Household Incomes in Tax Data

Using Addresses to Move from Tax-Unit to Household Income Distributions

Jeff Larrimore
Jacob Mortenson
David Splinter

ABSTRACT

A limitation of tax return data is the inability to identify members of separate tax units living in the same household. We overcome this obstacle and present the first set of entirely tax-based household income and inequality measures. We find using tax units as a proxy for households overstates household income inequality, as measured by Gini coefficients, by 13 percent. Consistent with previous findings, we also estimate that the CPS understates household income inequality by 5 percent. Compared to conventional tax-unit measures, the federal income tax code and earned income tax credit are less progressive when measured at the household level.

Jeff Larrimore is a section chief at the Federal Reserve Board of Governors. Jacob Mortenson is an economist at the Joint Committee on Taxation. David Splinter is an economist at the Joint Committee on Taxation. *The authors thank Katherine Arnold, Jesse Bricker, Richard Burkhauser, James Cilke, Jason DeBacker, Scott Winship, Gabriel Zucman, anonymous referees, and participants of presentations at the Federal Reserve Board, Drexel University, Ohio State University, the IRS–Census income measurement workshop, and the spring 2018 NBER public economics meeting for helpful comments. The results and opinions expressed in this paper reflect the views of the author and should not be attributed to the Federal Reserve Board. Additionally, this paper embodies work undertaken for the staff of the Joint Committee on Taxation, but as members of both parties and both houses of Congress serve on the Joint Committee on Taxation, this work should not be construed to represent the position of any member of the committee. The Joint Committee on Taxation reviewed this paper prior to its circulation. The authors did not receive any financial support for the work and have not received any supporting funds in the past three years in excess of $10,000. This paper uses confidential data from the Internal Revenue Service. The data can be obtained through the Internal Revenue Service Statistics of Income Joint Statistical Research Program call for proposals (https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program). The authors are willing to assist (jeff.larrimore@frb.gov), and programming codes are available on the author's website at www.davidsplinter.com.*
[Submitted July 2018; accepted May 2019]; doi:10.3368/jhr.56.2.0718-9647R1
JEL Classification: D31 and H24

Supplementary materials are freely available online at: http://uwpress.wisc.edu/journals/journals/jhr-supplementary.html

I. Introduction

Over the past decade, research using administrative Internal Revenue Service (IRS) tax return data has greatly expanded our understanding of incomes at the top of the U.S. income distribution (for example, Piketty and Saez 2003; Atkinson, Piketty, and Saez 2011). However, researchers have been forced to adapt their analysis to fit the limitations of IRS tax return data. The absence of nonfilers in tax return data has largely restricted analyses using tax records to the upper end of the income distribution. Additionally, tax returns provide information on those individuals appearing on the same tax return (a tax unit), but no information on others living in their household. Since households may contain multiple tax units or nonfilers, this situation has precluded household-level analysis, which is the standard unit of analysis in both national and cross-national distributional studies.

Using a new approach to link together tax units and nonfiling individuals, we overcome these limitations of IRS data and produce household identifiers for every individual in the United States, where households include all individuals listed on tax forms at a given address. These identifiers include individuals who file or appear on tax returns, as well as nonfilers who do not submit a tax return to the IRS. We have made these household identifiers available to researchers with access to IRS data. Consistent with the call from the Commission on Evidence-Based Policymaking (2017) to reduce barriers to effective use of existing administrative data, the creation of this data set can help improve the alignment of findings in IRS data with those from other data sources and allow for these data to be used more effectively by the research community.

We use these data to produce the first set of entirely tax-based income-distributional statistics analyzed at the household level rather than the tax-unit level. We then compare the distribution of income using these new tax-based household data with more traditional IRS tax-unit results and with survey-based household results from the Annual Social and Economic Supplement to the Current Population Survey (CPS) that is fielded every March. Finally, we use these data to provide the first tax-based measure of the distribution of the Earned Income Tax Credit (EITC) and overall tax burdens across U.S. households, as compared to tax units.

When comparing the income distributions of households in our new data with previous inequality estimates, household income inequality in tax data is roughly two Gini points (5 percent) higher than analogous estimates using CPS data. However, household income inequality is roughly six Gini points (13 percent) lower than analogous estimates using tax units as the unit of analysis, which is the standard approach in previous inequality research using tax data, including Piketty and Saez (2003). This finding suggests that researchers using tax units as proxies for households—taking advantage of more complete reporting of top incomes in tax data relative to surveys—may be fixing the downward inequality bias in the CPS data, while simultaneously introducing a notable positive bias by altering the sharing unit.

Finally, we estimate the progressivity of federal income taxes at the household level and compare these with analogous estimates at the tax-unit level. We find that federal income taxes are less progressive at the household level than is observed when focusing exclusively on tax units. We also find the distribution of EITC benefits at the household level contains significantly more mass in the top three quintiles than the tax-unit distribution. Both differences are due to households containing multiple tax units, including some tax units that appear low-income individually but have higher incomes when observed collectively. The income tax code targets the distribution of tax-unit income, not household income, and these multi-tax-unit households weaken the link between taxes and household income.

II. Background and Previous Literature

The concerns addressed in this work regarding IRS tax return data—the inability to observe households and the treatment of nonfilers—have long been recognized as important for inequality measurement and have been viewed as limitations of these data. This section considers previous research on these issues.

A central question when considering any income-distributional analysis, not just those using tax data, is the appropriate grouping of resources between people (that is, the "sharing unit"). In general, people do not consume out of their own income only, but instead consume based on the joint resources of their nuclear family, other relatives, and cohabiting partners. Hence, to avoid incorrectly classifying nonworking individuals living in a high-income household as having little or no income, inequality research typically assumes at least some resource sharing. This choice has been shown to greatly affect observed inequality trends (Burkhauser, Larrimore, and Simon 2012).

The U.S. tax system operates using a "tax unit" as the sharing unit, which groups together spouses who file a tax return together and those they claim as dependents for tax purposes (primarily children under age 19 and children under age 24 who are full-time students). This is distinct from grouping together all individuals living together at a physical address (a household sharing unit) or grouping together all individuals living together and related by blood or marriage (a family sharing unit). While there is some disagreement regarding whether the household or family sharing unit is preferable, numerous researchers have argued that the household is the sharing unit most closely resembling how individuals share economic resources (for example, Atkinson, Rainwater, and Smeeding 1995; Sheridan and Macredie 1999; Smeeding and Weinberg 2001; Congressional Budget Office 2018). The household is also the traditional sharing unit recommended by the Canberra Group for measuring income (United Nations Economic Commission for Europe 2011), and it is commonly used in analyses of national (Burkhauser et al. 2011) and cross-national inequality statistics, including Atkinson and Brandolini (2001) and the Luxembourg Income Study. We are unaware of any research suggesting that the tax unit is a preferable sharing-unit concept.

Because tax returns are submitted to the IRS at the tax-unit level, even researchers who prefer the household as the sharing unit have, out of necessity, focused on the tax unit as the sharing unit when using IRS tax records and treated it as a proxy for the household (for example, DeBacker et al. 2013; Chetty, Hendren, and Katz 2016; Chetty et al. 2018). As a result, researchers using tax return data have treated adult children who file their own tax returns but live with their parents as independent households. Similarly, they treat two cohabitating adults as independent households. This approach contrasts with the U.S. Census Bureau's official income statistics based on the CPS, where individuals residing together, but who file separate tax returns, are treated as a joint entity (Proctor, Semega, and Kollar 2016). As discussed by Atkinson, Piketty, and Saez (2011), this approach also can result in inconsistencies in international comparisons to countries that do not use the same tax-unit definition as the United States. While Alvaredo et al. (2016) remark that the difference between tax units and households is likely to most affect estimates for developing countries, our results suggest that these concerns remain for developed countries as well. Constructing data at the uniform household level will allow for more consistent estimates in cross-national comparisons.

Due to IRS data limitations, few researchers using tax records have attempted to create households with these data. Previous efforts to link tax units into households focused on statistical matches based on observable characteristics (Congressional Budget Office 2018) or direct links between Census Bureau survey data and administrative records (for example, Abowd and Stinson 2013; Wagner and Layne 2014). While a direct link between Census Bureau survey data and administrative records is a promising avenue, this form of matching is not covered under our current data-sharing agreements for the highly detailed address data we use for this paper. Additionally, previous research on such matches has found that this match is imperfect because between 8 and 12 percent of survey records cannot be matched to administrative data (Bond et al. 2014). These unmatched observations disproportionately occur among children, minorities, and low-income individuals. Furthermore, both the statistical matching and direct linking techniques using surveys may suffer from nonresponse error at both tails of the distribution, as demonstrated by Atkinson, Piketty, and Saez (2011); Bollinger et al. (2019); and Hokayem, Bollinger, and Ziliak (2014). Outside of these efforts to link administrative data to survey records, virtually all research based on tax return data assumes that resources are only shared within a tax unit, rather than among an entire household.

The second concern addressed in this paper—the representativeness of the sample population in IRS tax data due to nonfilers—is a well-known limitation of the IRS tax return data. The tax data used by most tax researchers, which sample from annual individual income tax returns (for example, Form 1040), excludes from the sampling frame the nearly 15 percent of adults and 13 percent of household heads who do not file a tax return and are not claimed as dependents each year (Auten and Gee 2009; Molloy, Smith, and Wozniak 2011). Were these nonfilers missing at random, the data would still be representative of the overall population, but this is not the case. Nonfilers are concentrated in the lower tail of the distribution, below the income threshold that legally requires filing a tax return (Langetieg, Payne, and Plumley 2017). Consequently, researchers using tax return data observe only a truncated version of the income distribution.

