THE MEASUREMENT OF INCOME GROWTH, MOBILITY, AND VOLATILITY IN THE U.S. BY RACE-ETHNICITY AND GENDER

INCOMPLETE DRAFT – NOT FOR DISTRIBUTION – DO NOT CITE OR CIRCULATE

Maggie R. Jones, Adam Bee, Amanda Eng, Kendall Houghton, Nikolas Pharris-Ciurej, Sonya R. Porter, Jonathan Rothbaum, and John Voorheis[1]

Abstract

Federal statistical agencies and policymakers have identified the need for integrated systems of household and personal income statistics. This interest marks a recognition that aggregated measures of income, such as GDP or average income growth, tell an incomplete story that may conceal large gaps in well-being between different types of individuals and families. Until recently, income data that are rich enough to calculate detailed income statistics, and that also include demographic characteristics such as race and ethnicity, have not been available. The MOVS project proposes to fill this gap using linked demographic and tax records on the population of U.S. working-age adults. We define households and calculate household income, applying an equivalence scale to create a personal income concept, and then trace the progress of incomes by individuals over time. We then output a set of intermediate statistics by race-ethnicity group, gender, year, and base-year state of residence and income decile. We selected our intermediate statistics as those most useful in developing more complex intragenerational income mobility measures, such as transition matrices, income growth curves, and variance-based volatility statistics. We provide these intermediate statistics as part of a publicly released data tool with downloadable flat files. This paper describes the data build process and the output files, and it showcases several use cases for the data tool from simple to complex, including an analysis of absolute income growth and upward and downward relative mobility.

[1]All authors: U.S. Census Bureau; corresponding author: Maggie R. Jones, *margaret.r.jones@census.gov*. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7511151: CBDRB-FY23-CES014-008, CBDRB-FY23-CES014-014, CBDRB-FY23-CES014-048)

1. Introduction

The estimation of statistics on income and earnings growth in the U.S. context has long presented multiple challenges for researchers. One core challenge stems from the lack of access to precise income information at the person or household level due to privacy and confidentiality concerns. While the U.S. is not necessarily unique in prioritizing privacy and confidentiality over social transparency when it comes to income, these priorities stand in contrast to those of some other industrialized countries, where information on income at the person level is more easily attained and occasionally even publicly reported.[2]

Previous research on U.S. income and earnings growth has made some headway through the use of repeated cross-sectional survey data. However, truly understanding mobility as it relates to personal and family well-being requires repeated measures on the same individuals over a long period. Key questions include: What does income growth look like at each point in an initial income distribution? How persistent is a person's position in the distribution? How do mobility patterns vary by demographic group? Until recently, data to answer these questions for the U.S. have been available only as survey data, which suffers from misreporting, small sample sizes, and attrition; in addition, much available survey data provides short time frames, multiple-year gaps, or both.

The increasing use of administrative records has led to improvements in population coverage and researchers' ability to incorporate repeated measures. In the U.S. context, however, administrative data that capture earnings and income often do not contain information on age, gender, and race-ethnicity. Because these characteristics are correlates of income—in terms of levels, inequality, and growth—any analysis of U.S. mobility would be incomplete without taking these characteristics into account.

The Mobility, Opportunity, and Volatility Statistics project (MOVS) uses administrative records and demographic data on a near population of working-age adults to address these challenges. The data-linkage infrastructure at the U.S. Census Bureau allows us to link persons over data sources and time, leading to annual populations of working-age persons whose income can

[2] Sweden is the best example of this; see, for example, Roine and Waldenström (2008).

be tracked over many years. For each person, these microdata contain age, gender, detailed race and Hispanic origin, household structure, and various levels of geography. MOVS is a publicly available suite of intermediate statistics, calculated within cells defined by income decile in each base year, gender, race-ethnicity, year, and geography. The resulting statistics disseminate valuable information to the public on patterns of income mobility and volatility. Moreover, released as a data tool and sets of flat files, MOVS allows researchers to make their own normative decisions on how to express aggregate income growth, volatility, and mobility measures; these intermediate statistics will also be valuable in their own right for geographic-level analyses.

The first available data year is base year 2005. The MOVS project plans to release state-level files continually for new base years with a static age definition (so that a new cohort enters and the oldest leaves). We also expect to release files back in time when key earnings data become available for years prior to 2005. In the meantime, development is underway for supplemental files, described in section 4, and metro-level files. MOVS expects to offer all files as part of the data tool, hosted on Census.gov, starting with base year 2005.

This paper describes the process of building these data for 2005, outlines the associated data sets that make up the intermediate statistics package, and introduces the data tool. We also provide some national- and state-level preliminary measures and graphs to demonstrate fitness for use of the intermediate statistics. In what follows, we describe the background and cover some of the important literature that has informed this project; we describe the data, covering the linkage process and sources of key information; we introduce the intermediate statistics file; and we produce and interpret some final mobility analyses created from the intermediate statistics.

2. Background and previous literature

Previous research discusses three key ideas in the realm of income measurement: inequality—the variation in income between the lower and upper parts of the distribution; mobility—how often or how far individuals or households move between the various parts of the distribution; and volatility—the number and magnitude of income changes, upward and downward, for a given household or individual over some period. These concepts may apply to an individual or household within the course of its existence (*intra*generational); or evaluated between a household or individual and their offspring (*inter*generational). Throughout this section, we briefly define these concepts for the sake of clarity, but note that considerable ongoing work exists seeking to precisely and thoroughly define them, the full nuances of which are discussed in other work (see Burkhauser and Couch (2011)).

Income inequality is a fairly straightforward concept, where some sort of distance metric (such as the 90/10 ratio) defines how "equal" or "unequal" a society is. Survey data has long been used to measure inequality (Reardon and Bischoff, 2011; Rose, 2016; Snipp and Cheung, 2016). A particularly robust literature using administrative records alone or in tandem with survey data or aggregates describes income inequality in the U.S. by documenting its increase over time (DeBacker et al., 2013; Piketty et al., 2018) after a period of decrease in the first part of the 1900s (Piketty and Saez, 2003; Kopczuk et al., 2010). Recent research demonstrates that income inequality and race-ethnic identity are intertwined, with Black non-Hispanic and American Indian/Alaska Native non-Hispanic adults faring especially poorly in terms of persistence at the bottom of the income distribution (Akee et al., 2019). Recent work suggests that accounting for the intricacies of the tax code may change these findings (Auten and Splinter, 2022), which has implications for the use of administrative records in our research.

Measuring income mobility presents researchers with conceptual challenges. In decisions over what "matters" when we think of income growth patterns, researchers have come to distinguish between shifts in market income that reflect changes in the overall structure of the distribution ("structural mobility") or changes in each individual's position in the distribution ("exchange mobility").[3] Although a swapping of position in a given income distribution can provide important information on how one person might be faring (to another person's detriment), it fails to provide information on absolute growth in income and how growth is shared out among individuals. For either measure, research suggests a persistence to disadvantage: the likelihood of an individual moving between income percentiles depends on percentile of origin—roughly half of those in the bottom quintile stay there (Auten and Gee, 2009; Auten et al., 2013). MOVS provides the tools to measure both types of mobility.

[3]Similarly, volatility may be reported using changes in absolute income or changes in rank within the distribution, both of which are best interpreted in the context of the growth of the overall economy and the various points in the distribution.

Researchers have long used tax data to examine patterns of income mobility and volatility, although these studies have necessarily been limited in their ability to examine correlates such as race and family structure (Kopczuk et al., 2010; Auten and Splinter, 2022). Recent work has expanded these analyses in innovative ways by examining pseudo-households (Larrimore et al., 2020) or superimposing demographic information onto administrative aggregates (Piketty et al., 2018). This research represents a considerable improvement over studies that use tax records alone, but may still tell an incomplete story due to the lack of individual-level information on crucial demographic characteristics.

Key papers that inform our research in this area include those that use either survey data alone (Bloome and Western, 2011; Bloome, 2014; Van Kerm, 2009) or in combination with administrative records (Akee et al., 2019). The recent ability to link administrative data, typically tax records, with demographic data at the individual level has generated a series of new papers studying these concepts. There appears to be substantial variation in mobility trends by racial and ethnic group. Black and American Indian/Alaska Native individuals have both higher rates of downward mobility and lower rates of upward mobility (Akee et al., 2019). Similarly, White and Asian individuals have a disproportionately large share of income in top quantiles and lower levels of representation in bottom quantiles (Akee et al., 2019). This work suggests that aggregate measurement of mobility, volatility, and growth do not capture the complete picture of income trends in the United States.

The presence of increasing or decreasing volatility in income has been a source of much debate (Dynan et al., 2012; Hardy and Ziliak, 2014; McKinney and Abowd, 2022; Moffitt and Zhang, 2018; Carr and Wiemers, 2018). Moffitt et al. (2022) explores how differing data sources form the source of this debate. Outside of trends over time, it is clear that household income is far from constant, as “4 in 10 people spent at least one year in poverty between 2007 and 2018” (Larrimore et al., 2020) and “half of all working-age adults—and 64 percent of low-income, working-age adults—have household income that for at least one month of the year will spike above or dip below 25 percent of their average monthly income” (Maag et al., 2017). The volatility statistics we produce, including arc percent changes and variance estimates, are intended to shed more light on this debate.

We highlight two additional features of mobility measurement that relate to the MOVS project. First, "mobility" may have a different meaning when data are cross-sectional versus repeated panels. Second, measures of mobility that do not take into account the direction of movement ("upward" or "downward") by different groups may miss patterns important to assessing overall well-being. Although, "greater mobility in the sense of less association between origins and destinations has long been linked with having a more open society and greater equality of opportunity" (Jäntti and Jenkins, 2015), if one group displays significant downward mobility and another displays significant upward mobility of the same magnitude, summarizing this information as though the two groups have equivalent mobility loses important nuance. In this paper, we center our provision of intermediate statistics as a way to explore how observed changes in mobility have complex implications when considering the direction of mobility for each demographic group.

Finally, our work contributes to the recent innovative work from other federal institutions estimating income statistics like year-to-year volatility and its time trend (Dahl et al., 2007), income inequality (Hungerford, 2011; DeNavas-Walt and Proctor, 2015; Fixler et al., 2020), and definitions of income class (Elwell, 2014). These new data products constitute a move to create official federal statistics where none have been available previously.

