The Accuracy of Tax Imputations: Estimating Tax Liabilities and Credits Using Linked Survey and Administrative Data

Bruce D. Meyer
University of Chicago, NBER, AEI, and U.S. Census Bureau

Derek Wu
University of Chicago

Grace Finley
University of Chicago

Patrick Langetieg
Internal Revenue Service

Carla Medalia
U.S. Census Bureau

Mark Payne
Internal Revenue Service

Alan Plumley
Internal Revenue Service

February 28, 2020

Abstract

This paper calculates accurate estimates of income and payroll taxes using a groundbreaking set of linked survey and administrative tax data that are part of the Comprehensive Income Dataset (CID). We compare our estimates to survey imputations produced by the Census Bureau and those generated using the TAXSIM calculator from the National Bureau of Economic Research. The administrative data include two sets of Internal Revenue Service (IRS) data: (1) a limited set of tax information for the population of individual income tax returns covering selected line items from Forms 1040, W-2, and 1099-R; and (2) an extensive set of population tax records processed by the IRS in 2011, covering nearly every line item on Form 1040 and most lines on a series of third-party information returns. We link these IRS records to the Current Population Survey Annual Social and Economic Supplement (CPS) for reference year 2010. We describe how we form tax units and estimate various types of tax liabilities and credits using these linked data, providing a roadmap for constructing accurate measures of taxes while preserving the survey family as the sharing unit for distributional analyses. We find that aggregate estimates of various tax components using the limited and extensive tax data estimates are close to each other and much closer to public IRS tabulations than either of the imputations using survey data alone. At the individual level, the absolute errors of survey-only imputations of federal income taxes and total taxes are on average 10% and 13%, respectively, of adjusted gross income. In contrast, the limited tax data imputations yield mean absolute errors for federal income taxes and total taxes that are about 2% and 3% of adjusted gross income, respectively. For the Earned Income Tax Credit, the limited tax data imputation is off by less than $20 on average for a typical family (compared to more than $500 using either of the survey-only imputations).

* This paper was prepared for the NBER-CRIW book, "Measuring and Understanding the Distribution and Intra/Inter-Generational Mobility of Income and Wealth." 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 other agency of the federal government, or the NBER. All results were approved for release by the U.S. Census Bureau, authorization numbers CBDRB-FY20-ERD002-014 and CBDRB-FY20-ERD002-038. We thank Matthew Stadnicki, Brian Curran, Alexa Grunwaldt, and Angela Wyse for excellent research assistance; Katie Genadek for expertly handling lengthy disclosure requests; Jim Davis and Maggie Jones for helping us access TAXSIM on the Census servers; and Dan Feenberg, Tom Hertz, Jonathan Rothbaum, David Splinter, Alex Yuskavage, and participants at the NBER-CRIW and NTA Conferences for helpful comments and discussions. We also appreciate the financial support of the Alfred P. Sloan Foundation, the Russell Sage Foundation, the Charles Koch Foundation, and the Menard Family Foundation.

1. Introduction

Accurately measuring tax liabilities and credits is important, as they play a pivotal role in our understanding of numerous research and policy topics. The Census Bureau has a long history of calculating features of the income distribution after accounting for taxes (see, e.g., U.S. Census Bureau 1988, 1993). In recent years, research has proliferated on topics ranging from the progressivity of the tax and transfer system in the United States (Piketty and Saez 2007, Congressional Budget Office 2018, Saez and Zucman 2019, Splinter 2019) and poverty rates after accounting for taxes and transfers (Burkhauser et al. 2019, Fox 2019, Meyer, Wu, and Medalia 2020) to the levels and trends in income inequality and distributional national accounts (Piketty, Saez, and Zucman 2018, Auten and Splinter 2019).

However, obtaining convenient and accurate measures of tax liabilities and credits can be difficult. In household surveys conducted by the U.S. Census Bureau, taxes are never explicitly asked about and must therefore be imputed. The imputation process makes many strong assumptions in forming tax units from family and household relationships and relies on income that is measured with substantial error in surveys. There are also shortcomings associated with administrative tax records from the Internal Revenue Service (IRS), even though these data may contain accurate measures of taxes. First, access to such data is often restricted and – even when access is granted – the data may not actually contain measures of taxes paid out and tax credits received. Furthermore, the IRS does not publish a single measure of net federal income tax liabilities (i.e., taxes owed net of non-refundable and refundable credits), which must be calculated from a combination of line items. IRS tax records also do not cover state income taxes, which must be imputed. Finally, the tax unit is not as natural a sharing unit for distributional analyses as the survey family.

In this paper, we use a groundbreaking set of linked survey and administrative tax data for reference year 2010 to calculate accurate estimates of federal and state income taxes and payroll taxes for families and unrelated individuals in the Current Population Survey Annual Social and Economic Supplement (CPS ASEC). We compare our estimates to the imputations from the Census tax calculator and those calculated using the National Bureau of Economic Research's TAXSIM program. Both of these calculators are extensively used by researchers, with the Census tax imputations constituting the default tax amounts on the public-use CPS ASEC file and the TAXSIM

calculator serving as perhaps the most well-known and convenient platform for simulating taxes.[1] Specifically, we use estimates from an extensive set of tax records (containing nearly every line item on a 1040) to assess the measurement error associated with imputations using survey data alone as well as imputations using survey data combined with a limited set of IRS tax records (containing certain 1040 line items but no tax liabilities or credits). This limited dataset is what the Census Bureau receives from the IRS under current regulations and maintains historically. Thus, we also assess the value to the U.S. Census Bureau (and potentially the broader research community) of more tax data being shared by the IRS.

We describe in detail how we form tax units and estimate commonly used measures of tax liabilities and credits using the linked survey and tax data. While we rely primarily on the TAXSIM calculator when using the limited tax data, we also modify TAXSIM in a number of ways to better utilize the advantages and accommodate the shortcomings of the data and TAXSIM. In doing so, we hope to provide a methodological roadmap for other researchers interested in using linked survey and administrative data to construct accurate measures of taxes while preserving the survey family as the sharing unit for distributional analyses. One should keep in mind that the incomes on the tax records are the incomes reported to the IRS, and therefore should not necessarily be considered the truth (as some individuals may not file tax returns and some filers may not accurately report their income, the tax benefits to which they are entitled, or even their filing status). But crucially, many of the variables we use are recorded consistently in taxpayer and business filings, and it is these reports that largely determine taxes ultimately paid.

While previous studies have compared the accuracy of various tax calculators using survey inputs to administrative targets at an aggregate level, this is one of the first studies to assess the quality of various tax calculators at the individual or family level. Jones and Ziliak (2020) uses linked survey and tax records to compare estimates of the Earned Income Tax Credit across different calculators, but no study – to the best of our knowledge – uses such linked microdata to evaluate the accuracy of a more comprehensive set of tax liabilities and credits. Another key contribution of this study is that it uses combined survey and administrative data to fill in holes associated with relying exclusively on values from the administrative 1040 data. Specifically, we bring in incomes from information returns including W-2s and 1099-Rs as well as survey reports (when information returns

[1] As of October 2020, Feenberg and Coutts (1993) – which provides an introduction to TAXSIM – has been cited nearly 1,000 times on Google Scholar.

are unavailable) to calculate taxes for those in the survey to whom a 1040 record cannot be attached. No administrative records can be linked to some of these individuals, as they are missing the Census identifiers that are necessary to link across data sources. In other cases, individuals may have had taxes withheld even if they did not file a 1040 or may have filed tax returns late.[2] Simulating taxes in this way also aligns our estimates better with aggregates published by the IRS Statistics of Income, which cover both taxes filed for the relevant tax years as well as taxes paid by some late filers.

We find that aggregate estimates of various tax components using the limited tax data imputations and extensive tax data calculations are close to each other and much closer to IRS SOI aggregates than any of the imputations using survey data alone. This pattern is particularly true for the EITC and Child Tax Credit. Looking at differences at the individual level, we find that the imputations of federal income tax liability and total tax liability using survey data alone have mean absolute errors of 10% and 13%, respectively, of average adjusted gross income, where we take the calculations based on the extensive tax data to be the benchmark. A key reason for these errors is that the mean absolute difference in the survey calculation of adjusted gross income is 44% of the mean amount from the extensive tax data. Moreover, looking at the bottom quartile of survey-reported income, the mean absolute errors in the survey calculations of the EITC alone have a magnitude of 5% of adjusted gross income.

In contrast, the limited tax data imputations have 22-23% of the absolute errors of the survey imputations for federal income tax liability and 19-20% of the absolute errors of the survey imputations for total tax liability, taking the extensive tax data calculations as the benchmark (the range is due to the choice of which calculator is used). The improvement in tax calculations using the limited tax data is especially noticeable in the top half of the survey income distribution. For a typical family, we also find that the limited tax data imputation is off by less than $20 for the EITC (compared to $550 and $569 using the Census and TAXSIM survey imputations, respectively) and by less than $50 for the Child Tax Credit (compared to $275 and $306 using the Census and TAXSIM survey imputations). All statements in the text are statistically significant at the 10% significance level, unless otherwise noted.