Many researchers partially overcome this problem by using tax return data to analyze only the top of the distribution and assume that all nonfilers have an income of 20–30 percent of average filer income (Piketty and Saez 2003; Auten and Splinter 2018). Yet, such an approach cannot be expanded to analyze lower-tail or distribution-wide inequality measures because it does not capture actual incomes for these nonfilers. Other researchers opt to ignore the nonfiler problem and only analyze the filing population despite the potential biases of missing lower-income individuals (Hungerford 2011; DeBacker et al. 2013; Congressional Budget Office 2019). A more sophisticated approach, which we build on, is that of Mortenson et al. (2009) and Chetty et al. (2014), who incorporate data from information returns (such as Forms W-2 and 1099) that the IRS receives for everyone with income from specific sources, even if that individual does not file a tax return. They use these data to construct the incomes of nonfilers. In 2010, we find that roughly 99.8 percent of the U.S. Census resident population has at least one information return or tax return filed to the IRS. Hence, in this approach, the data include nearly all people living in the United States. However, information returns are at the individual level and lack links to any other members of the household, including spouses and children. Because person-to-person links are not available for nonfilers, previous efforts to include nonfilers either focus exclusively on individual-level incomes (Larrimore, Mortenson, and Splinter 2016), base relationships on tax filing statuses in other years (Chetty et al. 2014), or use random pairings of nonfilers to simulate marriages and other relationships between nonfilers (Joint Committee on Taxation 2015).

Because the IRS data lack links for nonfilers to both relatives and others in the household, the two problems described above—the lack of information on nonfilers and the inability to organize individuals in tax records into true households—present overlapping challenges that need to be addressed simultaneously. Since nonfilers do not appear on a tax return and have no natural tax unit, any reasonable correction to the problem of nonfilers also requires determining with whom they share resources. By creating households using address fields from tax and information returns, we can incorporate these nonfilers into households and provide them an equivalent treatment to that given to filers. Hence, the approach taken here improves upon previous attempts to include nonfilers in tax-based analyses.

III. Data and Methods

The primary data for this paper are drawn from the universe of federal income tax data in 2010 collected by the IRS that have recently been used by Chetty et al. (2014) and Chetty, Hendren, and Katz (2016) to study income mobility questions. These data include both tax returns received on time and tax returns filed late but prior to our analysis of the data in 2018. In contrast to the public use and confidential versions of the Individual Income Tax Files produced by the Statistics of Income (SOI) division of the IRS, which have historically been used as the principal data sets of tax researchers (for example, Piketty and Saez 2003; Auten and Gee 2009), these data contain every individual who appears on a tax or information return. This universal coverage of tax data and near universal coverage of the U.S. population ensure that all individuals within households who are observed by the IRS appear in our data, which is necessary for aggregating observations to the household level.

The base IRS data contain annual income tax returns (for example, Form 1040 or Form 1040-EZ) and information returns, including Form W-2 (wage income), Form SSA-1099 (Social Security income), Form 1099-G (unemployment income), Form 1099-INT (interest income), Form 1099-DIV (dividend income), Form 1099-R (retirement savings distributions), Form 5498 (retirement savings rollovers), and Form 1099-MISC (miscellaneous income). Every tax form contains information on annual income for an individual or married couple from specified sources or, in the case of the annual income tax returns, income from all taxable sources. Each form also contains individual identifiers, such as the Taxpayer Identification Numbers (TINs, usually Social Security numbers), and mailing addresses. While annual income tax returns only exist for those who file a return, the IRS receives information returns on behalf of almost all adults without direct action from the taxpayer.

A. Calculating Income in Tax Data

Income reported to the IRS on both annual tax returns and on information returns is generally considered to be an accurate representation of individual incomes from taxable income sources. These taxable income sources include wages, self-employment income, interest, dividends, rents, certain business income, and taxable public transfer income. However, recognizing taxpayers' financial incentives to underreport their income, not all income will be reported to the IRS despite penalties for misreporting. Using IRS audit data, Johns and Slemrod (2010) estimate that approximately 11 percent of income that should appear in the adjusted gross income (AGI) on tax returns is not reported, although this has a neutral effect on the income distribution, as seen by the similarity between the distributions of estimated true AGI and reported AGI. These concerns are not limited to the IRS data, as Hurst, Li, and Pugsley (2014) observe that similar underreporting is found in survey data, potentially due to fears of self-incrimination by reporting different income amounts to the IRS and to household surveys.

An additional limitation of measuring income in tax data is that the IRS generally does not collect information on income from nontaxable sources. For this reason, several important sources of income to low-income households—including workers' compensation, supplemental security income, and public assistance welfare payments—are all excluded from these data (for a comparison of IRS and Census Bureau income concepts, see Henry and Day 2005). Furthermore, taxable income is defined based on current tax laws rather than economic income concepts, such as a Haig–Simons income definition (Slemrod 2016). Among other differences from a Haig–Simons definition, the IRS data exclude in-kind income sources—including food stamps, public housing, and until recently the value of government and employer-provided health insurance—which affects the distribution of observed incomes, especially among lower-income individuals (Burkhauser, Larrimore, and Simon 2012). This in-kind income is also excluded from the Census Bureau's income measure when computing official inequality statistics.

While acknowledging these limitations of the standard income definition used by tax researchers working with IRS tax data, we focus on *pre-tax cash income* excluding capital gains and excluding income sources not reported to the IRS, without attempting to impute nontaxable income sources that are excluded from IRS data collection. For annual tax returns, this definition starts with the total income from line 22 of IRS Form 1040, which includes income from wages, salaries, taxable interest, dividends, alimony, business income, rents and royalties, taxable Social Security, taxable private retirement income, and unemployment compensation. Five adjustments are made to this income from the Form 1040: (i) nontaxable interest reported on Form 1040 is added, (ii) realized capital gains (from Schedule D) are removed, (iii) taxable Social Security benefits are replaced by total Social Security benefits reported on Form SSA-1099, (iv) taxable private retirement income is replaced with gross private retirement income to reflect retirement savings distributions less rollovers from Forms 5498 and 1099-R, and (v) incomes are bottom-coded at $1 to limit the effect of business losses. This income measure is broader than the tax return income definition used by Piketty and Saez (2003), since it includes Social Security income and unemployment compensation. It also comes as close as possible to the pre-tax income measure from the CPS and used by the Census Bureau for their official income statistics. The primary difference between our income measure and the income measures used by the Census Bureau for their official income statistics is that we are not able to observe nontaxable cash transfer income such as public assistance and supplemental security income. As illustrated in Appendix Table A1, these nontaxable cash transfers represent approximately 2.5 percent of income reported by the Census Bureau—although they represent a larger share of income among low-income households.

For nonfilers, pre-tax cash income is calculated as the sum of income reported on information returns that would be included in the income definition for filers were they to file a tax return. Following Mortenson et al. (2009), who also derive income for nonfilers based on information returns, we include income from wages and salaries reported on Form W-2, unemployment benefits from Form 1099-G, Social Security and disability benefits from Form SSA-1099, interest income from Form 1099-INT, dividends from Form 1099-DIV, gross private retirement income as retirement savings distributions less rollovers from Forms 5498 and 1099-R, and self-employment income from Form 1099-MISC. This is a broader set of information return income than is used by Chetty et al. (2014), who also construct nonfiler income from information returns, but only use income found on Forms W-2, 1099-G, and SSA-1099. In contrast to these earlier papers, however, for Form 1099-MISC we offset reported income by 70 percent to reflect that gross income from self-employment activities appears on the 1099-MISC information returns, but the associated business expenses do not. This offset is necessary to convert gross self-employment income to net self-employment income. To determine the 70 percent offset, we observe that among low-income tax filers, income reported on tax returns and from information returns (after the offset) are nearly equal, as seen below. Hence, when using information returns to estimate the income of nonfilers, we assume a similar offset to Form 1099-MISC income while preserving all income from other sources that appear on information returns.[1]

This use of information returns for nonfilers implicitly assumes these forms accurately reflect the income they would report were they to file an annual tax return. To gain insight into the validity of this assumption, Figure 1 compares the tax-unit income on information returns for low-income tax filers with the amount reported on annual tax returns. This figure focuses on the lower half of the distribution, given previous findings, including that by Langetieg, Payne, and Plumley (2017) that nonfilers have substantially

1. Recognizing that the filing threshold for net earnings from self-employment is only $400, most self-employment income should appear on annual tax returns. Since we include nonfilers with apparent self-employment income above this threshold and nonfilers with total income above the general filing threshold (up to $100,000), this accepts that there is some degree of filing noncompliance among those both with and without self-employment income.

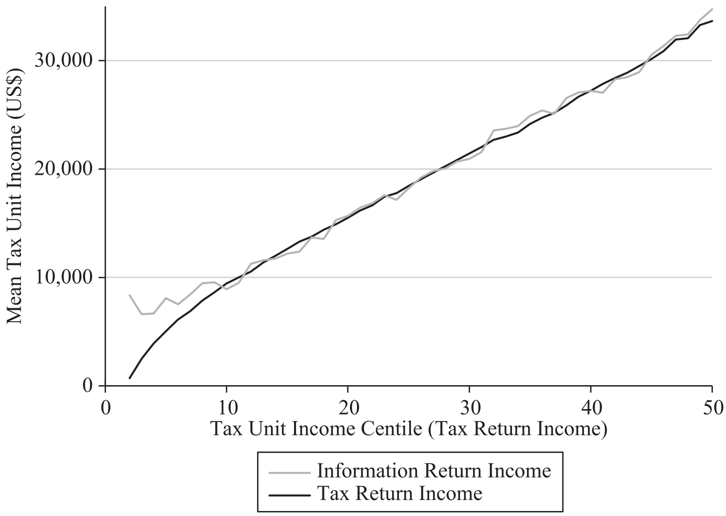

Figure 1

Comparing Income of Tax Filers from Information Returns and from Tax Returns, 2010

Source: Authors' calculations using IRS Statistics of Income data.

Notes: Tax return pre-tax income is total taxable income reported on tax returns, but adding nontaxable interest and nontaxable Social Security benefits, and excluding private retirement income and realized capital gains. Income is not adjusted for tax-unit size. Private retirement income is excluded to reflect that retirement income in this paper is gross private retirement income from information returns rather than coming from the tax return directly (see Section II for details). Information return income includes wages from Form W-2, dividends from Form 1099-DIV, interest from Form 1099-INT, unemployment benefits from Form 1099-G, benefits from Form SSA-1099, and 30 percent of earned income from Form 1099-MISC. Incomes are bottom-coded at $1. Centiles range from 1 to 100, each centile has an equal number of tax units, and ranks for both incomes are based on tax return income.

lower incomes than timely filers. If information return income accurately proxies the income of low-income filers, it should increase the confidence in using information returns to capture the income of nonfilers.

