HOUSEHOLD INCOMES IN TAX DATA: USING ADDRESSES TO MOVE FROM TAX UNIT TO HOUSEHOLD INCOME DISTRIBUTIONS

Jeff Larrimore
Federal Reserve Board

Jacob Mortenson
Joint Committee on Taxation

David Splinter
Joint Committee on Taxation

March 2018

Tax return data have been limited by their inability to identify members of separate tax units living in the same household. We overcome this obstacle and present the first set of tax-based household income and inequality measures. We find that using tax units as a proxy for households overstates household income inequality, as measured by Gini coefficients, by 10 percent. Consistent with previous findings, we also estimate that the CPS understates household income inequality by 6 percent. Compared to conventional tax unit level measures, the federal income tax code and EITC are less progressive when measured at the household level.

JEL Codes: D31, H24
Keywords: Household Income, Tax Unit, EITC, Income Inequality, Tax Data, Measurement

---

The results and opinions expressed in this paper reflect the views of the authors and should not be attributed to the Federal Reserve Board. 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 comprise the Joint Committee on Taxation, this work should not be construed to represent the position of any member of the Committee.

For helpful comments, we thank Jesse Bricker, Richard Burkhauser, Jim Cilke, Jason DeBacker, Scott Winship, Gabriel Zucman and participants of presentations at the Federal Reserve Board, Drexel University, The Ohio State University, and the annual conference of the Association for Public Policy Analysis & Management.

I. Introduction

Over the past decade, research using administrative IRS tax return data has greatly expanded our understanding of incomes at the top of the U.S. income distribution (see e.g. 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. In particular, the absence of non-filers in tax return data have 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 households may contain multiple tax units or non-filers. This has precluded household level analyses, which is the standard unit of analysis in both national and cross-national distributional studies.

This paper produces the first set of tax-based income distributional statistics analyzed at the household level rather than the tax unit level, using address fields on tax documents to combine non-filing individuals and tax units into households. In doing so, we compare the distribution of income using these new tax record-based household data with more traditional IRS tax unit results and with survey-based household results from the March Current Population Survey (CPS). Finally, we use these data to provide the first tax record-based measure of how the Earned Income Tax Credit (EITC) and overall tax burdens are distributed across U.S. households.

Standard tax return data excludes 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). These non-filers are not missing at random, and are instead concentrated in the lower-tail of the distribution – which means that researchers using tax return data observe only a truncated version of the income distribution. Most

researchers partially overcome this problem by using tax return data only to analyze the top of the distribution and assuming that all non-filers have an income of 20 to 30 percent of average filer income (Piketty and Saez, 2003; Auten and Splinter, 2017). However, such an approach cannot be expanded to analyze lower-tail or distribution-wide inequality measures because it does not capture observation-level incomes for these non-filers. A more sophisticated approach, which we build on, is that of Chetty et al. (2014), who use information returns (such as Forms W-2 and 1099) to observe income data for non-filers. But information returns are at the individual level and lack links to any other members of the household, including spouses and children.[1] This paper develops these links, allowing non-filers to be treated equivalently to individuals on tax returns.

A similarly important, but less discussed limitation of tax record data is that they fail to observe complete households. Instead, tax units – the set of individuals appearing on the same tax return as a primary or secondary filer or as a dependent – are typically used as proxies for households (see e.g. DeBacker et al., 2013; Chetty, Hendren, and Katz, 2015).[2] 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, two cohabitating adults are treated as independent households. This contrasts with the U.S. Census Bureau's official income statistics based on the CPS, where individuals who live together, but file separate tax returns, would be treated as a joint entity (Proctor, Semega and Kollar, 2016).

[1] More broadly, we are unaware of previous research that combines non-filers together in any way other than random pairing. The Joint Committee on Taxation Individual Tax Model uses a mostly random process to join male and female non-filing individuals together in order to create the correct number of resident married couples (Joint Committee on Taxation, 2015). Since they do not have access to non-filer information returns, Piketty and Saez (2003) estimate the total number of adult tax units and determine the number of non-filing tax units as the residual – although they do not correct for the large number of dependent filers younger than 20 years old.

[2] Atkinson, Piketty, and Saez (2011) outline some of the challenges of this assumption for cross-national comparisons.

Numerous researchers have argued that the household is the sharing unit most closely resembling how individuals share economic resources (see e.g. Atkinson, Rainwater and Smeeding, 1995; Sheridan and Macedrie, 1999; Smeeding and Weinberg, 2001; Congressional Budget Office, 2016). 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 used by researchers considering national and cross-national inequality statistics (Atkinson and Brandolini, 2001; Burkhauser et al., 2011). This distinction is made more important because the choice of sharing units can greatly impact observed inequality trends (see e.g. Burkhauser, Larrimore and Simon, 2012).

Due to data limitations few researchers using tax data have attempted to create households in the tax data. Previous efforts to link tax units into households focused on statistical matches based on observable characteristics (Congressional Budget Office, 2016), or direct links between Census Bureau survey data to administrative records (see e.g. Abowd and Stinson, 2013; Wagner and Layne, 2014). While a direct link between Census Bureau survey data and administrative records is a promising avenue, the match is imperfect, as 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 may suffer from non-response error in surveys at both tails of the distribution.[3] 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.

[3] See Atkinson, Piketty and Saez (2011) for a discussion of these concerns at the top end of the distribution. Also see Bollinger et al. (2015) and Hokayem, Bollinger and Ziliak (2014) for discussions of these concerns at the lower end of the distribution.

These two problems – the lack of data for non-filers and the inability to organize individuals in the tax data into true households – present overlapping challenges. Since non-filers do not appear on a tax return, they have no natural tax unit. Therefore, any reasonable correction to the problem of non-filers also requires determining with whom they share resources. In this paper, we supplement tax return data with information returns, which contain income data for individuals regardless of whether they file an annual tax return. Next, we create households using address fields from tax and information returns, allowing us to observe the incomes of complete households.

We draw a sample from this linked tax record-based data set called the Tax Household Sample (THS). The THS is a five percent sample of every household in the United States, where households include all individuals listed on tax forms at a given address. With these new data we observe household-level income distributions directly in the tax data. This includes inequality statistics that focus on the middle- and lower-end of the income distribution, which have not previously been explored in the tax return based inequality literature.

When comparing income distributions of households in our new THS data to previous inequality estimates, household income inequality in the tax data is 3 Gini points (6%) higher than analogous estimates using CPS data. However, household income inequality is roughly 5 Gini points (10%) lower than analogous estimates using tax units as the unit of analysis. This suggests that researchers who opt to use tax units as a proxy for households – taking advantage of more accurate top incomes in tax data relative to surveys – may be fixing the downward inequality bias in the CPS data by introducing a notable bias in the other direction. We also find the distribution of EITC benefits at the household level contains significantly more mass in the

top two quintiles than the analogous tax unit distribution. Finally, we estimate that federal income taxes are less progressive at the household level.

II. Data and Methods

The primary data for this paper are drawn from the universe of federal income tax data collected by the IRS and which have recently been used by Chetty et al. (2014) and Chetty, Hendren, and Katz (2016) to study income mobility questions. In contrast to the public use and confidential versions of SOI Individual Income Tax Files, which have historically been used as the principal datasets of tax researchers (see e.g. Piketty and Saez, 2003; Auten and Gee, 2009), these data contain tax returns for all tax returns and information returns for all individuals rather than subsamples. This universal coverage ensures that all individuals within households appear in our data, which is necessary for aggregating observations to the household level.[4]

The base IRS data contains annual income tax returns (e.g. 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

[4] One limitation of this dataset, however, is that it contains unedited data fields. To address the absence of editing, which could result in incorrect incomes for top households, we explored removing households with incomes larger than the largest tax unit income in confidential IRS Statistics of Income (SOI) annual files, which include the population of top earners (over about $7.5 million). While there are extreme outliers in other years that exceed the maximum income in the cleaned cases, in 2010, which is the focus of this paper, no such cases exist.

Social Security Numbers), and mailing addresses. While annual income tax returns only exist for those who file a return, information returns are generated on behalf of individual taxpayers without their direct action. Since information returns are generated for almost all adults, these forms capture nearly all U.S. residents and they are commonly used to observe information on the non-filing population (see e.g. Mortenson et al., 2009; Chetty, et al., 2014; Cilke, 2014; Heim, Lurie, and Pearce, 2014; and Larrimore, Mortenson, and Splinter, 2016).

a. Comparison of Population Counts to Census Bureau Results

The suitability of using IRS tax data in this fashion depends on whether these forms can accurately capture the entire U.S. population. In order to assess their capacity in this regard, we compare the population count and number of households in the THS with analogous estimates reported by the U.S. Census Bureau from the decennial census. In 2010, 307.9 million individuals living in the U.S. appear in the tax data. This includes 281.3 million individuals who appear on a tax return as a primary filer (132.2) or as a spouse or dependent (149.1), along with 26.6 million non-filers for whom there is at least one information return. The 307.9 million people observed in the tax data is comparable to the 308.7 million individuals in the United States observed in the 2010 decennial census. Hence, while only 91 percent of individuals appear on an annual income tax return, when including both the filing population and the non-filing population with information returns, the tax return data observe over 99.5 percent of the overall U.S. population in 2010. This is consistent with the findings of Cilke (2014) that 99.5 percent of the 2011 resident population was identified on either an annual tax return filing or an information return.

In addition to matching the aggregate count of individuals, the tax record data also produces a similar age distribution to that seen from the decennial census. This similarity, as well as the importance of incorporating non-filers in the analysis, can be observed in Figure 1. 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 our tax record based THS (solid black line), which includes all individuals for whom there is an information return, the age distribution closely mirrors that observed in the decennial census. To the extent that deviations exist between the THS results and the decennial census results, the THS observes more children under age 15, whereas it observes fewer teenagers ages 15 to 20 and middle-age adults ages 40 to 55.[5]

Comparing the spatial distribution of individuals across states in Figure 2, the population distribution across states is similar in the THS to that observed in the decennial census.[6] In most cases, the state population in the tax data are within 1 to 2 percent of that seen in the decennial census. States with substantial differences between the numbers of people in the two files may reflect residency preferences for tax purposes. In particular, Delaware and Alaska have substantially more people (in percentage terms) in the tax data than in the decennial census –

[5] The administrative tax data used only include up to four dependents on a tax return. There may be incomplete reporting of teenagers relative to children on tax returns because children over age 16 do not qualify for the child tax credit. These two effects will bias downward the number of dependents. A much larger effect, however, results in an upward bias in the number of dependents. As individuals filing tax returns may legally claim a child exemption for children living in Canada or Mexico, there was a surge in these children on tax returns coinciding with immigration leading up the 2008 recession (and expansions in the refundable child tax credit). As these children are not authorized to work in the U.S., they cannot receive Social Security Numbers, but instead receive Individual Taxpayer Identification Numbers (ITINs). To limit the number of these non-resident 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 that was observed by Cilke (2014).

[6] Appendix Table A-1 shows the number of individuals and households in each state in each of the two datasets.

which could be evidence of individuals selectively choosing their legal residencies for Delaware's low taxes or Alaska's Permanent Dividend Fund payments.

b. Forming households and cleaning addresses in tax data

After aggregating tax forms to the individual level, and having established that the population counts using these data are consistent with those from the Census Bureau, individuals are aggregated into households using the reported address and ZIP code.[7] To construct the household-level file, household addresses are recoded into a standard form (e.g. recoding "1ST ST" or "FIRST STREET to "FIRST ST") and individuals are considered to live together if their address and 5-digit ZIP code both match.[8] For individuals living in an apartment or multi-unit building, the unit number must match as well as the main address. Filer addresses come from tax returns. Non-filer street addresses, as opposed to PO Boxes, are chosen preferentially if an individual has multiple information returns.[9]

Even after our extensive standardization of common address abbreviations, misspelling of street names remain. In order to link records due to close misspellings, we implement near-year and fuzzy 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 2018).

[7] Spouses and dependents claimed on a tax return are all considered to have the same address as the primary filer. Some dependents included as part of their claimant's household may include college students, who for at least part of the year reside at an address different than the person claiming them. An individual up to the age of 23 years old may be claimed as a dependent if they are a full-time student for at least 5 months of the year, even if they list a different mailing address on their own tax form. As these dependent filers must receive at least half of their support from the taxpayer claiming them, they are not independent economic units and so we consider them part of the household claiming them.