3. Data

3.1. Sources and linkage

The process of the MOVS project is, first, to assemble annual, household size-adjusted market income information for the population of working-age adults in the U.S.; and, second, to report out the intermediate statistics that form the components of common mobility measures.[4] We bring a variety of data to the table to accomplish these goals, including the 2000 decennial census; the American Community Survey (ACS); Internal Revenue Service forms 1040, W-2, and 1099; files from the Department of Housing and Urban Development (HUD); the Census Bureau's master list of social security numbers (the Numident); the Master Address File (MAF); and further

[4]For example, rather than producing transition matrices, we produce origin deciles and destination quintiles, plus underlying counts, so that external researchers can produce their own.

address history and parent-child linkage files built from survey and administrative records. Appendix A details each data source and how we use each.

The Census Bureau processes each of these files via the Person Identification Validation System (PVS) (Wagner and Layne, 2014) in order to place a unique identifier, called a Protected Identification Key (PIK). This identifier is invariant within person over time, allowing us to match all data sources and years at the individual level. PVS takes key variables—social security numbers, names, dates of birth, and so forth—in each dataset and compares them against a master reference file to place the PIK.[5] Upon linking datasets, what results is a longitudinal panel of U.S. working-age adults for whom we have income and household information from 2005 to 2019.

We start by defining the analysis population. Depending on the data source and vintage, PVS succeeds in placing a PIK 84 percent to nearly 100 percent of the time in the data we use to build the population. First, using the 2000 decennial census, we collect all individuals who were born between 1955 and 1980, which gives us workers between the ages of 25 and 50 in 2005 (39 and 64 in 2019). Second, we link this population to the Numident and perform several adjustments. For observations that receive a PIK in the 2000 decennial census (about 84 percent of the analysis population),[6] we remove those who die before 2005. From the list of Numident observations in the age range who do not appear in the decennial census, we add individuals who received a social security number between 2001 and 2005 and are either citizens or legal residents authorized to work, thus updating the analysis population to 2005. Third, we link anyone in the age range from the Numident to tax year 1999 1040s filed from overseas, capturing citizens who happened not to be in the U.S. at the time of the 2000 enumeration.

Our choices leave us with a 2005 working-age population where approximately 88 percent of the workers have a person identifier that allows us to find them in future administrative records. In previous work using similar data, the analysis population was defined as either 1040 filers (the single-generation case in Akee et al. (2019), *or* the population of children claimed on a 1040 who also appear in the Numident (the intergenerational case in Chetty et al. (2020)). Neither choice

[5] These key matching variables are stripped from the data before researchers gain access. All analysis proceeds on anonymous data.

[6] More disadvantaged persons have a lower probability of receiving a PIK through PVS (Bond et al., 2014), a problem that can be partially remedied through inverse probability weighting to account for PIK placement.

is entirely satisfactory, because the issue of selection into the analysis population is problematic. In both cases, selection relies on adults' 1040 filing behavior, potentially biasing results due to missing information on individuals and families who do not file.[7] In the MOVS project, we retain both linkable and non-linkable observations to serve as a basis for developing inverse-probability-weighting (IPW) strategies that we employ using later years of combined data, described in detail below.

For our individual measure of income, we use an equivalized concept, where income is summed over a household, and each working-age household member is allotted total income divided by the square root of the number of individuals in the household. Such a calculation requires a measure of total income and a measure of household structure over time for each member of our analysis population.

At the individual level, the demographic data provide rich information and near-universal coverage on generally stable characteristics such as race and gender. Time-varying information on income and household structure derive from administrative records. The data appendix gives a full accounting of the business rules governing each source and how we prioritize information.

Table 2 provides a brief breakdown of household and income sources. Household members may be grouped into a physical address, or MAFID (Master Address File Identifier), or persons may be grouped into a household identifier when a physical location is absent or incomplete. Household structure information derives from, first, the 1040, where filing status and number of children claimed provides information on household members (approximately 73 percent of the analysis population have information attached this way in each year). Second, longitudinal HUD data provide this information for households when the 1040 is absent, but this accounts for less than 1 percent of cases. Next, in line with Larrimore et al. (2021), we use MAFID to collect into households those receiving a Form W-2 or 1099 at the same address in the same year. We treat these otherwise unassigned individuals as sharing units, but as in Larrimore et al. (2021), we restrict the number of individuals within a sharing unit to 10. As a last step, we employ two reference files that provide individual address histories or a parent-child link from combined survey

[7] It is less problematic in Chetty et al. (2020), where the analysis sample of children was based on parent claiming. Because of the tax advantages of claiming a child, a large portion of the U.S. child population was reliably captured (Gee et al., 2022).

and administrative data sources. This final collection process assigns about 8 percent of the sample to a housing unit. For the rest of the sample, who appear to be outside of any sharing unit, we assign a singleton household identifier (about 5 percent).

Each year's income values come from Form 1040 filings, W-2 reports, and income reported to HUD. Our primary measure of income is total money income, which includes most sources of income from the "income" section of a 1040 (Meyer et al., 2020), including wages, self-employment income, ordinary dividends, social security benefits, and rental income. Importantly, it does not include capital gains or losses. The definition thus aligns closely with the Census Bureau's definition of money income[8] that is used in the agency's official reports. We construct household income by summing total money income for all tax units in the household. If an individual is a non-filer, we instead add the wages reported on their W-2s to the incomes of other household members. If no one in a household files a 1040 or receives a W-2, then we define household income as the sum of all individual income reported to HUD. This last definition accounts for less than 1 percent of cases.

We find household income for approximately 90 percent of our working-age population in 2005. For the core set of intermediate statistics, we require persons to have an income report in the base year, but in a supplementary file we provide statistics for those whose base-year income is missing. Each year's income value, including the base year, is winsorized at the bottom and top of the distribution by a percentage that brings the lowest possible value to greater than zero. In later years, income values less than or equal to one and missing values are assigned one dollar of income for that year. We take this step so that researchers may easily calculate mobility measures that rely on the natural log of income. In a supplementary file, we provide suitable statistics, such as mean income within cell and income shares, using unwinsorized income.

3.2. Matching the population

We adopt an inverse probability weighting (IPW) strategy to make our linkable population (i.e., the set of individuals with PIKs) representative of the full working-age population. Although our underlying population does not change over time, we estimate the probability of receiving a PIK

[8]https://www.census.gov/topics/income-poverty/income/about/glossary/alternative-measures.html

in each year to account for improvements to the PVS process over time and the resulting increase in the probability of identifier placement. As a result of of these improvements, a person who appears in both the 2000 decennial census and, for example, the 2015 ACS may have a higher probability of receiving a PIK in the 2015 ACS than in the 2000 decennial census. To avoid double-counting, we retain our original linkable analysis population and, in every year, supplement this unchanging set of observations with working-age individuals in the ACS who do not receive a PIK. These unlinkable persons form the basis, in each year, of an IPW strategy that adjusts our statistics to account for selection into sample (i.e., the probability of receiving a PIK). Appendix B describes in detail how we calculate inverse probability weights using ACS data from 2005 to 2019.

Using our inverse probability weights, our baseline population looks remarkably similar over time to public-use ACS populations defined using the same age range and restricting to citizens. In Table 3, we report the gender composition and race-ethnicity for our weighted working-age population along with analogous statistics for individuals aged 25–50 in the 2005 ACS and aged 39 and 64 in 2019. The statistics closely match across the two datasets.

4. The intermediate statistics

The final step in the process involves calculating the suite of intermediate statistics for public consumption. We define a set of cells based on gender, race-ethnicity, decile of 2005 age-adjusted and equivalized income rank, and 2005 state of residence. We rank individuals by 2005 equivalized income within their birth cohort and assign their percentile position based on rank. The calculation of decile position and the cell-level statistics use the inverse probability weights to adjust for sample selection. Within each cell, we provide the statistics listed in Table 4. Mean income (logs and levels) and income changes in logs from the base year and year-to-year provide the opportunity for creating income growth curves. Because income within a cell may change due to movements of both income and family structure, we provide variables on base-year-to-year and year-to-year changes in the number of total family members, number of adults, and number of children. We also provide two measures that allow for separability of equivalized income into two parts: that due to changes in log unequivalized income and changes in the log of the scale (the square root of household size).

Using the deciles of base-year income, a researcher may calculate base-year Gini coefficients. We also provide the probability of ending up in a quintile for a later year for each quintile of the following-year distribution. We do not “restart” observations and calculate year-by-year quintile position, which would allow for development of a Shorrock’s index. However, further base-year releases (i.e., base year 2006, 2007, etc.) will allow researchers to build additional years of Ginis and calculate Gini-based mobility indices. The quintile probabilities we provide do allow for transition-matrix calculations, which we demonstrate below.

As an additional measure of inequality, we provide the variance of log income in each year. Previous research has focused on the extent to which changes in inequality are due to changes in persistent inequality versus transitory inequality (Gottschalk et al., 1994; Kopczuk et al., 2010; DeBacker et al., 2013). We provide two additional statistics that speak to this question: the variance of five-year average log income and the variance of deviations from five-year average log income. These two measures correspond, respectively, to the “permanent variance” and “transitory variance” measures used in Kopczuk et al. (2010). The “annual variance” measure used in the same study can be calculated by averaging the variance of log (annual) income across the surrounding five years.

While the transitory variance measure can be thought of as a measure of income volatility, we include three additional volatility measures in our intermediate statistics. We provide the percent of individuals who experience an increase in income greater than 25 percent of their prior-year income and the percent of individuals who experience a decrease in income greater than 25 percent. We also include the variance of the arc percent change in income, which is the percent change in income relative to the average of income for the current and previous year.

In each cell, we additionally provide the IPW-adjusted count of observations. An unreported statistic—the number of unique tax forms within a cell—was used to determine which cells had too few observations to meet IRS and Census disclosure thresholds. In all cases of small cells, a complementary cell was large enough that we could combine the information in one dimension. We provide higher-level aggregates for all cases regardless, including an overall national-level set of statistics. Thus cells that needed to be suppressed at a lower level of aggregation will be subsumed in a higher-level cell (e.g., race-by-decile-by-year-by-state for the two genders together). We note that the variance statistics are calculated separately for each aggregation level—for example, the variance of log income at the national level is calculated across all individuals in the population, while the variance of log income for women at the national level is only calculated across women (thus, the latter should be interpreted as a measure of inequality among women only).