In this figure, centiles are defined based on the income reported on the annual tax return form, so the tax units in each centile are the same across the two series. When aggregating income from information returns, with the 70 percent offset of 1099-MISC income to reflect their estimated business expenses associated with that income, the two sources of income data track closely, except for the bottom 5 percent of the distribution, where there is more income reported on information returns than on tax returns. This lower income on tax returns relative to information returns at the very bottom of the distribution may reflect either noncompliance among some of these tax filers or additional business deductions (which may lead to net business losses) that are not observed on the information returns. However, there is no evidence that the information returns are systematically missing substantial income among the low-income filing population.

B. Comparison of Population Counts to Census Bureau Results

The suitability of using IRS tax data to evaluate the entire U.S. income distribution depends on whether these forms can accurately capture the entire U.S. population. To assess their capacity in this regard, we compare the population count and number of households in the tax data with analogous estimates reported by the U.S. Census Bureau from the decennial census. In 2010, 308.1 million individuals living in the United States appear in these tax data. This includes 281.5 million individuals who appear on a tax return as a primary filer (132.3) or as a spouse or dependent (149.2), along with 26.6 million nonfilers for whom there is at least one information return.[2] The 308.1 million people observed in tax data is comparable to the 308.7 million individuals in the United States observed in the 2010 decennial census. Hence, while only 91.2 percent of individuals appear on an annual income tax return, when including both the filing population and the nonfiling population with information returns, these tax and information return data observe 99.8 percent of the overall U.S. population in 2010. This observation is consistent with the findings of Cilke (2014) that 99.5 percent of the 2011 resident population was on either an annual tax return filing or an information return. The small number of individuals who do not appear on any IRS tax forms consist of dependents not captured by our tax data (discussed below) and a small number with no reported income or taxable government benefits who cannot be claimed as a dependent by another filer to obtain a tax benefit.

In addition to nearly matching the aggregate count of individuals, tax record data also produce a similar age distribution to that seen from the decennial census. This similarity, as well as the importance of incorporating nonfilers in the analysis, is apparent in Figure 2. The dashed gray line represents the age distribution of the U.S. resident population from the 2010 decennial census. When considering only the resident tax-filing population (solid gray line), a sizeable number of individuals at almost every age are missing from these data. In contrast, in tax data including all individuals on a tax return or for whom there is an information return (solid black line), the age distribution closely mirrors that observed in the decennial census. To the extent that deviations exist between the tax data and the decennial census results, the tax data observe more children under age ten, whereas it observes fewer teenagers ages 15–20 and middle-age adults ages 40–55.[3] The underestimate of about one million teenagers aged 15–20 likely occurs because children older than 16 do not qualify for the child tax credit, which reduces the benefits of claiming these children as dependents, and those with no independent sources of income will neither file nor appear on information returns.

2. This primary filer count excludes filers claimed as dependents on other returns, which avoids double-counting these individuals.

3. As individuals filing tax returns may legally claim a child exemption for children living in Canada or Mexico (but not other foreign countries), there was a surge in these children on tax returns coinciding with both increased immigration in the early 2000s and with expansions in the refundable child tax credit in 2001 and 2004. Since these children are not authorized to work in the United States, they cannot receive Social Security numbers but instead receive Individual Taxpayer Identification Numbers (ITINs). To limit the number of these nonresident dependents, we remove a tax return's third and fourth dependents if they have ITINs. This adjustment corrects for the large overstatement of resident children in the IRS tax data observed by Cilke (2014).

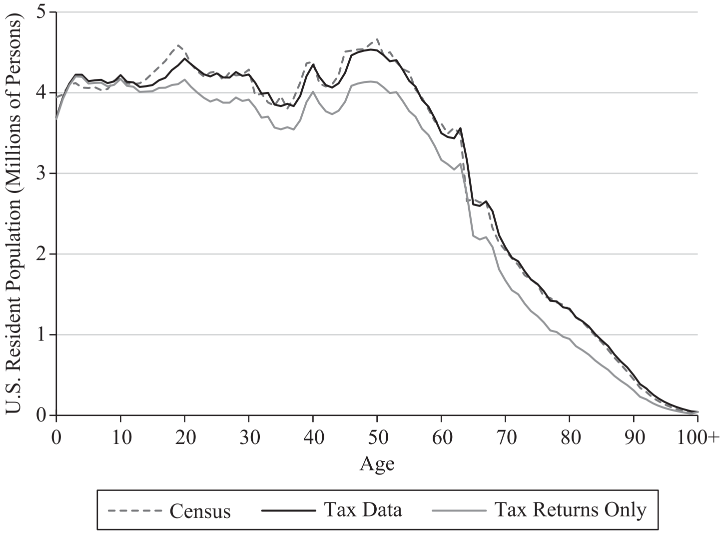

Figure 2
Number of Individuals by Age, 2010

Source: U.S. Census Bureau 2010 decennial census, Tax Household Sample, and authors' calculations.
Notes: Tax data include persons on tax returns and information returns for the 2010 tax year.

Comparing the distribution of individuals across states in Appendix Table A2, the population distribution across states is similar in the tax-based household data to that observed in the decennial census. In most cases, the state populations in tax data are within 1 to 2 percent of those seen in the decennial census. However, Alaska has around 4.6 percent more people in tax records than in the decennial census, which could be evidence of individuals selectively choosing their legal residence to access Alaska's Permanent Dividend Fund.

C. Forming Households and Cleaning Addresses in Tax Data

After aggregating tax forms to the individual level and establishing that the population counts using these data are consistent with those from the Census Bureau, individuals are aggregated into households using reported addresses and zip codes. Prior to linking tax returns by physical address, we link all individuals who appear together on the same tax return as either a primary filer, a spouse, or a dependent. Importantly, all dependents—even adult dependents—are considered to be part of the claimant's household and are not treated as separate economic units. This is consistent with the tax definition of a dependent, where an individual can only be claimed as a dependent if they fail to cover at least half of their own expenses. Consequently, all individuals in a tax unit are treated as being part of the primary filer's household and as having the same address as the primary filer. This is true even if one or more individuals list a separate address on their own tax forms. Most frequently, dependents with different addresses are likely children away at college for part of the year, who can still be claimed as a dependent if they are under age 24 and are a full-time student for at least five months of the year.

After constructing complete tax units, we turn to linking tax units and nonfiling individuals into households by physical address. We allow only one address per person. Filer addresses are always taken from tax returns if available. Nonfiler addresses come from information returns. If a nonfiler has multiple information returns that include both a street address and a P.O. box address, we use the street address. In the rare case of multiple street addresses on information returns, we sort the addresses numerically and alphabetically and choose the first address after sorting.

To construct the household-level file, these addresses are recoded into a standard form (for example, recoding "1ST ST" or "FIRST STREET" to "FIRST ST" and then removing all spaces), and individuals are considered to live together if their address and five-digit zip code both match (all address corrections included in this recoding are provided in the Online Appendix). For individuals living in an apartment or multi-unit building, the unit number must match, as well as the main address. To limit false matches for multi-unit buildings, we identify multi-unit building addresses that are missing unit identifiers and divide these tax units into separate households. For example, if at least three tax units have addresses of type "1 MAIN ST APT XX" with apartment numbers included, and two other tax units in the same zip code have addresses of "1 MAIN ST" but with no apartment number, we assume the two tax returns are simply missing unit information, and they are treated as separate households. This approach helps reduce the likelihood of false-positive merges from address-reporting errors on the tax forms.

Even after our extensive standardization of common address abbreviations, misspelling of street names remain. To link records due to close misspellings, we implement near-year matches. First, we identify misspelled street names by comparing our addresses to a master list of street names. This master list was provided by the address verification company SmartyStreets and includes 5,087,497 zip code–street name combinations (SmartyStreets 2019). Before making this comparison, uncleaned street names in the tax records are edited to be letter-only street names by: (i) converting number streets to letters as described in the address standardization process above, (ii) removing all remaining numbers including house and apartment numbers, and (iii) removing leading and trailing characters such as "APT" or "STREET." We then observe whether the street listed on each unmatched tax-unit household exists on the master list of street names in the taxpayer's zip code. For any unmatched tax unit with an invalid address, including a missing address or a P.O. box address, we first attempt to correct the address and zip code by replacing them with the next-year tax return or information return data.

The next year's address information is used if it meets specific criteria for its similarity to the current year (invalid) address. The next year's address must not be missing or a P.O. box. It also must have either the same first two digits of the house/apartment number and a different zip code where at least three of five digits are the same (to correct an apparently small number of misreported zip codes), or the same first two digits of their house/apartment number and the same zip code (to correct misspelled street names). Since these similarity tests are not possible when the current year's address is missing or is a P.O. box, the next-year physical address is also used if the current-year address is missing a street name or an unmatched P.O. box (to account for missing street addresses). This matching process is repeated with prior-year addresses. We then use these cleaned addresses to link individuals into complete households.

Finally, we recognize that some individuals with the same physical address live in group quarters, such as a dorm or a nursing home, and are not sharing a household in the traditional sense. These group quarters are flagged in the Census Bureau's CPS data and excluded from their household income statistics. We also drop individuals appearing to reside in group quarters. Since the IRS tax data do not classify the type of housing unit, addresses with 11 or more individuals and at least two tax units are treated here as group quarters, which captures nine million individuals at one-half million addresses. This approximates the eight million individuals listed as living in group quarters in the 2010 decennial census. Removing those in apparent group quarters also limits the extent to which erroneous links may affect our household income statistics—as could occur in cases such as a paid preparer using their business address on tax returns rather than taxpayers' addresses.

Our household identifiers for the population will be available to researchers with access to the IRS data files, including researchers outside of the federal government who access tax data through the IRS Statistics of Income Joint Statistical Research Program (for details on this research program, including application information, see Internal Revenue Service 2018). After creating household links for the population, we create the final Tax Household Sample (THS) by extracting a random 5 percent sample of households based on the last four digits of the TIN of one member of each household.[4]

The effect of each step described above on the number of observed households is outlined in Appendix Table A3. The vast majority of the difference between the number of households in the THS and the original number of tax units comes from linking by the unedited address data (and dropping a small number of group quarters). That is, without any additional data cleaning, there are almost 38 million fewer households in the tax data than there are tax units: 119.9 million versus 157.5 million. Splitting multi-unit building addresses that are missing unit identifiers increases the number of households by nearly one million. Standardizing abbreviations decreases the number of households by roughly four million, and cleaning based on near-year entries from the same taxpayer decreases it by roughly two million, yielding our final set of 115.3 million weighted households in the THS data set.