[8] See the online appendix for the SQL code to standardize addresses, along with information on all of the address corrections included.

[9] We sort non-empty addresses from information returns numerically and alphabetically and select the first address after sorting. This method preferentially selects street addresses starting with a number over PO Box addresses.

Before making this comparison, uncleaned street names in the tax data are edited to be letter-only street names by: (1) converting number streets to letters as described in the address standardization process above , (2) removing all remaining numbers including house and apartment numbers, and (3) removing leading and trailing characters such as "APT" or "STREET". We then observe whether the street listed on each single-person household exists on the master list of street names in the taxpayer's ZIP code.

For any single-person household with an invalid address, including a missing address, or a PO 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 (excluding PO Box and missing addresses). The next-year data is used if any of a number of criteria are met: the individual has the same first two digits of their house/apartment number and a different ZIP code (to correct an apparently small number of misreported ZIP codes), the same first digit of their house/apartment number and the same ZIP code (to correct misspelled street names and missing apartment numbers), a PO box address (to account for missing street addresses), or missing an address. This matching process is repeated with prior-year addresses.

For any remaining single-person households with an invalid address (excluding those with addresses replaced with near-year addresses or a PO Box), we then attempt to correct the street name to the most similar valid street name in the ZIP code using a fuzzy match based off of the Levenshtein distance method. The Levenshtein distance formula calculates the minimum number of characters that must be replaced, removed, or inserted to go from one character string to another (Levenshtein 1966). In performing this fuzzy match, we first observe if there are any valid street names in the ZIP code that are the same length and have the same first four letters as the reported street name. If there is only one such valid street name, and the Levenshtein distance

is less than or equal to 4, then we replace the reported street name with this valid street name. If there are 2 or more valid streets that meet this criteria, the valid street name with the lowest Levenshtein distance from that reported on the tax form is used. If no valid street names in the ZIP code have the same length and first four letters of the street name from the tax form (or those that do all have a Levenshtein distance greater than 4), a length deviation of 1 character is permitted, and again if there are multiple addresses meeting this criteria then the Levenshtein distance is used to determine the closer valid street name.

We subsequently repeat these steps, loosening the requirement that the first four letters match to require that just the first three and then two letters match in the addresses. If no valid address has been found in the ZIP code that meet these criteria, we then allow the length between the reported address and the addresses in the master address list to deviate by 2 or more characters, with the first four and then three letters matching sequentially. Once cleaning invalid addresses, we determine if other tax records in the zip code report the same, cleaned, household address and merge together any records that do. Finally, we form households using these cleaned addresses, and extract an annual random five-percent sample of households based on the last four digits of the TIN of one member of each household to arrive at the final sample for the THS.[10,11]

[10] The representative of each household is the household member with the largest TIN. All representative individuals whose TIN ends in one of 500 possible combinations is selected into the household sample. Since no TINs end with all zeros, there are technically 9,999 possible endings so the sample is just slightly over a 5 percent sample. 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 and Internal Revenue Service, 2015).

[11] A potential concern with a random sample, raised by Bricker et al. (2016a), is that at the very high-end of the distribution the thin tail leads to substantial sampling variability. Since the focus of this paper is not the very top of the distribution, and since the 5% probability of selection far exceeds the selection rate of any survey, including the Survey of Consumer Finances which oversamples high-wealth individuals, we do not view this as a substantial concern.

C. Calculating income of filers and non-filers

The income comparisons in this paper focus on pre-tax income excluding capital gains. 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: (1) non-taxable interest reported on Form 1040 is added, (2) realized Schedule D capital gains are removed, (3) taxable Social Security benefits are replaced by total Social Security benefits reported on Form SSA-1099, (4) taxable private retirement income is replaced with gross private retirement income, which reflects retirement savings distributions less rollovers from Forms 5498 and 1099-R, (5) income are bottom-coded at zero to limit the effect of business losses.[12] 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, and 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.[13]

Among non-filers, 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 Chetty et al. (2014), who also derive income for non-filers based on information returns, we include income from wages and salaries reported on Form W-2, unemployment

[12] The inclusion of gross private retirement income results in a difference from incomes measured by the Census Bureau definitions because 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.
[13] While this is the closest income measure to that from the CPS which is observable directly in the tax data, it excludes several non-taxable transfers such as public assistance and SSI income. These excluded income measures are discussed in greater detail in section IV of this paper and Appendix Table A-2. Our income measure, for both filers and non-filers, will also exclude some "off the books" income that is neither captured on tax returns or information returns.

benefits from Form 1099-G, and Social Security and disability benefits from Form SSA-1099. In addition to these income sources incorporated by Chetty et al., we include interest income from Form 1099-INT, gross private retirement income as retirement savings distributions less rollovers from Forms 5498 and 1099-R, and 30 percent of income from Form 1099-MISC, which captures net self-employment income. The offset of 70 percent of income reported on Form 1099-MISC reflects that gross income from self-employment activities appear on the 1099-MISC information returns, but the associated business expenses do not. As a result, an 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 non-filers, we assume a similar offset to Form 1099-MISC income while preserving all income from other sources that appear on information returns.[14]

This use of information returns for non-filers 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 3 compares the household income on information returns for low income tax-filers to the amount actually reported on annual tax returns. 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 non-filers.

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

In this figure, centiles are defined based on the income reported on the annual tax return form, so the individuals in each centile are the same across the two series. Both series also exclude any income in the households derived from non-filing individuals. When aggregating all 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, with the exception of the bottom 4 percent of the distribution where there is *more* income reported on information returns than on tax returns.[15] However, there is no evidence that the information returns are systematically missing substantial income among the low income filing population.

III. Comparing household and tax unit characteristics

In this section we compare the households formed from tax data as described in the previous section with households from the 2011 March CPS and the 2010 Decennial Census. The final row of Table 1 contains the number of households observed in each source.[16] 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.[17]

[15] The lower income reported on tax returns for these individuals relative to the information returns may reflect additional business deductions (which may lead to net business losses) that are not observed on the information returns. For some individuals, it may also reflect non-compliance in income reporting if they do not report their full income on the annual tax return.

[16] The 2011 March CPS is used here since it represents the 2010 income year. It also likely is the closest survey to the expected 2010 population from IRS data which should reflect addresses from the end of the 2010 calendar year.

[17] Since the IRS tax data do not classify the type of housing unit, addresses with 11 or more individuals are treated here as group quarters, which captures nine million individuals at half a million addresses. This approximates the eight million individuals listed as living in group quarters in the 2010 decennial census. Removing group quarters reduces total observed income by less than two percent.

In 2010 there are 113.5 million households in the THS data, roughly 3 million fewer than the 116.7 million households in the 2010 decennial census, and roughly 4 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 about 2 million of the fewer households in the tax data.

The impact of each step in the cleaning process, described in Section II.b., is outlined in Appendix Table A-2. Standardizing abbreviations and cleaning based on near-year entries from the same taxpayer have the largest effect on the total number of households. The fuzzy matches using the valid address list in the same ZIP code provides additional confidence on the quality of these data, but has a much smaller impact on the number of distinct households observed in the tax records.

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, then all households would consist of one filing tax unit and zero non-filing individuals. Instead, only 59 percent of households consist of just one filing tax unit and zero non-filers. In other words, 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 non-filing individual. The remaining 31 percent of households contain at least two separate filing tax units, two non-filing 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 has risen in recent years (see e.g. Dettling and Hsu, 2014; Lundberg, Pollak and Stearns, 2016) and likely comprise 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 non-dependent adult living with his or her parents, and 29 percent contain a cohabiting couple (3 percent of which also contain an adult living with his or her parents). The remaining 24 percent of households with multiple tax units have neither a non-dependent adult living with his or her parents nor a cohabiting couple – and therefore include either roommates or relatives besides parents/children who are living together.[18]

Figure 4 displays where tax units residing with other tax units fall in the tax unit income distribution. Figure 5 displays where households containing multiple tax units fall in the household income distribution. Figure 4 suggests that many of the tax units residing with others have relatively low incomes. Over half (60%) 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 5). In part, this reflects that multiple tax unit households have more earners, which can push their joint incomes further up

[18] Among households with multiple tax units in the top 5% of the income distribution in the CPS, 63 percent contain a non-dependent 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 cohabiting couple nor an adult child-parent relationship. Multiple tax unit households in the top 1% of the income distribution in the CPS have a similar distribution of relationships.

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.

IV. 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 the IRS and CPS data. Specifically, SSI, child support income, educational assistance, financial assistance, survivor's benefits, veteran's benefits, workers compensation, and public assistance income are removed from the CPS income definition since they cannot be observed on IRS tax forms.[19] In 2010, these excluded income sources represented 2.5 percent of total pre-tax income in the CPS (see Appendix Table A-3 for details).

Figure 6 compares the pre-tax household income distribution in the tax data and CPS data. 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. Non-filing individuals are paired semi-randomly to match the total number of married couples. This approach approximates the number of total tax units from the updates to Piketty and Saez (2003).[20]

[19] In order to create 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), one approach used by the Congressional Budget Office (2016) and by Larrimore et al. (2016) is to impute these sources into the tax data based on statistical matches. However, in order to focus solely on the income observed in tax records, we exclude these sources from both datasets while recognizing that including them would lower the observed levels of inequality.

[20] Were all dependent filers included and non-filers treated as single individuals, the difference between the tax unit series and the household series would be even greater.

Comparing the tax-based and Census household income series, it is apparent that the distribution of household incomes are similar across the two datasets with the exception of the top centiles of the distribution. Largely reflecting the more comprehensive coverage of private retirement income in the IRS data relative to the CPS, household incomes are slightly higher in each centile of the tax-based household income distribution than in the CPS data - but outside of the top decile of the distribution, this difference is always less than $10,000.[21] The primary differences between the IRS-based and CPS-based household income series occur in the top two percent of the household income distribution where household incomes in the CPS fall well below income reported in the tax data. In particular, relative to the tax-based household data the CPS understates the mean income of the 98th percentile of the distribution by 22 percent ($282,000 compared to $362,000) and the mean income of the top 1 percent of households by 52 percent ($524,000 compared to $1,093,000), as shown in Figure A-1. This is consistent with the view of many researchers – including Atkinson, Piketty, and Saez (2011) – that the CPS data fails 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 below that of tax-based households. For example, while the 50th tax unit centile has an average income of $31,987, the corresponding household centile has an income of $53,240. This pattern persists throughout the income distribution and suggests tax units are poor proxies for households when considering the distribution of income.[22]

[21] Were we to remove private retirement income from both series, CPS incomes are slightly *above* IRS incomes in the second through ninth deciles of the income distribution.

[22] In order to incorporate the additional individuals living in households who draw upon available resources, we also considered the size-adjusted incomes of households and tax units by dividing by the square root of the number of individuals in the sharing unit. When doing so, the large gap between the income of households and tax units remained.

The ability to observe households directly in tax return data also offers a refined perspective on income inequality. This is seen in Figure 7, which 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 more equal distributions. Reflecting the better ability of the IRS data to observe income at the top of the distribution, the tax record based household income series observes a higher concentration of income among the top centiles than does the CPS data. This 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.[23]

However, Figure 7 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 impact 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 three measures. Relative to the tax record 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 0.516, which is below the 0.570 Gini coefficient for tax units but above the Gini coefficient for households in the CPS of 0.483. Hence, using tax units as proxies for households will overstate the household income

[23] 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 record household income distribution, the gap between the Census and tax data household income Lorenz curves in Figure 6 would nearly disappear.

Gini coefficient by 10 percent, and using the CPS data will understate the household income Gini coefficient seen in the tax data by 6 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 ratio – 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: 14.0 percent versus 16.2 percent of income.[24] This is in spite of the top 1 percent of households having a slightly larger average income than tax units (Figure A-1). The lower top income share for households is partially a product of there being far fewer households than tax units (113.5 vs. 157.5 million), and fewer total households means fewer households in the top one percent. The increase in top shares from larger household average incomes is outweighed by the decrease from having fewer households than tax units in the top one percent.[25]

VII. Distribution of Earned Income Tax Credits

[24] Bricker et al. (2016b) find similar effects of switching from tax units to families when using the 2010 Survey of Consumer Finances, estimating a decrease in top 1 percent income shares by 2.4 percentage points.