4.1. Supplementary files

Our choices over ages, reported income, and treatment of missing information in later years are all intended to provide a tractable file of intermediate statistics where individuals represent the overwhelming majority of the U.S. working-age population. Several patterns of interest to researchers will be undetectable in this file, however. We therefore plan to produce four supplementary files intended to complete the picture of U.S. income mobility and volatility.

Unwinsorized data. It is a truth universally acknowledged that negative AGI reported on a 1040 indicates a taxpayer, possibly with high permanent income, who experiences business losses or negative capital gains. By winsorizing for negative or zero income, we also trim a portion of the highest-income taxpayers, many of whom may move between reporting income less than zero and very high income. We provide a file of unwinsorized information, where we retain zero and negative total money income values and do not winsorize top incomes. Because of the importance of unwinsorized values at the top of the distribution in the determination of income shares, we reserve the release of income share statistics to this file. The dimensions of the data (in terms of cell definition) will be determined.

Young adult file. Assessing the condition of younger workers in our full file was complicated by student status, which allows an 18- to 24-year-old to be claimed by parents. Our young adult file will collect those aged 18–24 and track their income over time, while taking into account their dependency status. We will assume that an 18- to 24-year-old who is claimed as a dependent, and then ceases to be a dependent at age 25, is a student during their dependency. The dimensions of the data (in terms of cell definition) will be determined.

Missing income file. For each base year we produce, we plan to restrict observations to those who are connected to reported income in the base year. For 2005, this means that approximately 8 million working-age adults are dropped from the analysis population (about 5.8 percent). However, many of these observations may be found in later years of income data. This supplementary file will provide statistics on mobility for this group when starting income is missing. The dimensions of the data (in terms of cell definition) will be determined.

Persistence of low income file. Using relative measures (based on median income) and an absolute measure that aligns as closely as possible with the U.S. official poverty measure, we will provide statistics on the depth and persistence of low income for working age adults. The dimensions of the data (in terms of cell definition) will be determined.

5. Mobility Measures

To demonstrate the usability of the intermediate statistics, we now provide some example mobility measures. We first provide evidence on differential growth in income between Black non-Hispanic and White non-Hispanic men. This can be presented as a set of maps, with mean income for a given race-year cell graphed for a specific starting decile (our example explores the 5th decile). These maps are informative for analyzing how income growth differs between races over U.S. states.

To move away from a point-in-time and point-in-distribution analysis, we next provide heat maps where we examine how income gains and losses evolve over the years for the full starting distribution. Following Fields and Ok (1999) and Jenkins and van Kerm (2011), we calculate the simplest case of an income growth curve, where individual-level change functions, $(\delta(\cdot))$, are calculated within and graphed against ranks of base-year income. Here, we show mean $log(income_{y+n}) - log(income_y)$ within base-year ranks, where $income_y$ and $income_{y+n}$ are equivalized income in year y and equivalized income in $y + 1$. We graph these values against base-year rank on the y-axis and year on the x-axis, creating heat maps of income growth. These maps uncover compelling patterns of income growth by race-ethnicity and gender that, over the years studied, show the differential impact of the Great Recession on income loss and recovery.

When examining mobility in relative terms—the average movement of persons of different demographic characteristics from one point in the income distribution to another—we use transition-matrix-based statistics. At the individual level, we calculate the probability of appearing in a given quintile of the income distribution in each future year. As with our other intermediate statistics, these probabilities are averaged within base-decile-gender-race-ethnicity bins, which can be further collapsed to base-year quintiles.

Following Apouey et al. (2022),[9] we use base-year-to-year transition matrices to draw out downward and upward mobility by race-ethnicity and gender. Unlike trace-based inequality measures, which express "stickiness" in the diagonals of a transition matrix, upward (downward) mobility can be expressed by the upper (lower) triangle of the matrix minus the diagonal. Here, upward mobility is defined as:

$$\sum_{i=1}^{K-1} \sum_{k=i}^{K} \left(\frac{v_{ik}}{\sum_{i=1}^{K} v_i} \right) \left(\frac{k-1}{K-i} \right) \tag{1}$$

where K indicates the number of categories of the matrix, i indexes the rows, and k indexes the columns. The element v_i is the sum of the cells on a row for cells where $j > i$ (i.e., $v_i = v_{ii} + \sum_{j>i}^{K}$). The second term in the double sum is a weighting factor that upweights moves of greater distance from the diagonal.

Downward mobility has a similar definition, with w_{ij} defined as $w_{ii} + \sum_{j=1}^{i-1}$.

$$\sum_{i=2}^{K} \sum_{j=1}^{i} \left(\frac{w_{ij}}{\sum_{i=1}^{K} w_i} \right) \left(\frac{i-j}{i-1} \right) \tag{2}$$

These mobility measures have a variety of nice properties outlined in Apouey et al. (2022); two of them are especially relevant in this context. The first is normalization—if no one moves to a higher (lower) quintile, but remains on the diagonal, then mobility is 0. When everyone moves to the highest (lowest) quintile, mobility is 1. The second is a weak Pareto principle: for two otherwise identical mobility matrices, if there is higher (lower) movement from the same initial condition in matrix 1 versus matrix 2, upward (downward) mobility is higher in matrix 1. These two conditions allow us to compare trajectories of upward and downward mobility by group.

[9] The authors point out that their proposed mobility statistics turn out to be a more general case of statistics proposed by Prais (1955), Bibby (1975, 1980) and Bartholomew (1967).

6. Results

Figure 1 provides an example of a simple analysis made possible by the intermediate statistics. For men whose 2005 equivalized income placed them in the middle of the income distribution (decile 5) we display the mean of 2010 equivalized income. We mapped this information by origin state and show the results for White non-Hispanic men (top panel) and Black non-Hispanic men (bottom panel). Several informative patterns emerge from this exercise. The first, and most obvious, is that regardless of location, Black men have less income in 2010 than White men, even conditional on having similar incomes in 2005. Second, patterns of differential income growth across the country were similar between White and Black men—areas with higher average incomes for White men, such as the lower Northeast, were also higher for Black men. At a micro level, North Dakota and its oil boom (which we will return to later) stands out as an especially strong location for income growth for White men (for every year of data), but Black men did not benefit as strongly.

Moving away from mean incomes to looking at income changes, we adopt a common measure of income growth: $log(income_{y+1}) - log(income_y)$, calculated over individuals within cells and averaged (Van Kerm, 2009). Figure 2 shows the change in income for White and Black men between 2005 and 2015. To set expectations and align with the previous literature, we show a version of the graph that calculates income growth over percentiles rather than deciles. Income growth follows patterns similar to that shown in Van Kerm (2009), with strong positive growth below the 10th percentile and a long tail of stagnant or negative growth. While Black and White men show similar patterns, we note that Black men cross the 0 growth point at a much lower position in the distrubtion (around the 30th percentile versus the 50th). Moreover, the line for Black men lies everywhere below the line for White men, indicating less positive growth.

Taking this concept, we graph income growth/losses as the z-axis of heat maps. Figure 3 demonstrates how lower income growth for Black men holds throughout the income distribution and over time. Our heat maps show year-over-year income growth for White men in the top panel and Black men in the bottom. We have standardized the categories of income loss and gain to be consistent between the two graphs, making the comparison between White men and Black men straightforward. Warm colors indicate income loss, while cool colors indicate income gain. Although there are some extreme values in the highest and lowest growth categories, most income changes are confined to between -0.15 and 0.15 log points, indicating changes of around -2 percent to +2 percent.

The graphs show losses for both groups throughout the distribution at the start of the period. Those in the very lowest part of the distribution (decile 1) show considerable volatility in income changes. From 2006 to 2013, Black men in general experienced income losses that were steeper than those experienced by White men. For both groups, 2008 and 2009 show especially steep income losses; while these losses appear to abate for White men beginning in 2010, Black men continue to experience steeper declines, especially at lower points of the initial distribution (an exception is seen in the first decile, but all other cells show declines between -2 and -5 percent). Incomes for both groups improved in 2015, although for White men, this was entirely concentrated in the upper half of the starting distribution. Incomes again saw largely across-the-board declines in 2016, 2017, and 2018, with some recovery—but only at the top—in 2019.

A couple of caveats apply regarding this analysis. First, recall that our income measure for the main data file is winsorized—thus these heat maps are not capturing very high incomes at the top of the distribution, which were trimmed to match the percent who were winsorized to get their income in the positive range. Second, in the interest of brevity, we have restricted our analysis to White and Black men, and patterns in other groups may display more positive growth.

The next set of figures, 4 and 5, move away from absolute concepts of mobility to relative. They show upward and downward mobility for White men versus Black men and Asian men. When we compare relative movements for Black men and White men, it is clear that Black men experience stronger downward mobility and weaker upward mobility over time. These results are consistent with our income growth heat maps in that Black men higher in the distribution experienced steeper losses and slower recovery from the Great Recession than did White men. White men display roughly equal downward and upward mobility over the period.

This leads us to wonder who is moving up strongly in the distribution. The answer is in Figure 4, which shows the trajectory for Asian men, again in relation to White men. Asian men experienced upward mobility that was considerably higher than that of White men, although they also experienced higher downward mobility. This suggests heterogeneous patterns for Asian men that is consistent with Akee et al. (2019), where income inequality and income transitions differed by country of ancestry.

It should be noted that while every statistic appears to strengthen over time, this is partly driven by a mechanical time element through calculating between the new quintile and the origin in every case—for each year, observations have an additional chance to enter a higher or lower quintile.

7. Suggestive evidence of income drivers

Our hope is to provide a resource for researchers to explore aspects of income growth, mobility, and volatility in ways that, alone or in combination with other data, "say" something about the well-being of U.S. workers and their families. In our exploration of this first year of data, we uncovered some suggestive patterns, two of which we present here.

7.1. Booms and busts

In a further examination of income at the 5th decile of the base-year distribution, we considered the spatial pattern of our income growth statistic—the average change in individual log income—for White non-Hispanic individuals. In the map of this measure for 2014 (Figure 5), a strong regional pattern emerges, with states in the Mountain and North Central Midwest showing especially positive growth. As a reminder, we are mapping based on the state in the base year—thus our target earners may have moved within these regions to states with higher growth. This pattern appears to be driven largely by the shale oil boom in North Dakota and Montana, which had an enormous impact on income growth in those states.[10]

The boom in oil production in the U.S. reached its peak in 2014, largely driven by efficiencies in technology and a glut of product from North Dakota in particular. The subsequent "bust" came in 2015 in the wake of collapsing prices. Our map for 2015 shows the change in fortunes for those from the central part of the country, with decreases in year-over-year income for both

[10] https://www.aei.org/carpe-diem/shale-oil-turned-north-dakota-from-one-of-the-poorest-us-states- into-the-second-most-prosperous-states-in-a-decade/

of the Dakotas and Montana during a year when neighboring states to the east and west experienced growth. The shale oil bust also impacted traditional oil-producing states, such as Texas and Louisiana, the latter of which experienced the highest negative income growth in the nation in 2015 (this negative experience was likely exacerbated by extreme flooding in that year).

