**Using Administrative Records to Support the Linkage of Census Data**

J. Trent Alexander, University of Michigan, jtalex@umich.edu
Katie R. Genadek, U.S. Census Bureau and University of Colorado, katie.r.genadek@census.gov

*Acknowledgement of funding sources*
This work was supported by the National Science Foundation under grant number 2023639.

**Using Administrative Records to Support the Linkage of Census Data**


## Abstract

The Decennial Census Digitization and Linkage project (DCDL) is expanding an infrastructure that already includes linked data from the censuses of 1940, 2000, 2010, and a broad range of administrative records and survey data, by adding the 1960-1990 Census data. When complete, the DCDL will serve as a multi-purpose longitudinal resource enabling a wide range of new discoveries and applications. Our approach to creating linkages between censuses will be to integrate routinely-collected administrative records files with one another, and then use the resulting administrative records composite to facilitate the assignment of linkage keys to the less-often collected records from the 1960-1990 Censuses. By using the administrative records that are proximate to census data in time, we reduce the risk of variables changing over time, and we maximize the use of the particular linkage variables available during each census year. This work builds on methods used to assign anonymous linkage keys to the most recent decennial census data, as well as insights gleaned from research focused on the linkage of historical census records from 1850-1940. We describe how we are building and integrating the composite file of routinely collected records, our linkage of the census records to the composite, expected coverage rates, and the novel methods we will use to document and evaluate error.

## Introduction

The longitudinal linkage of decennial census records involves challenges that will be familiar to anyone who has linked records collected at different times. High-quality personally identifying information (PII) is often more limited in the past, and the available PII can vary greatly from one file to the next. Even when a variable is present in multiple datasets for decades, linkage plans will need to account for changes in respondents' lives. People may have changed their names, their family composition, their addresses, and even the way they describe where they were born. These challenges have contributed to relatively low rates of coverage and high rates of error and bias in linked historical census data, especially when compared to linkages across modern surveys and administrative data (e.g. Massey 2017; Massey et al. 2018; Bailey et al. 2020).

We describe a large-scale effort to create linkages between twentieth-century decennial censuses, and to integrate that data into a linked infrastructure of survey data and administrative records held at the U.S. Census Bureau. Focused on census data from 1960-1990, the Decennial Census Digitization and Linkage project (DCDL) is expanding an infrastructure that already includes linked data from the censuses of 1940, 2000, 2010, and a broad range of administrative records and surveys. Our approach will be to creating linkages between censuses will be to integrate routinely-collected administrative records files with one another, and then use the resulting administrative records composite to facilitate the assignment of linkage keys to the non-routine records from the 1960-1990 Censuses. By using administrative records that are proximate to census data in time, we reduce the risk of variables changing over time, and we maximize the use of the particular linkage variables available during each census year. Our work builds on methods used to assign anonymous linkage keys to the most recent decennial census

data, as well as insights gleaned from research focused on the linkage of historical census records from 1850-1940.

We begin by describing the administrative records that are available during this period, and how we will integrate those data into a single composite file. We then describe how we will use this composite to facilitate the linkage of the lower-information census data from 1960 through 1990. We outline our expected coverage rates, steps we will take to maximize the coverage and minimize error, and how we will estimate and document error in our links. We conclude by emphasizing how proximate administrative records can serve as an invaluable tool for making high-quality linkages between files that would have been otherwise very difficult to link.

## Methods

### Creating an administrative records composite

The DCDL project will adapt and expand well-established techniques to conduct record linkage across multiple decennial censuses. The core of the record linkage strategy will use a modified version of the Census Bureau's production record linkage processes, the Person Identification Validation System (PVS) (Layne and Wagner 2014; Alexander et al 2017; Massey 2017; Massey et al. 2018). We will use probabilistic matching methods similar to those developed for the already-completed linkages for the 1940, 2000-2020 Censuses. As was done with those files, we will assign unique identifiers known as "Protected Identification Keys" (PIKs) to the individual records in each census. A respondent's PIK in any given file can be used to locate the same respondent in other files that have been assigned PIKs. Since the data

available in each decennial census varies over time, we will adapt these methods to accommodate each census.