[25] The difference between the total number of tax units and the number of households does not on its own imply a change in top income shares. It is only the non-random combination of tax units into households that causes this result. Since 1.14 million households make up each centile of the household income distribution, whereas 1.57 million tax units make up each centile of the tax unit distribution, the average income per household in the top centile would need to be 38 percent higher than is seen among tax units for there to be no change in the top 1 percent income share. Because the average household income in the top centile is only 24 percent higher than was seen for the top 1 percent of tax units, the top 1 percent income share among households is lower than among tax units.

Most analyses of the distributional effects of tax provisions focus on the distribution across tax units (see e.g. Joint Committee on Taxation, 2012; Tax Policy Center, 2017), as the underlying data used in these analyses are tax return data. However, in addition to affecting inequality measures, the distributional impacts 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 distributional impacts 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 that is 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 over $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 2 qualifying dependents in 2010 must have had less than $40,363 in adjusted gross income to be eligible, while a single child-less tax unit could only earn up to $13,460 to remain eligible. In tax year 2010, over 27 million tax filers claimed about $60 billion of EITC credits. This means the EITC has over 10 times the number of recipients and over 7 times the cash benefits than 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 can be seen in Figure 8. This figure shows the fraction of tax units in each centile of the pre-tax income distribution that claim the credit. While earnings requirements for the credit mean

that few tax units in the bottom 5 percent claim the EITC, claiming rates rise to about forty percent in the 2nd decile of the tax unit income distribution and 30 percent in the 3rd through 5th decile. Higher in the distribution, claiming rates fall sharply and nearly no tax units in the top three deciles of the tax unit distribution receive the EITC.

When using tax households as the unit of observation, the majority of claimants remain in the lower deciles of the distribution. However, 10 percent of households in the 8th and 9th deciles and 6 percent of those in the top decile receive EITC benefits. This is due to a non-trivial number of individuals in relatively low income tax units (thereby qualifying for the credits) residing in multi-unit households with aggregate incomes beyond the end of the EITC's phase-out.

Figure 9 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 the credits claimed. At the tax unit level, again, EITC benefits are well targeted at those in the bottom half of the distribution. Seventy-five percent of benefits are received by those in the bottom two quintiles of the pre-tax income distribution and just 1 percent go to those in the top two quintiles. But by linking the tax units into households, it is apparent that a non-trivial fraction of benefits go to those living in higher income households. At the household level, 17 percent of EITC benefits go to those in the top two quintiles. 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.

VIII. Distribution of Tax Burdens

Federal individual income taxes, as with the EITC, appear less progressive at the household than the tax unit level. Figure 10 compares average tax rates across the household and tax unit pre-tax income distribution. Federal income tax burdens are similar in the top two quintiles, but average tax rates are higher for households in the second and third quintiles. This suggests federal taxes are less progressive when considering households instead of tax units.

We estimate distribution-wide tax progressivity using the Kakwani index. This is estimated 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 0.355 for tax units to 0.316 for households, a decrease of 11 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.

IX. 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 non-filers. Using address fields on IRS tax records and the universe of tax forms, we 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 2 centiles of the income distribution. This limitation reduces observed pre-tax household income Gini coefficient in the CPS data by 3 Gini points in 2010, a 6 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 5 Gini points (10 percent). The inability of tax units to properly proxy for households reflects our finding that only 68 percent of households consist of a single tax unit or non-filing 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 notably less progressive when income is measured at the household level. In particular, the share of earned income tax credits going to the top two quintiles of the income distribution rises from 1 percent to 17 percent when we shift the unit of analysis from tax units to households. This 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 developed in this paper allows for an expansion of the research topics for which IRS tax data may be suitable. This 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 non-filers previously precluded the use of IRS data. 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.

References

Abowd, John, and Martha Stinson, "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics* 95:5 (2013), 1451-1467.

Atkinson, Anthony B., and Andrea Brandolini, "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 (2001), 771–799.

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

Atkinson, Anthony B., Lee Rainwater, and Timothy M. Smeeding, "Income Distribution in OECD Countries: The Evidence from the Luxembourg Income Study," Social Policy Studies No. 18. (Paris: Organization for Economic Cooperation and Development, 1995).

Auten, Gerald, and Geoff Gee, "Income Mobility in the United States: New Evidence from Income Tax Data," *National Tax Journal* 62:2 (2009), 301-328.

Auten, Gerald, and David Splinter, "Income Inequality in the United States: Using Tax Data to Measure Long-term Trends," Accessed Jan. 1, 2018 via http://davidsplinter.com/AutenSplinter-Tax_Data_and_Inequality.pdf (2017).

Bitler, Marianne, Hilary Hoynes, and Elira Kuka, "Do In-Work Tax Credits Serve as a Safety Net," *Journal of Human Resources* 52:2 (2017), 319–350.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak, "Trouble in the Tails? Earnings Non-Response and Response Bias across the Distribution," Working paper, (2015).

Bond, Brittany, J., David Brown, Adela Luque, and Amy O'Hara, "The Nature of the Bias when Studying only Linkable Person Records: Evidence from the American Community Survey," Census Bureau CARRA Working Paper 2014-08, (2015).

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus, "Measuring Income and Wealth at the Top Using Administrative Data," *Brookings Papers on Economic Activity* Spring (2016a), 261-312.

——, "Estimating Top Income and Wealth Shares: Sensitivity to Data and Methods." *American Economic Review* 106:5 (2016b), 641-645.

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

——, "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data," *The Review of Economics and Statistics* 44:2 (2012), 371-388.

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

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

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics* 129:4 (2014), 1553–1623.

Cilke, James. "The Case of the Missing Strangers: What we Know and Don't Know about Non-Filers." Proceedings of the 107th Annual Conference of the National Tax Association (2014).

Congressional Budget Office, "The Distribution of Household Income and Federal Taxes, 2013," *Congressional Budget Office Research Report* (2016).

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

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

Heim, Bradley T., Ithai Z. Lurie, and James Pearce, "Who Pays Taxes? A Dynamic Perspective," *National Tax Journal* 67:4 (2014), 755–778.

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

Internal Revenue Service, "Statistics of Income – 2013 Individual Income Tax Returns," Internal Revenue Service publication 1304 (Rev. 08-2015), (2015).

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

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

Larrimore, Jeff, Richard V. Burkhauser, Gerald Auten, and Philip Armour, "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 (2016).

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

Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet Physics Doklady. 10 (8): 707–710, (1966).

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

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak, "Internal Migration in the United States," *Journal of Economic Perspectives* 25:3 (2011), 173–196.

Mortenson, Jacob, James Cilke, Jonathan Zytnick, and Michael Udell, "Attaching the Left Tail: A New Profile of Income for Persons who do not Appear on Federal Income Tax Returns." Proceedings of the 102nd Annual Conference of the National Tax Association (2009).

Munnell, Alicia H. and Anqi Chen, "Do Census Data Underestimate Retirement Income?" Center for Retirement Research Report 14-19 (2014).

Proctor, Bernadette D., Jessica L. Semega, and Melissa A. Kollar, "Income and Poverty in the United States: 2014," Current Population Reports P60–256 (Washington DC: U.S. Census Bureau, 2015).

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

Sheridan M. and I. Macredie, "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* (Ottawa, Canada: Statistics Canada, June 1999), 305-316.

Slavov, Sita and Alan Viard, "Taxes, Transfers, Progressivity, and Redistribution: Part 1," *Tax Notes* Sept. (2016), 1437–1450.

Smartystreets, "Smartystreets.com Documentation," Retrieved from https://smartystreets.com/docs/methodology (2018), last accessed February 11, 2018.

Smeeding, Timothy M., and Daniel H. Weinberg, "Toward a Uniform Definition of Household Income," *Review of Income and Wealth* 47:1 (2001), 1–24.

Smith, Creston M., "The Social Security Administration's Continuous Work History Sample," *Social Security Bulletin* 52:10 (1989), 20–28.

Tax Policy Center, *The Tax Policy Center's Briefing Book: A Citizen's Guide to the Fascinating (Though Often Complex) Elements of the Federal Tax System.* Accessed Jan. 31, 2017 via http://www.taxpolicycenter.org/briefing-book/what-earned-income-tax-credit-eitc (2017)

United Nations Economic Commission for Europe, *Canberra Group Handbook on Household Income Statistics: Second Edition* (Geneva: United Nations, 2011).

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

Table 1. Number of households by household size, 2010 (thousands)

Size of Household	Decennial Census	March CPS	Tax Data
1	31,205	31,399	32,557
2	38,243	39,487	32,787
3	18,758	18,638	18,359
4	15,625	16,122	15,587
5	7,538	7,367	7,768
6	3,075	2,784	3,675
7 or more	2,272	1,739	2,799
Total	**116,716**	**117,538**	**113,532**

Notes: In the tax data, all dependents are included in the household 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.
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), THS and authors' calculations.

Table 2. Household composition by number of filing tax units and non-filing individuals in the household, 2010

		Non-filing individuals in the household		
		0	**1**	**2+**
Filing tax units in the household	**0**	--	9.7%	1.9%
	1	58.7%	3.9%	0.7%
	2+	22.8%	2.0%	0.4%

Notes: Dependent filers and dependent non-filers 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.
Source: Tax Household Sample (THS) and authors' calculations

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

	(1) Tax data (Household)	(2) Tax data (Tax Unit)	(3) March CPS (Household)	(4) % difference using tax units	(5) % difference using March CPS
Gini	0.516	0.570	0.483	10.4	-6.4
P90/P10	13.22	18.78	13.72	42.0	3.8
P80/P50	2.01	2.36	2.09	17.3	3.9
P50/P20	2.53	2.70	2.62	6.4	3.3
1st quintile share	2.69	1.99	2.70	-26.0	0.5
2nd quintile share	7.86	6.41	8.21	-18.4	4.5
3rd quintile share	13.53	11.69	14.51	-13.6	7.3
4th quintile share	21.39	20.39	23.56	-4.6	10.2
Top quintile share	54.55	59.52	51.02	9.1	-6.5
Top 5 percent share	27.86	31.37	21.76	12.6	-21.9
Top 1 percent share	13.99	16.20	---	15.8	---

Notes: See Figure 6 for details. March CPS data is 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.
Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, THS and authors' calculations.