The remainder of the paper is structured as follows. Section 2 describes the specific types of income and payroll taxes that we estimate. Section 3 discusses the different tax simulation models

[2] Langetieg, Payne, and Plumley (2017) find that late filers (defined as those filing within two years of the end of the tax year) constituted 4.5% of total required tax filers during the 2010 tax year.

used in the literature to calculate taxes. Section 4 describes the survey and administrative data and how they are linked. Section 5 discusses how we construct survey tax units in the linked sample and use inputs from the survey data and/or administrative tax records to calculate tax liabilities and credits. Section 6 presents results comparing tax estimates obtained using various tax calculators, both in aggregate and at an individual level across the income distribution. Section 7 concludes.

2. Background on Taxes

In this paper, we estimate several types of income tax liabilities and credits. Specifically, we estimate federal income taxes, state income taxes, and payroll taxes. With regard to tax credits, we principally estimate the Earned Income Tax Credit (EITC) and Child Tax Credit (CTC) – which are the two largest tax credits commonly calculated in tax simulation models – and bring in other tax credits where we can. Our goal is to calculate what taxes should be deducted from an individual's total income to obtain income available for consumption. As a result, we seek to calculate the tax liability that an individual accrued in a given calendar year, regardless of whether or not the individual paid any taxes or received any credits.

Federal Income Tax Liabilities and Credits

We define federal income taxes as total federal income taxes owed net of non-refundable and refundable federal tax credits. Federal income taxes can therefore be negative or positive, depending on whether or not federal income tax liabilities before credits exceed federal tax credits. The main components of federal income tax liabilities before credits are the tax directly calculated on taxable income and the alternative minimum tax; where we can, we also bring in smaller taxes like the additional taxes on Individual Retirement Accounts (IRAs) and other qualified retirement plans. Non-refundable tax credits are capped in the aggregate at the income tax amount before credits, while refundable tax credits can exceed the income tax amount before credits.

One of the most prominent refundable federal tax credits is the EITC, which is a benefit for families and individuals with positive but low earnings. For a family with two children, the maximum EITC amount in 2010 was $5,036. Eligibility for the EITC depends on several factors. First, eligible tax units must have members with valid Social Security Numbers (SSNs). Second, only those with positive earnings can receive the credit. The EITC initially increases proportionally with earnings, then remains at a maximum level for a range of earned income, and finally decreases

proportionally with additional earnings until it is completely phased out. Third, filers must have an adjusted gross income (AGI) below a given threshold to be eligible, with a higher threshold for joint (married) filers than single (unmarried) filers.[3] Lastly, the amount of the credit increases with the number of qualifying children in the tax unit.[4]

Like the EITC, the CTC is also a tax credit for families with positive but low earnings. The maximum CTC amount that a family could receive in tax year 2010 was $1,000 per eligible child, with this amount decreasing with AGI over $75,000 (for a single filer) and $110,000 (for a joint filer) until completely phasing out. Unlike the EITC, the CTC has both a non-refundable and refundable portion. The non-refundable portion of the CTC is capped at the federal income tax liability before credits. Meanwhile, the refundable portion – also known as the Additional Child Tax Credit (ACTC) – is capped at the remaining difference between the maximum CTC amount and the non-refundable portion of the CTC. Tax units earning less than $3,000 are generally ineligible for the refundable portion, as it is phased-in proportionally with earnings above the $3,000 threshold.[5]

Families and individuals can also claim a wide variety of non-refundable and refundable tax credits outside of the EITC and CTC. Other important non-refundable credits include: the foreign tax credit (for individuals who also pay income taxes to a foreign government), education credits (for families with qualifying students at institutions of higher education), and the child and dependent care credit (for families with certain expenses to care for qualifying children or other dependents). Other important refundable credits include: the Making Work Pay Credit (temporarily created as part of the American Recovery and Reinvestment Act of 2009 and providing a maximum of $400 or $800 for single and married earners, respectively) and the American Opportunity Tax Credit (also created as part of the 2009 Recovery Act to help cover qualified post-secondary education expenses).

State Income Tax Liabilities and Credits

State income taxes are administered by state governments and vary by state. Most states feature a progressive tax system like the federal income tax system, where higher levels of income are taxed at higher rates than lower levels. However, some states – including Colorado,

[3] In 2010, this threshold was $45,373 for a family with two qualifying children with the parents filing jointly.
[4] The definition of an EITC-qualifying child can be found on p. 46 of the IRS 1040 Tax Guide for 2010, and – in the cases when we simulate the number of EITC-eligible children from survey information only – we follow the rules as closely as we can. See https://www.irs.gov/pub/irs-prior/i1040gi--2010.pdf.
[5] The definition of a CTC-qualifying child can be found on p. 40 of the IRS 1040 Tax Guide for 2010. Once again, we follow the rules as closely as we can when calculating the number of children eligible for the CTC.

Massachusetts, and Michigan – impose a flat rate rather than a progressive tax rate. A few states – including Alaska, Florida, Texas, and Washington – do not have a state income tax. For states that do have a progressive tax system, the number of brackets and the marginal tax rates vary significantly (e.g., Hawaii had twelve brackets while Kansas only had three in 2010). Hawaii had the highest marginal state income tax rate in 2010, with a rate of 11% on incomes over $200,000. Nonetheless, even the highest marginal state income tax rate was only a fraction of the highest marginal federal income tax rate in 2010, which was 35% on incomes over $373,650 for both single and joint filers.

Additionally, states vary by the type of income that they tax. In most states, taxable income follows the federal definition of taxable income. However, some states only tax certain types of income (e.g., New Hampshire taxes interest and dividends but not earned income). Some states also allow federal income tax liabilities to be deducted from taxable income, while others do not. Lastly, while most states offer tax credits, the specific types of credits vary substantially by state. Twenty-one states (plus the District of Columbia) had enacted a state EITC program by 2010, with the state EITC amounts ranging from 5% to 85% of the federal EITC amount. Fewer than 10 states offered a state CTC program, with state amounts being only a fraction of the federal program amounts and state eligibility requirements sometimes being stricter than the federal requirements. Because of the variation in state tax laws, estimating state income tax liabilities can be more complex than estimating federal income tax liabilities. However, most tax calculators take this state-level variation into account.

Payroll Taxes

Payroll taxes fund the Social Security Administration's Old Age, Survivors, and Disability Insurance (OASDI) program and the Medicare program. Payroll taxes on wages and salaries – known as the Federal Insurance Contributions Act (FICA) taxes – are paid equally by employees and employers. The employee half is generally withheld by employers (and reported on a W-2) and must be paid regardless of whether or not the employee files a tax return, while the employer half is paid by the employer directly. Because our goal is to calculate what must be subtracted from gross income to obtain disposable income, we do not include taxes paid by the employer. In contrast, individuals pay payroll taxes on self-employment earnings – otherwise known as the Self-Employment Contributions Act (SECA) taxes – throughout the year and reconcile those payments

with the final liability directly on a 1040 when they file their federal income tax returns. For wage/salary earnings, the employee portion of payroll taxes in 2010 was 6.2% for OASDI taxes and 1.45% for Medicare taxes. Self-employed workers pay both the employer and employee portions of the payroll tax, amounting to 12.4% of self-employment earnings for OASDI taxes and 2.9% for Medicare taxes.[6] OASDI taxes are paid on only the first $106,800 of earnings, while Medicare taxes are uncapped.

3. Literature on Tax Calculators

Tax simulation models have been developed by several research organizations to impute federal and state income taxes and payroll taxes from income and demographic inputs. These calculators are necessary and useful for a number of reasons. First, taxes are never directly reported on any Census Bureau household surveys. Second, it can be difficult to access administrative microdata containing information on taxes paid and tax credits received. Third, even in datasets like the Public Use Tax File (PUF) from the Internal Revenue Service's Statistics of Income Division, state income taxes and payroll taxes are often not available. In this section, we discuss several tax calculators that have been frequently used by researchers to calculate taxes.

Description of Available Tax Calculators

One of the most widely used tax simulation models is the TAXSIM calculator from the National Bureau of Economic Research (Feenberg and Coutts 1993). Numerous studies have used the TAXSIM calculator to simulate taxes from various types of microdata, including survey data and samples of IRS tax returns (see, e.g., Gruber and Saez 2002, Meyer and Sullivan 2003, Piketty and Saez 2007, Dahl and Lochner 2012). The latest version of TAXSIM is designed to calculate taxes based on 27 variables provided by the user, and its outputs include – among other measures – federal income tax liability, state income tax liability, payroll taxes, and a variety of credits (including federal and state EITC amounts, the non-refundable and refundable portions of the CTC, and the child and dependent care credit). The TAXSIM model can be used to calculate taxes for all tax units, regardless of whether or not they are required to file tax returns. TAXSIM is also updated on an

[6] However, self-employed workers can deduct one-half of their total self-employment tax from their taxable income when calculating federal income tax.

annual basis to take into account changes in tax rules at the federal and state level, but it does not allow for the calculation of local income taxes.

The Bakija Income Tax Calculator – developed by Jon Bakija (Bakija 2014) – is another tax simulation model that researchers have used in recent years (see, e.g., Giertz 2007, Hoynes and Luttmer 2011, Meyer and Sullivan 2012, Jones and Ziliak 2020). Like TAXSIM, the Bakija model allows for the calculation of federal and state income taxes, payroll taxes, and a variety of credits (including federal and state EITC amounts, the non-refundable and refundable portions of the CTC, and the child and dependent care credit). The Bakija model, like TAXSIM, can be used to calculate taxes for all tax units, regardless of whether or not they are required to file taxes. And, like TAXSIM, the Bakija model is updated yearly to account for changes in federal and state income tax rules. Finally, both TAXSIM and the Bakija model are sufficiently flexible that they can run on data from any source, so long as the necessary inputs are provided.

