March 17, 2023

A more equal union? A new source on income disparities for U.S. communities defined by race, place, and ethnicity

Brandon Hawkins

Federal Reserve Bank of Minneapolis
brandon.hawkins@mpls.frb.org

Illenin Kondo

Federal Reserve Bank of Minneapolis
kondo@illenin.com

Kevin Rinz

Federal Reserve Bank of Minneapolis and U.S. Census Bureau
kevin.rinz@census.gov

John Voorheis

U.S. Census Bureau
john.l.voorheis@census.gov

Abigail Wozniak

Federeal Reserve Bank of Minneapolis
abigail.wozniak@mpls.frb.org

1 A new source for exploring U.S. group-level inequality

Earning power in the United States is significantly more unequal today than it was fifty years ago, when a multi-decade pattern of widely shared growth began to retreat. Since 1980, the pace of earnings growth has been uneven, and varying, for Americans throughout the earnings distribution.

The course of the widening earnings disparities has been well-documented, and a number of patterns and episodes emerge (Hoffmann et al., 2020). The highest earners have experienced consistent growth in their earning power since 1980, although this has slowed or become more concentrated among the very highest earners since about 2000 (Kopczuk et al., 2010; Blanchet et al., 2022). Relative real earnings for workers below the median have fallen and partially recovered more than once (Autor et al., 2008; Blanchet et al., 2022; Autor et al., 2023). Although these patterns are robust, there have been important differences across demographic groups, notably between women and men (Guvenen et al., 2022).

In spite of this significant body of evidence, there is still much we do not know about how earnings have changed for many Americans over the last five decades of stubbornly rising inequality. The open questions are particularly acute at the subnational level. For example, how has income inequality evolved for a broad range of race and ethnic groups relative to non-Hispanic White Americans? Does this evolution vary across places, and if so, what local factors are associated with more or less inequality either within or across these groups?

Answers to such questions has been limited in part by data availability. Even large surveys like the American Community Survey (ACS) sample shares of the population that are small in absolute terms, which makes it difficult to describe the full distribution of income for subgroups that may not be sampled in large enough numbers (either nationally or in a specific place) to support reliable estimates. Surveys also tend not to follow respondents over time, making it difficult to use them to understand income dynamics, or else trade off cross-sectional sample size against panel duration, limiting the granularity of any income dynamics estimates that may be producible. Moreover, measuring top incomes well is critical for understanding inequality, but measures based on surveys can be distorted by both disclosure avoidance procedures (e.g. topcoding) and survey misreporting (Burkhauser et al., 2011; Bee and Mitchell, 2017).

Such data issues are particularly important when it comes to understanding differences in income inequality and dynamics across groups defined by race and ethnicity, as these groups are unevenly distributed across space and differentially likely to be affected by measurement issues related to top incomes. It is especially important to be aware of differences in the prevalence of groups across space in this context, as state income distributions have become less similar to each other at the top in recent decades, and this phenomenon appears to be driven by the top of the White income distribution (Rinz and Voorheis, 2023).

In our paper, we present a new data source that fills a number of important needs in this developing area of statistics. Relative to related efforts to develop public sources for detailed measurements of U.S. inequality, the major advantages of our approach are threefold:

- Our data include a wide range of statistics for subnational geographic areas (currently states) as well as many subnational and sub-state populations.
- We provide statistics on earnings mobility and volatility in addition to statistics on income distributions for our populations of interest.
- Statistics in our data are computed directly from the universe of tax filers, which means we do not need to rely on statistical matching or projections to link micro and macro data sources.

These data are the result of a collaboration between researchers at the U.S. Census Bureau and the Minneapolis Fed's Opportunity & Inclusive Growth Institute. We plan to make the data publicly available by the time of the conference presentation so that they may serve as a new resource to researchers, interested policymakers, and the public at large.

In addition to introducing these data, our conference paper will explore sample applications that demonstrate the usefulness of the data for exploring questions related to the evolution of income inequality within and across communities defined by race, ethnicity, and place, as well as other demographics.