We provide year-over-years rates of changes in state of residence in the data, as well as the probability of changing state of residence from the base year. A slightly disconcerting finding using these variables shows that Black men whose origin states were North Dakota, South Dakota, or Montana in 2005 were nearly three times as likely to leave the state by the end of the period as White men (10.8 percent versus 3.7 percent). This contrasts with more equal leaving rates when looking at other states (5.1 percent versus 3.9). These patterns suggest that the financial benefits of the oil boom may not have been equally shared.

7.2. Family changes versus income changes

One of the challenges of using equivalized income as the value over which we calculate mobility relates to the simultaneous movement of income and persons. Two families starting off at the same size and the same income will look better off with either the departure of a non-earning family member or with an increase in income; a natural question is whether it is family changes or income changes that account for differential total income growth.

To explore this question further, we provide two variables in the intermediate file: the change in log *unequivalized* income and the change in the log of the equivalence scale. Following Fields and Ok (1999) and Van Kerm (2009) as before, we define income growth in a cell as a simple mean over cell members:

$$\frac{1}{N}\sum_{n=1}^{N}\left[log\left(\frac{inc_{n,t+1}}{\sqrt{famnum_{n,t+1}}}\right)-log\left(\frac{inc_{n,t}}{\sqrt{famnum_{n,t}}}\right)\right] \tag{3}$$

where $inc_{n,t}$ is the unequivalized (total) income of household n at time t and $famnum_{n,t}$ is the number of individuals in household n at time t. We can rewrite (3) as:

$$\frac{1}{N}\sum_{n=1}^{N}[log(inc_{n,t+1})-log(inc_{n,t})]+\frac{1}{N}\sum_{n=1}^{N}\left[log(\sqrt{famnum_{n,t}})-log(\sqrt{famnum_{n,t+1}})\right] \tag{4}$$

The first term in (4) is the average change in log unequivalized income and the second term is the average change in the log equivalence scale. Thus, we can see that these terms are informative components for uncovering how much of the difference over time in income growth patterns between groups is due to family changes versus income changes.

We ask the following question regarding these components: If Black women, for example, experienced the same family changes (or conversely, unequivalized income changes) as did White women over the years, would they have experienced better or worse income growth? Because of the additive nature of each component, we simply recalculate Black women's equivalized income changes over time, using the family changes experienced by White women added to Black women's unequivalized income changes within decile, and compare the average true income change over time to this counterfactual income change. We then perform the same exercise, but using White women's unequivalized income changes added to Black women's family changes. We repeat the exercise for each race-ethnicity group.

The results of this exercise appear in Figure 7. In each case, the comparison between the race or ethnicity group in question and White women indicates that the family change component has a smaller effect on income growth than the income component.[11] For Black women, and for all but those in the first decile, were they to have experienced the same changes in family number over time as White women, they would have experienced more negative income growth than what we observe. The difference is very small, amounting to no more than 0.001 log points for any point in the distribution. Had Black women experienced the same unequivalized income changes as White women did, depending on the part of the distribution, they would have had more negative (below the third decile) or more positive (above the third decile) overall income growth. The results suggest that at the lower end of the income distribution, Black women appear to experience both more beneficial family changes and more growth in unequivalized income than White women do. As we move up the distribution, White women's greater income growth drives the overall difference between the groups.

The fact that the counterfactual using White women's family changes is very similar to Black

[11] For this analysis we focused on women because they more often retain custody of children after the breakup of a marriage or cohabitation. Results for men showed similar, albeit less stark, patterns.

women's true income changes suggests that changes in family size alone are not the main drivers of the differences between White and Black women's income change patterns.

However, we cannot rule out a story where White women are more likely than Black women to experience a family change that involves individuals with earnings joining the household. Changes in unequivalized income may still capture the impact of changes in household members since total household income can change as individuals join or leave the household. Our definition of household members who earn income may include spouses, cohabiting partners, relatives, or roommates. The only requirement for household membership is an attachment to the same address in the same tax year (via 1040 filing, receipt of an information return, or a record in HUD data).

Although our change in scale variable allows us to directly decompose the components of total income, we also provide information on year-to-year changes in marital status and household size (number of persons, adults, and children). Table 5 shows the average marriage rates by race-ethnicity in 2005 and the cumulative probabilities of divorce and marriage over time. White women are considerably more likely to be married in 2005 than are Black women and to thus have slightly higher rates of divorce. The two groups' marriage rates are not different, however. White women have slightly more adults (2.2) in the household in 2005 compared to Black women (2.1), and also have fewer children (1.3 versus 1.7). In examining changes over time, we see that Black women actually experience a larger gain in adults than White women do (whether these new adults bring income in the household is a question for further exploration). All groups show a cumulative loss in the number of children over time, which should be expected due to children aging out of dependency.

Returning to Figure 7, Hispanic women show a similar pattern in both exercises as that seen for Black women. In contrast, across the distribution AIAN women would not have experienced any appreciable difference in income were they to have seen White women's changes in family structure. However, had they experienced White women's changes in income, their overall income growth would have improved by as much as 0.06 log points at the bottom of the distribution and 0.02 log points at the top. Meanwhile, Asian women would have been worse off, especially at the bottom of the distribution, had they experienced White women's income changes.

8. Conclusion

The Mobility, Opportunity, and Volatility Statistics project (MOVS) uses administrative records and population-level data on age, gender, race-ethnicity, and base-year state of residence to address the lack of regularly released statistics on income mobility in the U.S. The data-linkage infrastructure at the U.S. Census Bureau allows us to link persons to their income and household data over time, leading to annual populations of working-age persons who can be tracked for multiple years.

We publicly release suites of intermediate statistics, calculated within cells defined by income percentile in each base year, gender, race-ethnicity, and geography, that will inform the public on income growth by group and allow researchers to calculate their own final mobility measures. Released as a Tableau data tool, accompanied by sets of flat files, MOVS allows researchers to make their own normative decisions on how to express income growth, volatility, and mobility; these intermediate statistics will also be useful in their own right for state- or, in future, metro-level analyses.

A cursory look at a single intermediate statistic—average incomes in 2015 of White and Black men who were in the middle of the starting distribution—provides key information on differential patterns of income growth across the U.S. for these two groups. An "off-the-shelf" use of the tool should be accessible to anyone wishing to explore these patterns. More complex calculations are also possible using the flat files, as we show with heat maps of income growth over time for White and Black men and the development of upward and downward mobility indices.

Our "fitness for use" demonstrations show that, conditional on starting in the middle of the 2005 distribution, average incomes for Black men in 2015 were lower than average incomes for White men across the U.S. Meanwhile, patterns of higher or lower average incomes tended to follow similar geographic patterns regardless of group. Heat maps that show the entire distribution and time period indicate that Black men experienced earlier and more severe income loss over the Great Recession than White men and took longer to recover. These patterns are reflected in our relative mobility measures, which show that Black men are more downwardly than upwardly mobile, while White men have similar rates of upward and downward movement. Meanwhile, Asian men display much higher upward mobility than White men, and modestly higher downward mobility. Our relative mobility results correspond with much of what was observed in Akee et al. (2019) on the population of Form 1040 filers.

Two supplementary analyses show the possibilities for innovative research into the drivers of mobility using the intermediate statistics. We show suggestive evidence of how a shale-oil boom and bust played out in White men's income growth and loss. We use the components of total income changes to demonstrate how differences in total income between women of separate race-ethnicity groups are driven by differential changes in family size and unequivalized income.

This paper covers the first year of data release. More patterns and mobility measures will be possible when additional base years (and the four supplementary files) become available. We hope researchers will use these data to uncover more patterns in the U.S. income mobility story.

References

Akee, R., M. R. Jones, and S. R. Porter (2019). Race matters: Income shares, income inequality, and income mobility for all U.S. races. *Demography 56*(3), 999–1021.

Apouey, B., Z. Nissanov, and J. Silber (2022). Ordinal Variables and the Measurement of Upward and Downward Intergenerational Educational Mobility in European Countries. *Review of Income and Wealth*.

Auten, G. and G. Gee (2009). Income Mobility in the United States: New Evidence from Income Tax Data. *National Tax Journal 62*(2).

Auten, G., G. Gee, and N. Turner (2013). Income Inequality, Mobility, and Turnover at the Top in the U.S., 1987-2010. *American Economic Review 103*(3), 168–72.

Auten, G. and D. Splinter (2022). Income inequality in the United States: Using tax data to measure long-term trends. Technical report, Washington, DC: Joint Committee on Taxation.

Bartholomew, D. (1967). *Stochastic Models for Social Processes*. Chichester: John Willey & Sons.

Bibby, J. (1975). Methods of Measuring Mobility. *Quality and Quantity 9*, 107–36.

Bibby, J. (1980). Measures of Social Mobility Based on Income Inequality Measures. *Quality and Quantity 14*, 619–633.

Bloome, D. (2014). Racial inequality trends and the intergenerational persistence of income and family structure. *American Sociological Review 79*(6), 1196–1225.

Bloome, D. and B. Western (2011). Cohort change and racial differences in educational and income mobility. *Social Forces 90*(2), 375–395.

Bond, B., D. J. Brown, A. Luque, and A. O'Hara (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*.

Burkhauser, R. V. and K. A. Couch (2011). Intragenerational Inequality and Intertemporal Mobility. In *The Oxford Handbook of Economic Inequality*, pp. 522–546. Oxford University Press.

Carr, M. D. and E. E. Wiemers (2018). New evidence on earnings volatility in survey and administrative data. In *AEA Papers and Proceedings*, Volume 108, pp. 287–91.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter (2020). Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics 135*(2), 711–783.

Dahl, M., T. DeLeire, and J. Schwabish (2007). Trends in earnings variability over the past 20 years. Technical report, U.S. Congress, Congressional Budget Office.