However, there are some differences between the Bakija model and TAXSIM. First, the Bakija model allows more detailed inputs by calculating taxes based on 70 variables provided by the user (compared to 27 input variables for TAXSIM). For example, while TAXSIM tends to input income sources in a more aggregate form (e.g., total deductions, property income covering interest/rent/alimony/etc.) and sums them across the primary and secondary filers in a married tax unit, the Bakija model allows the input of disaggregated income sources within several categories and also by primary and secondary filer in a married tax unit. The Bakija model also calculates federal and state income taxes going back to 1913 and 1900, respectively, while TAXSIM only calculates federal and state income taxes going back to 1960 and 1977, respectively. On the other hand, TAXSIM has the capability to model tax rules for dependent filers, while the Bakija model currently does not.

The Census Bureau has also developed a tax model used to calculate tax liabilities and credits using inputs from the CPS ASEC and the IRS Statistics of Income PUF (O'Hara 2004, Webster 2011). The PUF is used through a statistical match with the CPS ASEC to impute certain variables, including itemized deductions and (until recently) capital gains, which the CPS does not ask about.[7] Like TAXSIM and the Bakija model, the Census model outputs federal and state income taxes, payroll taxes, and various tax credits (including the EITC, the non-refundable and refundable

[7] Specifically, while the Census tax model accounted for capital gains during reference year 2010 (our time period of interest), more recent changes to the model have resulted in capital gains being omitted in the model's current iteration.

portions of the CTC, and the child and dependent care credit). Unlike TAXSIM and the Bakija model, however, the Census model designates some tax units as nonfilers.

Finally, the Urban Institute has developed and maintained the Transfer Income Model Version 3 (TRIM3), a microsimulation model that – among various roles – calculates federal and state income taxes and payroll taxes (Zedlewski and Giannarelli 2015). Like the Census tax model, TRIM3 is specifically designed to run on the CPS ASEC by defining tax units, dependents and qualifying children for various tax credits, and income items using that survey. TRIM3, like the Census model, also relies on a statistical match with the IRS Statistics of Income PUF to impute itemized deductions and capital gains. Unlike the Census model, TRIM3 calculates taxes for all tax units, regardless of whether or not they are required to file taxes.

Comparisons of Tax Calculators

To the best of our knowledge, Wheaton and Stevens (2016) is the only study to have compared several outcomes of each of the tax simulation models described above. Using the CPS ASEC, they start by comparing income tax liabilities and credits calculated using TRIM3 to those calculated using the Census tax model, relative to administrative targets from the IRS Statistics of Income line item totals. They also compare the outputs of the TAXSIM and Bakija models using inputs generated separately by TRIM3 and the Census model. Although the Census tax model by default assigns some units to be nonfilers, Wheaton and Stevens simulate taxes for all units using the Census tax model regardless of whether they would actually file (following the approach used by TAXSIM, the Bakija model, and TRIM3).

Wheaton and Stevens find that TRIM3 captures a higher share of AGI and taxable income than the Census model, primarily because TRIM3 incorporates capital gains (based on a statistical match from the IRS PUF) while the Census model no longer captures capital gains. However, both models yield estimates of AGI and taxable income that fall short of administrative targets by 5-15%. When it comes to federal income tax liabilities, TRIM3 generally yields estimates that are closer to administrative targets throughout the income distribution than the Census tax model. For example, both TRIM3 and the Census tax model produce federal income tax estimates that are above administrative targets for tax units with AGIs between $10,000 and $200,000, with the Census model estimates generally being higher than TRIM3 estimates. The authors attribute this discrepancy to the Census model capturing fewer itemized deductions than TRIM3 in this portion of the income

distribution. Furthermore, TRIM3 and the Census tax model produce federal income tax estimates that are below administrative targets for tax units with AGI above $200,000, with Census model estimates generally being lower than TRIM3 estimates. The authors ascribe this difference to the Census model not accounting for capital gains.

When it comes to federal tax credits, estimates of the EITC obtained using TRIM3 continue to be closer to administrative targets than those calculated using the Census tax model. The difference in EITC estimates between TRIM3 and the Census tax model is partly due to TRIM3 identifying more EITC-qualifying children. Specifically, the Census model identifies EITC-eligible children as children under the age of 19 or between the ages of 19 and 23 and enrolled in school, while TRIM3 further accounts for adult disabled children, married children, and other relatives who meet the qualifying requirements for the EITC. TRIM3 also produces total state income tax estimates that are also closer to administrative targets (compiled from the Census of Governments) than the Census model, and estimates of state-level EITC amounts using TRIM3 are also closer to administrative targets than those obtained using the Census model.

Wheaton and Stevens also find that the TAXSIM and Bakija models yield tax estimates that are very similar to TRIM3 when inputs are defined using TRIM3 and very similar to the Census model when inputs are defined using the Census model. With regard to federal income tax liabilities, the TAXSIM and Bakija models yield similar estimates (conditional on inputs) throughout the entire income distribution except at the very bottom (i.e., for tax units with AGI amounts below $10,000). This difference comes from the Bakija model not accounting for dependent filers, most of whom have AGI amounts below $10,000. The TAXSIM and Bakija models also yield similar values to each other for tax credits and total state income taxes. However, there is some variation between these models when looking at state income taxes for individual states, with this difference being indicative of the complexities associated with state income tax modeling.

4. 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 CPS ASEC.[8] We also rely on two

[8] Technical documentation (CPS Technical Papers 66 and 77) with full discussion of CPS and CPS ASEC methodology can be downloaded from https://www.census.gov/programs-surveys/cps/technical-documentation/complete.2011.html.

IRS administrative datasets for the population: one with limited 1040 return information and another with extensive 1040 return information that includes all 1040 line items and most items on third-party information documents sent to the IRS. We focus on reference year 2010, since this is the only year for which the extensive tax data are available.

Survey Data

The CPS ASEC (hereafter referred to as CPS) collects demographic and income information for households representing the civilian non-institutionalized U.S. population. The 2011 CPS interviewed approximately 75,000 households (including approximately 90,000 families) between February and April 2011 about their annual incomes 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 tax inputs. These items include federal and state income tax, payroll tax, the Earned Income Tax Credit (EITC), Child Tax Credit (CTC), and adjusted gross income (AGI). While the CPS is a household-level survey, our unit of analysis is a family – defined as either a group of two or more related individuals residing together or an unrelated individual. The family is generally thought to better approximate the unit that shares resources and is the unit for the official poverty measure.

Administrative Tax Records

We rely on two different datasets provided by the Internal Revenue Service (IRS). The first dataset, hereafter referred to as the "limited tax data," is provided to the U.S. Census Bureau under Section 6103(j) of Title 26 of the U.S. Code, which allows the U.S. Census Bureau to use IRS tax data for Census survey improvement. The second dataset, hereafter referred to as the "extensive tax data," is provided to the U.S. Census Bureau under Section 6103(n) of Title 26, which gives access to the data for the purpose of tax administration. We also bring in information on taxable self-employment income from the Detailed Earnings Record of the Social Security Administration (SSA).

Limited Tax Data

The limited tax data include line items extracted from the universe of Forms 1040, W-2, and 1099-R submitted during the 2011 calendar year, primarily for the 2010 tax year. We start by

describing the information available in the limited 1040 data. These data first contain a set of variables covering tax unit structure, including an identifier assigned by the Census Bureau to every tax return, a variable identifying the type of 1040 form (e.g., 1040-EZ, 1040-NR, etc.), filing status, and the various types of exemptions claimed. We also have reported income amounts for the following 1040 line items: wages/salaries, taxable dividends, taxable and tax-exempt interest, gross rents and royalties, Social Security income, adjusted gross income (AGI), and total money income. Total money income incorporates most sources of income reported in the "income" section of the 1040 form (lines 7-21 of the 1040 form for tax year 2010), but it misses several key income sources – including capital gains – and is therefore not identical to total income on a 1040 (line 22 of the 1040 form for tax year 2010).

The limited 1040 data also include separate indicators each for whether a tax unit filed Schedule A (itemized deductions), Schedule C (profits or losses from self employment), Schedule D (capital gains and losses), Schedule E (supplemental incomes and losses, including rents and royalties), Schedule F (profits or losses from farming), and Schedule SE (self-employment taxes). Furthermore, the limited 1040 data include reported EITC earned income and the number of EITC-eligible children claimed (ranging from zero to three) for tax units that claim the EITC.[9] Finally, the limited 1040 data include a variable for posting date, which indicates the week that a 1040 return was posted to the IRS Individual Master File. As for the information returns in the limited tax data, the W-2 data contain amounts for taxable wages/salaries and deferred compensation associated with each employee-employer combination (with employers identified by their employer identification numbers – EINs), while the 1099-R data contain amounts of gross retirement distributions from both employer-sponsored plans and IRA withdrawals.