Ultimately, our new data source will enable a wide variety of novel analyses that can help researchers and policymakers understand the landscape of income inequality in the United States. This new data has the potential to support broad research agendas focused on the relationships

among income inequality/volatility/mobility, race/ethnicity, and places and their characteristics. The new data source also has the potential to provide significant value to policymakers and the public at large. Our goal with this project is to provide a broad set of data users with easy access to descriptive information about local income inequality and dynamics enabled by these new data.

2 Data description and construction methodology

In this project, we will construct a novel dataset of detailed statistics on income distributions, volatility, and mobility within granular cells defined by demographics, place, and time. This dataset is generated from underlying microdata that combine income and location information from the universe of 1040 and W-2 tax forms with demographic information available from various sources through the U.S. Census Bureau. These data have three key features that are difficult to collectively replicate in other settings: they provide detailed demographic information, high-quality measurements of top incomes, and large sample sizes.

These features are especially important for producing detailed measures characterizing experiences across income distributions within groups defined by race and ethnicity. These measures can be very demanding in terms of the number of observations required to produce them, and relying on samples generated by even large surveys like the ACS can leave some groups without estimates for most places and most groups without estimates for some places. Combining survey and administrative data as we do in this project substantially expands the set of groups and places that can be included, as well as the possible measures that can be produced. Akee et al. (2019) use these linkages at the US level to document differences across race/ethnicity groups in incomes, top income shares, and income mobility. Rinz and Voorheis (2023) also leverage these linkages at the subnational level for tax units to discuss large differences in income growth across race/ethnicity groups within a state, especially for top incomes.[1]

We produce a variety of statistics characterizing income inequality within years as well as income volatility and mobility across pairs of years. We produce these statistics nationally and at the state level within groups defined by race and ethnicity, as well as other demographic characteristics and

[1]See Partridge et al. (1996); Morrill (2000); Frank (2014); Sommelier and Price (2014) for existing work on subnational inequality in the U.S.

Statistics		Demographic characteristics defining groups
Inequality statistics	Volatility and mobility statistics	
Percentile values: p10, p25, p50, p75, p90, p95, p98, p99, p99.9, p99.99, etc. as sample size allows	**Transition probabilities**: $P(B_{y+k}\|B_y)$ for k={1,5} and income bins B={missing, p1-p25, p26-p50, p51-p75, p76-p90, p91-p100}	**Race/ethnicity**: NH White, NH Black, NH AIAN, NH Asian, NH NHOPI, Hispanic
Top income shares (within group): of all income earned by a given group, what share was earned by the top 10%, 5%, 2%, 1%, .1%, .01%, etc. of earners, as sample size allows	**Conditional income growth**: mean, p10, p25, p50, p75, and p90 of income growth for people starting in each income bin B above between y and y+k	**Age**: 16-24, 25-34, 35-44, 45-54, 55-64, 65+
Top income shares (across groups): of income earned by the top 10%, 5%, 2%, 1%, .1%, .01%, etc. of all earners (as sample size allows), the share earned by members of a given group		**Gender**: male, female
Share of top earners: The share of the top 10%, 5%, 2%, 1%, .1%, .01%, etc. of earners (as sample size allows) belonging to a given group		**Place of birth**: USA, other

Note: p# indicates percentile value. NH indicates non-Hispanic ethnicity. AIAN = American Indian and Alaska Native. NHOPI = Native Hawaiian or Other Pacific Islander. Top income bins in transition probability and conditional income growth are consolidated into a single p76-p100 bin for estimates at the state level. Top income percentiles used for state-level inequality statistics go up to the 98th percentile. Estimates are produced nationally and at the state level within the groups listed for each demographic characteristic, as well as for two-way interactions of select demographic characteristics.

Table 1: Summary of planned estimates

two-way interactions between demographic characteristics. Table 1 lists categories of statistics we plan to produce, as well as the demographic characteristics we plan to use to define groups.

