119	5.2	961	5.1
UK / Western Europe	575,348	1.3	855,309	1.5	1,966	1.8	1,329	1.7	380	2.0
Northern Europe	1,205,157	2.7	1,890,222	3.4	3,599	3.3	3,127	3.9	742	3.9
Canada	593,777	1.3	1,075,823	1.9	2,686	2.5	2,153	2.7	521	2.8
2+/3+ Non-Hispanic Other	88,035	0.2	149,460	0.3	294	0.3	355	0.4	79	0.4
2+/3+ Non-Hispanic Black	3,408,205	7.6	5,346,889	9.5	6,296	5.7	6,359	8.0	1,082	5.7
2+/3+ Non-Hispanic White	33,072,789	74.0	37,431,687	66.7	77,556	70.6	51,706	64.7	12,508	66.3
Mean Years of Education	8.0		9.2		12.1		11.4		14.0	
Highest Grade Completed										
Under Age 14	10,156	0.02	29,560,263	52.7	9	0.0	27,276	34.1	8	0.0
<= 8th grade	28,231,246	63.2	10,794,227	19.2	13,578	12.4	6,103	7.6	262	1.4
9-11th grade	7,202,167	16.1	8,245,645	14.7	15,711	14.3	17,608	22.0	906	4.8
12th grade	5,774,754	12.9	5,266,735	9.4	46,858	42.7	15,173	19.0	6,313	33.5
1-3 years college	2,060,715	4.61	1,544,536	2.8	15,511	14.1	10,537	13.2	4,517	23.9
4+ years college	1,421,420	3.18	697,414	1.2	18,175	16.6	3,200	4.0	6,861	36.4

Sources: 1940 Census; 1973, 1979, 1981-1990 Current Population Survey; Census 2000; 2001-2015 American Community Survey

Note: To prevent disclosure of individual identities, we insert a dash in cells that do not meet minimum record count requirements and round other cells in the same distribution.

Table 2. Weighted and Unweighted Years of Education of Respondent (G3), Respondent's Parent (G2) and Grandparent (G1).

	Weighted			*Unweighted*		
	ACS Respon dents (G3)	Parents in CPS (G2)	Grand-parents in 1940 Census (G1)	ACS Respon dents (G3)	Parents in CPS (G2)	Grand-parents in 1940 Census (G1)
NH White	13.9	12.5	8.5	14.1	12.9	9.0
NH African American	13.2	10.2	5.4	13.2	10.7	5.9
Mexican	12.6	9.7	4.1	12.9	10.2	4.6
Other Hispanic	13.6	12.1	7.4	13.2	12.0	7.4
Other	13.2	11.5	6.2	13.7	12.5	7.7

Sources: Authors calculations based on 1940 Census; 1973, 1979, 1981-1990 Current Population Survey; Census 2000; 2001-2015 American Community Survey

Sample: Grandchildren of 1940 1st Family Generation found in Census 2000 or the American Community Survey (N = 17,881)

Note: In cases in which we observed the educational attainment of more than 1 parent or grandparent, we took the average.

Table 3. OLS Regression Models Predicting the 3rd Family Generation Respondents' Educational Attainment (years).

	Model 1	Model 2	Model 3
Race/ethnicity			
African American	-0.72 ***	0.00	0.09
Mexican	-1.50 ***	-0.41 ***	-0.29 **
Other Hispanic	-0.50 **	-0.24	-0.19
Other	-1.00 ***	-0.59 **	-0.53 **
Number of foreign-born grandparents			
1	0.31 ***	0.22 ***	0.25 ***
2+	0.44 ***	0.35 ***	0.44 ***
Region			
Midwest	-0.17 **	-0.10 *	-0.10 *
South	-0.15 **	-0.10 *	-0.09
West	0.28 ***	-0.04	-0.06
Parental Education	---	0.34 ***	0.32 ***
Grandparents' education	---	---	0.04 ***
Intercept	14.05 ***	9.65 ***	9.50 ***
N	17,881	17,881	17,881
R-squared	0.021	0.175	0.177
Adj R-squared	0.021	0.174	0.176