Despite including all these items, there are several weaknesses associated with the limited tax data. First, even though the limited 1040 data cover enough line items to calculate tax liabilities and credits relatively accurately (especially for those low in the income distribution), they do not contain actual amounts for federal income taxes paid or tax credits received. The limited 1040 data also miss key line items, such as capital gains/losses and itemized deductions, that are necessary to generate accurate estimates of federal income tax liabilities for those high in the income distribution. Moreover, while the limited 1040 data indicate whether or not a tax unit filed a given Schedule, we

[9] These variables, along with filing status and AGI, allow us to calculate the exact amount of EITC received by tax units claiming the credit (not simply those tax units that the IRS thinks are eligible).

do not receive information on the contents of that Schedule. The limited tax data also do not cover information returns outside of Forms W-2 and 1099-R. Of particular interest to us, among the missing returns, are Forms 1099-G (covering unemployment compensation), 1099-MISC (covering self-employment earnings for independent contractors), and Schedule K-1 information returns (covering partnership earnings).

Extensive Tax Data

The extensive tax data comprise a set of over fifty data files containing information for nearly every line item corresponding to the universe of 1040 forms (along with the accompanying Schedules) and a wide set of information returns submitted for the 2010 tax year. The information returns in the extensive data pertain to only tax year 2010, while the 1040 forms include forms filed during calendar year 2011 for tax year 2010 as well as prior tax years. Crucially, the extensive 1040 data contain amounts reported for federal income tax liabilities, various tax credits (such as the EITC), and various deductions. The extensive 1040 data also contain two versions of nearly every line item on the 1040 – one containing raw values corresponding to what was filed and one containing computer-corrected values that correct for obvious errors (e.g., missing decimal point, too many zeros, etc.) but not for noncompliance. Wherever possible, we use the computer-corrected version of a variable.

With regard to information returns, the extensive tax data contain almost every line item for a large set of forms, including – but not limited to – Forms W-2, 1099-R, 1099-G, 1099-MISC, and Schedule K-1. A caveat associated with the information returns in the extensive tax data is that "payers" – such as employers on W-2s – are identified only by their five-digit zip codes (rather than by EINs, for example). Therefore, distinguishing between unique jobs in the extensive W-2 data may be more difficult for individuals working in multiple jobs within the same five-digit zip code. We discuss these issues in greater depth below.

Detailed Earnings Record

Our final source of administrative tax data is the Detailed Earnings Record (DER) database of the SSA. The DER contains wage and salary earnings derived from IRS W-2 Forms and a measure of self-employment earnings (namely, the Medicare-taxable portion of total net self-

employment earnings) derived from Schedule SE of Form 1040.[10] A caveat regarding the DER is that it contains records only for individuals with valid SSNs and therefore misses W-2s or 1040s filed using employees' Individual Taxpayer Identification Numbers (ITINs). In this paper, we primarily rely on the DER to calculate the self-employment portion of payroll taxes when we use the limited tax data, as those data do not contain any amounts corresponding to self-employment income.

Attaching Tax Records 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) (Wagner and Layne 2014). Approximately 92% of families in the CPS contain at least one member linked to a PIK. We limit our CPS sample to include survey families with at least one PIKed member and no individuals with entirely imputed income (often called whole imputes). These restrictions result in a sample size of 69,000 families containing 170,000 individuals. To correct for the bias arising from non-random missing PIKs and whole imputations, we divide individual survey weights by the predicted probability that at least one member of the family has a PIK and no member is a whole impute (conditional on observables in the survey). Doing so allows us to approximately match population totals using the re-weighted sample.

Almost all administrative tax records are linked to PIKs. In the limited tax data, PIKs for the primary filer, secondary filer, and up to four dependents are available for each 1040 return. In the extensive tax data, PIKs are available for the primary filer, secondary filer, and all dependents (uncapped) for each 1040 return. Nearly all of the information returns in the limited and extensive tax data also contain a PIK corresponding to the individual receiving the income relevant to that return (e.g., wage/salary earnings on a W-2, retirement income on a 1099-R, etc.). It is worth noting that we do not adjust for missing PIKs in the 1040s and other tax records, given that there are insufficient demographics in the tax data and it is not clear how one would reweight those data to account for missing PIKs. Consequently, we will slightly understate income amounts attached from the administrative tax data to the survey frame.

[10] Medicare-taxable self-employment earnings are defined as 92.35% of total net self-employment income minus health insurance costs.

Attaching 1040 Records to Survey Records

Because an individual should appear as a primary or secondary filer on only one tax return, we attach only one 1040 return to each survey individual.[11] However, there are cases when an individual may appear as a primary or secondary filer on multiple 1040 forms (e.g., when an individual has amended or corrected returns). If two or more 1040 forms can be attached to one survey individual, then we keep the return with the latest posting date, a filing status of married, filing jointly (if the returns have the same posting date), or higher AGI amount (if the returns have the same posting date and filing status).[12] When attaching 1040 returns to survey individuals, we do not bring in tax returns where only the dependent (and not the primary and/or secondary filers) appears in the CPS. We are able to attach at least one 1040 return to over 88% of CPS families in our sample, with 17% of families having multiple 1040 returns attached.[13]

Attaching Information Returns to Survey Records

Individuals may also have multiple valid forms for a given information return (unlike for a 1040). For example, a person may receive two W-2s (one from each employer) if she works two jobs. In this case, we attach both W-2 forms to that individual since each form represents a separate income stream. In rare cases, an employer may even file multiple valid W-2s for an employee. In the limited W-2 data (which contain EINs), we keep all W-2s pertaining to a "job" (corresponding to a PIK-EIN combination) if every return associated with that "job" is designated as an "original" return. When a "job" contains amended or corrected returns, we keep only the amended or corrected form.[14] For 1099-Rs, the limited tax data do not contain any 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.

[11] In the case of married individuals filing separately, an individual may appear on two returns. In this case, we still attach only one return to the individual because income and tax amounts on such returns only accrue to the filer and not the spouse.

[12] If a dependent appears on multiple forms, then we follow a similar methodology to keep only one 1040 form per dependent. While it is not legal for multiple people to claim the same dependent, it is reasonable to assume that some people do (e.g., divorced parents claiming the same child).

[13] Note that the percentages reported here – and in the rest of Sections 4 and 5 – are un-weighted.

[14] 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.

In the extensive tax data, we de-duplicate information returns slightly differently because "payers" are identified only by their five-digit zip codes rather than actual identification numbers. The presence of these payer zip codes and amendment codes across the information returns means that we can now de-duplicate all information returns (not just W-2s) in the extensive tax data. When a PIK-payer zip code combination has a corrected or amended form, we keep that form and drop at least one other original form (if one exists). In the case of the extensive W-2 data, if a PIK-payer zip code combination with a corrected/amended form contains either multiple corrected/amended forms or multiple original forms, then we use the EINs from the limited tax data to identify whether these returns correspond to a single employer or multiple employers within the same zip code.

5. Methods

In this section, we discuss how we construct survey tax units in the linked CPS sample and use inputs from the survey data and/or administrative tax records to calculate tax liabilities and credits. We examine four different ways of calculating tax liabilities and credits using the linked CPS sample. In the first approach, we use taxes calculated using the Census tax model relying only on CPS information. In the second approach, we continue to rely solely on CPS information but use TAXSIM to impute tax liabilities and credits from survey inputs, using information on filing status generated by the Census tax model.[15] In the third approach, we combine information from the CPS and limited tax data to form tax units and impute tax liabilities and credits using TAXSIM. In the fourth approach, we combine information from the CPS and extensive tax data to form tax units and, when needed, impute tax liabilities and credits using TAXSIM. Note that the first and second approaches represent two ways of imputing taxes using survey inputs only, and the third approach reflects our best attempt at simulating taxes using inputs from a combination of limited tax records and survey data.[16] The fourth approach represents the "gold standard" calculation, as it pulls tax liabilities and credits directly from an extensive set of 1040 line items whenever possible.

[15] We continue to rely on filing statuses generated by the Census tax model even when imputing tax liabilities using the TAXSIM calculator, as this allows for a more direct comparison of the tax calculators holding the inputs (including tax unit structure) constant. This also follows the methodology of Wheaton and Stevens (2016). In future work where our goal is no longer to specifically compare the calculators, we may consider using household and family relationships from the survey to construct tax units ourselves.

[16] While we refer to these calculations as "imputations" throughout the paper, one should think of them more as simulations based on survey reports and less as probabilistic imputations like hotdecking. While the Census tax model does rely on probabilistic statistical matches in some cases (e.g., capital gains, itemized deductions), many of these

Before laying out the methodology in detail, it is worth clarifying what we seek to calculate – specifically what should be deducted from an individual's observed income to obtain income available for consumption. The target for our tax calculations consists of the tax liability that an individual accrued in a given calendar year, regardless of whether or not the individual paid any taxes or received any credits. An implication of this conceptualization of taxes is that an individual may have a tax liability for tax year 2010 even if a 1040 filed in calendar year 2011 could not be attached to her survey record – specifically, linkage issues aside, an individual may have had taxes withheld even if they did not file a 1040 or may have filed taxes late. One implication of this goal is that the incidence of payroll taxes does not matter for the calculation of after-tax income. Our goal is just to calculate what is left over after paying taxes, and incidence would be reflected in pre-tax incomes. Of course, incidence does affect what income would be in the absence of taxes.