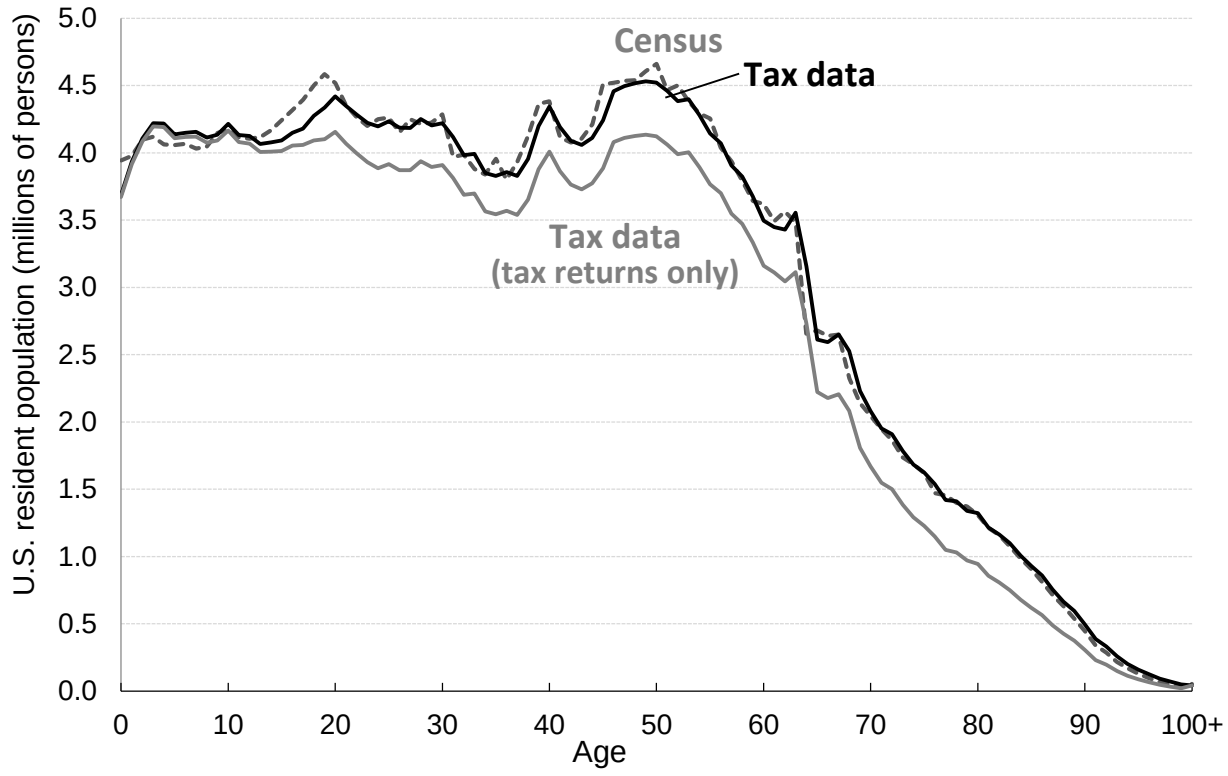


Figure 1. Number of individuals by age, 2010

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

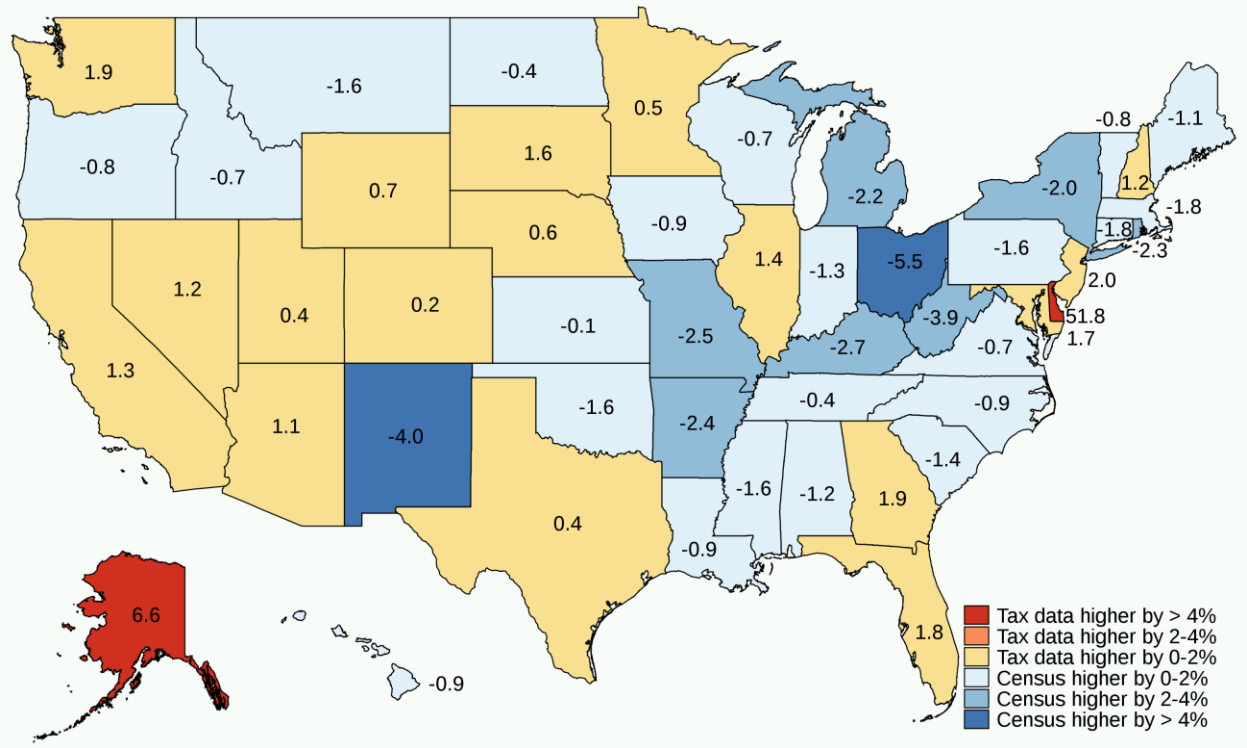


Figure 2. Map of percent population difference between tax data and Census, 2010

Notes: The Decennial Census population is based on March 2010 and tax data population on December 31. In the tax data, all dependents are included at the address of the person who claims them.
Source: U.S. Census Bureau 2010 decennial census, THS and authors' calculations.

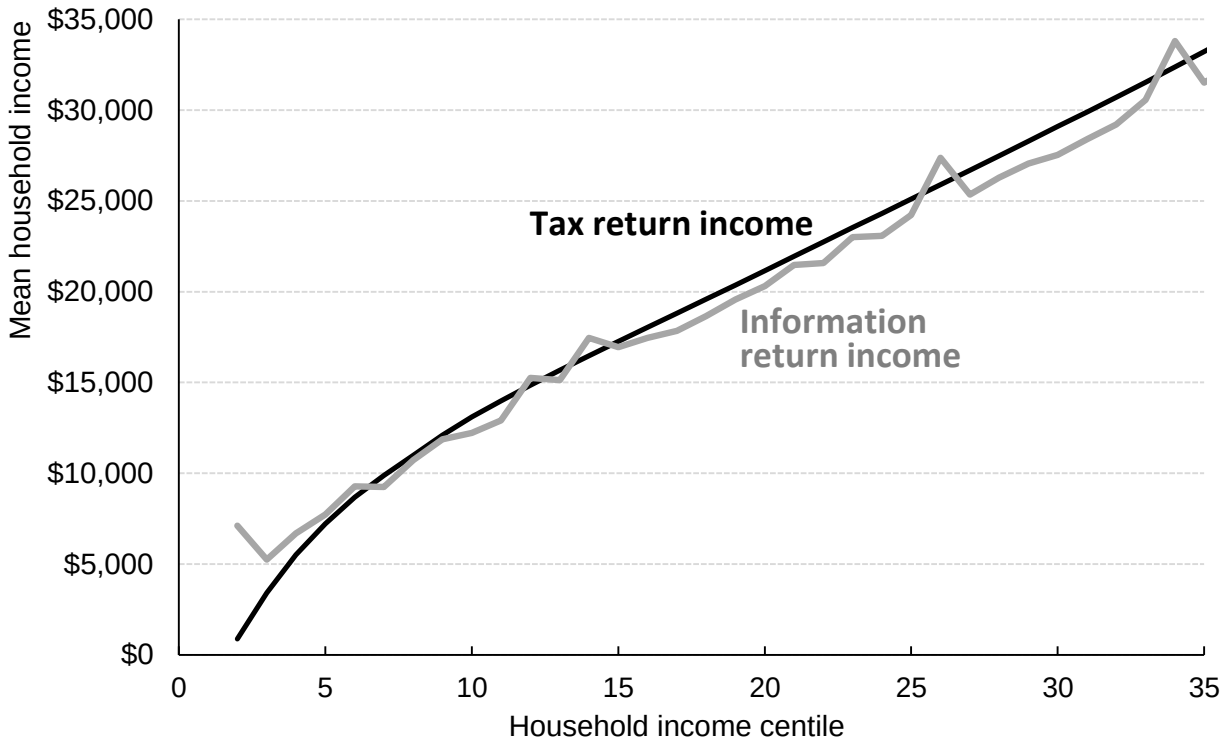


Figure 3. Comparing household income of tax filers from information returns and from tax returns, 2010

Notes: Tax return pre-tax income is total taxable income reported on tax returns, but adding non-taxable interest and non-taxable Social Security benefits, and excluding private retirement income and realized capital gains. 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 of the main text for details). Information return income includes wages from Forms 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 and each centile has an equal number of individuals. Thresholds for both incomes are based on tax return income and households with only non-filers are excluded from mean income estimates.
Source: THS and authors' calculations.

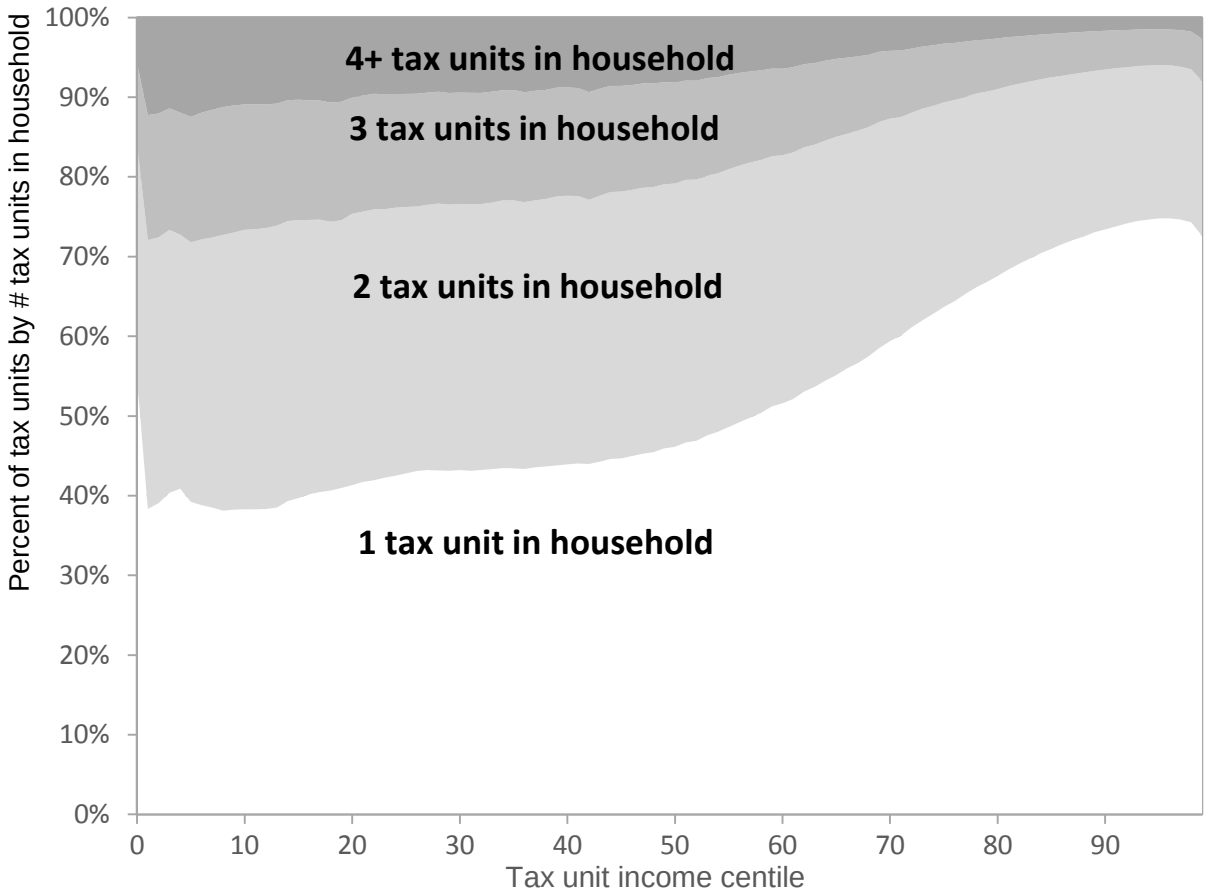


Figure 4. Number of filing tax units and non-filing individuals per household by tax unit income, 2010

Notes: As in Table 2, counts of tax units in this figure are based on the number of primary filers and non-filing 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 more than 10 tax units are excluded. For filers, pre-tax income is total taxable income reported on tax returns, but adding non-taxable interest, replacing taxable private retirement income with gross private retirement income, and excluding realized capital gains. For non-filers, 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.
Source: Tax Household Sample (THS) and authors' calculations.