D. Limitations of Using Households Constructed from Address Fields in IRS Data

Although the IRS tax records data have many advantages, there are some limitations of constructing households in IRS data, in addition to the general limitations of measuring income in IRS data discussed in Section III.A. First, while the Statistics of Income

4. The representative of each household is the household member with the largest TIN. All representative individuals whose four-digit TIN ending is 500 of 9,999 possible combinations are selected into the household sample (no TINs end in all zeros). Sampling on four-digit TIN endings is an established random sampling method, regularly used by both the Social Security Administration and the IRS Statistics of Income division for the creation of their random samples (Smith 1989; Internal Revenue Service 2015).

Division (SOI) at the IRS produces a cleaned data file for a subset of tax returns, the universe level IRS tax files that are necessary to link all records into households do not undergo editing by the IRS. Consequently, there is the potential for data entry errors. To alleviate the risk of extreme outliers altering distributional results, we examined the data for any households with incomes larger than the largest tax-unit income in the cleaned tax records file produced by SOI. Since the SOI file includes the full population of top earners making more than about $7.5 million, any values in the unedited file above those seen in the SOI file must be erroneous. Such extreme outliers exist in other years, but in 2010, which is the focus of this paper, no such cases exist, and no records were removed for this reason.

A second limitation is the potential for false positive matches resulting from erroneously linking individuals into a single household who live separately. These can occur for several reasons, including fraudulent returns, paid preparers using their business address on tax returns they file on behalf of others, outdated addresses for individuals who move, or data entry errors for address information.

Although erroneous links from fraudulent returns sending refund checks to a single address are a potential concern, the IRS devotes substantial resources to identifying fraudulent returns, and returns initially rejected are not processed and therefore excluded from the population tax return files that we use. Although paid preparers using the same address for multiple returns may be in our initial file, we failed to find evidence that paid preparers are systematically using a single address when filing tax returns. Additionally, if the returns they file with that address contain at least 11 individuals, the return would be dropped through our group-quarters flag—as would any other false positive matches that erroneously combine at least 11 people into a single household. Hence, paid preparers could result in some individuals being dropped from our file, but would not result in the erroneous linking of large numbers of tax returns for our inequality statistics.

The more substantial concerns are false positives due to movers and erroneous data entry. We attempt to reduce these matches by disallowing merges of tax records in apparent apartment buildings for individuals who do not provide an apartment number. Additionally, while outdated addresses are a concern for nonfilers and those who receive their refund as a direct deposit (or owe the IRS an outstanding balance), for taxpayers who receive a paper check their current address is necessary to receive their refund. Nevertheless, we cannot fully eliminate the potential for false positive matches in the data.

Finally, a third potential limitation is false negatives, where we fail to link individuals who live together. As is the case with false positives, false negatives can occur due to outdated addresses for some in the household or through data entry errors in the address fields. Our additional cleaning procedures, which address typos in the text portions of the address fields, are intended to reduce the potential for false negatives that come from data entry errors. However, as with false positives, we cannot fully eliminate the potential for false negatives in our data.

The similarity of the number of households we observe using address data in IRS records and the number observed by the Census Bureau, as discussed in Section IV below, suggests that the limitations of using address data do not substantially alter the number of households observed. But since it is possible that the effects of false positives and negatives offset one another, readers should be aware of these potential limitations when considering household-level results based on the address fields in IRS data.

IV. Comparing Household and Tax-Unit Characteristics

In this section, we compare the households formed from tax data as described in Section III with the number of tax units in the tax data and the number of occupied households from the 2010 decennial census and the 2011 CPS (which represents the 2010 income year and is the closest data available to the 2010 tax forms, which are filed at the beginning of 2011). In all three household data sets, including subsequent analyses, we remove individuals living in group quarters from the sample since these are usually not economic sharing units and are typically excluded from results using CPS and census data.

As seen in Table 1, in 2010 there are 115.3 million weighted households in the THS data, as compared to 157.5 million tax units. As discussed below, the larger number of tax units is due to multiple tax units living in one household. Compared to the number of households in survey data, the THS has roughly one million fewer households than the 116.7 million in the 2010 decennial census, and roughly two million fewer than in the 2011 CPS. In particular, as displayed in the household-size distribution in Table 1, the gap results from the THS having fewer households with two individuals. There are several potential reasons for this difference, including that dependent college students living in off-campus housing will typically be counted as part of their parents' household in the THS data but as part of their household near campus in the Census data. This difference explains almost all of the fewer households in the tax data.

Table 2 provides a first look at the substantial difference between households and tax units in our THS data. If households and tax units were the same, and every individual appeared on a tax return, all households would consist of one filing tax unit and zero nonfiling individuals. Instead, only 59 percent of households consist of just one filing tax unit and zero nonfilers. In other words, filing tax units are not direct proxies for 41 percent of households according to these data. Ten percent of households contain no tax filers and one nonfiling individual. The remaining 31 percent of households contain at least two separate filing tax units, two nonfiling individuals, or one of each.

While we cannot precisely identify the type of relationships in these multi-tax-unit households, both adult children living with their parents and cohabitation of unmarried partners have risen in recent years (for example, Dettling and Hsu 2014; Lundberg, Pollak, and Stearns 2016) and likely constitute a sizeable portion of these households. We can compare the relationships of those living in multiple-tax-unit households as captured by the CPS, which contains relationship information, and create tax units within households through the procedure from Burkhauser et al. (2012). When doing so, we observe that 49 percent of CPS households with multiple tax units in 2010 contain a nondependent adult living with his or her parents, and 29 percent contain a cohabiting couple (3 percent of which also contain an adult living with their parents). The remaining 24 percent of households with multiple tax units have neither a nondependent

Table 1
Number of Households or Tax Units by Size, 2010 (Thousands)

Size of Household or Tax Unit	Tax Data (Tax Units)	Decennial Census (Household)	March CPS (Household)	Tax Data (Household)
1	73,811	31,205	31,399	35,173
2	43,017	38,243	39,487	32,254
3	18,184	18,758	18,638	18,081
4	14,259	15,625	16,122	15,506
5	5,741	7,538	7,367	7,745
6	1,752	3,075	2,784	3,698
7 or more	752	2,272	1,739	2,868
Total	157,515	116,716	117,538	115,325

Source: American FactFinder (Table H13) from the U.S. Census Bureau 2010 decennial census, Census Bureau Families and Living Arrangements Historical Data (Table HH-4), IRS Statistics of Income data, Tax Household Sample, and authors' calculations.
Notes: In the tax data, all dependents are included in the household or tax unit of the person who claims them. This includes children who are away at college, who would be treated as living at their college address in either the decennial census or the March CPS. Individuals living in group quarters are excluded, which is defined in the tax data as households with 11 or more individuals.

adult living with their parents nor a cohabiting couple—and therefore include either roommates or relatives besides parents and children who are living together.

The relationships in multiple-tax-unit households for those at the top of the income distribution are similar to those for all households, but with somewhat more nondependent adults living with parents and somewhat fewer cohabiting couples. Among households with multiple tax units in the top 5 percent of the income distribution in the CPS, 63 percent contain a nondependent adult living with his or her parents, 22 percent contain a cohabiting couple (3 percent of which also contain an adult living with his or her parents), and 18 percent contain neither a nondependent adult living with his or her parents nor a cohabiting couple. Multiple-tax-unit households in the top 1 percent of the income distribution in the CPS have a similar distribution of relationships as those in the top 5 percent.

Figure 3 displays where tax units residing with other tax units fall in the tax-unit income distribution. Figure 4 displays where households containing multiple tax units fall in the household income distribution. Figure 3 suggests that many of the tax units residing with others have relatively low incomes. Two-thirds of tax units in the bottom quintile of the tax-unit income distribution live in a household with at least one other tax unit. The likelihood of a tax unit living with others declines as the tax-unit income increases. Many tax units living with others, however, fall into relatively high-income households (Figure 4). In part, this reflects that multiple-tax-unit households have more earners, which can push their joint incomes further up the household income distribution. But it also reflects that some low-income tax units are living in the same household as tax units whose income is much higher.

Table 2
Household Composition by Number of Filing Tax Units and Nonfiling Individuals in the Household, 2010

		Nonfiling Individuals in the Household		
		0	1	2+
Filing tax units in the household	0		10.6%	1.8%
	1	58.8%	3.7%	0.6%
	2+	22.5%	1.8%	0.3%

Source: Tax Household Sample and authors' calculations.
Notes: Dependent filers and dependent nonfilers are included as part of the tax unit of those who claim them as a dependent. In constructing households, all dependents are included in the household of the person who claims them.

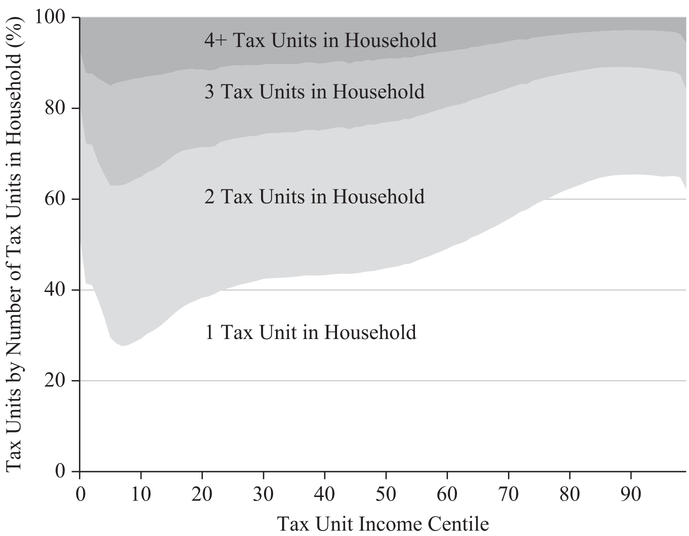

Figure 3
Number of Filing Tax Units and Nonfiling Individuals per Household by Tax-Unit Income, 2010

Source: Tax Household Sample and authors' calculations.
Notes: As in Table 2, counts of tax units in this figure are based on the number of primary filers and nonfiling individuals not claimed as a dependent. Individuals claimed as dependents, whether filing or not, and spouses on joint returns are not counted as separate tax units. Households with 11 or more individuals are excluded. For filers, pre-tax income is total taxable income reported on tax returns, but adding nontaxable interest, replacing taxable private retirement income with gross private retirement income, and excluding realized capital gains. For nonfilers, pre-tax income is wages from Form W-2, dividends from Form 1099-DIV, interest from Form 1099-INT, unemployment benefits from Form 1099-G, benefits from Form SSA-1099, gross private retirement income from Forms 5498 and 1099-R, and 30 percent of earned income from Form 1099-MISC. Pre-tax income excludes cash and in-kind transfer income that is not reported on individual tax returns. Income is at the tax-unit level and not size-adjusted.