Forming Survey Tax Units

Using Survey Information Only

When calculating taxes using only CPS information (the first two approaches above), we follow the Census tax model's process for assigning survey individuals to tax units based on marital status and household relationship. Every survey tax unit has a primary filer. If the filing status of the survey tax unit is married filing jointly, then there is also a secondary filer. Specifically, married couples in primary families are assigned a filing status of married filing jointly, and unmarried family heads in primary families are assigned a filing status of head of household if they can claim a dependent (and a filing status of single otherwise). Dependents are assigned to family heads if they are either children under the age of 19, children under the age of 24 and simultaneously enrolled in school, or other relatives with incomes below the filing threshold ($3,650 in 2010).[17] All other individuals are labeled as single filers. Taxes are calculated only for units who meet the filing requirement based on the following criteria: having total income exceeding the filing threshold,

matches are likely relevant only for the high end of the income distribution and would therefore not affect the vast majority of tax units.

[17] Previous studies – including Jones and O'Hara (2016) and Splinter, Larrimore, and Mortenson (2017) – have shown that dependents are sometimes strategically reassigned within multi-family households to minimize tax liabilities or maximize EITC benefits. These reassignments may serve as a channel through which survey imputations of the EITC – which assign dependents based on family structure – might be understated relative to administrative values. Another related reason for why survey values of the EITC might be understated is that the EITC-qualifying dependents during a given tax year may not appear in a family or household when it is subsequently interviewed for the CPS ASEC.

being eligible for the EITC, having self-employment income above $400, or having negative gross or self-employment income. Tax units that do not meet these requirements would not have a tax liability and are treated as not having a reason to file.

Using Survey Information and Tax Records

When using a combination of survey information and tax records, we create survey tax units in one of two ways. 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. Second, for all individuals to whom we cannot attach a 1040 return, we use survey information on family relationship to assign them to survey tax units. Note that we use survey reports to calculate taxes for individuals to whom we cannot attach a 1040 return, as they may have had taxes withheld and/or paid taxes late. We thus assign everyone in our CPS sample to a survey tax unit with the role of primary filer, secondary filer, or dependent. Our methodology for forming survey tax units generally holds when attaching forms from both the limited and extensive tax data, though we discuss small differences between the data sources below.

For survey individuals to whom we can attach a 1040 return, we identify their roles as primary filers, secondary filers, and dependents based on their status on the 1040 returns. We designate a dependent filer as a dependent to whom we can also attach a separate 1040 return on which he/she is a primary or secondary filer. Because the limited 1040 data identify at most four PIKed dependents and a small number of 1040 dependents may also not link to a PIK, we use the CPS to assign un-attached individuals qualifying as dependents (based on survey characteristics) to tax units for which the number of 1040 dependent exemptions exceeds the number of PIKed dependents.[18] We undertake this process because the limited 1040 data do not contain amounts or even the number of qualifying children for certain child-related tax credits, such as the CTC and the child and dependent care credit. We therefore manually calculate the number of qualifying children for these credits by first identifying all dependents and then linking birth dates from the SSA Numident file to calculate the number of dependents with ages falling below the specific thresholds for relevant credits (age 17 for the CTC and age 13 for the child and dependent care credit). We assign un-attached dependents in this way only when attaching the limited tax records and not the extensive tax records, as the extensive 1040 data provide amounts for all tax credits.

[18] We do this for only 1% of tax units, because in 99% of tax units the number of dependent exemptions equals the number of PIKed dependents.

Moreover, for a small number of survey individuals (in 4.3% of tax units), we attach joint 1040 returns even though one of the primary or secondary filers on those returns does not appear in the CPS. We deal with four such cases:

1. No spouse is recorded in the CPS family, but a spouse is noted as absent in the CPS. We use the full amount of tax liabilities and credits calculated for the joint 1040 return. This case covers 0.76% of tax units.
2. No spouse is recorded in the CPS family, and the individual 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. This case covers 1.54% of tax units.
3. An un-PIKed spouse is recorded in the CPS family. We assume that the un-PIKed spouse is the primary or secondary filer on the 1040 return and use the full amount of tax liabilities and credits calculated for the joint 1040 return. This case covers 0.99% of tax units.
4. There is a PIKed spouse interviewed in the CPS family who is not the other spouse on the joint 1040 return. We assign half the amount of tax liabilities and credits calculated for the joint 1040 return to the individual to whom this return is attached, and we classify the other PIKed spouse as a single filer or head of household (depending on whether he/she has dependents). This case covers 1.03% of tax units.

In a very small number of cases, a 1040 return attaches to multiple primary and secondary filers split across different families within a household. Because we cannot determine to which family we should attach the 1040 return, we do not bring in the 1040 return information in these extremely rare cases. However, we continue to calculate taxes using dependents listed on a 1040 return even if they do not appear in the survey family (as they may have moved away for college, been reassigned, etc.).

Among primary families to whom we attach at least one 1040 return, 62% contain exactly one tax unit that files as married, filing jointly. Primary families with related subfamilies are highly 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, while unrelated subfamilies often contain tax units for whom we attach a single 1040 return filing as a head of household.

To construct tax units out of the remaining individuals to whom we cannot attach or assign a 1040 return, we rely on family members' relationships to the family head. 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.

Calculating Tax Liabilities and Credits

Using Census Tax Model and Survey Data

For our first method of calculating tax liabilities and credits, we rely on estimates produced by the Census tax calculator using CPS data statistically matched to the IRS Statistics of Income PUF. The Census tax calculator provides a number of tax outcomes on the CPS file, including federal and state income taxes (before and after credits), payroll taxes, various tax credits (including the EITC and the CTC), and some intermediate outcomes (including AGI and taxable income). Many of the inputs into the imputation process come directly from demographic and income information on the CPS. However, the Census tax model relies on a statistical match to the PUF to impute certain tax items missing from the CPS, including itemized deductions, IRA contributions, and various self-employment retirement deductions and health insurance premiums.[19] For further information on the Census tax model, see O'Hara (2004) and Webster (2011).

Using TAXSIM and Survey Data

In our next method, we calculate tax liabilities and credits by once again relying on CPS information but this time inputting survey income and other variables into TAXSIM. We continue to generate tax units based on filing statuses used in the Census tax model, but – unlike the Census tax model – we do not bring in the PUF to fill in tax items that are missing in the CPS. Instead, we rely only on the demographic and income variables available in the CPS to construct each of the TAXSIM inputs. As a result, certain inputs that are not asked about in the CPS – namely capital gains, itemized deductions, and other deductions factoring into AGI – take values of zero using this

[19] The statistical match between the CPS and PUF relies on the following variables: income, filing status, presence of earned income, presence of self-employment income, presence of unearned income, presence of Social Security income, presence of mortgage, presence of pension income, number of child exemptions, state of residence, and whether or not a person is a dependent on another return.

approach.[20] This decision likely results in a more substantial bias in tax liabilities for higher-income units – for whom these missing tax items are more relevant – than lower-income units.

We rely on the TAXSIM output for federal and state income tax liabilities and credits, but we calculate payroll taxes manually (i.e., outside of TAXSIM) using survey-reported amounts for wages and self-employment earnings. We modify TAXSIM because it does not distinguish between wage/salary and self-employment earnings when calculating payroll taxes, even though wage/salary employees pay only the employee portion of the payroll tax while self-employed individuals pay the entire amount of the payroll tax.[21] We follow the rules on withholding limits for tax year 2010 by setting the maximum level of earnings subject to the Social Security tax at $106,800, while placing no cap on the maximum level of earnings subject to the Medicare tax.

Combining Survey and Limited Tax Data

We now discuss how we calculate tax liabilities and credits using a combination of survey data and limited tax records entered into TAXSIM. We must simulate taxes ourselves because the limited 1040 and W-2 data do not contain amounts for federal income and payroll taxes, and we also do not have amounts corresponding to state income taxes. For survey tax units to whom we can attach a 1040 return, we rely on inputs from the limited tax data wherever possible. Specifically, we bring in wages, interest income, and dividends (all at the tax unit level) from the limited 1040 data extract as well as self-employment earnings from the DER.[22] We then input the difference between AGI (for which we have amounts in the limited 1040 data) and the sum of the aforementioned income sources into the "other non-property income" field in TAXSIM, which supports negative values.[23] This effectively "tricks" TAXSIM into always generating the AGI amount that we observe

[20] In future work, we may consider imputing itemized deductions based on average amounts by AGI bracket that are publicly available in the IRS Statistics of Income summary data. One caveat here is that we would have to impute not only itemized deduction amounts but also whether or not a tax unit itemizes its deductions.

[21] However, self-employed individuals are allowed to deduct one-half of their self-employment tax from total income for calculating federal income tax liability. In essence, the SSA obtains the full amount of self-employment tax, but part of the total amount can be thought of as being transferred from the personal income tax by the tax rules.

[22] Because the DER records only Medicare-taxable self-employment earnings, we divide the DER self-employment amount by 0.9235 to obtain our measure of total self-employment earnings. This will understate actual self-employment earnings, both because we do not have the information in the limited tax data to add back in health insurance deductions and because individuals with sufficiently low self-employment amounts (below $400 in tax year 2010) do not have to pay the self-employment tax.

[23] Note that our limited tax data also include values for gross (rather than taxable) Social Security income, but we do not separately input gross Social Security income into the relevant TAXSIM field. This is because TAXSIM automatically converts the gross amount to the taxable amount, meaning we would need to subtract taxable Social Security income

on the 1040. Note, however, that our calculation of state income taxes may be slightly biased even though we anchor to the correct federal AGI amount, as certain income sources are taxed differently by certain states. Furthermore, while we know which tax units itemized their deductions (i.e., filed a Schedule A), we do not know the corresponding amounts. We therefore impute itemized deduction amounts for these tax units by matching them to average amounts by AGI bracket that are publicly available in the IRS Statistics of Income summary data.