We also produce statistics using several different measures of income. Three are obtained from form 1040 and aggregated to the household level using the Census Bureau's Master Address File ID: adjusted gross income (AGI), wage and salary income, and non-wage income, which is defined as the difference between AGI and wage and salary income. We assign the household values of these variables to each individual when constructing estimates and think of these measures as representing the resources available to members of the household.[2] We produce estimates based on these measures for 1998 through 2019.

Two additional individual-level measures are obtained from form W-2: wage/salary income

[2]We also plan to report statistics for equivalence scale-adjusted versions of select household-level statistics.

and total compensation, which is the sum of wage/salary income and deferred compensation. We produce estimates based on these measures for 2005 (the first year in which the W-2 data are available to us) through 2019.

3 Key features and sample applications

The dataset features three novel dimensions compared to currently available statistics on income inequality and income dynamics.

First, we produce granular subnational statistics on top income inequality and income dynamics by demographic group including minority race/ethnicity groups. As discussed above, this is a particularly important feature especially when it comes to minority racial and ethnic groups and their intersections with gender or age. As in Akee et al. (2019), we leverage the linkage from individual tax records to Census Bureau unique, protected identification key (PIK) in order to obtain race, ethnicity, gender, age, foreign born status, and other demographic characteristics. This deliberate focus on race and ethnicity will help understand the anatomy of income inequality, income volatility, and income mobility along the income distribution by type of income, by demographic group, and across states. We build on Akee et al. (2019) by constructing yearly state-level statistics on income inequality by race, by race and age, and by race and gender for different types of income. We build on the contribution of McKinney et al. (2022) in the Global Repository of Income Dynamics (GRID) project by documenting percentiles of the conditional growth distribution at the subnational level by type of income, by demographic group, and across states.

Second, we produce new **household-level** statistics on income inequality, income volatility, and income mobility in the context of U.S. administrative tax records. By leveraging the linkage to the Census Bureau Master Address File ID (MAFID), we derive a household construct for each individual in a given year. We circumvent issues associated with the notion of a 'household head' or the assumptions required for the apportionment of household income among its members. Instead, we report the total household income for our 1040 derived income variables. While this concept does not neatly lend itself to distributional national account applications, it allows us to preserve the rich variety of demographic characteristics within a household. In doing so, we are able to document

top incomes at the household level and to contrast their levels and fluctuations by race/ethnicity, its intersections with gender and age, and by state.[3]

Finally, we report income inequality, income dynamics, and income mobility statistics for different types of income as reported in the tax records. For the 1040-derived incomes, we study not just the household AGI but also its split between earnings from 'wages, salaries, and tips' (WSI) and 'non-wage income' (NWI). We define 'non-wage income' as the difference between AGI and earnings 'wages, salaries, and tips'. For the W2-derived incomes, we not only report inequality measures for individual 'wages, salaries, and tips' but also 'total compensation' for U.S.-level statistics.

It is important to note that along these novel dimensions of the dataset, unique characteristics of the underlying tax records data are critical. First, by using the universe of tax records, albeit only for a couple of decades and a few income variables, we are able to more systematically explore the upper tail of income levels for groups other than non-Hispanic Whites at the subnational level. Second, we use the panel dimension of this large data to document the conditional distribution of income growth and income mobility along the income distribution. For these reasons, we choose to preserve the population coverage of the dataset as much as possible. However, when studying W2-derived income variables, we also report statistics when using a secondary sample restricted to prime-age workers with earnings above a threshold.

We think the dataset will be extremely useful in applications where it is important to have granular group-specific and local measures of income inequality, income volatility, or income mobility. While these applications are potentially numerous and cut across fields from applied microeconomics to macro labor economics, public finance, political economy, and social science disciplines outside of economics, we outline two sample applications here that align with the 2024 conference focus. We will develop these and report them in the full paper.

Considering income inequality and dynamics within granular demographic and geographic cells can help reveal important facts about the functioning of the broader economy. Our new dataset will yield insights on the racialized and geographical nature of income inequality, income volatility, and income mobility in the U.S. We can already show some preliminary results from a limited initial

[3]See for example Guvenen and Kaplan (2017) for a discussion on units the issue of individual records, tax-filing-unit records, and household units when it comes to top income and inequality.