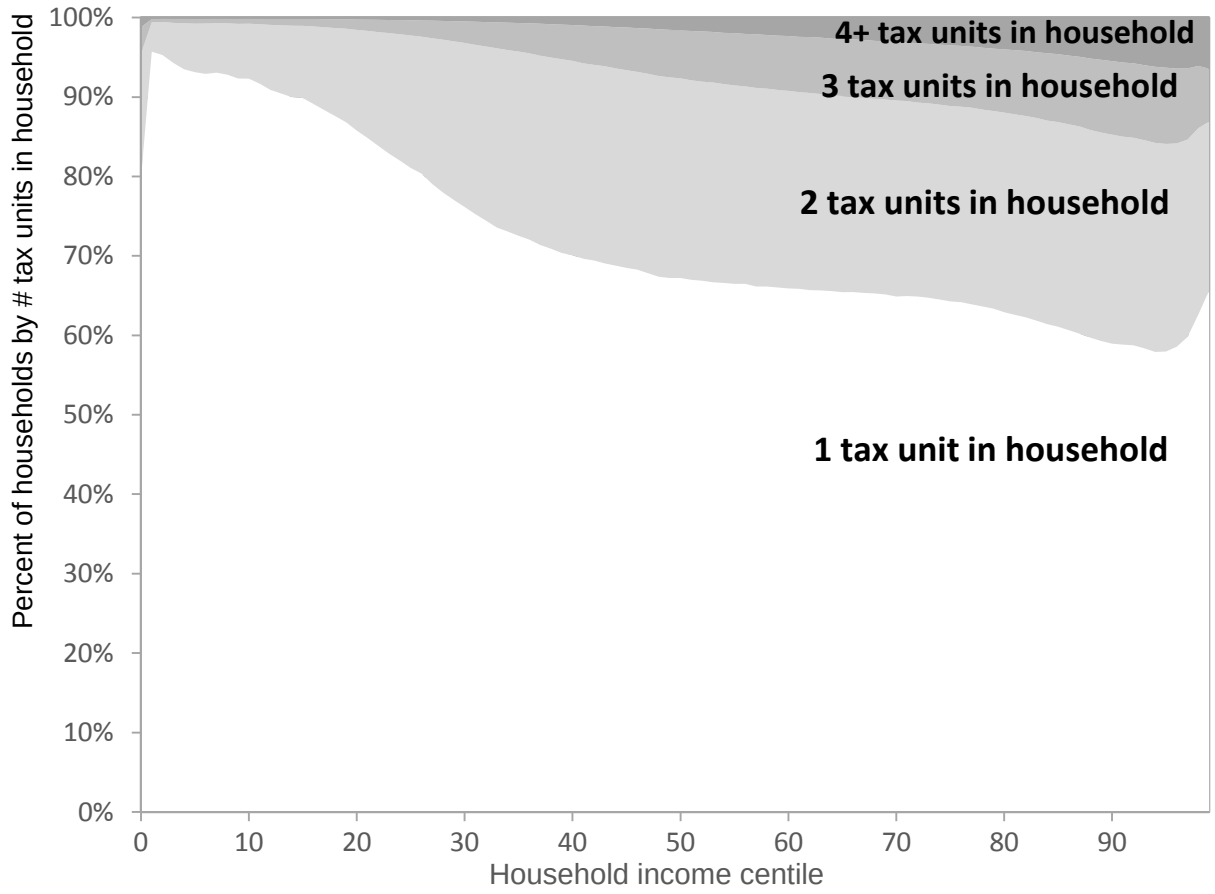


Figure 5. Number of filing tax units and non-filing individuals per household by household income, 2010

Notes: As in Table 2, counts of tax units in this figure are based on the number of primary filers and non-filing 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 more than 10 tax units are excluded. For filers, pre-tax income is total taxable income reported on tax returns, but adding non-taxable interest, replacing taxable private retirement income with gross private retirement income, and excluding realized capital gains. For non-filers, 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.
Source: Tax Household Sample (THS) and authors' calculations.

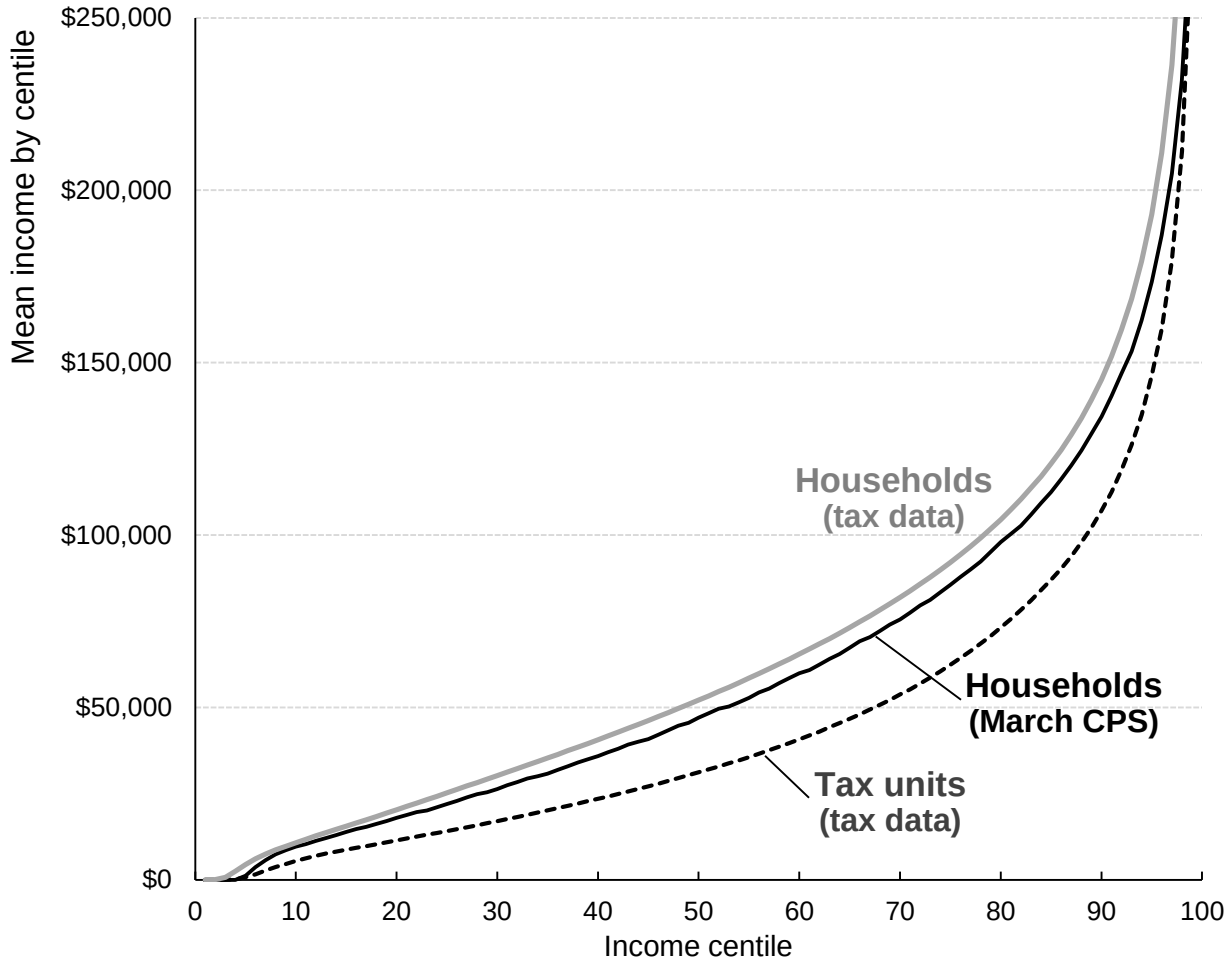


Figure 6. Distribution of pre-tax income by unit of observation and data source, 2010

Notes: For filers, pre-tax income is total taxable income reported on tax returns, but adding non-taxable interest and non-taxable Social Security benefits, replacing taxable private retirement income with gross private retirement income, and excluding realized capital gains. For non-filers, 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. Incomes are bottom-coded at $1. For the households series, individuals living in group quarters are excluded, which is defined in the THS as households with 11 or more individuals. Tax units include 135.0 million non-dependent filers using IRS Statistics of Income (SOI) tax return sample and 22.5 million non-dependent non-filer tax units using information returns. For the tax unit series, in order to match the overall marriage rate among tax units, about 40% of non-filer tax units are assumed to be married, where many pairings are based on actual 2007 filing status. All points are the mean income within the specified centile of the distribution.
Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, THS and authors' calculations.