We rely on TAXSIM outputs for most federal and state income tax variables, with the exception of the EITC. We calculate the EITC separately because TAXSIM requires the number of EITC-qualifying children to exceed the number of CTC-eligible children, even though some tax units that are eligible for the CTC are ineligible for the EITC.[24] Since we have both EITC earned income and the number of EITC-qualifying children from the 1040 extracts, we are able to calculate the EITC directly. We again calculate payroll taxes manually to estimate just the employee portion of the payroll tax for wage/salary employees and the full amount of the payroll tax for self-employed individuals. In particular, for FICA taxes, we multiply Social Security wages (capped at $106,800 per individual) from the W-2 by 6.2% to calculate Social Security taxes and the uncapped sum of taxable wages and deferred compensation from the W-2 by 1.45% to calculate Medicare taxes. For SECA taxes, we multiply self-employment earnings in the DER (derived from Schedule SE of the 1040) by 12.4% (once again up to the individual cap of $106,800) to calculate Social Security taxes and the uncapped amount of self-employment earnings by 2.9% to calculate Medicare taxes.

We also simulate tax liabilities – using information from the DER, IRS information returns, and the CPS – for survey tax units to whom we cannot attach a 1040. There are several reasons for why tax units may pay taxes even if they are not attached to a 1040. First, not all 1040 returns can be properly linked to the CPS sample (e.g., 1040s without PIKs or with incorrect PIKs, 1040 returns attaching to individuals split across multiple survey families, etc.). Second, within our CPS sample, there may be un-PIKed family members in PIKed families (where at least one CPS member links to a PIK) to whom we cannot attach a 1040 return. Third, individuals typically have had taxes withheld

---

from AGI to accurately construct the "catch-all" term. Since the process of converting gross to taxable amounts is not straightforward and appears to rely on some additional income components that we do not have in the limited tax data, we choose not to use gross Social Security income as an input into TAXSIM and also not to subtract it from AGI in constructing the "catch-all" term.

[24] TAXSIM requires the number of EITC children to exceed the number of CTC children, which creates a conflict – for example – in cases where we would like to enter zero EITC children but positive CTC children for tax units that are eligible for the CTC but ineligible for the EITC.

even if they did not file a 1040 return. And fourth, individuals may have filed returns at a later date. For survey tax units to whom we cannot attach a 1040 return, we bring in taxable wages, retirement income, and self-employment from W-2s, 1099-Rs, and the DER, respectively. When information from these administrative sources is not available, we rely on CPS income and demographic variables. We once again rely on TAXSIM to calculate federal and state income taxes and calculate payroll taxes outside of TAXSIM. For a detailed mapping of survey and administrative variables to TAXSIM inputs for the limited tax data imputation, see Appendix Table A.1.

Combining Survey and Extensive Tax Data

Finally, we discuss how we calculate tax liabilities and credits using a combination of survey data and extensive tax records. While the extensive 1040 data contain actual amounts for federal income tax liabilities and credits (as well as nearly every other line item on the 1040), the data do not contain a single variable equal to the net federal income tax liability. We therefore calculate the net federal income tax liability as the sum of federal income tax after non-refundable credits (line 55) and additional tax on IRAs (line 58) minus a series of refundable credits, including the Making Work Pay Credit (line 63), the EITC (line 64a), the refundable portion of the Child Tax Credit (line 65), the American Opportunity Credit (line 66), the First-Time Homebuyer Credit (line 67), the Credit for Federal Tax on Fuels (line 70), and a set of smaller credits listed on line 71.[25] An equivalent alternative calculation would take the sum of federal income taxes withheld and taxes due (net of refundable and non-refundable tax credits) and subtract payroll taxes on self-employment earnings, although additional adjustments may be needed for some tax items that pertain to other tax years. We continue to use TAXSIM to calculate state income taxes, relying on inputs from the extensive tax data wherever possible. We also calculate payroll taxes outside of TAXSIM using the same methodology as that used for the limited tax data, with one exception: instead of calculating SECA taxes on self-employment earnings in the DER, we take self-employment taxes directly from the 1040 (line 56). Finally, we continue to calculate tax liabilities and credits for survey tax units to whom we cannot attach a 1040, drawing from IRS information returns when available and relying on CPS demographics and incomes otherwise. For a detailed mapping of survey and administrative variables to TAXSIM inputs for the extensive tax data calculation, see Appendix Table A.2.

[25] Line numbers in this sentence and throughout the subsection correspond to the 1040 form for tax year 2010.

6. Results

In this section, we discuss three sets of results.[26] We begin by showing aggregate estimates of income and tax components obtained using various tax calculators. We then analyze estimates of income and tax components across deciles of survey-reported family income, focusing on differences between various tax calculators. We finally assess the mean absolute errors in tax imputations from the various tax calculators, relative to estimates obtained using a combination of survey and extensive tax data. For reference, Appendix Tables A.3, A.5, A.6a, and A.6b contain the standard errors of the estimates in the main tables, and Appendix Table A.7 contains the results of statistical tests used to assess additional comparisons we make in the text.

Aggregate Comparisons

Table 1 shows aggregate dollar amounts for various tax liabilities and credits (as well as certain intermediate outputs) for our four estimates of taxes calculated on the linked CPS sample. These include the Census tax model imputation using CPS data – hereafter referred to as the "CPS tax imputation" (Column 3), the imputation using CPS data entered into TAXSIM – hereafter referred to as the "CPS-TAXSIM imputation" (Column 4), the limited tax data imputation (Column 5), and the extensive tax data calculation (Column 6). Recall that the limited tax data imputation and extensive tax data calculation use some non-1040 IRS data and/or survey reports of income to simulate taxes for units to whom we cannot attach a 1040. Columns 7 and 8 show alternative estimates that simulate taxes only for those units to which a 1040 can be attached, and Columns 9 and 10 additionally use non-1040 data to simulate taxes only for families containing an individual who does not link to a PIK (and to whom we therefore cannot attach a 1040). Note that the estimates in Columns 7-10 are compiled to better understand the estimates in Columns 5-6 and – on their own – are not designed to match population totals.

We compare these estimates to two sets of independent aggregates. The first set of benchmarks comes from publicly available sources, including IRS SOI line item totals for federal income tax items, the Census Bureau Survey of State Governments (specifically covering the

[26] Our results are subject to error arising from various sources, including our use of sample data, misreporting of certain variables in our survey data sources, processing errors, and others.

Quarterly Summary of State and Local Tax Revenues) for state income tax liabilities, and the SSA for payroll tax liabilities (Column 1).[27] Note that the SOI aggregates cover 1040s filed during calendar year 2011 for tax year 2010 as well as any 1040s filed for the previous two tax years (2008 and 2009).[28] The second set of aggregates is calculated from the extensive 1040 universe data (Column 2). In contrast to the SOI aggregates, the extensive tax data tabulations include 1040s filed during calendar year 2011 that we filter only for tax year 2010. As a result, we can see that the SOI aggregates – containing late filers from previous tax years – uniformly exceed the extensive tax data aggregates for all items.

Focusing first on the CPS tax imputation in Column 3, we see that it estimates a total of $808 billion in federal income tax liabilities and $217 billion in state income tax liabilities, both of which fall short of the independent aggregates. Interestingly, the underestimation of federal income taxes (defined as federal income taxes paid net of federal tax credits) persists despite the CPS tax imputation underestimating total EITC amounts by nearly one-third and total CTC amounts by nearly one-fifth. In fact, much of this underestimation can be attributed to the CPS underestimating AGI and therefore taxable income, resulting in much lower estimates of federal incomes taxes before credits relative to the independent aggregates. In contrast, the CPS estimates a total of $461 billion in payroll tax liabilities, which exceeds the SOI aggregate by approximately 6%. Thus, even though the CPS underestimates AGI, it appears to overestimate the earnings on which payroll taxes are paid.

Interestingly, the CPS-TAXSIM imputation in Column 4 yields estimates of federal income tax liabilities ($859 billion) that are higher than the estimates from the CPS tax imputation and closer to the independent aggregate despite continuing to rely on CPS income reports and CPS-constructed tax unit structures.[29] The primary reason for this is that the CPS-TAXSIM imputation accounts for too few itemized deductions. While the CPS tax imputation in Column 3 brings in itemized deductions using a statistical match to the PUF, the CPS-TAXSIM imputation relies only on the limited information available in the CPS to construct itemized deductions. To see this, note that the

[27] The benchmarks for payroll taxes are calculated as half of the total OASDHI (Old-Age, Survivors, Disability, and Hospital Insurance) FICA tax and the total amount of SECA tax. Calculating payroll tax aggregates in this way aligns with our definition of payroll taxes as being what the employee pays out to SSA. Alternatively, we could make an additional adjustment to account for the payment of FICA tax over the taxable maximum by a combination of employers of a given individual that are not refunded to the firms.

[28] Those filed for tax years 2008 and 2009 may be thought of as roughly approximating returns for tax year 2010 that will be filed after calendar year 2011.