The Census Bureau's PVS process assigns PIKs to modern data by probabilistically linking the census or survey records with PII to a composite database created from trusted federal agency administrative records that also contains PII as well as each person's unique PIK value. Each PIK is associated with a Social Security Number (SSN) or an Individual Taxpayer Identification Number (ITIN). When a census or survey record is linked to the composite database, the record is assigned the relevant PIK from the composite. Respondent names are removed from the data and researchers are provided with a "PIK Crosswalk", which includes the survey's record identification numbers (such as a household ID number and a person ID number) and a PIK for each case that was successfully linked to the composite. Census and survey data with PIK crosswalks are made available to researchers on Census Bureau approved research projects within a restricted environment for research. We will follow an adapted version of PVS to assign PIKs to the 1960-1990 Censuses, and to make those data available in a restricted data environment.

Table 1 shows the administrative data sources to be integrated and used for the composite reference file for linking the 1960-1900 Censuses, as well as the variables available to support the linkage and assignment of PIKs in each data set. The core components of the composite include the Social Security Administration (SSA) Numerical Identification file (the "Numident" file), the SSA Master Beneficiary Record (MBR), and the Internal Revenue Service (IRS) 1040 tax return data. The three primary administrative data sources in the composite will be linked to one another using the SSNs on each file. Though the SSN links are expected to be high quality, names will be used to validate the links when possible.

Table 1. Data sources used to construct the administrative records composite file

| | Census Numident 1936-present | IRS 1040 forms 1969 | IRS 1040 forms 1974, 1979, 1984, 1989, 1994 | SSA Master Beneficiary Record 1962-present |
|---|---|---|---|---|
| Universe | all SSN holders | all tax filers | all tax filers and spouses | SSN holder who earned benefits, child or spouse receiving benefits |
| Name | x | | x | |
| Alternate Name | x | | | |
| Date of Birth | x | | | |
| Date of Death | x | | | |
| Sex | x | | | |
| Marital Status | | x | x | x |
| State or country of birth | x | | | |
| Parents' names | x | | | |
| Street address | | x | x | |
| County | | x | x | |
| SSN | x | x | x | x |
| Spouse's SSN | | | x | x |
| Child's SSN | | | | x |

Most of the composite's demographic information comes from the SSA Numident, which is built from applications for a SSN. Since the SSA Numident is provided as a quarterly transactions file that includes all new applications as well as changes to existing SSNs, the Census Bureau has standardized this flow of data into a single snapshot file called the "Census Numident." The Census Numident contains basic information on all SSN-holders going back to the beginning of the program in 1936, also including any updates the SSA received, such as new names and even dates of death (when reported to the SSA). The variables that we include in the composite include SSN, name, alternate/married name, sex, date of birth, state or country of birth, parents' names and—to exclude ineligible candidate links—date of death.

The composite's information on addresses comes from the IRS 1040 tax returns. These data include the complete universe of IRS 1040 tax return files, obtained by the Census Bureau from the IRS shortly following each tax year. The Census Bureau holds these data for every fifth year from 1969 through 1994, and every year from 1995 to the present. The files contain the

original data transferred from the IRS and have information from the IRS 1040 files including SSN, marital status, mailing address, and—for 1974 forward—all of the same information as well as individual filers' and spouses' names. The IRS 1040 data files prior to 1995 were modified at the time by the Census Bureau to include a county variable. Our use of these files will focus on 1969, 1979, and 1989, because those files were created at almost exactly the same time as each census was conducted. For instance, 1969 IRS 1040 file includes addresses from tax returns that were likely filed on and before April 15, 1970, which is nearly the same time that the 1970 Census was conducted. Unlike the modern IRS 1040 data, the 1969-1994 IRS 1040 data do not include identifying information for household dependents. For 1969, there are only SSNs for the filer, or head of household, with no names provided. For 1974-1994 there are SSNs *and* names for the filer and—for joint returns—the filer's spouse. None of these files include information for dependents or children in the filers' households.