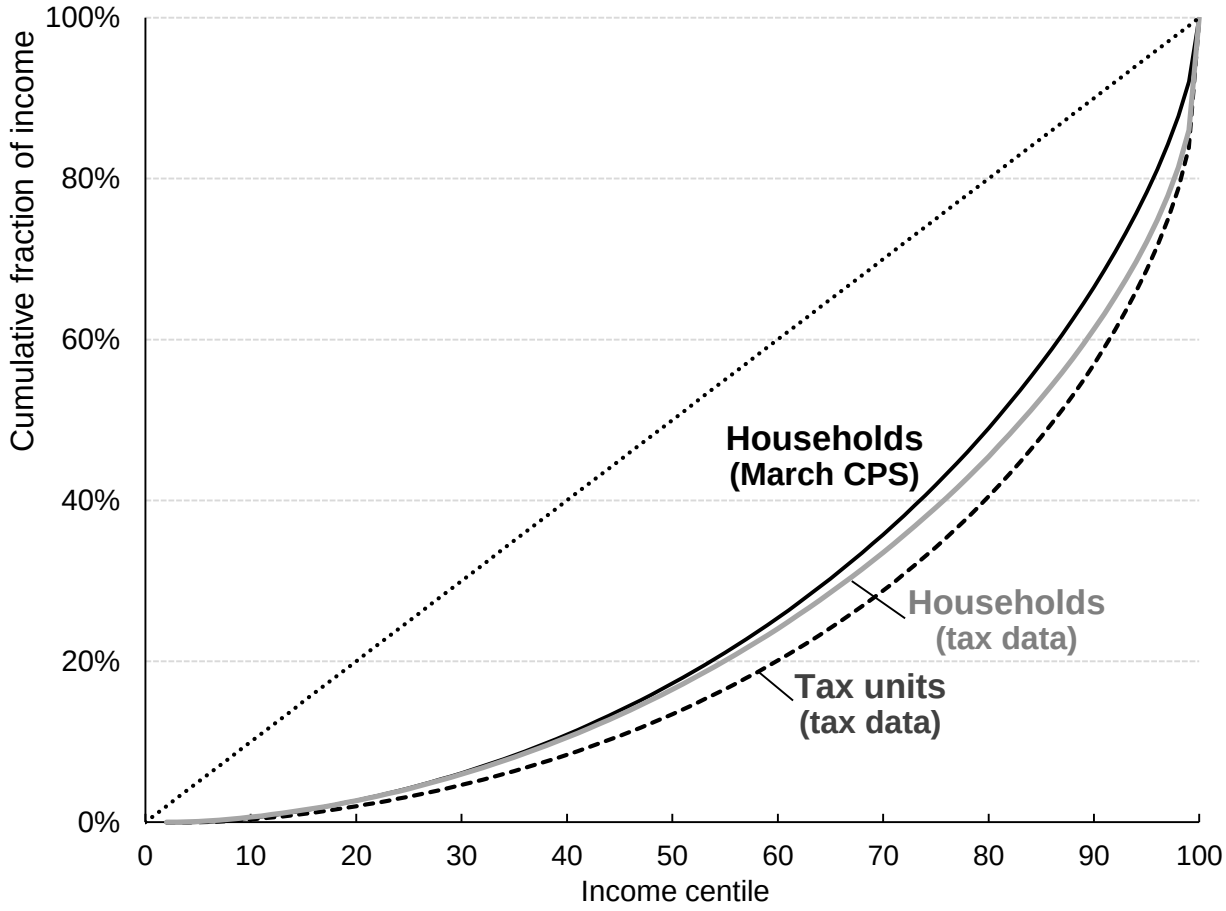


Figure 7. Lorenz Curve by unit of observation and data source, 2010

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

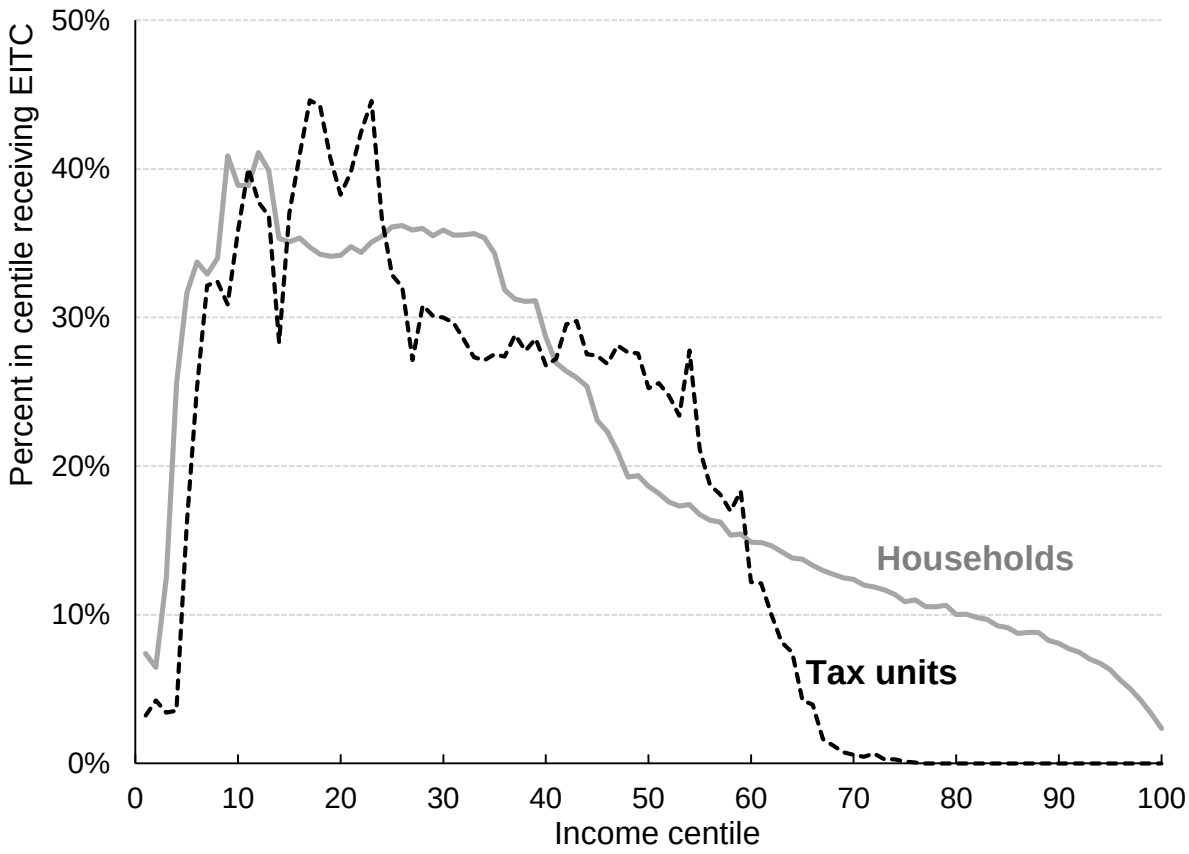


Figure 8. Share of tax units and households claiming the EITC, 2010

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

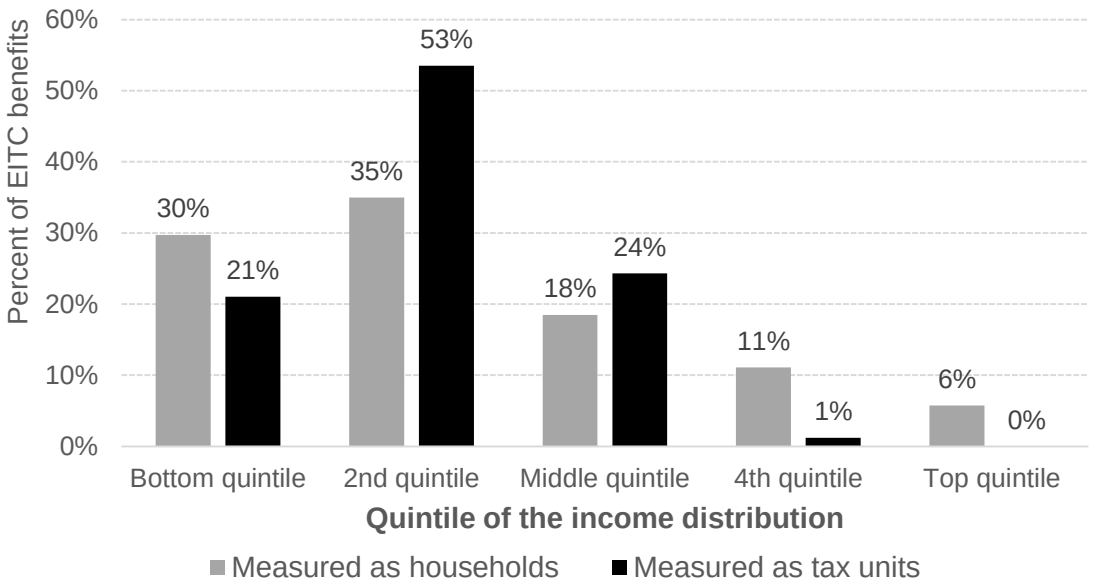


Figure 9. Distribution of the EITC, 2010

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

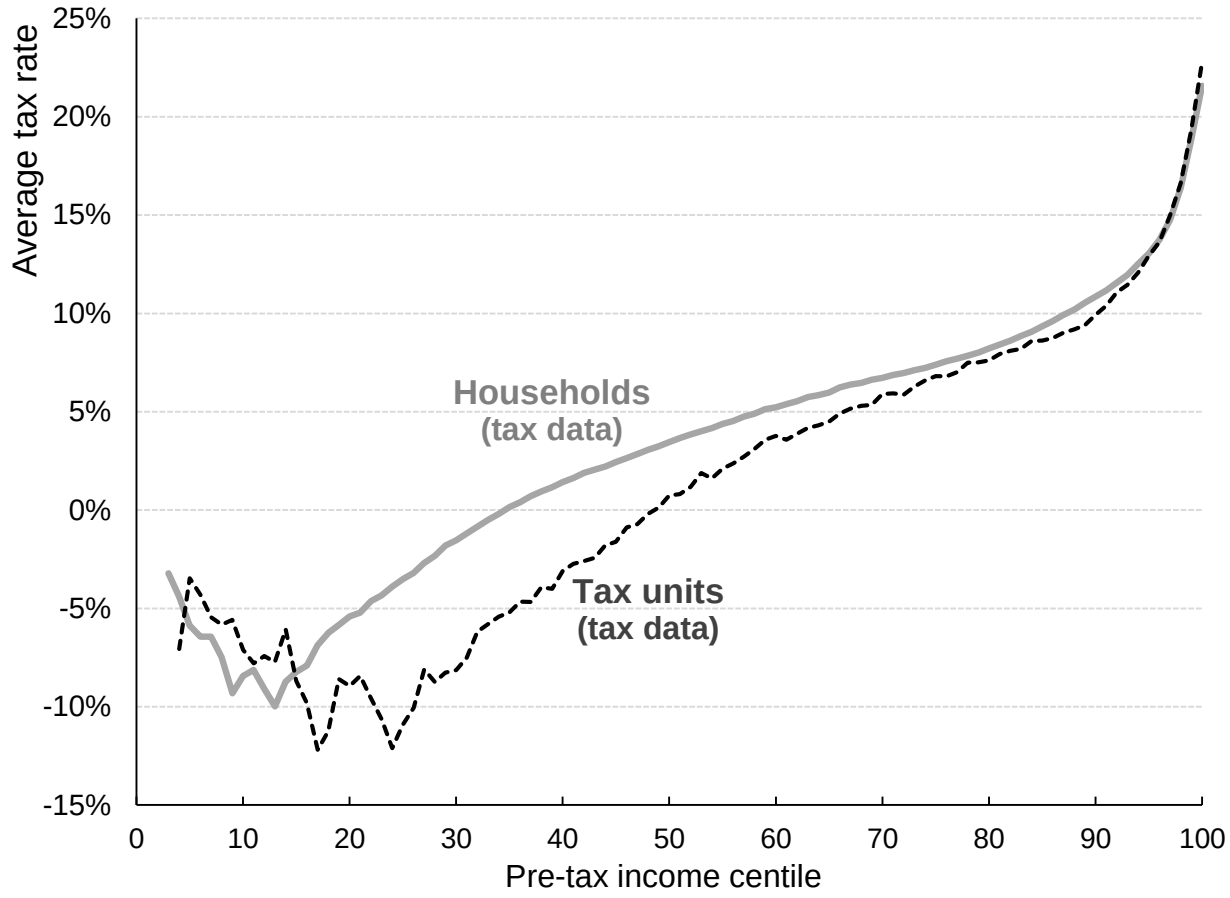


Figure 10. Average tax rates, 2010

Notes: Pre-tax income is defined as in Figure 6, 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 non-filers are assumed to be zero.
Source: THS and authors' calculations.

Appendix Tables and Figures

Table A-1. State populations in IRS and Census Data, 2010

	Individuals			Individuals	
State	Decennial Census	Tax Data	State	Decennial Census	Tax Data
AK	710	757	MT	989	973
AL	4,780	4,722	NC	9,535	9,453
AR	2,916	2,847	ND	673	670
AZ	6,392	6,462	NE	1,826	1,837
CA	37,254	37,746	NH	1,316	1,332
CO	5,029	5,040	NJ	8,792	8,971
CT	3,574	3,510	NM	2,059	1,976
DC	898	602	NV	2,701	2,732
DE	602	914	NY	19,378	18,992
FL	18,801	19,135	OH	11,537	10,900
GA	9,688	9,870	OK	3,751	3,689
HI	1,360	1,348	OR	3,831	3,802
IA	3,046	3,018	PA	12,702	12,499
ID	1,568	1,557	RI	1,053	1,029
IL	12,831	13,004	SC	4,625	4,561
IN	6,484	6,398	SD	814	827
KS	2,853	2,850	TN	6,346	6,320
KY	4,339	4,223	TX	25,146	25,240
LA	4,533	4,494	UT	2,764	2,774
MA	6,548	6,428	VA	8,001	7,947
MD	5,774	5,874	VT	626	621
ME	1,328	1,313	WA	6,725	6,851
MI	9,884	9,668	WI	5,687	5,648
MN	5,304	5,330	WV	1,853	1,781
MO	5,989	5,838	WY	564	568
MS	2,967	2,918	**TOTAL**	**308,746**	**307,860**

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. *Source*: U.S. Census Bureau 2010 decennial census, THS and authors' calculations.

Table A-2. Number of households by household size, 2010 (thousands)

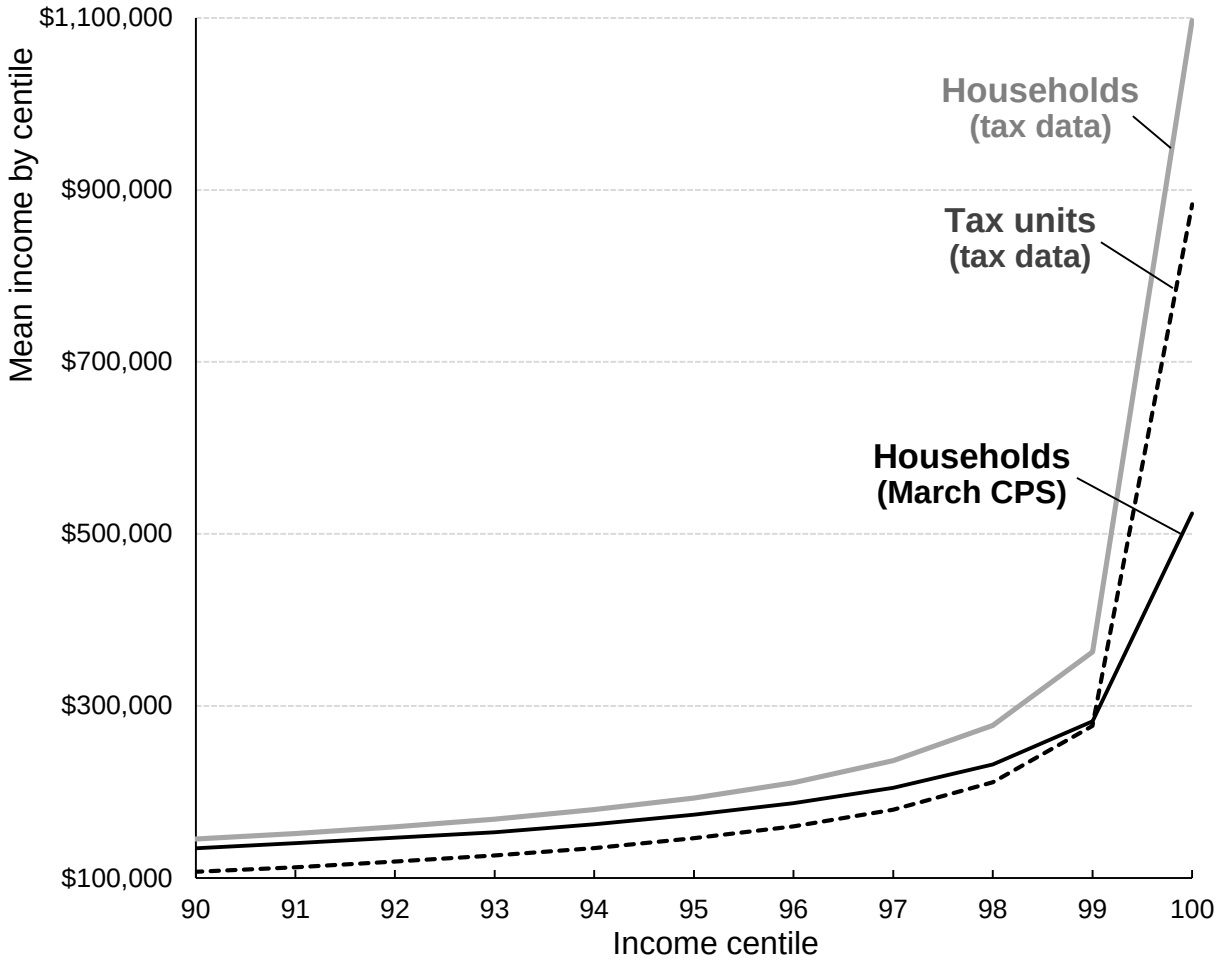
Size of Household	Unedited addresses	Standardize abbreviations	Next-year match	Prior-year match	Fuzzy match
1	40,804	36,016	33,121	32,607	32,555
2	32,575	32,630	32,772	32,766	32,771
3	17,899	18,087	18,277	18,311	18,314
4	15,227	15,379	15,484	15,504	15,505
5	7,429	7,600	7,690	7,709	7,711
6	3,477	3,586	3,635	3,646	3,647
7 or more	2,537	2,697	2,756	2,769	2,771
Total	**119,948**	**115,995**	**113,735**	**113,312**	**113,272**