[29] The difference between the CPS-TAXSIM estimate of total federal income taxes and the SOI aggregate is not statistically significant at the 10% level.

gap in taxable income between the CPS tax imputation and the CPS-TAXSIM imputation is nearly three times the gap in AGI, with much of the conceptual difference between AGI and taxable income due to itemized deductions. The CPS-TAXSIM imputation also yields estimates of state income tax liabilities ($236 billion) and payroll tax liabilities ($470 billion) that are higher than those obtained using the CPS tax imputation (and therefore closer to the independent aggregates for state income taxes and farther from those aggregates for payroll taxes), while estimates of the EITC and CTC are slightly underestimated relative to those from the CPS tax imputation and the SOI aggregates.[30]

Moving on to Columns 5 and 6, we see that aggregate estimates for many of the tax items between the limited tax data imputation and extensive tax data calculation are strikingly close to each other and also to SOI aggregates.[31] Starting with federal income tax liabilities, we find that the limited tax data imputation yields $853 billion and the extensive tax data calculation yields $845 billion, with the former within 1% and the latter within a tenth of 1% of the SOI aggregate.[32] Approximately 95% of all federal income tax liabilities are associated with tax units to whom a 1040 can be attached. However, the similarity of these estimates may be a result of some offsetting errors associated with the limited tax data imputation. On one hand, the limited tax data imputation appears to overstate itemized deductions, as the estimate of taxable income – and therefore federal income taxes before credits – is smaller using the limited tax data imputation. On the other hand, there are certain tax credits – including the foreign tax credit, education credits, and the first-time homebuyer credit – that are captured in the extensive tax data but are not simulated by TAXSIM and are therefore missing in the limited tax data imputation.

We also find that the limited and extensive tax data estimates in Columns 5 and 6 are comparable to each other for payroll tax liabilities while the extensive tax data estimate is slightly smaller for state income tax liabilities.[33] The comparability in payroll taxes, even when separating out FICA and SECA taxes, suggests that the different de-duplication strategies used across the

[30] The difference between the CPS-TAXSIM estimate of total state income taxes and the SOI aggregate is not statistically significant at the 10% level.
[31] We believe that the SOI aggregates in Column 1 are conceptually closer to the limited and extensive tax data estimates in Columns 5 and 6, while the extensive tax data benchmarks in Column 2 are conceptually closer to the estimates in Columns 7-10. This is because Columns 5 and 6 attempt to calculate taxes for late filers (who are covered by the SOI aggregates) by simulating estimates using non-1040 information, while Columns 7-10 calculate taxes only for those to whom we can attach a 1040 (and for families containing individuals for whom a key to attach a 1040 is missing).
[32] In each case, the difference with the SOI aggregate is not significant at the 10% level.
[33] The difference in total payroll taxes between the limited tax data imputation and the extensive tax data calculation is not statistically significant at the 10% level.

limited and extensive W-2 data and the different sources for taxable self-employment earnings (DER for limited tax data and 1040 for extensive tax data) yield almost equivalent results.[34] Finally, the estimates for the EITC and CTC in Columns 5 and 6 are remarkably close to each other and to their respective independent aggregates.[35] One potential concern with these estimates in Columns 5 and 6 is that an individual should not be able to receive the EITC or CTC if they did not file a 1040. However, out of the $59 billion estimated for the EITC in Columns 5 and 6, it appears that $55 billion are associated with units to whom a 1040 can be attached (Columns 7 and 8). Approximately half of this difference can be explained by individuals that have missing PIKs and therefore cannot be attached to a 1040 (Columns 9 and 10). The remaining difference can likely be attributed to an assortment of other reasons – e.g., a small share of the IRS records cannot be linked to PIKs, 1040 returns may occasionally attach to individuals split across multiple survey families (in which case we ignore the 1040 information), and there might be slight biases associated with the IPW adjustment to survey weights or the original CPS weights themselves.

Comparisons of Means by Decile

Table 2 shows how the differences between the various ways of estimating taxes break down across the income distribution. We assign families in Table 2 to deciles of survey-reported family pre-tax money income.[36] Note that the survey-reported income distribution should not be interpreted as the *true* income distribution. In fact, we find a striking non-monotonicity in taxable income (and thus federal income taxes before credits) calculated using the extensive tax data along the survey-reported income distribution. Specifically, we find that families in the bottom decile of survey-reported income do not have lower levels of taxable income and federal income tax before credits

[34] One potential reason for why the SOI aggregates for payroll tax liabilities on wages (FICA tax) is slightly higher than the estimates in Columns 5 and 6 is that the federal government does not refund employers when a worker has too much withheld as a result of multiple employers withholding under the Social Security cap. If this is the case, then it may not make sense to assume a completely even split between the employee and employer portions of the FICA tax benchmark – instead, the weight for the employee portion should likely be slightly below 50%.

[35] The difference in total EITC between the limited tax data imputation and the SOI aggregate is not statistically significant at the 10% level. The difference in total EITC between the extensive tax calculation and the SOI aggregate is also not statistically significant at the 10% level.

[36] We adjust survey incomes according to the equivalence scale recommended in National Academy of Sciences (1995) of the form $(A + PK)^F$, where A and K respectively designate the number of adults and children in the family. Following Meyer and Sullivan (2012), we set $P = F = 0.7$.

(per the extensive tax data) than those in the second decile.[37] This finding – while puzzling at face value – is consistent with the presence of income under-reporting at the bottom of the survey-reported income distribution, a result that a number of other studies have found (see, e.g., Brewer, Etheridge, and O'Dea 2017, Meyer et al. forthcoming).

Given mean values from the extensive tax data as our basis for comparison in Column 1, we report results for each of the imputation methods in Columns 2-4 as dollar differences from the extensive tax data counterpart. Consequently, a negative value in Columns 2-4 suggests that the imputation method leads to an underestimate (and vice-versa). Although Table 2 reports estimates for both the CPS tax imputation (Column 2) and CPS-TAXSIM imputation (Column 3), the estimates between the two survey-only imputations are comparable to each other. Thus, we mostly focus on the CPS tax imputation (Column 2) and discuss how it compares to the limited tax data imputation (Column 4). Appendix Table A.4 shows the mean amounts (rather than the dollar differences from the extensive tax data calculations) for each of the imputation methods.

We start by examining differences in federal income tax liabilities. We find that the CPS tax imputations understate federal income tax liabilities in the bottom seven deciles of survey-reported income, before overstating federal income tax liabilities in the top three deciles of survey-reported income.[38] This suggests that the overall underestimate of federal income taxes by the CPS tax calculator in Table 1 is due not so much to the CPS understating incomes for the richest individuals (as defined by the survey), but rather to the CPS systematically understating incomes for at least the bottom two-thirds of the survey-defined income distribution. Indeed, we find that the CPS tax imputation understates the extensive tax data calculation for AGI in the bottom nine deciles of survey income, for taxable income in the bottom seven deciles, and for federal income tax before credits in the bottom nine deciles.[39] Likewise, the CPS-TAXSIM imputation understates the extensive tax data calculation in six of the bottom seven deciles for federal income tax liabilities, all

[37] Specifically, using the extensive tax data calculations, we find no statistically significant difference at the 10% level in either average taxable income or average federal tax before credits between the first and second deciles.
[38] However, the differences in most of the survey income deciles – specifically, the 1st, 4th, 5th, 7th, 8th, and 9th deciles – are statistically insignificant at the 10% level.
[39] The differences in some of the survey income deciles – specifically, the top three deciles for average AGI, the 7th and 9th deciles for average taxable income, and the top two deciles for average federal tax before credits – are statistically insignificant.

ten deciles for AGI, the bottom seven deciles for taxable income, and the bottom nine deciles for federal income tax before credits.[40]

In contrast, the limited tax data imputation statistically overstates federal income tax liabilities in the majority of survey income deciles.[41] This bias is likely due to the limited tax data imputation missing a number of tax credits (e.g., education credits, foreign tax credit) that are available in the extensive tax data and are accessible more broadly across the income distribution. Interestingly, compared to the CPS tax imputation, the limited tax data imputation yields estimates of federal income tax liabilities that are closer to the extensive tax data estimates at many of the bottom and top deciles of survey income (but not in the middle).[42] We suspect this is a result of errors in both tax liabilities and credits offsetting in the CPS tax imputation. Looking however at AGI, taxable income, and federal income tax before credits, the limited tax data estimates are closer to the extensive tax data counterparts throughout most of the survey income distribution.[43]

For state income tax liabilities, we continue to find that both the CPS tax imputation and the CPS-TAXSIM imputation understate the extensive tax data means throughout most of the survey-reported income distribution and that the limited tax data imputation overstates the extensive tax data means throughout the distribution.[44] For payroll taxes, we find that the limited tax data imputation yields estimates that are much closer to the extensive tax data counterparts than the CPS tax imputation, which understates payroll taxes in most of the bottom half of the survey income distribution and overstates payroll taxes in the top half.[45] Comparably, the CPS-TAXSIM imputation understates payroll taxes in the bottom six deciles of the survey income distribution.[46]

---

[40] The differences in some of the survey income deciles – specifically, the 1st, 4th, 5th, 7th, and 9th deciles for average federal income tax, the 10th decile for average AGI, the 7th and 9th deciles for average taxable income, and the top three deciles for average federal tax before credits – are statistically insignificant at the 10% level.

[41] The differences in the 8thand 9th deciles are not statistically significant at the 10% level.

[42] The differences in the 1st, 8th, and 9th deciles are not statistically significant at the 10% level.

[43] The differences in some of the survey income deciles – specifically, the top two deciles for average AGI, the 7th and 9th deciles for average taxable income, and the 9th decile for average federal tax before credits – are statistically insignificant at the 10% level.