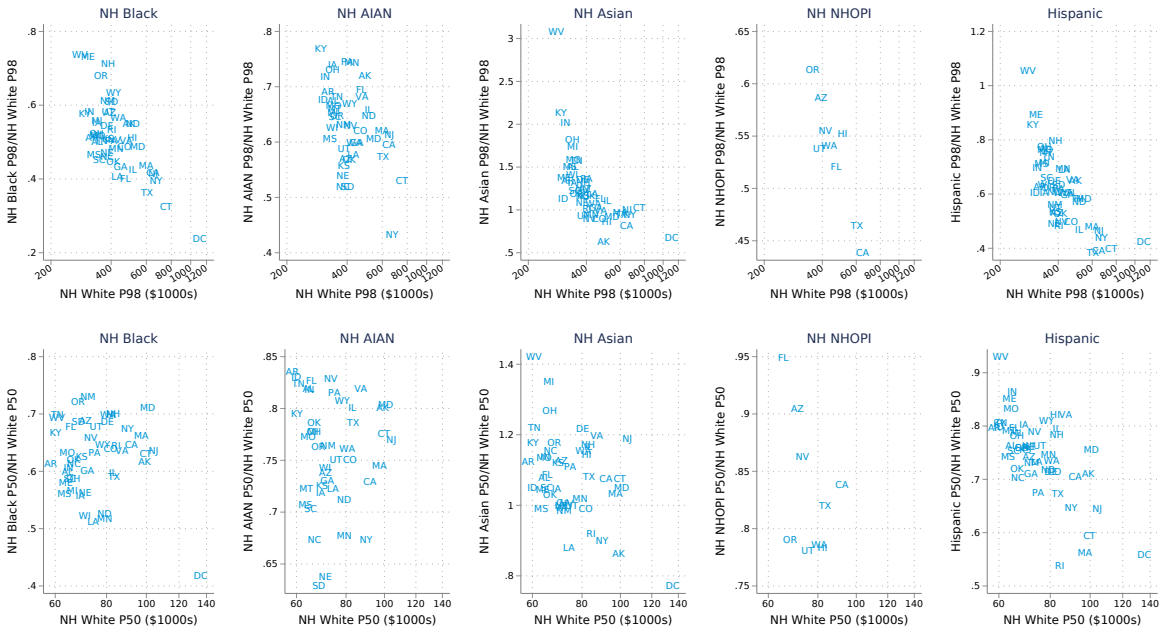


Figure 1: Cross-race/ethnicity income ratios by non-Hispanic White income levels, 2012

Source: IRS Form 1040, Best Race and Ethnicity Administrative Records Composite file, and authors' calculations.
Note: Panels show scatter plots of ratios of select percentiles of adjusted gross income (AGI) for the indicated race/ethnicity groups to the same percentiles of AGI for non-Hispanic Whites on the y-axis against the level of that same percentile of AGI for non-Hispanic Whites (on a log scale) on the x-axis. The top row makes this comparison for the 98th percentile, while the bottom row uses the median. NH indicates non-Hispanic ethnicity. AIAN = American Indian and Alaska Native. NHOPI = Native Hawaiian or Other Pacific Islander. Disclosure Review Board release authorization numbers CBDRB-FY23-CES014-019.

data release from the project.

In our first example, we analyze how top incomes for other groups relate to top incomes for non-Hispanic Whites at the subnational level. Our preliminary analysis suggests that top incomes in other race/ethnicity groups are closest to those of non-Hispanic Whites in states where top non-Hispanic White incomes are lowest. The panels in the top row of Figure 1 show scatter plots of the ratio of the 98th percentile of AGI for other race/ethnicity groups to the 98th percentile of income for non-Hispanic Whites on the y-axis against the level of the 98th percentile of AGI for non-Hispanic Whites (on a log scale) in 2012 on the x-axis. While the levels on the y-axis vary, there is a clear downward slope to these scatter plots across all groups presented. The panels in the bottom row repeat this exercise using the median rather than the 98th percentile and show no such relationship.