DeBacker, J., B. Heim, V. Panousi, S. Ramnath, and I. Vidangos (2013). Rising inequality: transitory or persistent? New evidence from a panel of U.S. tax returns. *Brookings Papers on Economic Activity 2013*(1), 67–142.

DeNavas-Walt, C. and B. D. Proctor (2015). Income and poverty in the United States: 2013. Technical report, U.S. Department of Commerce, U.S. Census Bureau.

Dynan, K., D. Elmendorf, and D. Sichel (2012). The evolution of household income volatility. *The BE Journal of Economic Analysis & Policy 12*(2).

Elwell, C. K. (2014). The Distribution of Household Income and the Middle Class. Technical report, U.S. Congress, Congressional Research Service.

Fields, G. S. and E. A. Ok (1999). The measurement of income mobility: An introduction to the literature. *Measurement 1*, 1–1999.

Fixler, D. J., M. Gindelsky, and D. S. Johnson (2020). Measuring inequality in the national accounts. Technical report, U.S. Department of Commerce, Bureau of Economic Analysis.

Gee, G., J. Goldin, J. Hancuch, L. Ithai, and V. Vohra (2022). The Claiming of Children on U.S. Tax Returns, 2017-2019. *Available at SSRN 4066787*.

Gottschalk, P., R. Moffitt, L. F. Katz, and W. T. Dickens (1994). The growth of earnings instability in the u.s. labor market. *Brookings Papers on Economic Activity 1994* (2), 217–272.

Hardy, B. and J. P. Ziliak (2014). Decomposing trends in income volatility: The "wild ride" at the top and bottom. *Economic Inquiry 52* (1), 459–476.

Hungerford, T. L. (2011). Changes in the distribution of income among tax filers between 1996 and 2006: The role of labor income, capital income, and tax policy. Technical report, U.S. Congress, Congressional Research Service.

Jäntti, M. and S. P. Jenkins (2015). Income mobility. In *Handbook of income distribution*, Volume 2, pp. 807–935. Elsevier.

Jenkins, S. P. and P. van Kerm (2011). Trends in individual income growth: measurement Methods and British evidence. Technical report, IZA, Institute of Labor Economics.

Kopczuk, W., E. Saez, and J. Song (2010). Earnings inequality and mobility in the United States: evidence from social security data since 1937. *The Quarterly Journal of Economics 125* (1), 91–128.

Larrimore, J., J. Mortenson, and D. Splinter (2020). Presence and persistence of poverty in U.S. tax data. Technical report, National Bureau of Economic Research.

Larrimore, J., J. Mortenson, and D. Splinter (2021). Household incomes in tax data using addresses to move from tax-unit to household income distributions. *Journal of Human Resources 56* (2), 600–631.

Maag, E., H. E. Peters, A. Hannagan, C. Lou, J. Siwicki, K. Holub, and D. Williams (2017). Income Volatility: New Research Results with Implications for Income Tax Filing and Liabilities. Technical report, Tax Policy Center, Urban Institute, Brookings Institution.

McKinney, K. L. and J. M. Abowd (2022). Male Earnings Volatility in LEHD before, during, and after the Great Recession. *Journal of Business & Economic Statistics 0* (0), 1–8.

Meyer, B. D., D. Wu, G. Finley, P. Langetieg, C. Medalia, M. Payne, and A. Plumley (2020). The accuracy of tax imputations: Estimating tax liabilities and credits using linked survey and administrative data. Technical report, National Bureau of Economic Research.

Moffitt, R., J. Abowd, C. Bollinger, M. Carr, C. Hokayem, K. McKinney, E. Wiemers, S. Zhang, and J. Ziliak (2022). Reconciling Trends in U.S. Male Earnings Volatility: Results from Survey and Administrative Data. *Journal of Business & Economic Statistics*, 1–11.

Moffitt, R. and S. Zhang (2018). Income volatility and the PSID: Past research and new results. In *AEA Papers and Proceedings*, Volume 108, pp. 277–80.

Piketty, T. and E. Saez (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics 118*(1), 1–41.

Piketty, T., E. Saez, and G. Zucman (2018). Distributional national accounts: methods and estimates for the united states. *The Quarterly Journal of Economics 133*(2), 553–609.

Prais, S. (1955). Measuring Social Mobility. *Journal of the Royal Statistical Society, Series A* (118), 56–66.

Reardon, S. F. and K. Bischoff (2011). Income inequality and income segregation. *American Journal of Sociology 116*(4), 1092–1153.

Roine, J. and D. Waldenström (2008). The evolution of top incomes in an egalitarian society: Sweden, 1903–2004. *Journal of Public Economics 92*(1-2), 366–387.

Rose, S. (2016). The growing size and incomes of the upper middle class. Technical report, Urban Institute, Income and Policy Benefits Center.

Snipp, C. M. and S. Y. Cheung (2016). Changes in racial and gender inequality since 1970. *The Annals of the American Academy of Political and Social Science 663*(1), 80–98.

Van Kerm, P. (2009). Income mobility profiles. *Economics Letters 102*(2), 93–95.

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

Table 1: Linkage rates for baseline population

Source	Count	PIKed percent	Final baseline count
Decennial census	137,900,000	84.20%	116,100,000
Numident post 2000	5,350,000	100%	5,350,000
Foreign address 1040	154,000	100%	154,000
		Total	122,000,000

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-008.

Table 2: Source of household information at baseline

Source	Count	Percent of total
MAFID from Form 1040	88,470,000	72.75
MAFID from HUD	336,000	0.28
MAFID from information returns	14,870,000	12.23
MAFID from other source	1,773,000	1.46
Other household information	9,918,000	8.16
Insufficient or no household info	6,230,000	5.12
Assigned singleton Household ID	18,720,000	15.59
Source of income information at baseline		
Form 1040	98,380,000	80.90
Form W-2	9,494,000	7.81
Missing Income	13,730,000	11.29

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-008.

Table 3: Mean characteristics of individuals in our data versus the 2005 and 2019 American Community Survey

	2005		2019	
Characteristic	MOVS	ACS	MOVS	ACS
Male	48.68	49.45	48.36	48.96
Race-ethnicity				
Non-Hispanic, White	65.43	65.49	66.78	68.11
Non-Hispanic, Black	11.35	11.82	11.52	12.74
Hispanic, Any Race	15.50	15.42	14.05	16.07
Non-Hispanic, AIAN	0.71	0.73	0.71	0.73
Non-Hispanic, Asian	4.88	5.11	4.81	4.79
Other Race/Ethnicity	2.14	1.43	2.20	1.85

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK; ACS 2005 PUMS. The Hispanic group includes individuals of any race. The Other Race/Ethnicity group includes individuals who are non-Hispanic Native Hawaiian or Other Pacific Islander, non-Hispanic other race, non-Hispanic more than one race, or non-Hispanic missing race in addition to individuals missing ethnicity information. Columns 2 and 4 show estimates from the public use American Community Survey. DRB approval number: CBDRB-FY23-CES014-048.

Table 4: Intermediate statistics: variable name and definition

Variable	Definition
Dimension Variables	
rank	Age-adjusted income decile, 2005*
sex	Sex: Male or Female*
raceeth	Race and ethnicity: Hispanic, NH White, NH Black, NH AIAN, NH Asian, or Other*
state	Name of 2005 state of residence*
year	Year of income and household data*
fips	FIPS code of 2005 state of residence
Annual Variables	
minc	Mean income, level
mlinc	Mean income, log
mpq1	Mean probability that individual is in first quintile
mpq2	Mean probability that individual is in second quintile
mpq3	Mean probability that individual is in third quintile
mpq4	Mean probability that individual is in fourth quintile
mpq5	Mean probability that individual is in fifth quintile
vlinc	Variance of log income
count	IPW-weighted count of individuals in cell
mfnum	Mean number of individuals in household
madults	Mean number of adults in household
mkids	Mean number of children in household
mmar	Percent of individuals who are married
mage2005	Mean age in 2005
Difference Variables, Relative to Prior Year	
mdiffinc	Mean change in log income
munscinc	Mean change in log household income (not equivalized)
mscalediff	Mean change in log equivalence scale
mltq25diff	Percent with decrease in income > 25% of prior year income
mgtq25diff	Percent with increase in income > 25% of prior year income
varcpdiff	Variance of arc percent change in income
mdiffstate	Percent with change in state of residence
mprobdiv	Percent of individuals who change from married to single
mprobmar	Percent of individuals who change from single to married
Difference Variables, Relative to Base Year	
mdiffincb	Mean change in log income from 2005
mdiffstateb	Percent with change in state of residence from 2005
Five-Year Window Variables	
vperminc	Variance of log five-year average income
vtransinc	Variance of deviations from five-year average income

A * indicates a variable that defines cells. Cells are unique in rank, sex, raceeth, state, and year.

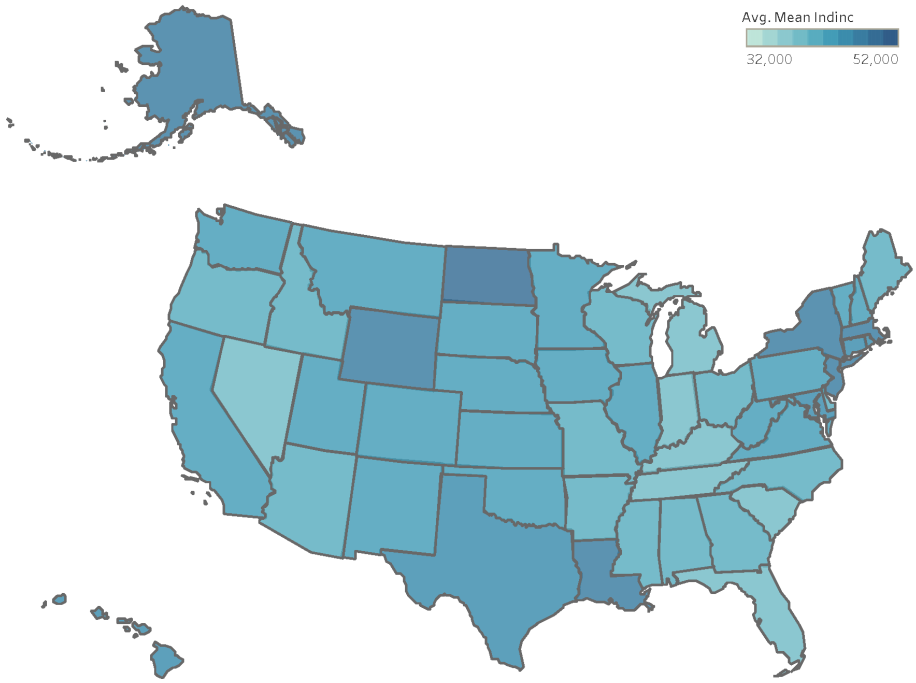
Table 5: Marriage and divorce patterns for women, 2005–2019