The third file we will use to construct the administrative records composite is the SSA's Master Beneficiary Record (MBR). The MBR is particularly useful for specifying family relationships. The MBR provides information on Social Security benefit recipients, including the SSN of the person who originally earned the benefit providing the basis for payment. Since the earner and beneficiary are usually related, this file often effectively provides an SSN-to-SSN link between two spouses or between a parent and child. These relationships are explicitly indicated by a separate code on the record. These data provide an additional record connecting potential couples within a household or parent and children pairs in a household. The MBR contains information on individuals who received benefits since 1962, which will reveal family relationships for many older respondents in the 1960-1990 Censuses.

As a single integrated file, the composite will contain all the information available for each person, as well as additional constructed variables to make the information comparable to each year's decennial census data. For example, the composite's street address and county variables are converted to census geographies, specifically census tract, for each of the censuses using a variety of methods. The composite's date of birth variable will be converted to year of birth and quarter of birth, again to maximize comparability with the census data which does not include exact birth date. The SSN information in the MBR will be enriched with individual names from the Census Numident, including names for benefit recipients, as well as their spouses and parents who are specified MBR. Finally, the variables will be harmonized so all data have the same codes and formats to facilitate the linkage. The resulting file will contain one row for each person ever observed in the administrative records, with all available information about that person and any observed family members.

**Linking the 1960-1990 Censuses to the administrative records composite**

Table 2 shows the linkage variables that can be used to probabilistically link census records to the administrative records composite file. The variables available vary only slightly by census year. For all records across all years, the linkage to the composite file will primarily rely on name, age, sex, and place of residence. For children living with a parent, parents' names will be used to link to the composite. For couples, the marital status variable and the spouse's name will be used for the linkage. In all years other than 1990, the use of quarter of birth will be investigated for matching, though birthdays are often misreported in survey data (Larsen et al. 2019). The 1990 Census contains full addresses for most respondents. In the 1960-1980, census

geographies (such as census tract) and county information will be used to aid in the individual level linkage to the composite.

For each census, about one-sixth to one-fourth of households were asked to complete additional questions via the long form survey. For the subset of census respondents who received the long form questionnaire, the linkage to the composite will also use the state or country of birth variable. This variable is also available in the Census Numident file, so it is a useful variable for linking the census respondents to the composite file, even though it will only be available for some cases.

Table 2. Linkage variables in 1960-1990 Censuses and the administrative records composite file

| | Decennial Censuses Short Forms 1970-1990 | Decennial Censuses Long Forms 1960-1990 | Administrative Records Composite |
|---|---|---|---|
| Universe | all U.S. residents | 16-25% sample of U.S. residents | Varies by field |
| Name | x | x | x |
| Age/Year of Birth | x | x | x |
| Quarter of birth | | 1960-1980 | x |
| Sex | x | x | x |
| Marital Status | x | x | x |
| State or country of birth | | x | x |
| Parents' names | x* | x* | x |
| Street address | 1990 | 1990 | x |
| Census Tract | x | x | x |
| County | x | x | x |
| SSN/PIK | | | x |

*available for respondents living with parents

The 1960-1990 Censuses have a large number of linkage variables, and it may seem that we would have success linking one census to another without the use of an administrative records composite. With censuses that are 10 years apart, however, the power of the composite made from administrative data is that it includes year-specific versions of information that often

changes over time. For instance, the place of residence variables available in each census will be linked using the composite's time-specific location information from the IRS 1040 data. Without this information, the censuses' geographic variables would only be useful for linking those who had not moved within a 10-year period. Similarly, the composite contains all name variants that were ever reported to the SSA, including names that were changed upon marriage or for other reasons. Without the composite, the censuses' name variables would only be useful for those who had never changed their names.

