# Online Appendix

For "Estimating Tax Liabilities and Credits

Using Linked Survey and Administrative Data"

by Bruce D. Meyer, Derek Wu, Grace Finley,

Patrick Langetieg, Carla Medalia, Mark Payne, and Alan Plumley*

*– For Online Publication Only –*

* Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau, the Internal Revenue Service, any agency of the federal government, or the NBER. All results were approved for release by the U.S. Census Bureau, authorization number sCBDRB-FY20-ERD002-014 and CBDRB-FY20-ERD002-038.

## Detailed Description of Data

To calculate tax liabilities and credits, we rely on two types of data: survey data and administrative tax records. Our survey data come from the 2011 Current Population Survey Annual Social and Economic Supplement. We also rely on two IRS tax datasets: one with limited 1040 return information and another with extensive 1040 return information that includes all 1040 line items. We focus on reference year 2010, since this is the only year for which the extensive 1040 return information is available.

### *Current Population Survey*

The CPS ASEC (hereafter referred to as CPS) collects demographic and income information for households representing the civilian non-institutionalized U.S. population.[1] The 2011 CPS interviewed households between February and April 2011 about their economic characteristics for the previous calendar year. Furthermore, using reported incomes and information on family structure, the CPS uses an in-house calculator to impute amounts for various tax liabilities, credits, and inputs. These items include federal and state income tax, payroll tax, Earned Income Tax Credit (EITC), Child Tax Credit (CTC), and adjusted gross income (AGI). While the CPS is a household-level survey, we focus on families and unrelated individuals as our units of analysis. CPS family

[1] For more information on CPS's technical documentation, see https://www.census.gov/programs-surveys/cps/technical-documentation.html.

units are contained within CPS households. Following the CPS definition of families, we combine primary and related subfamilies.[2]

### *Administrative Tax Data (Overview)*

We rely on two different datasets provided by the IRS. The first dataset, hereafter referred to as the "limited tax data," is provided to the U.S. Census Bureau under U.S. code 6103(j), which allows the U.S. Census Bureau to use IRS tax data for Census survey improvement.[3] The second dataset, hereafter referred to as the "extensive tax data," is provided to the U.S. Census Bureau under U.S. code 6103(n), which gives access to the data for the purpose of tax administration.[4]

### *Limited Tax Data*

The limited tax data include data provided on Forms 1040, W-2, and 1099-R. While we have data for tax years 1969, 1974, 1979, 1984, 1989, 1994-1995, and 1998-2016, we focus on tax year 2010 for our analysis. The limited tax data for 2010 include 1040s, W-2s, and 1099-Rs submitted during calendar year 2011 for tax year 2010. Note that the limited tax data contain 139.1 million 1040 forms, while the extensive tax data contain 139.1 million 1040 forms submitted for the 2010 tax year – the similarity of these numbers further confirms that the limited tax data only contain 1040 forms submitted for the 2010 tax year.

We start by describing the information available on 1040s in the limited tax data. We first have a set of variables covering tax unit structure – these include a unique tax unit identifier (generated by the Census Bureau), a variable identifying the type of 1040 form (e.g., 1040EZ, 1040NR), filing status, and various types of exemptions claimed. The numbers of exemptions claimed are broken down by the type of exemption (primary exemptions, secondary exemptions, exemptions for children living at home, exemptions for children not living at home, exemptions for children in excess of nine, exemptions for parents of taxpayer/spouse living at home, exemptions for

[2] For primary families, total family income is the sum of individual incomes across all members of a primary and related subfamily. For related subfamilies, total family income is only the sum of individual incomes across the related subfamily. Most income information is at the individual level, meaning that we can combine income across individuals to estimate family-level income. We assign income sources received at the family level to the family head when summing across individual income.

[3] See https://www.census.gov/content/dam/Census/programs-surveys/center-for-economic-studies/IRS_Criteria_Document.pdf.

[4] See page 3180, https://www.govinfo.gov/content/pkg/USCODE-2011-title26/pdf/USCODE-2011-title26-subtitleF-chap61-subchapB-sec6103.pdf.

individuals other than primary taxpayer, spouse, children, and parents of taxpayer/spouse, and exemptions for other children on the return eligible for EITC).