Race-ethnicity	Married in 2005	Cum. Pr(Divorce)	Cum. Pr(Marriage)	
White NH	0.629	0.384	0.335	
Black NH	0.242	0.316	0.338	
AIAN NH	0.405	0.459	0.404	
Asian NH	0.694	0.333	0.323	
Other NH	0.474	0.454	0.401	
Hispanic any race	0.511	0.413	0.419	
	No. Adults in HH (2005)	No. Children in HH (2005)	Cum. Change in Adults	Cum. Change in Children
White NH	2.177	1.330	0.071	-0.558
Black NH	2.073	1.654	0.114	-0.727
Asian NH	2.573	1.461	0.173	-0.684
AIAN NH	2.100	1.622	-0.016	-0.439
Other NH	2.346	1.407	-0.099	-0.485
Hispanic any race	2.416	1.958	-0.002	-0.886

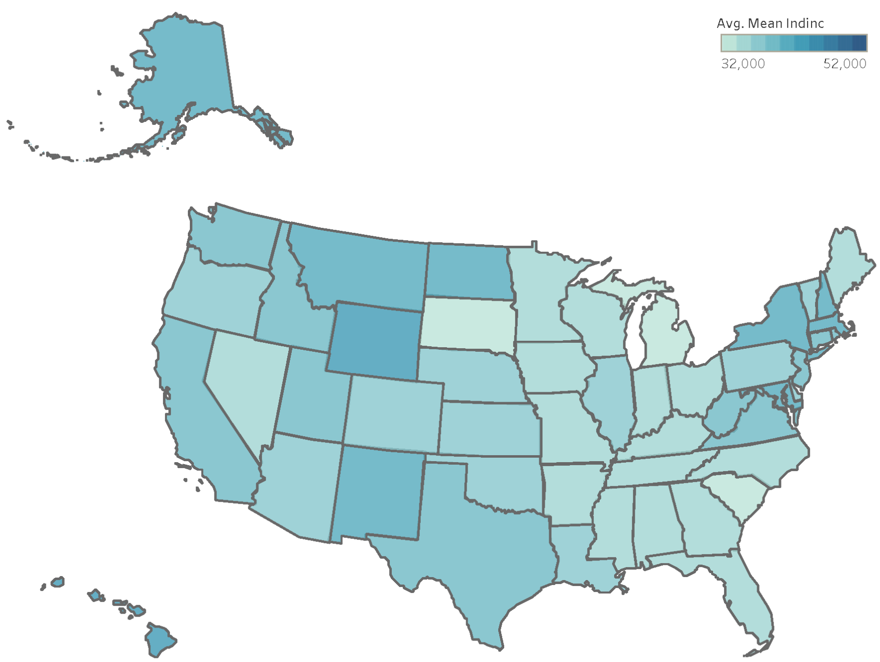
Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. The Hispanic group includes individuals of any race. The Other Race/Ethnicity group includes individuals who are non-Hispanic Native Hawaiian or Other Pacific Islander, non-Hispanic other race, non-Hispanic more than one race, or non-Hispanic missing race in addition to individuals missing ethnicity information. DRB approval number: CBDRB-FY23-CES014-048.

Figure 1: Mean equivalized income in 2010 of men in 5th decile of the income distribution in 2005, by state of residence in 2005

Non-Hispanic White Men

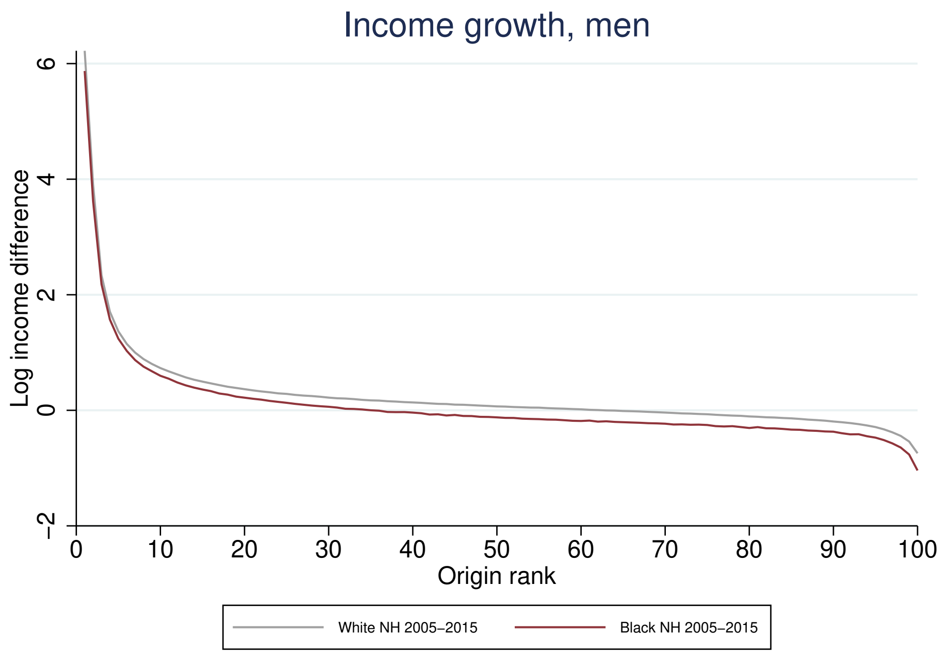


Non-Hispanic Black Men



Note: Equivalized income is calculated for individuals and collapsed into cells defined by age-adjusted 2005 income deciles, race/ethnicity, gender, and state in 2005. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-048.

Figure 2: One year (2005–2015) income growth curves, White and Black non-Hispanic men

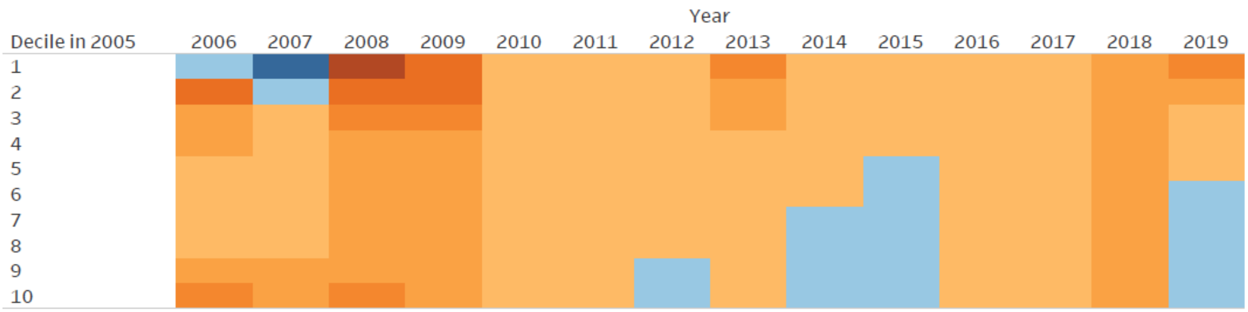


Note: Income growth function is $log(income_{y+1}) - log(income_y)$. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-048.

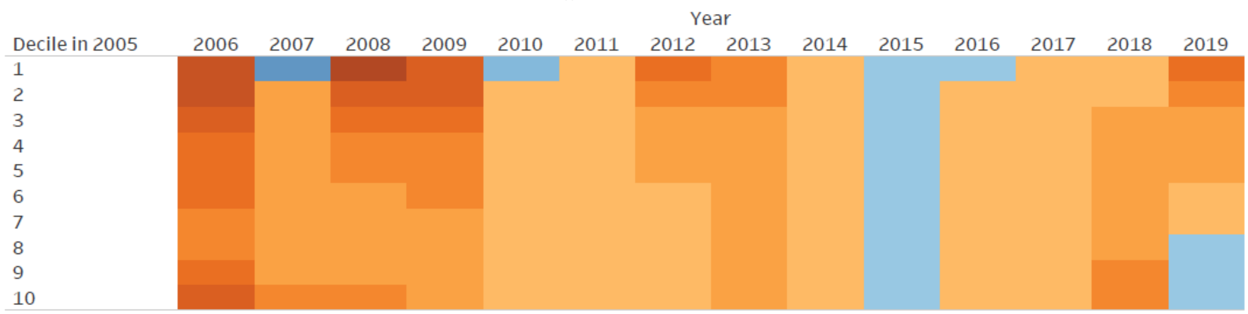
Figure 3: Heat maps of year-over-year income growth, 2006–2018

Orange shades indicate negative income growth and blue shades positive growth.