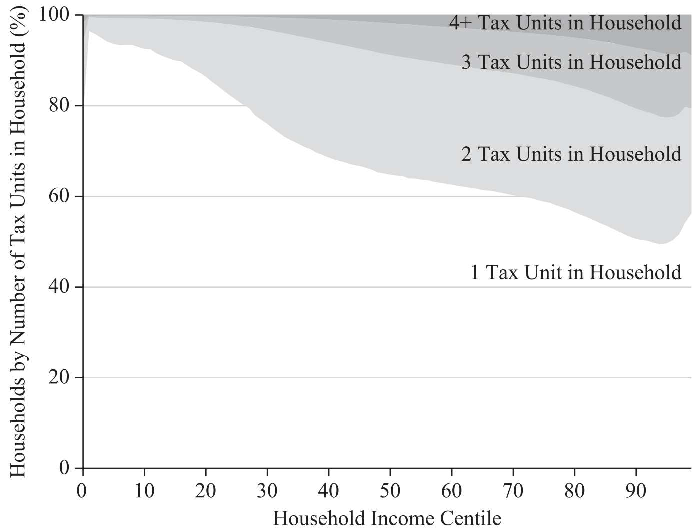

Figure 4

Number of Filing Tax Units and Nonfiling Individuals per Household by Household Income, 2010

Source: Tax Household Sample and authors' calculations.
Notes: See Figure 3 for details. Income is at the household level and is not size-adjusted.

V. Comparison of Income Distributions to Census Bureau Results

In this section, we compare the THS household income distribution with the tax-unit income distribution and the household income distribution in the 2011 March CPS (which covers income year 2010). While the income types in the tax unit and THS data are the same, there are several differences between how income is captured in IRS and CPS data. Specifically, supplemental security income (SSI), child support income, educational assistance, financial assistance, survivors' benefits, veterans' benefits, workers' compensation, and public assistance income are removed from the CPS income definition since they cannot be observed on IRS tax forms.[5] For the rest of this paper, we size-adjust incomes and set the person (rather than the tax unit or the household) as the unit of observation. Hence, each percentile of the distribution contains the same number of individuals. Specifically, we divide tax unit or household

5. One approach for creating a tax-based household income measure that matches the full pre-tax, post-transfer income definition used by the Census Bureau for their official income statistics in Proctor, Semega, and Kollar (2016) is to impute these sources into the tax data using statistical matches (Congressional Budget Office 2018; Larrimore et al. 2016). However, in order to focus solely on the income observed in tax records, we exclude these sources from both data sets while recognizing that including them would lower the observed levels of inequality.

income by the square root of the number of individuals in the tax unit or household.[6] This approach accounts for economies of scale and sharing within a household. As discussed by Citro and Michael (1995), a four-person household does not require twice the resources of a two-person household to obtain the same standard of living, and size-adjusting reflects those differences in needs. Our equivalence scale is similar to that used to estimate the official poverty thresholds and follows the conventional approach in the household-level inequality literature (Gottschalk and Smeeding 1997; Atkinson and Brandolini 2001).

Figure 5 compares the pre-tax household income distribution in the tax data and CPS data (additional detail on the top 10 percent is available in Online Appendix Figure B2). It also compares the THS and tax-unit distributions. For the tax-unit series in this paper, dependent filers are considered part of the return on which they are claimed and are not treated as independent tax units. Nonfiling individuals are paired semirandomly to match the total number of married couples. This approach reduces the difference between tax units and households relative to treating all nonfilers as single individuals. It also approximates the number of total tax units from the updates to Piketty and Saez (2003).

Comparing the tax-based and census household income series, it is apparent that the distribution of household incomes is similar across the two data sets, with the exception of the top centiles of the distribution. Household incomes are slightly higher in each centile of the tax-based household income distribution than in the CPS data, and, outside of the top decile of the distribution, the difference in per-person size-adjusted income is always less than \$5,000. This difference can largely be attributed to the IRS data better capturing retirement income than the CPS (see Munnell and Chen 2014; Bee and Mitchell 2017).[7]

The primary differences between the tax-based and CPS-based household income series occur in the top 2 percent of the household income distribution, where household incomes in the CPS fall well below income reported in the tax data. Relative to the tax-based household data, the CPS understates the mean size-adjusted income of the 98th percentile (P98–99) of the distribution by 24 percent (\$172,300 compared to \$227,000) and the mean size-adjusted income of the top 1 percent of households by 53 percent (\$322,900 compared to \$692,800). This finding is consistent with the view of many researchers—including Atkinson, Piketty, and Saez (2011)—that the CPS data fail to fully capture the income of high-income individuals.

There are more substantial differences between the distribution of household incomes and tax-unit incomes. The income of tax units in a given centile is typically well

6. We considered non-size-adjusted incomes and setting income groups by the number of tax units and households in an earlier version of this paper and our relative inequality results were similar. However, because households may have more people than tax units, the difference in levels of income between tax units and households were greater when not size-adjusting incomes.

7. While both data sets include retirement income, the CPS asks respondents about regular payments from IRA, 401(k), and Keogh accounts, whereas the IRS includes all withdrawals. Munnell and Chen (2014) observe that in 2012 the CPS captured \$18 billion of income from defined contribution plans, whereas IRS data observed \$229 billion from these plans. Additionally, Bee and Mitchell (2017) match CPS respondents in 2012 with tax data, and find substantial underreporting of pension income in the CPS. This results in the median income for those 65 and older being understated by 30 percent in the CPS data relative to their income reported in tax data.

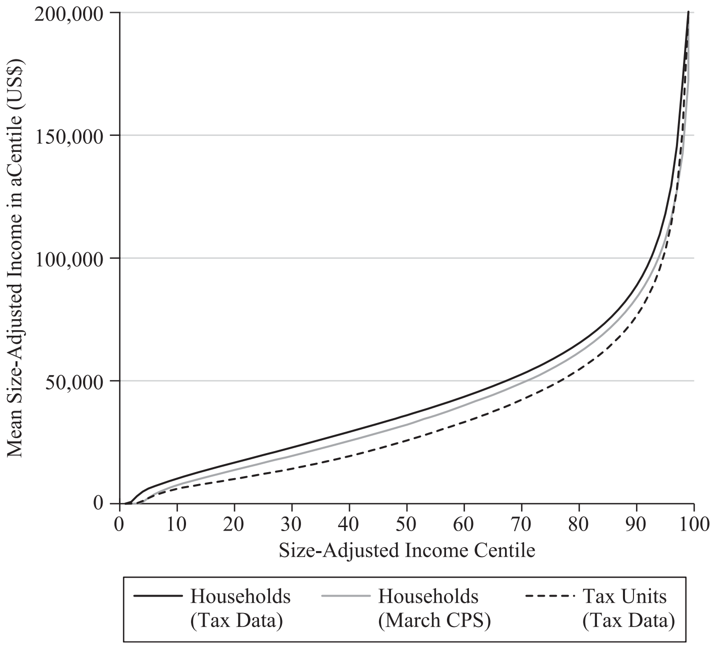

Figure 5

Distribution of Size-Adjusted Pre-Tax Income, 2010

Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, Tax Household Sample, and authors' calculations.

Notes: Incomes are size-adjusted, and income groups set by the number of individuals. For filers, pre-tax income is total taxable income reported on tax returns, but adding nontaxable interest and nontaxable Social Security benefits, replacing taxable private retirement income with gross private retirement income, and excluding realized capital gains. For nonfilers, pre-tax income is wages from Form W-2, dividends from Form 1099-DIV, interest from Form 1099-INT, unemployment benefits from Form 1099-G, benefits from Form SSA-1099, gross private retirement income from Forms 5498 and 1099-R, and 30 percent of earned income from Form 1099-MISC. It excludes cash and in-kind transfer income that is not reported on individual tax returns and is bottom-coded at $1. For the households series, individuals living in group quarters are excluded, which is defined in the Tax Household Sample as households with 11 or more individuals. Tax units include both nondependent filers and nonfilers. For the tax-unit series, in order to match the overall marriage rate among tax units, about 40 percent of nonfiler tax units are assumed to be married. All points are the mean income within the specified centile of the distribution.

below that of tax-based households. For example, while the median size-adjusted tax-unit income is $26,100, the median tax-based size-adjusted household income is $36,300. This pattern persists throughout the income distribution and suggests tax units are poor proxies for households when considering the distribution of income.

The ability to observe households directly in tax return data also offers a refined perspective on income inequality. Figure 6 displays Lorenz curves for each income series. The Lorenz curve represents the share of income held by those at or below each centile of the distribution: curves closer to the 45-degree line indicate distributions

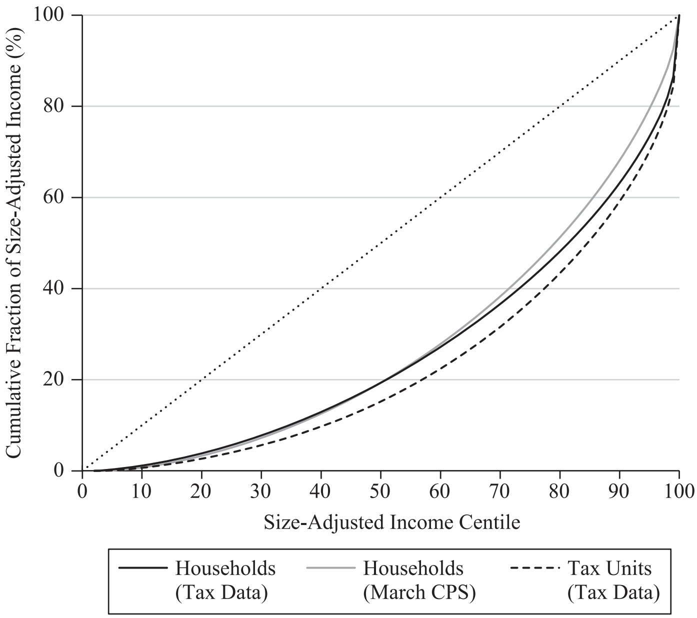

Figure 6

Lorenz Curve for Size-Adjusted Pre-Tax Income, 2010

Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, Tax Household Sample, and authors' calculations.
Notes: See Figure 5 for details.

that are more equal. Reflecting the better capacity of the IRS data to observe income at the top of the distribution, the tax-based household income series observes a higher concentration of income among the top centiles than does the CPS data. This higher concentration provides further evidence that household incomes are less equally distributed than is observed in the official income statistics released by the Census Bureau based on the CPS data.[8]

However, Figure 6 also illustrates the extent to which researchers using tax units to proxy for households will overstate the true level of household income inequality. Outside of the top 1 percent, the tax-unit series shows substantially lower shares of income relative to when income is aggregated to the household level.