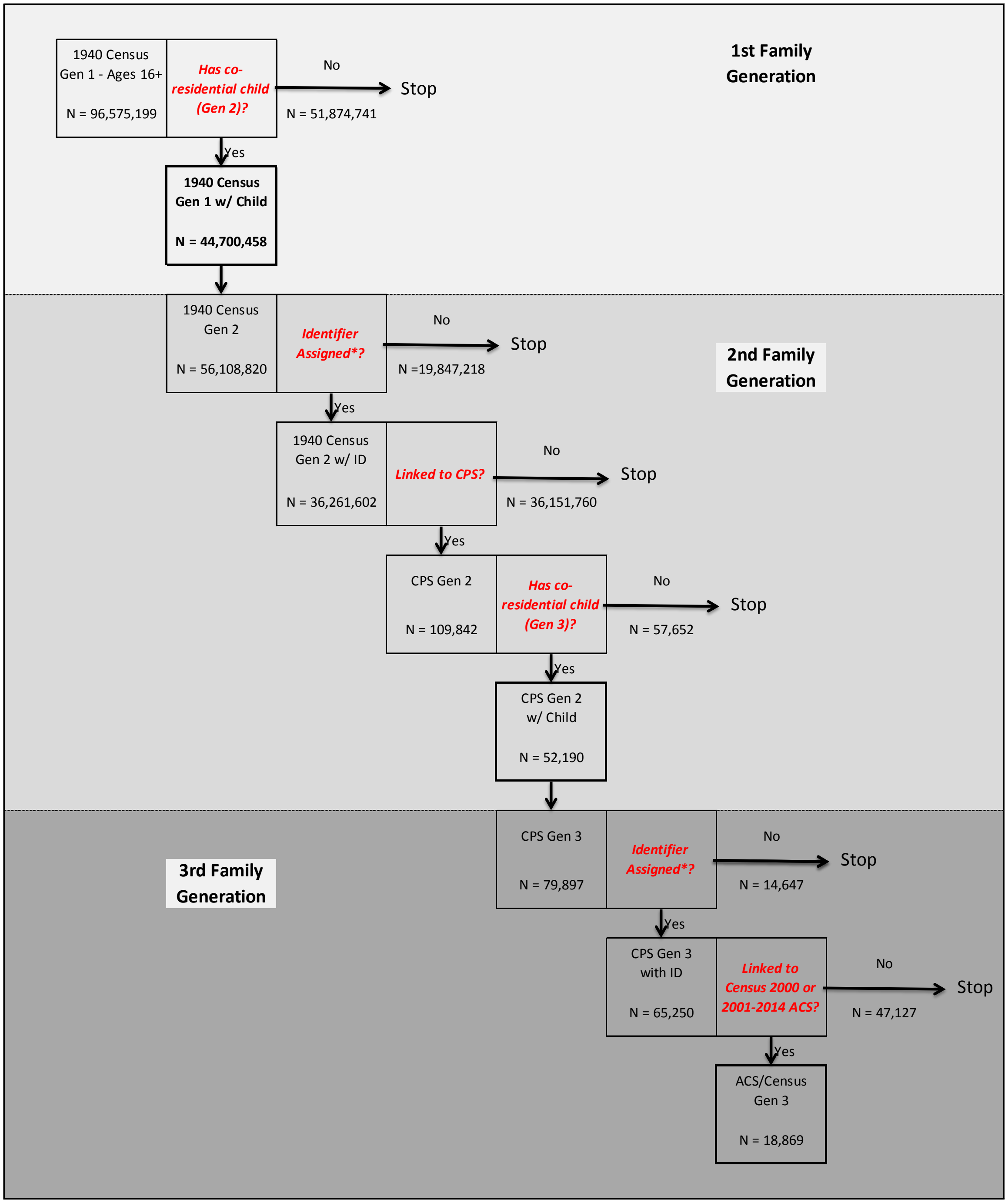
Sources: Authors' calculations based on 1940 Census; 1973, 1979, 1981-1990 Current Population Survey; Census 2000; 2001-2015 American Community Survey

Sample: Grandchildren of 1940 1st Family Generation found in Census 2000 or the American Community Survey (N = 17,881)

Note: In cases in which we observed the educational attainment of more than 1 parent or grandparent, we took the average.