We also have income amounts for the following 1040 line items: wages/salaries, taxable dividends, taxable interest, tax-exempt interest, gross rents and royalties, social security income, adjusted gross income, and total money income.[5] Total money income includes most sources of income reported in the "income" section of the 1040 form, but it misses several key income sources – including capital gains – and is therefore not identical to total income on a tax return (line 22 of the 2010 1040 form). Specifically, total money income includes wages/salaries, total interest income (taxable and tax-exempt), taxable dividends, alimony received, business income, pensions and annuities, net rents/royalties, farm income, unemployment compensation, and total social security benefits. The limited tax data also include separate indicators each (equal to 1 if filed a given schedule and 0 otherwise) for whether a tax unit filed Schedules A, C, D, E, F, and SE.

Furthermore, the limited tax data include EITC earned income and the number of EITC-eligible children (ranging from 0 to 3) for tax units that claim the EITC. In other words, these variables have non-missing values only for tax units claiming the EITC. Location-related variables (e.g., street, zip code, state) are also available in the limited tax data for all 1040 forms that are not submitted from international addresses.[6] The post office name for forms submitted from international addresses will be the country from which a tax return is filed rather than the U.S. city. Lastly, the limited tax data for 1040s include a variable for posting date, which indicates the date (in particular, the week) that a 1040 return was posted to the IRS individual master file.[7]

Moving on to information returns in the limited tax data, the W-2 data contain amounts for wages and salaries (Box 1) and deferred compensation (Box 12) for the universe of formal sector jobs.[8] The 1099-R data include information on gross retirement distributions (Box 1) from employer-sponsored plans and Individual Retirement Account withdrawals.

Nevertheless, the limited tax data contain a number of significant holes. In particular, even though they cover enough line items to calculate tax liabilities and credits relatively accurately (especially for those lower in the income distribution), these data do not contain actual amounts for federal income taxes paid or tax credits received. The limited tax data also miss key line items, such

[5] For the 2010 1040 form, see https://www.irs.gov/pub/irs-prior/f1040--2010.pdf.

[6] The post office name for forms submitted from international addresses will be the country from which a tax return is filed rather than the U.S. city

[7] For information on the IRS master file, see https://www.irs.gov/pub/irs-pia/imf_pia.pdf.

[8] For the 2010 W-2 form, see https://www.irs.gov/pub/irs-prior/fw2--2010.pdf.

as capital gains/losses and itemized deductions, which are necessary to generate an accurate estimate of federal income tax liabilities for those higher in the income distribution.[9] While the limited tax data note whether a tax unit files a schedule, we do not receive data on the contents of the schedule that could fill in missing data on certain types of income, such as self-employment income (from Schedules C, F, SE, and K-1).[10] Other than Forms W-2 and 1099-R data, we are missing all other information returns. Of particular interest to us are 1099-Gs (which cover unemployment compensation), 1099-MISCs (which cover self-employment earnings for independent contractors), and Schedule K-1 information returns (which cover partnership earnings). Without these information returns, we are unable to accurate estimate tax liabilities for survey individuals to whom we attach a 1040 form but not an information return (as they may have had federal and/or state income taxes withheld).

*Extensive Tax Data*

The extensive tax data comprise a set of over 50 files comprising information for 1040 returns, certain schedules (namely, Schedules A, C, D, E, F, and SE), and information returns submitted to the IRS during the tax filing period for tax year 2010. While the information returns are only for tax year 2010, the 1040 forms include forms filed during 2010 for tax year 2010 and prior tax years (late filers). Specifically, 3.4% of 1040 forms filed in calendar year 2011 were filed for tax years prior to 2010.

The extensive data 1040 return information is comprised of five separate files: the income file, a secondary filer file, and three dependent files. The secondary filer file and dependent files provide the identifying information for the secondary filers and dependents. The income file includes variables covering nearly even line on a 1040 return, such as filing status, wages, taxable interest, ordinary dividends, and unemployment compensation. Crucially, this file contains amounts for federal income tax liabilities, credits (such as the Earned Income Credit), and various deductions (such as the health savings account deduction). Importantly, the extensive tax data contain two versions of nearly every line item of the 1040 – one containing raw values corresponding to what

[9] Specifically, the limited tax data do not directly include income amounts for alimony received (Line 11), self-employment income (Line 12 + Line 18), capital gains/losses (Line 13), pensions and annuities (Line 15b + Line 16b), and unemployment compensation (Line 19). However, besides capital gains/losses, these income types are captured in total money income.