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

Appendix Table A-3. Household income by source, 2010 (millions of dollars)

	Tax data	Census
Earnings		
Wages and salaries	5,888,401	6,132,916
Self-employment and farm income (minus loss)	292,842	374,998
Other private income		
Partnership, S corporation, rent, royalty, estates/trusts (minus loss)	437,522	---
Rent/royalty/estates/trusts (minus loss)	---	68,374
Interest and Dividends	379,029	255,850
Pensions, annuities, and IRA distributions	929,886	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	140,607	97,361
Social Security and disability benefits	695,307	593,855
Total pre-tax income on tax returns	**8,859,661**	**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

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 THS, but interest and dividends are based on total income plus tax-exempt interest less other sources.
Source: U.S. Census Bureau's March CPS, IRS Statistics of Income data, THS and authors' calculations.



Appendix Figure A-1. Top decile distribution by unit of observation and data source, 2010

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

ONLINE APPENDIX

HOUSEHOLD INCOMES IN TAX DATA: USING ADDRESSES TO MOVE FROM TAX UNIT TO HOUSEHOLD INCOME DISTRIBUTIONS

Jeff Larrimore
Jacob Mortenson
David Splinter

Address abbreviation standardization SQL code:

```
update #filers1 set address=replace(address, '.', '');
update #filers1 set address=replace(address, ',', '');
update #filers1 set address=replace(address, '/', '');
update #filers1 set address=replace(address, '-', '');
update #filers1 set address=replace(address, '#', '');
update #filers1 set address=replace(address, '&', '');
update #filers1 set address=replace(address, ':', '');
update #filers1 set address=replace(address, '(', '');
update #filers1 set address=replace(address, ')', '');
update #filers1 set address=replace(address, '''', '');

update #filers1 set address=replace(address, ' NUMBER ', ' ');
update #filers1 set address=replace(address, ' UNIT ', ' ');
update #filers1 set address=replace(address, ' APART ', ' ');
update #filers1 set address=replace(address, ' APT ', ' ');
update #filers1 set address=replace(address, ' SUITE ', ' ');
update #filers1 set address=replace(address, ' STE ', ' ');
update #filers1 set address=replace(address, ' NO ', ' ');
update #filers1 set address=replace(address, ' NM ', ' ');

update #filers1 set address=replace(address, 'PO BOX', 'BOX');
update #filers1 set address=replace(address, 'POBOX', 'BOX');
update #filers1 set address=replace(address, 'PO ', '');
update #filers1 set address=replace(address, 'P O ', '');
update #filers1 set address=replace(address, 'POST OFFICE ', '');
update #filers1 set address=replace(address, 'PMB ', 'BOX ');
update #filers1 set address=replace(address, 'POB ', 'BOX ');
update #filers1 set address=replace(address, 'BX ', 'BOX ');

update #filers1 set address=replace(address, 'NORTH', 'N');
update #filers1 set address=replace(address, 'SOUTH', 'S');
update #filers1 set address=replace(address, 'EAST', 'E');
update #filers1 set address=replace(address, 'WEST', 'W');

update #filers1 set address=replace(address, 'FIRST', '1ST');
update #filers1 set address=replace(address, 'FRST', '1ST');
update #filers1 set address=replace(address, 'FST', '1ST');
update #filers1 set address=replace(address, 'SECOND', '2ND');
update #filers1 set address=replace(address, 'THIRD', '3RD');
update #filers1 set address=replace(address, 'FOURTH', '4TH');
update #filers1 set address=replace(address, 'FIFTH', '5TH');
update #filers1 set address=replace(address, 'SIXTH', '6TH');
update #filers1 set address=replace(address, 'SEVENTH', '7TH');
update #filers1 set address=replace(address, 'EIGHTH', '8TH');
update #filers1 set address=replace(address, 'NINTH', '9TH');
update #filers1 set address=replace(address, 'TENTH', '10TH');
```

```
update #filers1 set address=replace(address, ' OFFICE', ' OFC');
update #filers1 set address=replace(address, ' ALLEY', ' ALY');
update #filers1 set address=replace(address, ' ALLY', ' ALY');
update #filers1 set address=replace(address, ' ANNEX', ' ANX');
update #filers1 set address=replace(address, ' ARCADE', ' ARC');
update #filers1 set address=replace(address, ' AV ', ' AVE ');
update #filers1 set address=replace(address, ' AVENUE', ' AVE');
update #filers1 set address=replace(address, ' AVENU', ' AVE');
update #filers1 set address=replace(address, ' AVEN', ' AVE');
update #filers1 set address=replace(address, ' BAYOU', ' BYU');
update #filers1 set address=replace(address, ' BEACH', ' BCH');
update #filers1 set address=replace(address, ' BEND', ' BND');
update #filers1 set address=replace(address, ' BLUFFS', ' BLFS');
update #filers1 set address=replace(address, ' BLUFF', ' BLF');
update #filers1 set address=replace(address, ' BOTTOM', ' BTM');
update #filers1 set address=replace(address, ' BOULEVARD', ' BLVD');
update #filers1 set address=replace(address, ' BOULEV', ' BLVD');
update #filers1 set address=replace(address, ' BLVDAPT', ' BLVD APT');
update #filers1 set address=replace(address, ' BRANCH', ' BR');
update #filers1 set address=replace(address, ' BRIDGE', ' BRG');
update #filers1 set address=replace(address, ' BROOKS', ' BRKS');
update #filers1 set address=replace(address, ' BROOK', ' BRK');
update #filers1 set address=replace(address, ' BURG', ' BG');
update #filers1 set address=replace(address, ' BYPASS', ' BYP');
update #filers1 set address=replace(address, ' CAMP', ' CP');
update #filers1 set address=replace(address, ' CANYON', ' CYN');
update #filers1 set address=replace(address, ' CAPE', ' CPE');
update #filers1 set address=replace(address, ' CAUSEWAY', ' CSWY');
update #filers1 set address=replace(address, ' CENTER', ' CTR');
update #filers1 set address=replace(address, ' CENTRE', ' CTR');
update #filers1 set address=replace(address, ' CNTR', ' CTR');
update #filers1 set address=replace(address, ' CIRCL', ' CIR');
update #filers1 set address=replace(address, ' CIRC', ' CIR');
update #filers1 set address=replace(address, ' CLIFFS', ' CLFS');
update #filers1 set address=replace(address, ' CLIFF ', ' CLF ');
update #filers1 set address=replace(address, ' COMMONS', ' CMNS');
update #filers1 set address=replace(address, ' COMMON', ' CMN');
update #filers1 set address=replace(address, ' CORNERS', ' CORS');
update #filers1 set address=replace(address, ' CORNER', ' COR');
update #filers1 set address=replace(address, ' COURSE', ' CRSE');
update #filers1 set address=replace(address, ' COURTS', ' CTS');
update #filers1 set address=replace(address, ' COURT', ' CT');
update #filers1 set address=replace(address, ' CR ', ' CREEK ');  // reverse abbr.
update #filers1 set address=replace(address, ' CRES ', ' CRESCENT '); // reverse abbr.
update #filers1 set address=replace(address, ' CRST', ' CREST');   // reverse abbr.
update #filers1 set address=replace(address, ' CROSSING ', ' XING');
update #filers1 set address=replace(address, ' CROSSROADS', ' XRD');
update #filers1 set address=replace(address, ' CROSSROAD', ' XRD');
update #filers1 set address=replace(address, ' CURVE', ' CURV');
update #filers1 set address=replace(address, ' DALE', ' DL');
update #filers1 set address=replace(address, ' DAM ', ' DM');
update #filers1 set address=replace(address, ' DIV', ' DV');
update #filers1 set address=replace(address, ' DRIVE', ' DR');
update #filers1 set address=replace(address, ' DRIV', ' DR');
update #filers1 set address=replace(address, ' DRV', ' DR');
update #filers1 set address=replace(address, ' ESTATES', ' ESTS');
update #filers1 set address=replace(address, ' ESTATE', ' EST');
update #filers1 set address=replace(address, ' EXPRESSWAY', ' EXPY');
update #filers1 set address=replace(address, ' EXPRESS', ' EXPY');
update #filers1 set address=replace(address, ' EXPR', ' EXPY');
update #filers1 set address=replace(address, ' EXPW', ' EXPY');
update #filers1 set address=replace(address, ' EXP', ' EXPY');
```

```
update #filers1 set address=replace(address, ' EXTENSION', ' EXT');
update #filers1 set address=replace(address, ' FALLS', ' FLS');
update #filers1 set address=replace(address, ' FERRY', ' FRY');
update #filers1 set address=replace(address, ' FIELDS', ' FLDS');
update #filers1 set address=replace(address, ' FIELD', ' FLD');
update #filers1 set address=replace(address, ' FLATS', ' FLTS');
update #filers1 set address=replace(address, ' FLAT', ' FLT');
update #filers1 set address=replace(address, ' FLOOR', ' FL');
update #filers1 set address=replace(address, ' FOREST', ' FRST');
update #filers1 set address=replace(address, ' FORGE', ' FRG');
update #filers1 set address=replace(address, ' FORKS', ' FRKS');
update #filers1 set address=replace(address, ' FORK', ' FRK');
update #filers1 set address=replace(address, ' FORT', ' FT');
update #filers1 set address=replace(address, ' FREEWAY', ' FWY');
update #filers1 set address=replace(address, ' FRWY', ' FWY');
update #filers1 set address=replace(address, ' GARDENS', ' GDNS');
update #filers1 set address=replace(address, ' GARDEN', ' GDN');
update #filers1 set address=replace(address, ' GARDN', ' GDN');
update #filers1 set address=replace(address, ' GATEWAY', ' GTWY');
update #filers1 set address=replace(address, ' GLENS', ' GLNS');
update #filers1 set address=replace(address, ' GLEN', ' GLN');
update #filers1 set address=replace(address, ' GREENS', ' GRNS');
update #filers1 set address=replace(address, ' GREEN', ' GRN');
update #filers1 set address=replace(address, ' GROVES', ' GRVS');
update #filers1 set address=replace(address, ' GROVE', ' GRV');
update #filers1 set address=replace(address, ' HARBOR', ' HBR');
update #filers1 set address=replace(address, ' HARB', ' HBR');
update #filers1 set address=replace(address, ' HAVEN', ' HVN');
update #filers1 set address=replace(address, ' HEIGHTS', ' HTS');
update #filers1 set address=replace(address, ' HEIGHT', ' HTS');
update #filers1 set address=replace(address, ' HT', ' HTS');
update #filers1 set address=replace(address, ' HIGHWAY', ' HWY');
update #filers1 set address=replace(address, ' HIWAY', ' HWY');
update #filers1 set address=replace(address, ' HILLS', ' HLS');
update #filers1 set address=replace(address, ' HILL', ' HL');
update #filers1 set address=replace(address, ' HOLLOW', ' HOLW');
update #filers1 set address=replace(address, ' ISLES', ' ISLE');
update #filers1 set address=replace(address, ' JUNCTION', ' JCT');
update #filers1 set address=replace(address, ' KEYS', ' KYS');
update #filers1 set address=replace(address, ' KEY', ' KY');
update #filers1 set address=replace(address, ' KNOLLS', ' KNLS');
update #filers1 set address=replace(address, ' KNOLL', ' KNL');
update #filers1 set address=replace(address, ' LAKES', ' LKS');
update #filers1 set address=replace(address, ' LAKE', ' LK');
update #filers1 set address=replace(address, ' LANDING', ' LNDG');
update #filers1 set address=replace(address, ' LANE', ' LN');
update #filers1 set address=replace(address, ' MANOR', ' MNR');
update #filers1 set address=replace(address, ' MEADOWS', ' MDWS');
update #filers1 set address=replace(address, ' MEADOW', ' MDW');
update #filers1 set address=replace(address, ' MILL', ' ML');
update #filers1 set address=replace(address, ' MOUNTAIN', ' MTN');
update #filers1 set address=replace(address, ' MOUNT', ' MT');
update #filers1 set address=replace(address, ' ORCHARD', ' ORCH');
update #filers1 set address=replace(address, ' PRK', ' PARK');
update #filers1 set address=replace(address, ' PARKS', ' PARK');
update #filers1 set address=replace(address, ' PARKWAY', ' PKWY');
update #filers1 set address=replace(address, ' PKWAY', ' PKWY');
update #filers1 set address=replace(address, ' PKY', ' PKWY');
update #filers1 set address=replace(address, ' PIKES', ' PIKE');
update #filers1 set address=replace(address, ' PINES', ' PNES');
update #filers1 set address=replace(address, ' PINE', ' PNE');
update #filers1 set address=replace(address, ' PLACE', ' PL');
update #filers1 set address=replace(address, ' PLAINS', ' PLNS');
```

```
update #filers1 set address=replace(address, ' PLAIN', ' PLN');
update #filers1 set address=replace(address, ' PLAZA', ' PLZ');
update #filers1 set address=replace(address, ' POINTS', ' PTS');
update #filers1 set address=replace(address, ' POINT', ' PT');
update #filers1 set address=replace(address, ' PORTS', ' PRTS');
update #filers1 set address=replace(address, ' PORT', ' PRT');
update #filers1 set address=replace(address, ' PRAIRIE', ' PR');
update #filers1 set address=replace(address, ' RANCH', ' RNCH');
update #filers1 set address=replace(address, ' RAPIDS', ' RPDS');
update #filers1 set address=replace(address, ' RAPID', ' RPD');
update #filers1 set address=replace(address, ' REST', ' RST');
update #filers1 set address=replace(address, ' RIDGE', ' RDG');
update #filers1 set address=replace(address, ' RIVER', ' RIV');
update #filers1 set address=replace(address, ' RVR', ' RIV');
update #filers1 set address=replace(address, ' ROADS', ' RDS');
update #filers1 set address=replace(address, ' ROAD', ' RD');
update #filers1 set address=replace(address, ' ROUTE', ' RTE');
update #filers1 set address=replace(address, ' SHOALS', ' SHLS');
update #filers1 set address=replace(address, ' SHOAL', ' SHL');
update #filers1 set address=replace(address, ' SHORES', ' SHRS');
update #filers1 set address=replace(address, ' SHORE', ' SHR');
update #filers1 set address=replace(address, ' SKYWAY', ' SKWY');
update #filers1 set address=replace(address, ' SPRINGS', ' SPGS');
update #filers1 set address=replace(address, ' SPRING', ' SPG');
update #filers1 set address=replace(address, ' SQUARE', ' SQ');
update #filers1 set address=replace(address, ' SQU', ' SQ');
update #filers1 set address=replace(address, ' STATION', ' STA');
update #filers1 set address=replace(address, ' STRAVENUE', ' STRA');
update #filers1 set address=replace(address, ' STRAVEN', ' STRA');
update #filers1 set address=replace(address, ' STRAVN', ' STRA');
update #filers1 set address=replace(address, ' STREAM', ' STRM');
update #filers1 set address=replace(address, ' STREET', ' ST');
update #filers1 set address=replace(address, ' STAPT', ' ST APT');
update #filers1 set address=replace(address, ' STR ', ' ST ');
update #filers1 set address=replace(address, ' SUMMIT', ' SMT');
update #filers1 set address=replace(address, ' TERRACE', ' TER');
update #filers1 set address=replace(address, ' TERR', ' TER');
update #filers1 set address=replace(address, ' TRACE', ' TRCE');
update #filers1 set address=replace(address, ' TRACK', ' TRAK');
update #filers1 set address=replace(address, ' TRAILS', ' TRL');
update #filers1 set address=replace(address, ' TRAIL', ' TRL');
update #filers1 set address=replace(address, ' TRAILER', ' TRLR');
update #filers1 set address=replace(address, ' TUNNEL', ' TUNL');
update #filers1 set address=replace(address, ' TURNPIKE', ' TPKE');
update #filers1 set address=replace(address, ' UNION', ' UN');
update #filers1 set address=replace(address, ' VALLEYS', ' VLYS');
update #filers1 set address=replace(address, ' VALLEY', ' VLY');
update #filers1 set address=replace(address, ' VALLY', ' VLY');
update #filers1 set address=replace(address, ' VIEW', ' VW');
update #filers1 set address=replace(address, ' VILLAGE', ' VLG');
update #filers1 set address=replace(address, ' VILLAG', ' VLG');
update #filers1 set address=replace(address, ' VILLE', ' VLG');
update #filers1 set address=replace(address, ' VILL', ' VLG');
update #filers1 set address=replace(address, ' VISTA', ' VIS');
update #filers1 set address=replace(address, ' VIST', ' VIS');
update #filers1 set address=replace(address, ' WY ', ' WAY ');
update #filers1 set address=replace(address, ' WELLS', ' WLS');
update #filers1 set address=replace(address, ' WELL', ' WL');
```