A second sample insight from our initial data release comes from analyzing our measures of non-wage income. Our current analysis shows that non-wage income is a more unequal type of

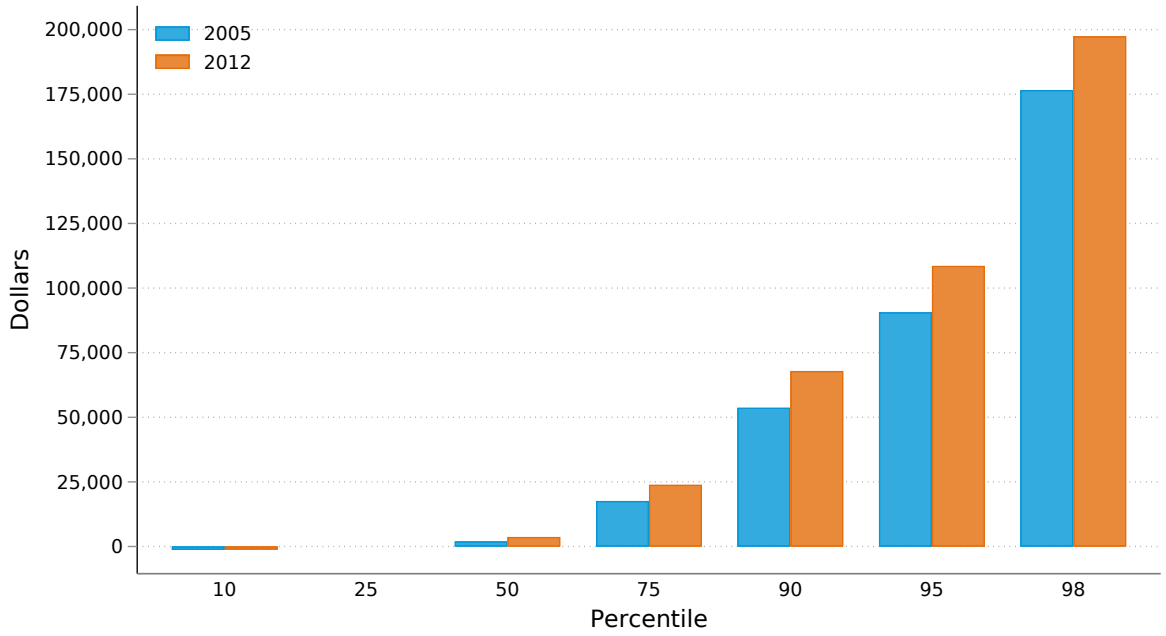


Figure 2: Percentiles of non-wage income by year

Source: IRS Form 1040 and authors' calculations.
Note: Bar charts show the value of non-wage income (NWI) in 2005 (blue bars) and in 2012 (orange bars) for select percentiles of NWI income on the x-axis. Disclosure Review Board release authorization number CBDRB-FY23-CES014-016.

income at both the extensive margin and the intensive margin. Figure 2 shows percentiles of the distribution of non-wage income for all households in our dataset across the U.S. Non-wage income is uncommon as most households have little or no income of this type: the median household had a couple of thousand dollars in non-wage income in 2005 and in 2012. Non-wage income also seems more unequal conditional on positive values. For brevity, we do not report the WSI or AGI counterparts of this graph. However, top income ratios for non-wage income appear substantially higher than standard values in the literature for AGI or WSI income. For example, the p90-p50 ratio for WSI is between 2 and 3 in the literature (see for example Guvenen et al., 2022). Given the significant disparities documented in the literature for AGI or WSI across demographic groups, the more pronounced inequality in non-wage income motivates our study of the distribution of non-wage income and its dynamics at a granular and subnational level.