[10] We are also missing the payroll tax on self-employment income that appears on Line 55 of the 1040 form.

was originally filled out and one containing computer-corrected values that correct for obvious errors (e.g., missing decimal point, too many zeros, etc.).[11] When possible, we use the computer-corrected version of a variable. We also have a taxpayer information file associated with the 1040s, which contains the primary filer's PIK and address information. Lastly, another file contains occupation information for the primary and secondary filer.

The extensive tax data also contain a data file covering most lines on Schedules A, C, D, E, F, and SE. The data for Schedule A, which covers itemized deductions, include amounts on total itemized deductions and some line items that serve as inputs into total itemized deductions.[12] The data for Schedule C, which covers profit or loss from business, contain net profit/loss.[13] The data for Schedule D, which covers capital gains and losses, contain four variables: long-term and short-term sales and net long-term and short-term gains/losses.[14] The data for Schedule E, which covers supplemental income and loss, include a wide variety of income sources, including estate and trust income/loss, farm/fishing income, mortgage income, non-passive income/loss, rental income, and royalty income.[15] The data for Schedule F, which covers profit or loss from farming, include net farm profit from either the accrual method or cash.[16,17,18] Finally, the data for Schedule SE, which covers self-employment tax, include total self-employment earnings (provided they are above $400 as well as their components: nonfarm profit/loss, net farm profit/loss, and Conservation Reservation Program Payments.[19]

Lastly, the extensive tax data cover nearly all information returns – including, most notably, Form 1099-G, Form 1099-MISC, Form 1099-R, and Schedule K-1. We also have access to a series of files labeled "tax info," each containing address information and a list of the tax forms

[11] The corrected amounts are generated by a computer program that catches if individuals make obvious mistakes in filling out their tax forms. For example, we discovered a CTC amount that had been inputted twice (e.g., 10,001,000 instead of 1000) into the data field. The computer program caught this mistake and only listed the CTC amount once in the corrected variable. Other mistakes caught by the computer program include addition and subtraction mistakes.

[12] Schedule A line items include Lines 1, 4, 6, 8, 9, 10, 14, 15, 18, 19, 21, 27, 28, and 29. See Schedule A for line item descriptions, https://www.irs.gov/pub/irs-prior/f1040sa--2010.pdf.

[13] Schedule C line items include Lines 1d, 3, 4, 6, 23, and 30. We also have a computer-calculated Line 31. See Schedule C for line item descriptions, https://www.irs.gov/pub/irs-prior/f1040sc--2010.pdf.

[14] See Schedule D for more information, https://www.irs.gov/pub/irs-prior/f1040sd--2010.pdf.

[15] Schedule E line items include Lines 23a, 23b, 24, 29a, 29b, 30, 31, 35, 36, 39, and 42. See Schedule E for line item descriptions, https://www.irs.gov/pub/irs-prior/f1040se--2010.pdf.

[16] See Schedule F for line item descriptions, https://www.irs.gov/pub/irs-prior/f1040sf--2010.pdf.

[17] Unlike Schedule C, Schedule F does not provide net farm income, which is necessary to calculate self-employment income.

[18] Self-employed workers can choose the accounting method by which their income is estimates for taxation purposes. The cash method follows cash into and out of a worker's bank account. The accrual method focuses on when expenses were incurred. For example, a worker may finish a project prior to receiving payment.

[19] See Schedule SE for line item descriptions, https://www.irs.gov/pub/irs-prior/f1040sse--2010.pdf.

corresponding to every individual. In these files, “payer” refers to the entity generating the form – usually a bank, employer, or trust depending on the form. These files generally contain two addresses – one presumably for the taxed individual (payee) and the other for the payer.