**Increasing PIK assignment in the 1960-1990 Censuses**

As described above, the linkage strategy consists of linking one file (the census) to an integrated administrative records composite, using well-established probabilistic matching techniques with low-tolerance differences between variables in the two files. These techniques produce robust linkages that have already been used to produce research in top-tier journals across the social sciences, as well as in production record linkage applications within the Census Bureau (e.g., Alexander et al. 2017; Song et al. 2020; Ferrie et al. 2021; Lowrey et al. 2021; Miller et al. 2021; Stevenson et al. 2021). For a large majority of the population, this approach will continue to work well. However, to reach an even greater proportion of the population, we will need to use new techniques.

The first approach for broader coverage will take advantage of known family links from available data. For instance, consider a married couple in the tax data where we link one spouse to the census but not the other. In this case, we will make another pass at the census family, considering the unlinked spouse from the tax data as a candidate link for the unlinked spouse in the census. We will carry out similar processes with other available data where we observe spousal links, child-parent links, and sibling links, such as the Medicare Enrollment Database,

and data such as censuses preceding and following the census being linked and proximate household survey data.

The second approach to improve coverage in the linkage of DCDL records will use new techniques for entity resolution to assign PIKs to respondents who were not initially linked to the administrative records composite. About 10% of respondents in 2000 were not assigned PIKs, and we expect a larger percent of census respondents in the earlier census years not to have PIKs assigned through the processes described. Considering only unlinked individuals from the censuses, as well as those from administrative records who were never linked, we will look for links over time and between censuses and the other records. Rather than a file-to-file linkage, we will approach these sets as a large deduplication problem (e.g. Steorts 2015, Steorts et al. 2016). When we observe duplicate records in this set of data, we will create a new, unique PIK value that is not associated with an SSN or ITIN.

## Results

### Expected linkage rates

For the 2000 and 2010 Censuses, about 90% of cases were linked to the administrative records composite via PVS. For the 1850-1940 Censuses, linkage rates across 10-year periods are typically around 10-45% for native-born men, and much less for the whole population (Goeken et al., 2011; Helgertz et al. 2022; Massey et al. 2018). In order to predict what rate of coverage we can expect to see in the 1960-1990 Censuses, we consider work that we and others have done with the entire 1940 Census as well as pilot projects from the 1960 and 1990 Censuses. Drawing on these examples, we expect linkage rates of close to 90% for the 1990 Census, and 70% or more for the 1960-1980 Censuses.

The assignment of PIKs to the 1940 Census, which used the primary linkage variables available for the 1960 through 1990 Censuses but without geographic information, produced PIKs for 40% of adults and 70% of children (Massey et al. 2018). Each of the 1960-1990 Censuses has additional features that will enable even higher linkage rates than those obtained for 1940. The use of the residential geographic variables for the 1970-1990 Censuses should increase the linkage rates to the administrative records composite. For the 1990 Census, we will directly match street address from the 1989 tax data to street address from the 1990 Census records. The 1970 and 1980 Censuses do not have street addresses. For these years, we will use the low-level geographic identifiers available on the census files (tracts and county subdivisions) to match to the census geography information generated from the address information from tax records in the composite.

We also gain insights from a pilot project conducted with about 1,500 hand-entered records from the 1960 Census. In that work, about 75% of records were successfully linked to administrative records (Massey 2014). That work also suggested that the quarter of birth variable significantly improved linkage results beyond what would have been achieved using just the age variable (Massey 2014). Quarter of birth is available in the 1960-1980 files, and it was not available in 1940. Used in conjunction with age and year of birth, this variable will likely allow for a finer-grained match to the exact date of birth variable in the administrative records composite file.