Fuzzy Address Match SQL code:

```
/* Levenshtein distance function, counts number of characters different
www.kodyaz.com/articles/fuzzy-string-matching-using-levenshtein-distance-sql-server.aspx */
CREATE TEMPORARY FUNCTION Ldist(@s1 nvarchar(3999), @s2 nvarchar(3999))
RETURNS int
AS
BEGIN
 DECLARE @s1_len int, @s2_len int
 DECLARE @i int, @j int, @s1_char nchar, @c int, @c_temp int
 DECLARE @cv0 varbinary(8000), @cv1 varbinary(8000)

 SELECT
  @s1_len = LEN(RTRIM(@s1)),
  @s2_len = LEN(RTRIM(@s2)),
  @cv1 = 0x0000,
  @j = 1, @i = 1, @c = 0

 WHILE @j <= @s2_len
  SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1

 WHILE @i <= @s1_len /* outer loop: through s1 one character at a time */
 BEGIN
  SELECT
   @s1_char = SUBSTRING(@s1, @i, 1),
   @c = @i,
   @cv0 = CAST(@i AS binary(2)),
   @j = 1

  WHILE @j <= @s2_len  /* inner loop: through s2 one character at a time */
  BEGIN
   SET @c = @c + 1
   SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
    CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0 ELSE 1 END
   IF @c > @c_temp SET @c = @c_temp
   SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
   IF @c > @c_temp SET @c = @c_temp
   SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1
 END

 SELECT @cv1 = @cv0, @i = @i + 1
 END

 RETURN @c
END;

// MERGE TO VALID STREET NAMES IF INVALID (fuzzy match): inner join if invalid address and
minimum street criteria in same zip and HHsizes=1 (deal with multiple merges deal next)
select a.tin, ltrim(rtrim(a.alphonly)) as a1, ltrim(rtrim(b.alphonly)) as a2, a.address
    , case when SUBSTR(a1,1,4)=SUBSTR(a2,1,4) and LEN(ltrim(rtrim(a1)))
=LEN(ltrim(rtrim(a2)))                                                    then 1
           when SUBSTR(a1,1,4)=SUBSTR(a2,1,4) and
(LEN(ltrim(rtrim(a1)))+1)>=LEN(ltrim(rtrim(a2))) and (LEN(ltrim(rtrim(a1)))-
1)<=LEN(ltrim(rtrim(a2))) then 2
           when SUBSTR(a1,1,3)=SUBSTR(a2,1,3) and LEN(ltrim(rtrim(a1)))
=LEN(ltrim(rtrim(a2)))                                                    then 3
           when SUBSTR(a1,1,3)=SUBSTR(a2,1,3) and
(LEN(ltrim(rtrim(a1)))+1)>=LEN(ltrim(rtrim(a2))) and (LEN(ltrim(rtrim(a1)))-
1)<=LEN(ltrim(rtrim(a2))) then 4
           when SUBSTR(a1,1,2)=SUBSTR(a2,1,2) and LEN(ltrim(rtrim(a1)))
=LEN(ltrim(rtrim(a2)))                                                    then 5
           when SUBSTR(a1,1,2)=SUBSTR(a2,1,2) and
(LEN(ltrim(rtrim(a1)))+1)>=LEN(ltrim(rtrim(a2))) and (LEN(ltrim(rtrim(a1)))-
1)<=LEN(ltrim(rtrim(a2))) then 6
```

```
           when SUBSTR(a1,1,4)=SUBSTR(a2,1,4)   then 7
           when SUBSTR(a1,1,3)=SUBSTR(a2,1,3)   then 8
           when SUBSTR(a1,1,2)=SUBSTR(a2,1,2)   then 9
           when SUBSTR(a1,1,1)=SUBSTR(a2,1,1) and
LEN(ltrim(rtrim(a1)))=LEN(ltrim(rtrim(a2)))
then 10
           when SUBSTR(a1,1,1)=SUBSTR(a2,1,1) and
(LEN(ltrim(rtrim(a1)))+1)>=LEN(ltrim(rtrim(a2))) and (LEN(ltrim(rtrim(a1)))-
1)<=LEN(ltrim(rtrim(a2))) then 11
           when SUBSTR(a1,1,1)=SUBSTR(a2,1,1) then 12 else 500 end as arank
     into #badadd3
     from #badadd2 a join #addr_data b
       on a.zipcode=b.zipcode and a.alphonly<>b.alphonly
        and a.HHsize=1
        and SUBSTR(a.alphonly,1,1)=SUBSTR(b.alphonly,1,1)
        and (LEN(ltrim(rtrim(a.alphonly)))+3)>=LEN(ltrim(rtrim(b.alphonly))) and
(LEN(ltrim(rtrim(a.alphonly)))-3)<=LEN(ltrim(rtrim(b.alphonly)));

// rank with duplicates
select a.tin, a.address, a1, a2, arank, Ldist(a1,a2) as Ldis into #badadd4a
     from (select z.tin, z.address, z.a1, z.a2, z.arank
       from (select tin, address, a1, a2, arank, rank() over (partition by tin order by arank
asc) as brank
             from #badadd3) z where z.brank=1) as a;

// rank by Ldist to remove duplicates (false positives with arank 9 and over)
select a.tin, a.address, a1, a2, arank into #badadd4b
     from (select z.tin, z.address, z.a1, z.a2, z.arank
       from (select tin, address, a1, a2, arank, row_number() over (partition by tin order by
Ldis asc) as crank
             from #badadd4a where Ldis<=4) z where z.crank=1 and arank<9) as a;

update #badadd4b set address=replace(ltrim(rtrim(address)), ltrim(rtrim(a1)),
ltrim(rtrim(a2)));


// Merge fuzzy matches back to full list to replace address:
select a.tin, a.zipcode as zip, a.state, case when (b.address IS NOT NULL) then b.address
else a.address end as address into #filers20  // 543766 obs 1sec
   from #filers11 a left outer join #badadd4b b on a.tin=b.tin;
```