## Attaching Tax Data to Survey Data

### *Protected Identification Keys*

We attach administrative tax records to survey data using Protected Identification Keys (PIKs) created by the U.S. Census Bureau’s Person Identification Validation System (PVS). 91.7% of families in the CPS contain at least one member linked to a PIK. Because administrative records cannot be attached to CPS families where no member has a PIK, we limit our CPS sample to only include survey families with at least one PIKed member and no whole imputed individuals. Compared to an original sample size of 205,000 individuals forming 84,500 families in the CPS, our analysis sample contains 170,000 individuals forming 69,000 families. To correct for the bias arising from non-random missing PIKs and whole imputations, we divide individual CPS weights by the predicted probability that at least one member of the family has a PIK and no member is whole imputed (conditional on observables in the survey). Inverse probability weights are based on a probit model that determines the probability of a family being PIKed based on observable characteristics in the survey.

Almost all administrative tax records are linked to PIKs. In the limited tax data, more than 99.9% of 1040 forms have at least a PIKed primary or secondary filer, more than 99% of W-2 forms are PIKed, and 98.9% of 1099-R forms are PIKed.[20] In the extensive tax data, 99.9% of 1040 forms have at least a PIKed primary or secondary filer, 99.5% of W-2 forms are PIKed, and 99.8% of 1099-R forms are PIKed. Furthermore, in the extensive tax data, 98.8% of 1040 forms have the same number of dependent exemptions as PIKed dependents, while 1.1% have fewer PIKed dependents than dependent exemptions and 0.09% have fewer dependent exemptions than PIKed dependents.[21]

In the limited tax data, identifiers for the primary filer, secondary filer, and up to four dependents are available for each tax return.[22] In the extensive tax data, each of the 1040 return files

[20] Shares are based on the original dataset before the data are cleaned, including before removing duplicate observations.
[21] Once again, note that shares are based on the original dataset prior to removing tax units containing duplicate PIKs.
[22] In addition to the actual PIKs, the limited tax data also include PVS verification flags equal to a character indicator for the type of PIK (e.g., unPIKed). Unlike PIKs, which may be missing for some individuals in certain tax units, the PVS verification flags are non-missing for all primary filers and appear to be non-missing for tax units who list secondary filers or dependents since one potential value for the flag reflects not being PIKed.

is each identified by the primary filer's PIK, posting date, and 1040 tax return ID. The secondary filer file also contains the secondary filer's PIK, and the dependent files contain the PIKs of the dependents. Unlike the limited 1040 tax data, the extensive 1040 tax data are divided into several different components that have to be merged together to form a final 1040 data file. Rather than merging these files together based on only a PIK, we instead use the 1040 tax return ID, primary filer PIK, and posting date. The 1040 tax return ID is created by the Census Bureau, while the posting date is the date the tax return is posted to the IRS master file. Together, the tax return ID, primary filer PIK, and posting date uniquely identify more than 99.9% of tax units. If a tax unit cannot be uniquely identified by the tax return ID, primary filer PIK, and posting date, we drop the tax return since we cannot identify which components belong together.  We also drop tax returns that are missing a tax return ID and primary filer PIK (encompassing 0.1% of tax returns) since it is not possible to merge the different components of the 1040 return without at least one of these.

While each of the extensive 1040 schedule files contains the primary filer's PIK, certain schedules (namely C and SE) also contain the PIK corresponding to the individual (either the primary or secondary filer) receiving the income. For these schedules (C and SE), we attach the schedule to the survey individual that reported the income. For individuals that file more than one Schedule C or SE (which is common for Schedule C but rare for Schedule SE), we sum incomes across the schedules. If this person's PIK is missing (which is the case for 1.2% of Schedule C and SE forms), then we sum income from the schedule across the primary filer PIK, tax return ID, and posting date which identifies the tax return and attach the information to the primary filer regardless of who earned the income. All other schedules are attached to all survey individuals in a tax unit rather than a single individual. Once again, we merge based on tax return ID, primary PIK, and posting date, dropping tax returns that are missing a primary PIK and tax return ID or cannot be uniquely identified by the tax return ID, primary PIK, and posting date.

With regard to information returns, the W-2s and 1099-Rs in the limited tax data and extensive tax data each contain a PIK corresponding to the individual receiving that information return. 1099-Gs in the extensive tax data also contain a PIK corresponding to the individual receiving that information return.