The effect on the observed level of inequality from aggregating tax records into households is further apparent in Table 3, which presents key inequality statistics across the

8. Highlighting the importance of the top 2 percent of the distribution to the Lorenz curve, were you to replace the top 2 percent of the Census Bureau household income distribution with the top 2 percent of the tax-based household income distribution, the gap between the CPS and tax data household income Lorenz curves in Figure 6 would nearly disappear. This finding supports the mixed CPS/tax-return approach of distributing personal income used by Fixler, Gindelsky, and Johnson (2019).

Table 3
Comparison of Income Inequality Statistics for Pre-tax Income, 2010

	Tax Data (Household) (1)	Tax Data (Tax Unit) (2)	March CPS (Household) (3)	% Difference Using Tax Units (4)	% Difference Using March CPS (5)
Gini	0.477	0.538	0.453	12.9	–4.9
P90/P10	8.61	12.48	10.85	44.9	26.0
P80/P50	1.82	2.13	1.92	16.7	5.6
P50/P20	2.13	2.55	2.32	19.5	8.8
1st quintile share	3.84	2.66	3.30	–30.9	−14.1
2nd quintile share	9.07	7.07	9.25	–22.1	1.9
3rd quintile share	14.25	12.70	15.29	–10.9	7.3
4th quintile share	21.02	20.96	23.43	–0.3	11.5
Top quintile share	51.82	56.63	48.76	9.3	–5.9
Top 5 percent share	26.69	29.92	20.57	12.1	–22.9
Top 1 percent share	13.53	15.56		15.0	

Source: U.S. Census Bureau's March CPS and authors' calculations using IRS Statistics of Income data and the Tax Household Sample.
Notes: Incomes are size-adjusted, and income groups are set by the number of individuals. See Figure 5 for details. March CPS data are not available for the top 1 percent due to top-coding of the public-use CPS data. Column 4 is the percent difference using tax units instead of tax households, a comparison between Columns 3 and 1. Column 5 is the percent difference using the March CPS household income distribution instead of tax households, a comparison between Columns 4 and 1.

three measures. Relative to the tax-based household income series, using tax units overstates the level of inequality, and using the CPS data understates the level of income inequality. For example, the Gini coefficient for the new household series in tax data is 0.477, which is below the 0.538 Gini coefficient for tax units but above the 0.453 Gini coefficient for households in the CPS. Hence, using tax units as proxies for households will overstate the household income Gini coefficient by 13 percent, and using the CPS data will understate the household income Gini coefficient seen in the tax data by 5 percent.

The relative gap in inequality measures between tax households and tax units, or tax households and census households, varies across the income distribution. For income and inequality metrics that are not influenced by the very top of the distribution—such as the 90/10, 80/50, and 50/20 ratios—the tax household income inequality results are much more closely aligned with the household income inequality results in the CPS than the tax-unit series. However, looking at the top 5 percent income share results, where the known deficiencies in the CPS income data are greatest, the shares for tax units in the tax data are closer to the tax household results. Moving higher up the income distribution, we find the top 1 percent of households constructed from tax data earn smaller income shares than the top 1 percent of tax units: 13.5 percent versus 15.6 percent of income. This decline of two percentage points is consistent with the Bricker et al. (2016) estimate that shifting from tax units to households decreased top 1 percent income shares by 2.4 percentage points in the 2010 Survey of Consumer Finances.

VI. Distribution of Earned Income Tax Credits

Thus far, we have focused on the distributional effect of analyzing income at the household level, but the sharing unit is also important for other public policy questions, including the distribution of tax burdens and tax credits. Most analyses of the distributional effects of tax provisions focus on the distribution across tax units, as the underlying data used in these analyses are tax return data. Tax-unit distributions are often used in estimates by government agencies, think tanks, and others using tax data (Joint Committee on Taxation 2012; Tax Policy Center 2019; Hoynes and Patel 2018). An important exception to this approach for distributional burdens of tax provisions is the Congressional Budget Office (2018), which presents results at the household level. Yet, while the Congressional Budget Office prefers to present household-level distributional analyses, they are unable to observe actual households and instead create synthetic households through statistical matches with Census Bureau data. In contrast to their work, we consider the distribution of household income without relying on a statistical matching approach.

The distributional consequences of specific tax provisions may differ depending on whether households or tax units are used as the unit of analysis. Here we consider how the observed distribution of one of the most important social safety net programs—the Earned Income Tax Credit (EITC)—differs when focusing on household incomes rather than tax-unit incomes.

The EITC is a refundable credit intended to increase the after-tax incomes of low-income workers. This credit is available to tax filers with earned income and is substantially more generous for tax units with dependent children. For example, in 2010 a tax unit with two children was eligible for a maximum EITC of more than $5,000, while a childless tax unit could only receive around $450. The maximum income under which a tax unit remains eligible for the credit also varies by filing status and number of dependents: a single tax filer with two qualifying dependents in 2010 must have had less than $40,363 in adjusted gross income to be eligible, while a single childless tax unit could only earn up to $13,460 to remain eligible. In tax year 2010, more than 27 million tax filers claimed about $60 billion of EITC credits. This means the EITC has more than ten times the number of recipients and more than seven times the cash benefits of the traditional cash welfare program Temporary Assistance for Needy Families (Bitler, Hoynes, and Kuka 2017).

The importance of the unit of analysis when evaluating the distributional impacts of the EITC is apparent in Figure 7. This figure shows the fraction of tax units in each centile of the pre-tax size-adjusted income distribution that claim the credit. While earnings requirements for the credit mean that few tax units at the very bottom of the distribution claim the EITC, claiming rates rise to about 70 percent in the second decile of the size-adjusted tax-unit income distribution and 50 percent in the third through

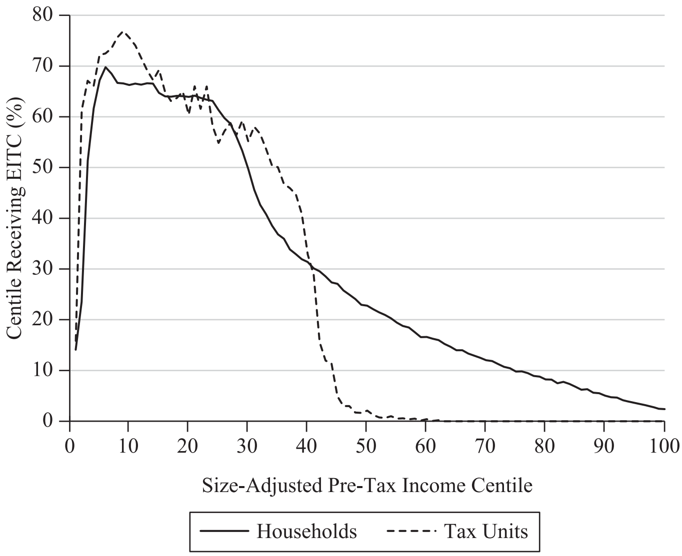

Figure 7

Share of Tax Units and Households Claiming the EITC by Size-Adjusted Income, 2010

Source: IRS Statistics of Income data, Tax Household Sample, and authors' calculations.
Notes: See Figure 5 for details.

fourth decile. Higher in the distribution, claiming rates fall sharply, and nearly no tax units in the top half of the tax-unit distribution receive the EITC.

When using households as the unit of observation, most claimants remain in the lower deciles of the distribution. However, 8 percent of households in the eighth and ninth deciles and 3 percent of those in the top decile receive EITC benefits. This result is due to a nontrivial number of individuals in relatively low-income tax units (thereby qualifying for the credits) residing in multi-tax-unit households with aggregate incomes beyond the end of the EITC's phaseout.

Figure 8 shifts from considering the fraction of individuals who take credits to the fraction of total credits claimed by each pre-tax income quintile. This distribution of benefits incorporates the number of individuals claiming credits and the amount of credits claimed. At the tax-unit level, again EITC benefits are well targeted at those in the bottom half of the distribution. Those in the bottom two quintiles of the pre-tax income distribution receive 99 percent of benefits, whereas just 1 percent goes to those in the top three quintiles. But by linking the tax units into households, it is apparent that a nontrivial fraction of benefits flow to those living in higher-income households. At the household level, 18 percent of EITC benefits go to those in the top three quintiles, which is similar to that observed by the Congressional Budget Office (2013) using synthetic households. If we ranked households based on non-size-adjusted incomes, as is common in distributional tax analyses such as those provided by the Joint Committee on Taxation and the Tax Policy Center, an even larger 37 percent of EITC benefits would

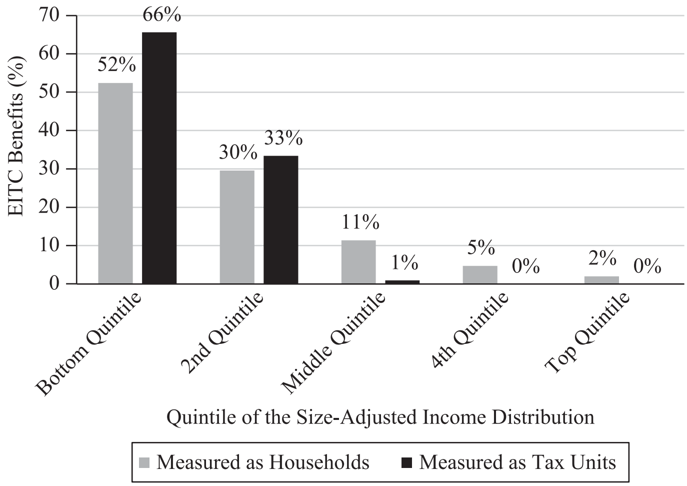

Figure 8

Distribution of the EITC, 2010

Source: IRS Statistics of Income data, Tax Household Sample, and authors' calculations.
Notes: See Figure 5 for details.

go to the top three quintiles of the household income distribution. Hence, while the benefits still appear to be targeted at those with lower incomes—even at the household level, a majority of EITC benefits go to the bottom two quintiles—the redistributive impacts are less pronounced than when the unit of analysis is a tax unit.