Finally, a pilot project from the 1990 Census also suggests we may expect relatively high linkage rates. In that study, PIKs were assigned to a broadly representative sample of 28,000 records collected as part of the 1990 Content Reinterview Survey (CRS), the only 1990 Census records for which digitized respondent names were available (Johnson et al. 2015). The 1990

CRS linkage team assigned PIKs to 84% of cases, including 87% of adults and 73% of children. We expect higher rates of PIK assignment to the full 1990 Census file, because the CRS project did not use the residential addresses from the 1990 Census file or the 1989 1040 tax data in the linkage.

**Estimating the quality and error in the 1960-1990 PIK assignment**

Research based on census data from 2000 forward suggests that error in PIK assignment is minimal; about 90% of cases have PIKs assigned, and nearly all assigned PIKs seem to be accurate (Layne et al. 2014). Even with this high rate of coverage, PIK coverage varies significantly by key population characteristics. Since PIKs are currently based on SSNs and ITINs, respondents who do not have SSNs or ITINs cannot receive PIKs. The PIKs assigned to recent censuses and surveys have been shown to under-represent the highly mobile population, those with limited English proficiency, non-citizens, and other population sub-groups that are less likely to appear in federal government administrative records. Linkage rates also vary by geography, and are lowest in the Southwest, and highest in the upper Midwest (Rastogi and O'Hara 2012, Bond et al. 2014, Layne et al. 2014).

Linkage quality issues are likely to be greater in the 1960-1990 Census data than in the more recent decennial census data. The historical linkages will be made with less precise linkage variables. For example, the 1960-1990 Census do not include exact date of birth. Since the administrative records composite contains full date of birth, we need to convert (and reduce) this information into variables for age (at the time of the decennial census), year of birth, and quarter of birth. By contrast, the censuses of 2000-2020 contained exact date of birth and are thus able to take advantage of the full birthdate information from the composite. The PIK assignment for

1960-1990 Censuses will also likely rely on less precise name data, since names recovered using OCR in 1960-1990 may be lower-quality than the names recovered using OCR with clerical edits in 2000-2020 (Alexander et al. 2021; U.S. Census Bureau 2003, 2009a, 2009b, 2017). Thus, the PIKs assigned to the 1960-1990 Censuses are likely to cover fewer records and have more errors than those from the 2000-2020 Census.

We will carry out two strategies for documenting linkage quality. Our first strategy will be to replicate each year's linkage approach with the 2000 Census data, and then compare PIKs generated with the full information available in the 2000 Census to PIKs generated with the more limited information that was available in the 1960-1990 Censuses. This will follow a similar strategy used for assessing the quality of PIKs assigned to the 1940 Census (Massey et al. 2018). While the 2000 Census PIKs were assigned probabilistically and are therefore not real "truth data," they were assigned with more information than the 1960-1990 PIKs will be. This comparison will help us understand the cost of making these lower-information linkages, and to quantify differences in coverage and error between the 2000 Census PIK assignment and the pre-2000 PIK assignment.

For example, to estimate the quality of the 1960 links, we will conduct a new linkage with the 2000 Census file. We will begin by removing the exact date of birth from the 2000 data, replacing it with the level of information available in 1960 (age, year of birth, and quarter of birth). If our analysis of 1960's OCR-based name data shows that those names were of lower quality than name data from 2000 (when compared to the hand-entered "truth data" from each year), we will degrade the quality of the 2000 Census names accordingly. Finally, we will create an administrative records composite consisting only of records from the Numident file, since that is the only administrative records file available in 1960. Using this modified base of 2000 Census

data and administrative records, we will carry out the record linkage processes we used for the 1960 Census. We will then compare the "limited information" links made with the 2000 Census data to the "full information" links we had originally made for the 2000 Census. We will use similar assessments for all years to generate new information on linkage quality in the lower-information methods we use for the older censuses.

Our second strategy for documenting linkage quality will use a new "ground truth" file based on the Panel Study of Income Dynamics (PSID). The PSID is the longest-running longitudinal household survey in the world and has been conducted at the University of Michigan's Institute for Social Research since its inception. The survey's leadership team is in the process bringing these data into the Census Bureau's Data Linkage Infrastructure. Initiated with 18,000 individuals in 5,000 families in 1968, PSID respondents were interviewed every year through 1997, and then every two years after that. The PSID has collected SSNs for about 5,000 respondents who have either died or who receive Medicare.