### *Attaching Limited Tax Data to Survey Data*

When attaching administrative 1040 returns to survey data, we want to attach only one 1040 return to each survey individual. This is because an individual can theoretically only appear as a primary or secondary filer on one tax return, except for an individual who files as married filing separately. Note that we only attach information from 1040 forms filed as married, filing separately to the filer and not the spouse of the filer.[23] This is because income and tax amounts on married, filing separately returns are only representative of the filer and not the spouse, even though the Social Security Numbers of both spouses appear on the return.

For all other individuals, we have to choose which 1040 tax return to attach to the CPS when an individual is a primary or secondary filer on multiple returns (e.g., has amended or corrected returns). If two or more forms can be attached to one survey individual, we start by attaching the form with the latest posting date. If the forms have the same latest posting date, then we attach the form with the filing status of married, filing jointly.[24] If the forms have the same filing status, we attach the form with the higher AGI amount.[25,26,27] When attaching 1040 returns to survey individuals, we only bring in tax returns whose primary or secondary filers appear in the CPS.[28] If a dependent appears on multiple forms as a dependent, then we keep the dependent on only one 1040 form, choosing the form with the latest posting date, filing status of married, filing jointly, or highest AGI. Ideally, we would not drop any dependent because while it is not legal for multiple people to claim the same dependent, it is reasonable to assume that people do (e.g., divorced parents claiming the same child). However, to simplify the process of assigning CPS individuals to tax units, we require that every person attach to only one tax unit.

In some very rare circumstances, a joint 1040 return may have a primary and secondary filer appearing in different families in the CPS. Because our unit of analysis is the CPS family or an unrelated individual, we require tax units to be contained within a CPS family. If the primary and secondary filers from a 1040 return are attached to members of different families, then we do not use

[23] 1.8% of limited 1040 records (and 1.9% of extensive 1040 records) are filed as married, filing separately.

[24] If married filing separately, then keep return where person is the primary filer.

[25] A form that we do not allow to be attached to one individual (i.e. it does not have the latest posting date) cannot be attached to any other survey individuals.

[26] If the forms have the same AGI, then we take the one with the highest wage, taxable income, total tax, or total payments. If the forms have the amounts of each of these 1040 line items, we then choose a form randomly.

[27] If dependents appear on multiple returns, ideally we would attach all returns claiming a dependent to that dependent, but to simplify, we attach the return to the dependent that has latest posting date, is a joint return, or has the highest AGI.

[28] We do not attach 1040 returns to CPS individuals with duplicate PIKs.

the 1040 return information for that tax unit.[29] We are able to attach at least one 1040 return to approximately 89% of CPS families in our sample, with 17% of families attaching to more than one 1040 return.

Individuals may also have multiple valid forms for a given information return (unlike for a 1040). For example, if a person works two jobs, they may receive two W-2s, one from each employer. In this case, we would want to attach both W-2 forms to that person since they represent separate income streams. In rare cases, an employer may even file multiple valid W-2s for an employee. Therefore, for a given combination of PIK and employer identification number (which we can think of as a "job"), we keep all W-2s pertaining to that "job" if every return is designated as an "original" return. When a "job" contains amended or corrected returns, we keep only the amended or corrected form. Specifically, if more than one amended or corrected form appears for a given job, then we keep the form with the latest posting date. If the forms have the same posting date, then we keep the one with the higher income amount.

For 1099-Rs, the limited tax data do not contain amendment codes or information on payers, so we sum over the retirement distributions for all 1099-Rs received by an individual to calculate the total retirement income associated with that individual.

### *Differences Using Extensive Tax Data*

As with the limited tax data, we only want to attach one extensive tax data record of a given type (e.g., 1040 return) to a CPS individual, requiring us to de-duplicate the extensive tax data in a similar manner. Specifically, when de-duplicating the 1040 returns, we follow the same methodology as used for the 1040 returns in the limited tax data, choosing the form with the latest posting date, filing status of married, filing jointly, or the higher AGI (in that order). As with the limited tax data, we attach all extensive data forms using PIKs, and we only attach 1040 returns if the primary or secondary filer on the 1040 form attach to a CPS family and the primary and secondary filers are not split across CPS families.

[29] If dependents are in different families than the primary and secondary filer, then they are treated as single filers with tax liabilities estimated using only CPS information.