VII. Distribution of Tax Burdens

Federal individual income taxes, of which the EITC is one component, also appear less progressive at the household than the tax-unit level. As with the EITC, it is common to present tax-based tax distribution estimates using tax units, although Congressional Budget Office (2018) is a notable exception. Figure 9 compares average tax rates across the household and tax-unit pre-tax income distributions, where income groups are again set such that there is an equal number of individuals in each percentile. Federal income tax burdens are similar in the top half of the distribution but higher at the household level in the bottom half, suggesting federal taxes are less progressive when considering household incomes and tax burdens instead of tax-unit incomes and tax burdens.

A related measure of the distribution of tax liabilities is the share of the population that is paying no federal individual income taxes (see Splinter 2019 for additional discussion of this measure). Using this metric, only the bottom 32 percent of individuals are in households paying no federal individual income tax, compared to 37 percent who

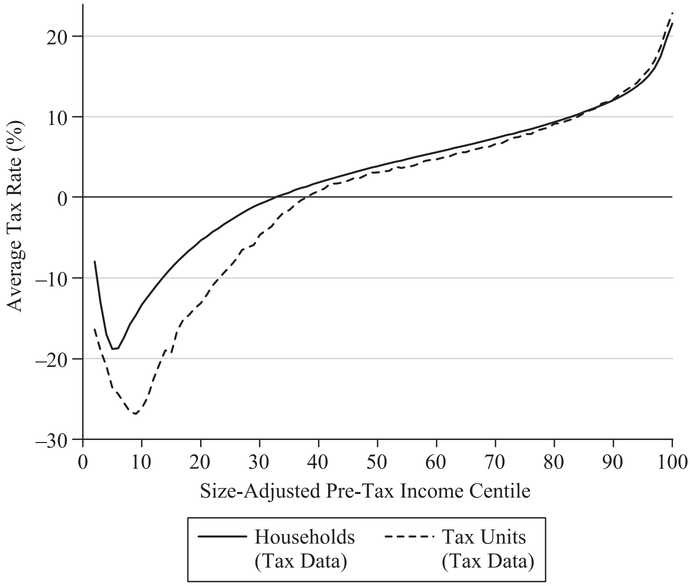

Figure 9

Effective Tax Rates, 2010

Source: Tax Household Sample and authors' calculations.
Notes: Pre-tax income is defined as in Figure 5, except realized capital gains are added to filer incomes. Only federal individual income taxes are considered and for filers are defined as taxes paid less refundable earned income and child tax credits received, and for nonfilers, are assumed to be zero.

are in tax units paying no federal income tax. Hence, the share of the population with no federal income tax liabilities is smaller when observed at the household level than at the tax-unit level.

We estimate distribution-wide tax progressivity using the Kakwani index. This index is computed as the tax concentration coefficient—a Gini coefficient-type measure of tax burdens where tax units or households are ranked by pre-tax income—less the Gini coefficient of pre-tax income (Slavov and Viard 2016). Tax progressivity falls from 0.416 at the tax-unit level to 0.358 at the household level, a decrease of 14 percent. This decrease in tax progressivity is unsurprising: taxes are allocated progressively by tax-unit level income; aggregating multiple tax units into one household weakens the link between taxes and income.

VIII. Discussion

Advances in administrative tax data have provided an increasingly detailed picture of the tax-unit income distribution but have not described income at the household level or fully incorporated the income of nonfilers. Using address fields on

IRS tax records and the universe of tax forms, we create the Tax Household Sample, which aggregates IRS tax records up to the household level. We then use these data to produce the first household-level income distribution constructed entirely from IRS tax records. In doing so, we confirm the failure of CPS data to fully capture the incomes of households in the top two centiles of the income distribution. This limitation reduces the observed pre-tax household income Gini coefficient in the CPS data by 2.3 Gini points in 2010, a 5 percent understatement of inequality relative to that observed in the tax data. However, we also observe that using tax units as proxies for households leads to an overstatement of household income inequality of 6.2 Gini points (13 percent). The inability of tax units to proxy properly for households reflects our finding that only 69 percent of households consist of a single tax unit or nonfiling individual.

The difference between tax units and households is also important for understanding the distributional impacts of the income tax system as a whole, as well as that of specific tax provisions such as the EITC. This tax credit is concentrated among lower-income individuals irrespective of the unit of analysis, although it is less progressive when income is measured at the household level. In particular, the share of earned income tax credits going to the top three quintiles of the income distribution rises from 1 percent to 18 percent when we shift the unit of analysis from tax units to households. This notable difference demonstrates the importance of the unit of analysis when estimating the progressivity of tax provisions.

Beyond its application to distributional analyses, the new tax-based household data set developed here allows for an expansion of the research topics for which IRS tax data may be suitable. This expansion includes topics for which household-level information is important, as well as those focused lower in the income distribution, for which the lack of information on nonfilers previously precluded the use of IRS data. For example, these data can be used for analyses of coordination of financial decisions within households. Research using household-level data shows that multi-tax-unit household appear to coordinate who claims a child for tax purposes (Splinter, Larrimore, and Mortenson 2017), and these data can be used for other questions regarding how individuals coordinate within their household. Other potential topics for which household-level information may enhance the analyses include research on the effects of living arrangement decisions, such as cohabitation, on subsequent financial outcomes, which previously would not have been possible with IRS data. The detailed address data can also be used to obtain better estimates of how geographic mobility relates to financial outcomes. Moreover, including income of cohabiting partners and resident family members, whether they file a tax return or not, can mitigate measurement error in studies using tax-unit income as a proxy for household income, including many studies of intergenerational income mobility (Chetty et al. 2014). Household-level links also provide a step towards producing tax-based measures of poverty. Additionally, there is a wealth of information that the IRS observes—including college attendance, health insurance coverage, and employer characteristics—which can be combined with the Tax Household Sample in order to address a broader range of policy questions.

Appendix

Table A1
Household Income by Source, 2010 (Millions of Dollars)

	Tax Data	Census
Earnings		
Wages and salaries	5,896,195	6,132,916
Self-employment and farm income (minus loss)	398,042	374,998
Other private income		
Partnership, S corporation, rent, royalty, estates/trusts (minus loss)	440,455	
Rent/royalty/estates/trusts (minus loss)		68,374
Interest and dividends	381,422	255,850
Pensions, annuities, and IRA distributions	930,257	369,166
Alimony	8,796	5,061
Other private income		7,625
Other income in Form 1040 total income	87,272	
Transfer income included on tax forms		
Unemployment compensation	140,671	97,361
Social Security and disability benefits	695,542	593,855
Total pre-tax income on tax returns	8,978,652	7,859,358
Cash transfer income in the March CPS that is not included on tax forms and excluded from this analysis		
Public assistance and SSI		47,111
Child support		26,422
Education assistance and financial assistance		80,000
Veteran's income and worker's compensation		47,831
Total nontaxable cash income in the March CPS excluded from this analysis		201,364

Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, Tax Household Sample, and authors' calculations.
Notes: Tax data amounts for alimony and other income (state tax refunds, gambling earnings, and other income less loss) are based on aggregate tax return data from IRS. Other tax data amounts are from the Tax Household Sample, but interest and dividends are based on total income plus tax-exempt interest less other sources.

Table A2
State Populations in IRS and Census Data, 2010

	Individuals			Individuals	
State	Decennial Census	Tax Data	State	Decennial Census	Tax Data
AK	710	743	MT	989	968
AL	4,780	4,730	NC	9,535	9,465
AR	2,916	2,849	ND	673	670
AZ	6,392	6,501	NE	1,826	1,838
CA	37,254	37,765	NH	1,316	1,326
CO	5,029	5,039	NJ	8,792	8,981
CT	3,574	3,524	NM	2,059	1,972
DC	602	607	NV	2,701	2,739
DE	898	910	NY	19,378	19,000
FL	18,801	19,157	OH	11,537	10,932
GA	9,688	9,887	OK	3,751	3,690
HI	1,360	1,350	OR	3,831	3,804
IA	3,046	3,018	PA	12,702	12,510
ID	1,568	1,558	RI	1,053	1,030
IL	12,831	13,020	SC	4,625	4,559
IN	6,484	6,395	SD	814	826
KS	2,853	2,854	TN	6,346	6,327
KY	4,339	4,225	TX	25,146	25,268
LA	4,533	4,496	UT	2,764	2,749
MA	6,548	6,430	VA	8,001	7,968
MD	5,774	5,887	VT	626	619
ME	1,328	1,311	WA	6,725	6,852
MI	9,884	9,677	WI	5,687	5,653
MN	5,304	5,351	WV	1,853	1,773
MO	5,989	5,846	WY	564	563
MS	2,967	2,915	Total	308,746	308,126

Source: U.S. Census Bureau 2010 decennial census, Tax Household Sample, and authors' calculations
Notes: Units are thousands of individuals. Census populations are calculated in March, and tax data population is based on the population on December 31. Individuals living in group quarters are excluded, which is defined in the tax data as households with 11 or more individuals. In the tax data, all dependents are included in the household of the person who claims them.

Table A3
Number of Households by Household Size, 2010 (Thousands)

Size of Household	Tax Data (Tax Units)	Unedited Addresses	Split Multi-Unit Addresses	Standardize Abbreviations	Next-Year Match	Prior-Year Match
1	73,811	40,802	41,653	37,075	36,257	35,173
2	43,017	32,520	32,595	32,689	32,485	32,254
3	18,184	17,877	17,925	18,128	18,088	18,081
4	14,259	15,262	15,262	15,416	15,428	15,506
5	5,741	7,447	7,428	7,597	7,640	7,745
6	1,752	3,490	3,471	3,578	3,617	3,698
7 or more	752	2,550	2,506	2,657	2,729	2,868
Total	157,515	119,947	120,839	117,140	116,243	115,325

Source: Tax Household Sample and authors' calculations.
Notes: Individuals living in group quarters are excluded, which is defined in the tax data shown here as households with 11 or more individuals.