For PSID cases with known SSNs, we will assign PIKs by linking respondents' SSNs to the administrative records composite file and validating link with name and date of birth. These PIKs will present us with near-certain truth data on individuals who, in many cases, were interviewed consistently for 50 years. Using the PSID interviews that are most proximate to the 1970-2010 Censuses, we will remove the PII information (SSN and exact date of birth) from the PSID to make the PSID have the same PII as the proximate census data. We will create "Evaluation PIKs" on the PSID by carrying out our census linkage procedures on these lower-information PSID records. When we compare a PSID respondent's truth PIK (based on the SSNs and additional PII in the PSID) to the same individual's Evaluation PIK (based on lower-information PII from PSID interviews proximate to census years), we will be observing the

quality of the 1970-1990 linkages. We will also benefit from knowing that the PIKs on PSID data should not change over time, since they were verifiably the same individual who was interviewed by the PSID in (approximately) 1970, 1980, 1990, 2000, and 2010. When we observe an Evaluation PIK on a PSID respondent changing over time, we will be observing errors in the late-twentieth century linked data.

Our investigations of linkage quality will have a feedback effect that informs our core linkage techniques. For instance, when we identify subsets of the low-information links from 1960-1980 that are subject to significant error, we will attempt to fine-tune tolerances in the linkage process, such as the string comparator scores or age differences. Our linkage experiments using the 2000 Census and the PSID will thus serve both a documentation purpose—describing the procedures we followed and the error we observed—as well as providing us with feedback and strategies to improve the record linkage processes for Censuses of 1960-1990.

## Conclusion

Survey data and other non-routine collections are relatively expensive to create and rich with content, whereas routine administrative records are routinely collected and are therefore relatively inexpensive, but also usually narrow in content. In addition, administrative records have ample high-quality PII that is used to administer social programs, whereas survey records have less complete PII. We leverage another distinction between these data types in the late-twentieth century United States. We leverages these differences to support the linkage of the substantively rich files from the 1960-1990 Censuses. In order to mitigate the impact of change over time in key variables such as name and place of residence, we have proposed to use proximate administrative records to facilitate the linkages.

Through the DCDL project, the Census Bureau is currently in the process of digitizing respondent names from the 1960-1990 Censuses (Genadek and Alexander, 2019). While all other information from these censuses was digitized shortly after each census was conducted, respondent names were never digitized and are only available on microfilm images of the original census forms. Without the 1960-1990 linked into the Census Bureau's data linkage infrastructure, there is a sizable gap in the 20th century where longitudinal research cannot take place. When the DCDL digitization work is complete, we will apply the described methods of combining these censuses to administrative data to create a massive longitudinal infrastructure of census data with significantly more coverage and less error than if we simply tried to link each census to the others.

**References**

Alexander, J. Trent, Jonathan Fisher, and Katie Genadek (Forthcoming). Digitizing Hand-Written Data with Automated Methods: A Pilot Project Using the 1990 U.S. Census. *Journal of Economic and Social Measurement*. Also available as Alexander, Fisher, and Genadek, *Associate Director for Economic Programs Working Paper Series,* 2021-06.

Alexander, J. T., Leibbrand, C., Massey, C., & Tolnay, S. (2017). Second-Generation Outcomes of the Great Migration. *Demography*, *54*(6), 2249-2271.

Bailey, M. J., C. Cole, M. Henderson, and C. Massey. (2020). How well do automated methods perform in historical samples? Evidence from new ground truth. *Journal of Economic Literature* 58 (4):997–1044.

Bond, B., Brown, J. D., Luque, A., & O'Hara, A. (2014). The nature of the bias when studying only linkable person records: Evidence from the American Community Survey. *Center for Administrative Records Research and Applications Working Paper*, *8*.