However, we de-duplicate information returns in the extensive tax data in a slightly different manner than the limited tax data. This is because “payers” (e.g., employers on a W-2) in the extensive tax data are only identified by their five-digit zip code rather than by their actual IDs – though we have these payer zip codes for all information returns and not just W-2s. When de-duplicating information returns in the extensive tax data, we therefore identify unique sources of income by PIK, payer zip code, and reported income. When a single income source has a corrected or amended form, we keep that form and drop all other original forms. If a payer zip code and PIK combination contains at least one amended/corrected W-2, then we use the limited tax data to identify employers based on the employer identification number (EIN) and keep only one amended W-2 per PIK and EIN combination.

### Generating Survey Tax Units

We assign everyone in the CPS to a survey tax unit with the role of primary filer, secondary filer, or dependent. We create survey tax units in one of two ways.[30] First, for individuals to whom we can attach a 1040 return, we rely on the 1040 tax unit structure to assign them to survey tax units. All other individuals are assigned to survey tax units based on survey family relationship information. Every tax unit has a primary filer. If the filing status of the tax unit is married filing jointly, then there is also a secondary filer.[31] Dependents are children under the age of 19, children under the age of 24 and enrolled in school, individuals that are permanently and totally disabled, or other relatives with income below $3,650.[32] We start by describing how we construct tax units by attaching forms from the limited tax data, and then we discuss any differences in methodology between the limited and extensive tax data.

#### *Individuals with an Attached or Assigned 1040 Form*

For CPS individuals to whom we attach a 1040 return in the limited tax data, we identify their role as primary filers, secondary filers, and dependents based on their status on the 1040 return. A dependent to whom we can also attach a separate 1040 return where he/she is a primary or

[30] These roles are assigned in order to generate the variables for TAXSIM described later.
[31] In very rare cases, some 1040 tax returns attached to CPS individuals contain secondary filers even though these units do not file as married, filing jointly. We do not change the filing status of these tax units.
[32] Other criteria include living with the filer for more than half the year, depending on filer for at least half of his support, and not filing a joint return. For a full list of dependent criteria, see Tax Guide 2011, pg. 26. https://www.irs.gov/pub/irs-prior/p17--2011.pdf

secondary filer is treated as a dependent filer (who therefore does not receive the personal exemption). Since the 1040 returns in the limited tax data only identify at most four PIKed dependents (and a small number of 1040 dependents may not link to a PIK), we use the CPS to assign unattached individuals who qualify as dependents to tax units for which the number of dependent exemptions on the 1040 return exceeds the number of PIKed dependents merged with the Numident.[33] The purpose of this is to accurately estimate both the number of CTC-eligible children and the number of children eligible for the child and dependent care credit.

Specifically, we use the Numident data to identify the ages of dependents on a 1040 return (to help us accurately calculate relevant tax credits). Since we only use the number of dependents to estimate the CTC and the child and dependent care credit, we only assign unattached dependents who are less than 17 years of age, which is the age requirement for the CTC. We assign dependents until the total number of assigned child dependents and dependents to whom we attach a 1040 return equals the number of 1040 dependent exemptions (excluding parent exemptions) – i.e., the sum of the 1040 extract's number of exemptions claimed for the following mutually-exclusive dependent types: children living at home, children not living at home, children in excess of 9, other children on the return eligible for EITC, and dependents other than primary taxpayer, spouse, children. We first assign un-PIKed and unattached dependents in order of age.[34] If the total number of attached and assigned dependents falls short of the number of dependents, then we assign PIKed, unattached dependents in order of age until the total number of attached dependents and assigned dependents equals the total number of dependents.

Moreover, for a small number of survey individuals, we attach joint 1040 returns where either the primary or secondary filer does not appear in the CPS. We deal with four such cases:[35]

1. There is no spouse interviewed in the CPS, but a spouse is noted as absent in the CPS family. We use the 1040 return information to estimate tax liabilities and credits.
2. There is no spouse in the CPS, and the person to whom we attach the joint 1040 return is identified as unmarried in the CPS. We use half the amount of tax liabilities and credits calculated for the joint 1040 return.[36]

[33] We do not attach CPS dependents to whom we do not attach a 1040 return to a 1040 tax unit if the family contains a 1040 tax unit split across a primary family and subfamily.