Non-Hispanic White Men

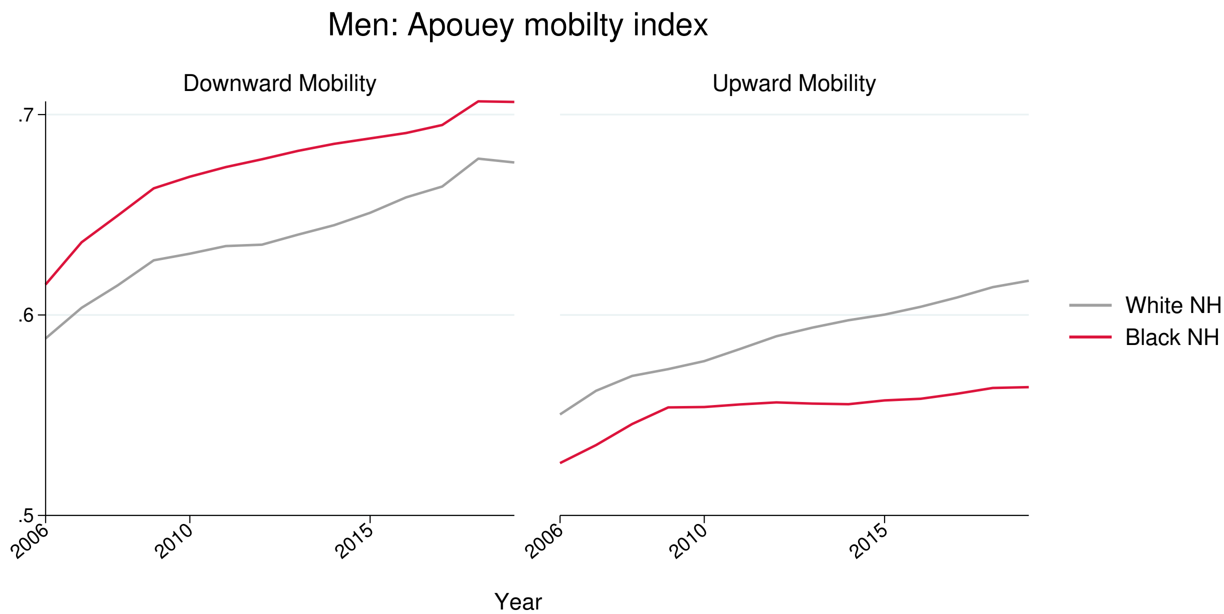


Non-Hispanic Black Men



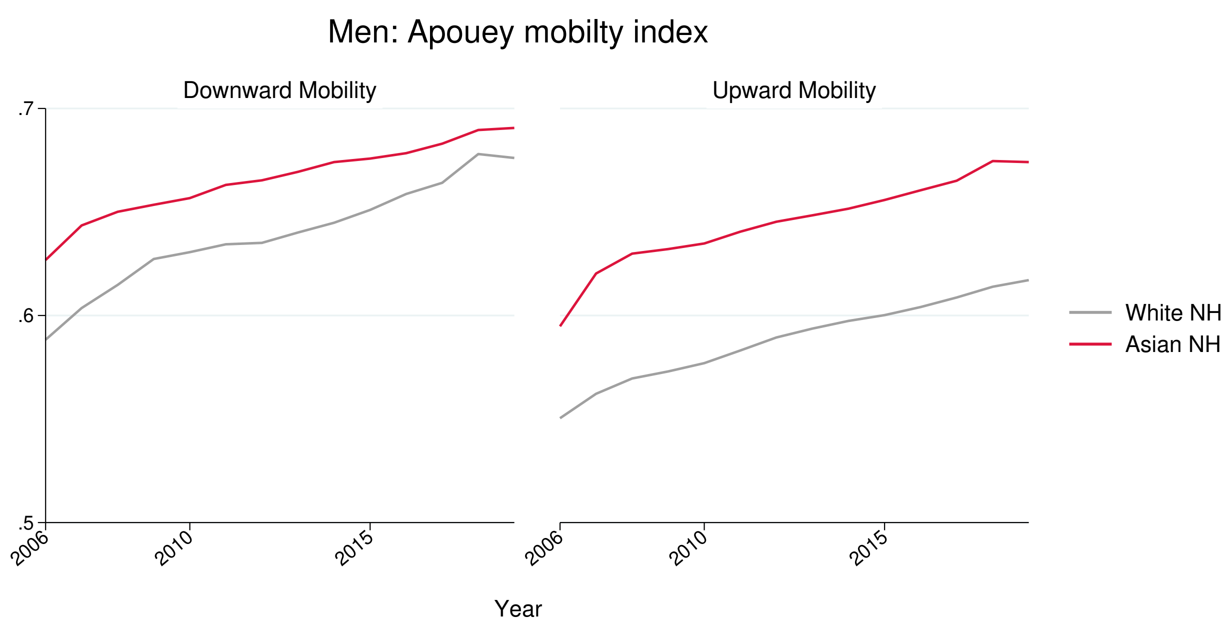
Note: Year-over-year income growth, 2006–2018, for White non-Hispanic men (top panel) and Black non-Hispanic men (bottom panel). Log income t minus log income $t-1$, calculated for individuals and collapsed into cells defined by age-adjusted 2005 income deciles, race/ethnicity, and gender. Orange shades indicate negative growth and blue shades positive. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-048.

Figure 4: Downward and upward mobility, 2006–2015, Black and White non-Hispanic men



Note: See text for calculation. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-008.

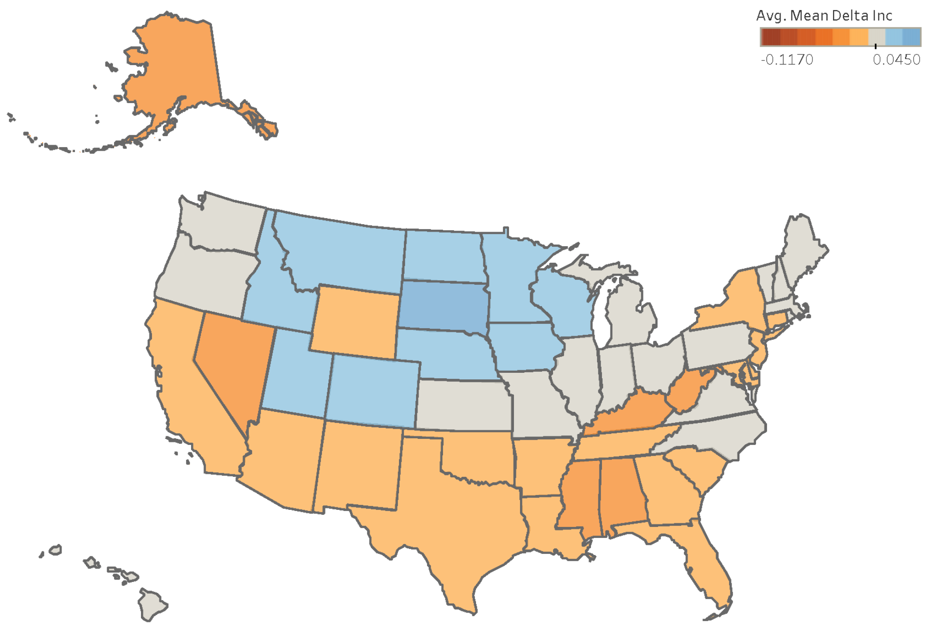
Figure 5: Downward and upward Mobility, 2006–2015, Asian and White non-Hispanic men



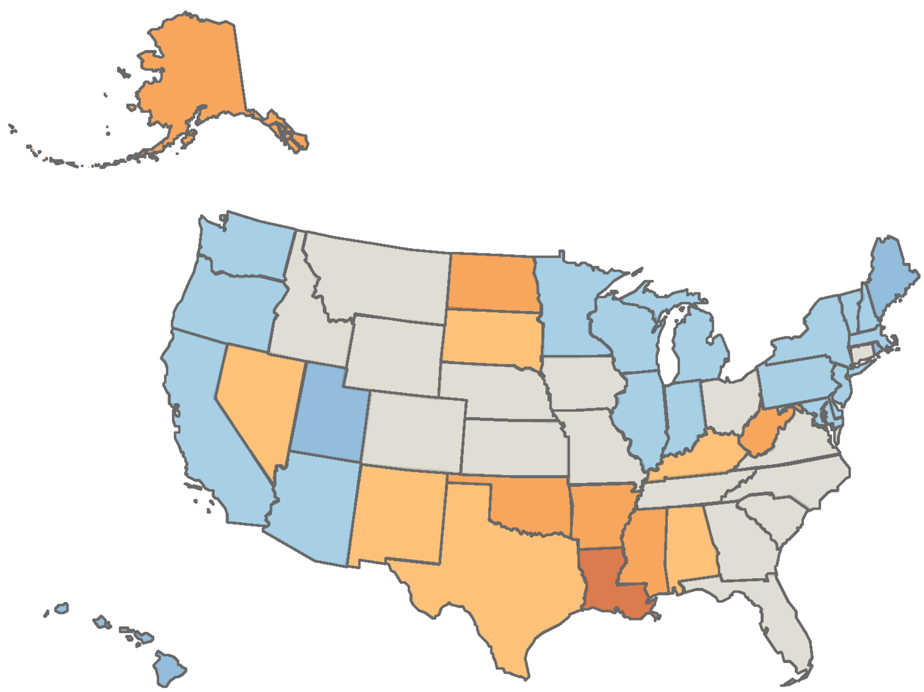
Note: See text for calculation. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-008.

Figure 6: Change in income growth for White non-Hispanic individuals at the 5th decile of the 2005 income distribution in 2014 and 2015

2014

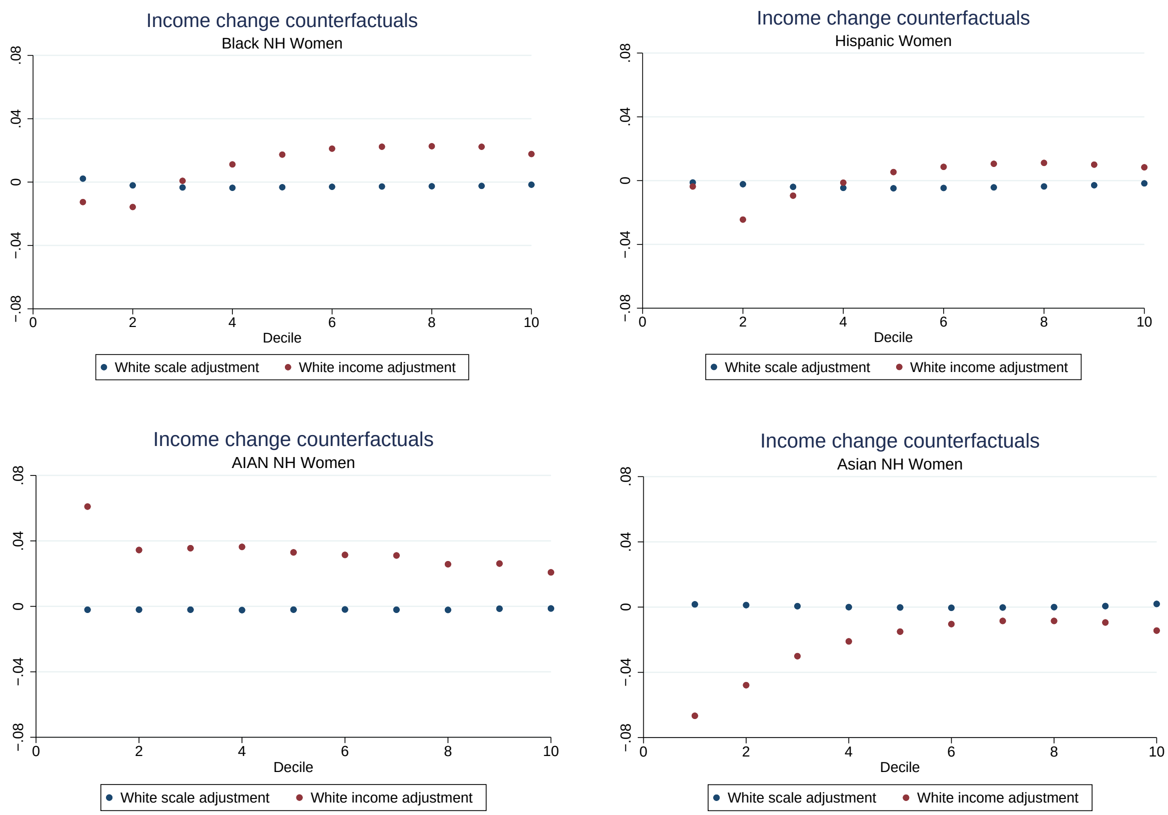


2015



Note: The pattern of income growth and loss is suggestive of the rapid boom in shale oil and its equally rapid bust. Equivalized income calculated for individuals and collapsed into cells defined by age-adjusted base-year rank deciles, race-ethnicity, and gender. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-048.