* p-value<0.1; ** p<0.05; *** p<0.01

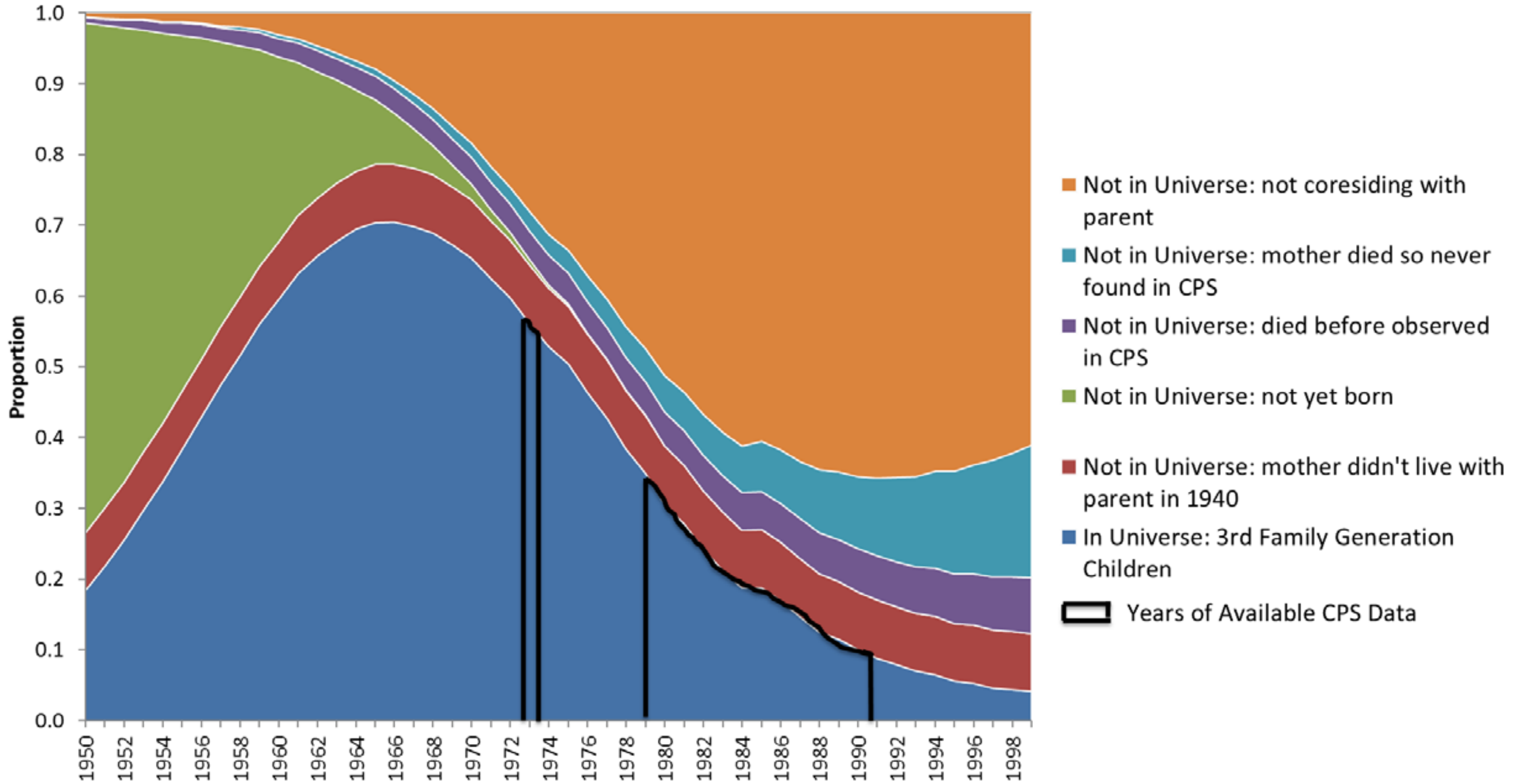
Figure 1. Data Sources and Number of Individual Linkages by Family Generation, 1940 to 2015.


1st Family Generation
1940 Census Gen 1 - Ages 16+ N = 96,575,199
Has co-residential child (Gen 2)?
No
Stop
N = 51,874,741
Yes
1940 Census Gen 1 w/ Child N = 44,700,458
2nd Family Generation
1940 Census Gen 2 N = 56,108,820
Identifier Assigned*?
No
Stop
N =19,847,218
Yes
1940 Census Gen 2 w/ ID N = 36,261,602
Linked to CPS?
No
Stop
N = 36,151,760
Yes
CPS Gen 2 N = 109,842
Has co-residential child (Gen 3)?
No
Stop
N = 57,652
Yes
CPS Gen 2 w/ Child N = 52,190
3rd Family Generation
CPS Gen 3 N = 79,897
Identifier Assigned*?
No
Stop
N = 14,647
Yes
CPS Gen 3 with ID N = 65,250
Linked to Census 2000 or 2001-2014 ACS?
No
Stop
N = 47,127
Yes
ACS/Census Gen 3 N = 18,869

Sources: 1940 Census; 1973, 1979, 1981-1990 Current Population Survey; Census 2000; 2001-2015 American Community Survey

* Census Bureau is able to assign a unique, protected identifier to an individual.

Figure 2. Simulation of Universe of 3rd Family Generation Children Born to Mothers in 1921-1940 Birth Cohorts.



Sources: Authors' calculations based on 1940 Census; *Vital Statistics Rates in the United States: 1900 – 1940* (National Center for Health Statistics 1947); *Human Mortality Database* (University of California, Berkeley and Max Planck Institute for Demographic Research 2016); *Integrated Public Use Microdata Series: Version 6.0* (Ruggles, Genadek, Goeken, Grover, and Sobek 2015).