[34] We start with un-PIKed individuals because we likely were not able to attach 1040 return to them because they were missing a PIK.

[35] 4.3% of attached tax units (using the limited tax data) have a primary or secondary filer missing in the survey (i.e., missing PIK in survey or not in survey frame).

3. There is a spouse present in the CPS who is not PIKed.[37] We assign the unPIKed spouse to the tax unit based on the 1040 return and use the 1040 return information to estimate tax liabilities and credits. In this case, we assume that the unPIKed spouse is the primary or secondary filer on the 1040 return.
4. There is a spouse present in the CPS who is PIKed but is not the spouse on the 1040 return. We use half the amount of tax liabilities and credits calculated for the 1040 return for the individual to whom this return is attached, and we categorize the other PIKed spouse as a single filer or head of household (depending on whether this spouse has dependents).

Thus, if a joint return attaches to only one survey spouse and one un-PIKed spouse is present, we assume they are the spouse on the joint return (0.99% of attached tax units). Only in this case will we assign a spouse to a tax unit as defined by a 1040 return. In all other cases, we do not assign a spouse to whom we cannot attach a 1040 return to a tax unit formed from a 1040 return.

When assigning un-attached dependents and spouses to tax units in the aforementioned cases, we rely on family members' relationships relative to the family head.[38] A caveat of this approach is that in the rare case when a spousal or dependent relationship is between individuals identified as "other relatives" or "nonfamily members," we are unable to identify the spousal or dependent relationship. In other words, we can only assign dependents and spouses to tax units formed from a 1040 return where the primary filer or secondary filer is the reference person.[39,40]

Among primary families to whom we attach to at least one 1040 tax return, 62% contain exactly one tax unit that files as married, filing jointly. varied in their filing status configurations – for example, only 15% of such families contain exactly one tax unit that files as married, filing jointly. Non-family householders and secondary individuals file often as single non-dependents,

[36] For example, if a spouse dies after a married couples taxes are filed, the spouse would appear on the 1040 tax return but may not appear in the survey.

[37] 4.9% of families for whom we attach a tax return contain at least one individual who is not PIKed.

[38] As a result, someone can be designated as a reference person (head of family), spouse, child, or other relative.

[39] If a family contains multiple tax units that attach to the reference person and spouse, then all remaining members of the family are treated each as their own tax units because we cannot identify which tax unit to which they belong.

[40] Since combined primary and related subfamilies will have two heads of families, we treat combined primary and related subfamilies slightly differently than all other families. For combined families, those with 1040 tax units split between the primary and related subfamily are formed into tax units at the combined family level, while families with 1040 tax units fully contained in either the primary or related subfamily are formed into tax units at the subfamily level. For example, suppose the primary family is composed of reference person A, spouse A, and child A, and the related subfamily is composed of reference person B, spouse B, and child B. If a 1040 tax form places reference person A, spouse A, and reference person B in the same tax unit, then 1) they will all form one tax unit, 2) spouse B and child B will form another tax unit, and 3) child A will form his/her own tax unit. For families other than combined primary and related subfamilies, all remaining individuals are each treated as having their own tax units.

while unrelated subfamilies often contain tax units for whom we attach a single 1040 return filing as a head of household.

*Remaining Individuals*

In order to construct tax units out of the remaining individuals, we also rely on family members' relationships to head of family. We start with primary families, assuming that any married couple among the remaining family members files a joint return. We calculate taxes assuming any dependents of theirs not on a 1040 form are claimed as their dependents. For the sake of simplicity, we assume that individuals who meet the dependent criterion but do not have incomes above the filing threshold do not file separate returns themselves. We then assume that any other related individuals are single filers.

*Differences Between Limited Tax Data and Extensive Tax Data*

For CPS individuals to whom we attach or assign a 1040 return, we follow the same methodology as with the limited tax data except for how we treat dependents. Since the extensive tax data provide information on all tax credits, we no longer need to estimate the number of dependents eligible for dependent-related tax credits by assigning additional dependents to CPS tax units defined by a 1040 return beyond those defined by the 1040 return. We then form survey tax units for remaining individuals, using the same methodology as used with the limited tax data.