Figure 7: Counterfactual income changes by race-ethnicity, women



Note: Counterfactual income growth rates, applying the income changes or family size changes of White women. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-048.

A. Household and Income Variable Construction

The main measure of income that we use in our data is equivalized household income. In order to create this measure, we need to group individuals into households and sum income across individuals in the household. We identify households using tax data augmented with the Census Numident, Department of Housing and Urban Development (HUD) files, individual address information from the Composite Person Record (CPR) or MAF-ARF (Master Address File Auxiliary Reference File), or combined administrative records on parent-child linkages (the Census Household Composition Key, or CHCK file). Our income data come from IRS 1040s, W-2s, and HUD.

A.1. Description of data sources

Census Numident. The Census Numident is derived from the Social Security Administration (SSA) Numident file, which records all transactions related to social security numbers (SSNs). The Census Numident file is a person-level file covering all individuals who have received an SSN. Observations are uniquely identified by a PIK, and include information on individuals' date of birth, place of birth, sex, citizenship, and date of death.

IRS Form 1040. Census receives an extract of information from the universe of individual income tax returns, also known as Form 1040. We receive information related to filing status, income, address at the time of filing, and PIKs for the primary filer, secondary filer, and up to four dependents. To identify an individual's location in a given year, we use the address information on the Form 1040 filed for the prior tax year.

IRS 1099. Census also receives an extract of information from information returns. These data cover information returns related to wages, interest, dividends, Social Security benefits, retirement and pension distributions, unemployment benefits, miscellaneous income, mortgage interest payments, and real estate transactions. In this extract, we observe whether individuals received these types of information returns and the address to which the information returns were sent, but we receive no information on amounts reported on the returns. As with the Form 1040 data, we use prior tax year forms for the purposes of identifying addresses.

IRS W-2. Census receives an extract of information from the universe of wage and salary information returns. These data include information on wages, tips, and other compensation along with employer identification numbers (EINs).

Housing and Urban Development Public and Indian Housing Information Center and Tenant Rental Assistance Certification Systems Longitudinal File (HUD PIC-TRACS). The HUD PIC-TRACS data contain information on individuals residing in public housing, participating in the Housing Choice Voucher Program, or receiving project-based rental assistance. The data are derived from forms used to determine households' eligibility for benefits and include information on households' locations and income and the individuals living within the households.

Composite Person Record (CPR). The CPR is a file constructed from various administrative records held at Census. It represents an attempt to compile a "best" record for individuals each year from 2004-2009. It contains information on address, date of birth, gender, and race.

Master Address File Auxiliary Reference File (MAF-ARF). The MAF-ARF is a file constructed from various administrative records held at Census. In most years, it contains a single address for each individual for whom Census can assign a MAFID in the underlying administrative records.

Census Household Composition Key (CHCK). The CHCK data lists PIKs of individuals aged 0-19 in the relevant year along with their parents' PIKs. The file is derived from SSA Numident records. Census uses the names of parents listed on children's Numident records to assign the PIKs of parents. Information on parents' names come from SSN applications, which for most children are based on the information on birth certificates.

A.2. Household construction

In each year, we begin with the set of individuals who appear in the Census Numident and are alive for at least one day during the focal year. We search across our data sources for information on individuals' locations in the year of interest, focusing on address (identified by MAFID), state, and ZIP code of residence. We prioritize location information from IRS 1040s, followed by information from (in order of preference) HUD, IRS 1099s, and the CPR. For years with no CPR data available, we search for location information in the MAF-ARF. Finally, if a child does not appear in any other data source and their parent's address comes from the 1099s or CPR (i.e., they are neither Form 1040 filers nor receiving HUD assistance), then we assign the child the address of their parent, favoring the mother's address over the father's address, if they are different.

Once we have constructed preliminary files for each year, we try to fill in missing information using data from other years. If an individual is missing geographic information in a given year, we check if they have geographic information in the year before and year after the focal year. If this geographic information matches in both the year prior and the year after, then we assign the individual that geographic information in the focal year.

Once we have collected available location and household information for individuals in our data, we assign individuals to households. We use both MAFID and household information from the data sources to construct our households. Households may contain combinations of individuals observed in any of the data sources. If we do not observe a MAFID for an individual, then we use pseudo-address identifiers from 1040 returns, HUD, or the CPR. Individuals with no MAFID or other household information are designated as singleton households.

We apply the following business rules for determining whether to treat individuals observed in the same MAFID as members the same household or as separate households. The steps are applied in order. Within each step, we check whether adding more individuals to a household would create a household with more than 10 individuals. If this is the case, we do not add more individuals to the household and do not carry the household forward to later steps. For example, in step 2, if combining a Form 1040 household with a HUD household in the same MAFID would create a household with 11 individuals, we do not group the Form 1040 and HUD household together and also do not allow either of the households to be grouped with individuals with MAFIDs from the 1099 or CPR data in later steps.

1. When we see multiple tax units in a single MAFID in the Form 1040 data, we group these tax units into one household. Tax filing generally occurs between February and April of a single year, so we assume that individuals with 1040 forms listing the same MAFID are co-residing.

2. When we see individuals in a MAFID in the Form 1040 data and individuals in the same

MAFID in the HUD data for the same year, we do the following: If at least one person in the Form 1040 tax unit shares a HUD household ID with at least one person only observed in HUD, then we group them together. Otherwise, we create separate units.

3. When we see individuals in a Form 1040 MAFID and individuals in the same MAFID in the 1099, CPR, MAF-ARF, or CHCK data in the same year, we group the 1099/CPR/MAF-ARF/CHCK individuals with the Form 1040 household if the 1099/CPR/MAF-ARF/CHCK individuals appear in the same MAFID as the 1040 primary filer(s) in either the year preceding or following the relevant year. Note that the MAFID in the preceding or following year need not be the same MAFID as the current year. In other words, we only group Form 1040 and 1099/CPR/MAF-ARF/CHCK individuals together if they appear to reside together in more than one year. We impose this rule because we want to avoid grouping together individuals who lived at the same address at different times during the same year.

4. When we see multiple individuals in the same MAFID in the 1099, CPR, MAF-ARF, or CHCK data, we combine them into one household. Given that we do not have additional household information about individuals whose MAFIDs come from these datasets, we default toward assuming co-residence. We limit the size of these households to 10 following Larrimore et al. (2021).

5. When we see multiple households with the same MAFID in the HUD data, we keep these households separate. HUD household units are supposed to include all individuals residing together. Thus, we assume that groups appearing as separate households in the HUD data do not reside together.

6. When we see individuals in a MAFID in the 1099, CPR, MAF-ARF, or CHCK data and individuals in the same MAFID from HUD, we group them separately. As with the multiple HUD household rule above, we default towards assuming that these households are more likely to live in the MAFID at different times during the year rather than being part of the same household. Unlike Form 1040 filing, we cannot place 1099/CPR/MAF-ARF/CHCK recipients at a location at a specific time in the year. It also seems unlikely that individuals living in a HUD household would not be included when the household was certified. Thus we do not include individuals whose 1099/CPR/MAF-ARF/CHCK MAFID matches a HUD household MAFID as part of the HUD unit unless they are specifically listed in the unit.

7. When individuals appear in the Form 1040, HUD, or CPR data but do not have a MAFID, we use their unit identifiers to group them into households. All individuals appearing in the same tax unit are included in the same household, and all individuals appearing in the same HUD household ID are grouped into the same household. If a CPR household ID (HUID) contains ten individuals or less, then we group the individuals into the same household. Otherwise, we keep the CPR individuals as singleton households.

A.3. Equivalized Income

Once we have constructed households, we calculate the number of adults, number of children, and number of individuals living in the household.

We then merge in information on income for the year. Our income data come from IRS 1040s, W-2s, and HUD. To construct household income, we add together all income from 1040s for the tax units contained within the household plus income reported on W-2s for any non-filers in the household. If no one in a household has 1040 or W-2 income, then we add together all individual income reported to HUD.

Our equivalized income measure is this household income divided by the square root of the number of individuals living in the household

B. Inverse Probability Weights

Our underlying data consist of working-age individuals in 2005 who have PIKs. To make our population more representative of the entire working-age population, we use inverse probability weights to give greater weight to individuals whose characteristics are similar to individuals who do not receive PIKs. Because there were improvements to the PVS process during our time period, we also need to account for the changing probability of receiving a PIK over time. For a given individual in our data, we will have a greater probability of linking their household and income information in later years than in earlier years because they have a higher probability of receiving a PIK in the household and income data sources. Thus, we estimate inverse probability weights for every year.

In each year, we combine our data with the set of individuls in the relevant year of American Community Survey (ACS) data who do not receive a PIK. We use the ACS because it contains similar demographic variables to our population data, which allows us to model the probability of receiving a PIK in our population. We assume that the individuals without PIKs in the ACS for a given year are representative of the individuals who would not receive a PIK in our administrative data sources for the same year since the files were likely processed using similar PVS technology.

We use a logit model to estimate the probability of receiving a PIK. Our model includes indicators for household size, number of children in household, bins of age, state of residence, and interactions of sex, race, and marital status. Our final weights are the inverse of the predicted probability of receiving a PIK in a given year.