In addition to the analysis of our sample applications, our conference paper will also document the usefulness of the new data compared to alternative sources. We are performing the following comparisons to highlight the value of unique aspects of our data: a large underlying sample size,

quality top income data, dynamic linkages, and household unit concepts.

Comparisons with restricted-use survey data We merge our dataset with restricted-use ACS and Current Population Survey Annual Social and Economic Supplement (CPS-ASEC) data to document 1) the value of administrative income data for capturing the top of the income distribution, and 2) the implications of using different income-sharing units (e.g. family, tax unit, household). We will also describe how the implications of using tax data differ by race and ethnicity.

Comparisons with the Panel Study of Income Dynamics (PSID) We construct conditional income growth statistics in the PSID by race and ethnicity and the associated standard errors. We compare these estimates and their precision along the income distribution by race and ethnicity to estimates from this project.

Comparisons with the Statistics of Income (SOI) Finally, we produce household-level counterparts of the IRS SOI Individual Income Tax State Data and augment these statistics with a breakdown by race and ethnicity.

Overall, we propose a new source to understand the nature of income disparities across communities defined by race, place, age, and ethnicity in the U.S. We explore the dynamic components of these disparities by characterizing the conditional distribution of one- and five-year income growth within these groups. We also highlight key differences across different types of income such wage income and non-wage income. We think this systematic exploration of income disparities will fill some important gaps in the vast literature on income inequality in the U.S.

References

Akee, Randall, Maggie R Jones, and Sonya R Porter, "Race matters: Income shares, income inequality, and income mobility for all US races," *Demography*, 2019, *56* (3), 999–1021.

Autor, David, Arindrajit Dube, and Annie McGrew, "The Unexpected Compression: Competition at Work in the Low Wage Labor Market," Technical Report, National Bureau of Economic Research 2023.

Autor, David H, Lawrence F Katz, and Melissa S Kearney, "Trends in US wage inequality: Revising the revisionists," *The Review of economics and statistics*, 2008, *90* (2), 300–323.

Bee, Adam and Joshua Mitchell, "Do Older Americans Have More Income Than We Think?," *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, 2017, *110*, 1–85.

Blanchet, Thomas, Emmanuel Saez, and Gabriel Zucman, "Real-Time Inequality," Technical Report, National Bureau of Economic Research 2022.

Burkhauser, Richard V, Shuaizhang Feng, Stephen P Jenkins, and Jeff Larrimore, "Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring," *The Journal of Economic Inequality*, 2011, *9*, 393–415.

Frank, Mark, "A new state-level panel of annual inequality measures over the period 1916-2005," *Journal of Business Strategies*, 2014, *31* (1), 241–263.

Guvenen, Fatih and Greg Kaplan, "Top income inequality in the 21st century: Some cautionary notes," Technical Report, National Bureau of Economic Research 2017.

_ , _ , Jae Song, and Justin Weidner, "Lifetime earnings in the United States over six decades," *American Economic Journal: Applied Economics*, 2022, *14* (4), 446–79.

Hoffmann, Florian, David S Lee, and Thomas Lemieux, "Growing income inequality in the United States and other advanced economies," *Journal of Economic Perspectives*, 2020, *34* (4), 52–78.

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

McKinney, Kevin L., John M. Abowd, and Hubert P. Janicki, "U.S. long-term earnings outcomes by sex, race, ethnicity, and place of birth," *Quantitative Economics*, 2022, *13* (4), 1879–1945.

Morrill, Richard, "Geographic variation in change in income inequality among US states, 1970–1990," *The Annals of Regional Science*, 2000, *34*, 109–130.

Partridge, Mark D, Dan S Rickman, and William Levernier, "Trends in US income inequality: evidence from a panel of states," *The Quarterly Review of Economics and Finance*, 1996, *36* (1), 17–37.

Rinz, Kevin and John Voorheis, "Re-examining Regional Income Convergence: A Distributional Approach," February 2023. CES Working Paper 23-05.

Sommelier, Estelle and Mark Price, "The increasingly unequal states of America," *Washington: Economic Policy Institute*, 2014.