Ferrie J, Massey C, Rothbaum J. Do Grandparents Matter? Multigenerational Mobility in the United States, 1940–2015. *Journal of Labor Economics* [Internet]. 2021 Jul 1 [cited 2022 Apr 26];39(3):597–637.

Genadek, K. R., & Alexander, J. T. (2019). The Decennial Census Digitization and Linkage Project. *Census Bureau Working Paper, No ADEP-2019-01*

Goeken, Ron, Lap Huynh , T. A. Lynch & Rebecca Vick (2011) New Methods of Census Record Linking, *Historical Methods*, 44:1, 7-14.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles & Catherine A. Fitch (2022) A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel, *Historical Methods* 55:1, 12-29,

Johnson, D. S., Massey, C., & O'Hara, A. (2015). The opportunities and challenges of using administrative data linkages to evaluate mobility. *The ANNALS of the American Academy of Political and Social Science*, *657*(1), 247-264.

Larsen AF, Headey D, Masters WA. Misreporting Month of Birth: Diagnosis and Implications for Research on Nutrition and Early Childhood in Developing Countries. *Demography*. 2019;56(2):707-728.

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

Linkage Software. *Center for Administrative Records Research and Applications Working Paper, 1*.

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

Lowrey K, Van Hook J, Bachmeier J, Foster T. Leapfrogging the Melting Pot? European Immigrants' Intergenerational Mobility across the Twentieth Century. *SocScience* [Internet]. 2021 [cited 2022 May 20];8:480–512.

Massey, C. G. (2014a). *Creating linked historical data: An assessment of the Census Bureau's ability to assign protected identification keys to the 1960 Census* (No. 2014-12). Center for Economic Studies, US Census Bureau.

Massey, C. G. (2017). Playing with matches: An assessment of accuracy in linked historical data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *50*(3), 129-143.

Massey, C. G., Genadek, K. R., Alexander, J. T., Gardner, T. K., & O'Hara, A. (2018). Linking the 1940 US Census with modern data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(4), 246-257.

Miller S, Johnson N, Wherry LR. Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data. *The Quarterly Journal of Economics* [Internet]. 2021 Jun 30 [cited 2022 Apr 26];136(3):1783–1829.

Rastogi, S. and Amy O'Hara. (2012). 2010 Census Match Study. *2010 Census Planning Memoranda Series*, #247.

Song X, Massey CG, Rolf KA, Ferrie JP, Rothbaum JL, Xie Y. Long-term decline in intergenerational mobility in the United States since the 1850s. *Proc Natl Acad Sci U S A*. 2020 Jan 7;117(1):251–258.

Steorts, R. (2015). Entity Resolution using Empirically Motivated Priors, *Bayesian Analysis*, 10(4) 849–875.

Steorts, R., R. Hall, and S.E. Fienberg (2016). A Bayesian Approach to Graphical Record Linkage and Deduplication, *Journal of the American Statistical Association*, 111(516): 1660-1672.

Stevenson AJ, Genadek KR, Yeatman S, Mollborn S, Menken JA. The impact of contraceptive access on high school graduation. *Sci Adv.* 2021 May;7(19):eabf6732.

U.S. Census Bureau (2003) “Census 2000 Data Capture.” Census 2000 Testing, Experimentation, and Evaluation Program: Technical Report Series, No. 3.

U.S. Census Bureau (2009a) History: 2000 Census of Population and Housing (Volume 1). U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2009b) History: 2000 Census of Population and Housing (Volume 1). U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2017) 2020 Census Detailed Operational Plan for: 10. Paper Data Capture (PDC) Operation. U.S. Government Printing Office, Washington, DC.

Wagner, D., & Layne, M. (2014). The person identification validation system: Applying the Center for Administrative Records and Research and Applications’ record linkage software. *Center for Administrative Records Research and Applications Report Series (# 2014-01).*