References

Abowd, John, and Martha Stinson. 2013. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data." *Review of Economics and Statistics* 95(5):1451–67.

Alvaredo, Facundo, Tony Atkinson, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. "Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world." WID.world Working Paper 2016/1.

Atkinson, Anthony B., and Andrea Brandolini. 2001. "Promises and Pitfalls in the Use of Secondary Data Sets: Income Inequality in OECD Countries as a Case Study." *Journal of Economic Literature* 39(3):771–99.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. 2011. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49(1):3–71.

Atkinson, Anthony B., Lee Rainwater, and Timothy M. Smeeding. 1995. "Income Distribution in OECD Countries: The Evidence from the Luxembourg Income Study." Social Policy Studies 18. Paris: Organization for Economic Cooperation and Development.

Auten, Gerald, and Geoff Gee. 2009. "Income Mobility in the United States: New Evidence from Income Tax Data." *National Tax Journal* 62(2):301–28.

Auten, Gerald, and David Splinter. 2018. "Income Inequality in the United States: Using Tax Data to Measure Long-term Trends." Unpublished.

Bee, Adam, and Joshua Mitchell. 2017. "Do Older Americans Have More Income Than We Think?" Census Bureau SESHD Working Paper 2017-39. Washington, DC: U.S. Census Bureau.

Bitler, Marianne, Hilary Hoynes, and Elira Kuka. 2017. "Do In-Work Tax Credits Serve as a Safety Net?" *Journal of Human Resources* 52(2):319–50.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch." *Journal of Political Economy* 127(5):2143–85.

Bond, Brittany, J., David Brown, Adela Luque, and Amy O'Hara. 2014. "The Nature of the Bias when Studying only Linkable Person Records: Evidence from the American Community Survey." Census Bureau CARRA Working Paper 2014-08. Washington, DC: U.S. Census Bureau.

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus. 2016. "Estimating Top Income and Wealth Shares: Sensitivity to Data and Methods." *American Economic Review* 106(5):641–5.

Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 2011. "Trends in United States Income Inequality Using the March Current Population Survey: The Importance of Controlling for Censoring." *Journal of Economic Inequality* 9(3):393–415.

———. 2012. "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data." *Review of Economics and Statistics* 44(2):371–88.

Burkhauser, Richard V., Jeff Larrimore, and Kosali I. Simon. 2012. "A 'Second Opinion' on the Economic Health of the American Middle Class." *National Tax Journal* 65(1):7–32.

Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "Race and Economic Opportunity in the United States: An Intergenerational Perspective." NBER Working Paper 24441. Cambridge, MA: NBER.

Chetty, Raj, Nathaniel Hendren, and Lawrence Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106(4):855–902.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129(4):1553–623.

Cilke, James. 2014. "The Case of the Missing Strangers: What We Know and Don't Know about Nonfilers." In *Proceedings of the 107th Annual Conference of the National Tax Association.* Washington, DC: National Tax Association.

Citro, Constance F., and Robert T. Michael, eds. 1995. *Measuring Poverty: A New Approach.* Washington, DC: The National Academies Press.

Commission on Evidence-Based Policymaking. 2017. *The Promise of Evidence-Based Policymaking*. Washington, DC: Commission on Evidence-Based Policymaking.

Congressional Budget Office. 2013. "The Distribution of Major Tax Expenditures in the Individual Income Tax System." Congressional Budget Office Research Report.

———. 2018. "The Distribution of Household Income, 2014." Congressional Budget Office Research Report.

———. 2019. "Marginal Federal Tax Rates on Labor Income: 1962 to 2028." Congressional Budget Office Research Report.

DeBacker, Jason, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos. 2013. "Rising Inequality: Transitory or Permanent? New Evidence from a Panel of U.S. Tax Returns 1987–2006." *Brookings Papers on Economic Activity* 2013:67–122.

Dettling, Lisa, and Joanne Hsu. 2014. "Returning to the Nest: Debt and Parental Co-residence among Young Adults." Federal Reserve Board Working Paper 2014-80.

Fixler, Dennis, Marina Gindelsky, and David Johnson. 2019. "Improving the Measure of the Distribution of Personal Income." *American Economic Association Papers and Proceedings* 109:302–6.

Gottschalk, Peter, and Timothy M. Smeeding. 1997. "Cross-National Comparisons of Earnings and Income Inequality." *Journal of Economic Literature* 35(2):633–87.

Henry, Eric L., and Charles D. Day. 2005. "A Comparison of Income Concepts: IRS Statistics of Income, Census Current Population Survey, and BLS Consumer Expenditure Survey." Internal Revenue Service Research Report.

Hokayem, Charles, Christopher Bollinger, and James P. Ziliak. 2014. "The Role of CPS Nonresponse on the Level and Trend in Poverty." University of Kentucky Center for Poverty Research Discussion Paper Series 2014-05. Lexington, KY: University of Kentucky.

Hoynes, Hilary W., and Ankur J. Patel. 2018. "Effective Policy for Reducing Poverty and Inequality? The Earned Income Tax Credit and the Distribution of Income." *Journal of Human Resources* 53(4):859–90.

Hungerford, Thomas 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." Washington DC: Congressional Research Service.

Hurst, Erik, Geng Li, and Benjamin Pugsley. 2014. "Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-Employed." *Review of Economics and Statistics* 96(1):19–33.

Internal Revenue Service. 2015. "Statistics of Income—2013 Individual Income Tax Returns." Internal Revenue Service Publication 1304 (Rev. 08-2015).

———. 2018. "Statistics of Income Joint Statistical Research Program." https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program

Johns, Andrew, and Joel Slemrod. 2010. "The Distribution of Income Tax Noncompliance." *National Tax Journal* 63(3):397–418.

Joint Committee on Taxation. 2012. "Overview of the Definition of Income Used by the Staff of the Joint Committee on Taxation in Distributional Analyses." Joint Committee on Taxation Report JCX-15-12.

———. 2015. "Estimating Changes in the Federal Income Tax: Description of the Individual Tax Model." Joint Committee on Taxation Report JCX-75-15.

Langetieg, Patrick, Mark Payne, and Alan Plumley. 2017. "Counting Elusive Nonfilers using IRS Rather than Census Data." Internal Revenue Service Statistics of Income Working Paper.

Larrimore, Jeff, Richard V. Burkhauser, Gerald Auten, and Philip Armour. 2016. "Recent Trends in U.S. Top Income Shares in Tax Record Data Using More Comprehensive Measures of Income Including Accrued Capital Gains." NBER Working Paper 23007. Cambridge, MA: NBER.

Larrimore, Jeff, Jacob Mortenson, and David Splinter. 2016. "Income and Earnings Mobility in U.S. Tax Data." In *Economic Mobility: Research & Ideas on Strengthening Families, Communities & the Economy*, ed. Federal Reserve Bank of St. Louis and the Board of Governors of the Federal Reserve System, 481–516. Saint Louis, MO: Federal Reserve Bank of Saint Louis.

Lundberg, Shelly, Robert A. Pollak, and Jenna Stearns. 2016. "Family Inequality: Diverging Patterns in Marriage, Cohabitation, and Childbearing." *Journal of Economic Literature* 30(2):79–102.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States." *Journal of Economic Perspectives* 25(3):173–96.

Mortenson, Jacob, James Cilke, Michael Udell, and Jonathan Zytnick. 2009. "Attaching the Left Tail: A New Profile of Income for Persons Who Do Not Appear on Federal Income Tax Returns." In *Proceedings of the 102nd Annual Conference of the National Tax Association.* Washington, DC: National Tax Association.

Munnell, Alicia H., and Anqi Chen. 2014. "Do Census Data Underestimate Retirement Income?" Center for Retirement Research at Boston College Report 14-19. Boston, MA: Boston College.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1):1–39.

Proctor, Bernadette D., Jessica L. Semega, and Melissa A. Kollar. 2016. "Income and Poverty in the United States: 2015." Current Population Reports P60–256(RV). Washington DC: U.S. Census Bureau.

Sheridan, M., and I. Macredie. 1999. "Revisiting Statistical Units: Concepts, Definitions and Use, in International Expert [Canberra] Group on Household Income Statistics." In *Third Meeting on Household Income Statistics: Papers and Final Report, 305-16.* Ottawa, Canada: Statistics Canada.

Slavov, Sita, and Alan Viard. 2016. “Taxes, Transfers, Progressivity, and Redistribution: Part 1.” *Tax Notes* Sept. 2016:1437–50.
Slemrod, Joel. 2016. “Caveats to the Research Use of Tax-Return Administrative Data.” *National Tax Journal* 69(4):1003–20.
Smartystreets. 2019. “Smartystreets.com Documentation.” https://smartystreets.com/docs/methodology (accessed May 28, 2019).
Smeeding, Timothy M., and Daniel H. Weinberg. 2001. “Toward a Uniform Definition of Household Income.” *Review of Income and Wealth* 47(1):1–24.
Smith, Creston M. 1989. “The Social Security Administration’s Continuous Work History Sample.” *Social Security Bulletin* 52(10):20–8.
Splinter, David. 2019. “Who Pays No Tax? The Declining Fraction Paying Income Taxes and Increasing Tax Progressivity.” *Contemporary Economic Policy* 37(3):413-26.
Splinter, David, Jeff Larrimore, and Jacob Mortenson. 2017. “Whose Child Is This? Shifting of Dependents among EITC Claimants within the Same Household.” *National Tax Journal* 70(4):737–58.
Tax Policy Center. 2019. *The Tax Policy Center’s Briefing Book: A Citizen’s Guide to the Fascinating (Though Often Complex) Elements of the Federal Tax System.* http://www.taxpolicycenter.org/briefing-book/what-earned-income-tax-credit-eitc (accessed May 28, 2019).
United Nations Economic Commission for Europe. 2011. *Canberra Group Handbook on Household Income Statistics*. Second Edition. Geneva: United Nations.
Wagner, Deborah, and Mary Layne. 2014. “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software.” Census Bureau CARRA Working Paper 2014-01. Washington, DC: U.S. Census Bureau.