[44] The differences between the CPS imputations and the extensive tax data calculations are statistically insignificant at the 10% level in the top four deciles. The differences between the CPS-TAXSIM imputations and the extensive tax data calculations are statistically insignificant at the 10% level in the 7th, 8th, and 9th deciles. The difference between the limited tax data imputation and the extensive tax data calculation is statistically insignificant in the top income decile.

[45] The difference between the CPS imputation and the limited tax data imputation is statistically insignificant at the 10% level in the 4th decile.

[46] The difference between the CPS-TAXSIM imputation and the extensive calculation is statistically insignificant at the 10% level in the 4th and 6th deciles.

For the EITC, the limited tax data imputation yields estimates that are within 1% of the extensive tax data means at every decile of survey income.[47] In contrast, the CPS tax imputation understates those means for EITC in nearly every decile of survey income (except for the second), with these differences being most pronounced for families in the fourth and fifth deciles. The CPS-TAXSIM imputation understates EITC means in every decile of survey income. Finally, the limited tax data imputation also yields estimates of the CTC that are on average closer to the extensive means than the CPS tax imputation, both on average and throughout most of the survey income distribution.[48]

Mean and Median Absolute Errors in Tax Imputations (Tables 3a and 3b)

Finally, Table 3a shows mean absolute differences in tax liabilities and credits (and other tax components) at the family level, with estimates from each imputation calculated relative to the extensive tax data estimates. While the previous table analyzed net mean differences (with understatements and overstatements canceling each other out), this table analyzes mean absolute deviations (with understatements and overstatements each contributing to the error). Columns 2, 4, and 6 show the mean absolute dollar differences for the CPS tax imputation, the CPS-TAXSIM imputation, and the limited tax data imputation, respectively. Columns 3, 5, and 7 report these mean differences as a percentage of the extensive tax data estimates in Column 1. We once again examine these differences across the survey-reported income distribution, assigning families to quartiles of survey-reported family pre-tax money income.[49] In comparison, Table 3b shows median absolute differences in tax liabilities and is structured similarly as Table 3a. Since the patterns between the imputation methodologies are by and large similar in Tables 3a and 3b, we primarily discuss our results in terms of mean absolute errors.

We start by discussing the mean absolute differences for the CPS tax imputation (Column 2) relative to the extensive tax data estimates (Column 1). Again, we focus our discussion on the CPS tax imputation rather than the CPS-TAXSIM imputation, although the estimates from the two survey models are close to each other. For federal income tax liabilities, the mean absolute error for the CPS tax imputation is $7,202 for all families, which is greater than the average federal income tax amount

[47] The differences in the 5th, 6th, 9th, and 10th deciles are statistically insignificant at the 10% level.
[48] The differences in the 2nd, 3rd, 5th, and 6th deciles are statistically insignificant at the 10% level.
[49] Again, we adjust survey incomes using the equivalence scale described earlier.

($6,824) and nearly 10% of average AGI ($73,680). Analogously, the mean absolute error for the CPS-TAXSIM imputation of federal income taxes is $7,434 over all families.[50]

The mean absolute errors in state income tax liabilities and payroll tax liabilities are much smaller than the mean absolute error in federal income taxes as a share of AGI (each approximately 2% of AGI), even though these errors are sizeable as a share of their mean values from the extensive tax data (65% for state income taxes and 37% for payroll taxes). Looking at total tax liabilities, the mean absolute error is $9,385 for all families and about 13% of average AGI for the CPS tax imputation (and $9,870 over all families for the CPS-TAXSIM imputation). Once again, the mean absolute errors as a share of AGI are highest in the bottom and top quartiles. It is worth contextualizing the magnitudes of these mean absolute errors in taxes against errors in other income sources calculated in the literature. For example, Duncan and Hill (1985) find that the average absolute difference between survey and administrative values of earnings in their 1982 sample was $2,123 – amounting to approximately 7% of mean earnings. Compared to this difference, the average absolute difference that we calculate for total taxes (13% of AGI, which is typically a larger income base than earnings) is considerably greater.

A key reason for the considerable errors in tax liabilities for the survey-only imputations is that AGI is measured with substantial error. The mean absolute difference in AGI for the CPS tax imputation is $32,050 for all families, which is about 44% of the mean AGI amount. Likewise, the mean absolute difference in AGI for the CPS-TAXSIM imputation is $34,670, which is about 47% of the mean AGI amount. However, outliers appear to drive at least part of the mean absolute error in AGI; the median absolute error in AGI is $10,090 (about 14% of the mean AGI amount), suggesting that the distribution of absolute errors is skewed to the right. As a share of the mean amount, the mean absolute error in AGI is largest in the bottom quartile (65%). Measurement error in AGI naturally translates to measurement error in taxable income – which consists of AGI minus itemized/standard deductions and exemptions – and federal income tax before credits. The mean absolute difference in taxable income for the CPS is $27,280, which is 57% of the mean taxable income amount and among the highest (as a share of the mean amount) in the bottom quartile. Analogously, the mean absolute difference in federal income tax before credits for the CPS is

[50] Appendix Table A.6a also contains estimates of the mean absolute errors of the CPS-TAXSIM imputation of federal income taxes by quartile, although we do not perform statistical tests for differences of significance between quartile estimates.

$6,366, which is 69% of the mean taxable income amount and among the highest in the bottom quartile (as a share of the mean amount). However, the median absolute differences for taxable income and federal income tax before credits are only about 18% and 15% of their mean amounts, respectively.

We also observe substantial biases in the estimation of tax credits using the survey-only tax imputations. Specifically, the mean absolute errors for the EITC are $550 and $569 for all families using the CPS tax and CPS-TAXSIM imputations, respectively, with these errors amounting to approximately three-quarters of the mean EITC amount.[51] The mean absolute error in the CTC using the CPS tax imputation is smaller at $275 for all families, although this figure is still 37% of the mean CTC amount.

In contrast to either of the CPS imputations, the limited tax data imputation yields much smaller mean absolute differences relative to the extensive tax data estimates. For federal income tax liabilities, the mean absolute error between the limited tax data imputation and extensive tax data calculation is 24% for all families. This is approximately one-quarter of the absolute errors for the CPS tax imputation (106% of the extensive data mean) and CPS-TAXSIM imputation (109% of the extensive data mean). Similarly, the mean absolute error between the limited and extensive tax data estimates is a mere 1% for AGI (compared to 44% for the CPS tax imputation), 10% for taxable income (compared to 57% for the CPS tax imputation), and 16% for federal income tax before credits (compared to 69% for the CPS tax imputation). The improvement in tax calculation using the limited tax data is particularly noticeable in the top half of the survey income distribution.

The similarity in absolute terms between the limited and extensive tax data estimates – relative to the CPS imputations – also holds for other tax calculations. The mean absolute difference between the limited and extensive tax data estimates is 16% for state income taxes (compared to 65% for the CPS tax imputation) and 3% for payroll taxes (compared to 37% for the CPS tax imputation). Taken together, the mean absolute error between the limited tax data imputation and extensive tax data calculation for total tax liabilities is about 15%, which is approximately 20% of the absolute errors for the CPS tax imputation (72% of the extensive data mean) and CPS-TAXSIM imputation (76% of the extensive data mean). Furthermore, the limited tax data lead to particularly accurate estimates of tax credits. For a typical family, the limited tax data imputation of the EITC is

[51] Appendix Table A.6a also contains estimates of the mean absolute errors for the EITC by quartile, although we do not perform statistical tests for differences of significance between quartile estimates.

off by less than $20 (compared to more than $500 using the CPS tax and CPS-TAXSIM imputations) and the limited tax data imputation of the CTC is off by less than $50 (compared to $275 and $306 using the CPS tax and CPS-TAXSIM imputations, respectively). Finally, when looking across all families, the median absolute differences between the limited and extensive tax data estimates are approximately zero for most tax calculations and very small for the others.

7. Conclusions

This paper calculates estimates of income and payroll taxes using two different sets of administrative tax records linked to the 2011 CPS ASEC. By describing how to form tax units and estimate various types of tax liabilities and credits using these linked data, this paper provides a roadmap for constructing accurate measures of taxes while preserving the survey family as the sharing unit for distributional analyses. We find that aggregate estimates of various tax components (particularly tax credits like the EITC and CTC) calculated using the limited tax data imputations and extensive tax data calculations are similar to each other and much closer to IRS SOI aggregates than any of the imputations using survey data alone. Across the deciles of the reported income distribution, the limited tax data imputations tend to give us a picture of the distribution of income, taxes, and their components that better match what we see in the extensive tax data, but this pattern is far from generally true. At the individual level, the CPS tax and CPS-TAXSIM imputations have substantial errors, with each having mean absolute errors for federal taxes and total taxes equal to approximately 10 and 13 percent of mean AGI, respectively. The limited tax data imputations have 22-23% of the absolute errors of the survey-only imputations for federal income tax liability and 19-20% of the absolute errors of the survey-only imputations for total tax liability (relative to the extensive tax data calculations). For the EITC, the limited tax data imputation is off by less than $20 for a typical family (compared to more than $500 using the survey-only imputations).

In summary, this paper emphasizes the impacts of errors in tax imputations for three types of statistics: overall means, means by income decile, and family-level values. These results likely apply to many uses of the imputations, with larger impacts on some than others. The differences in overall means, which are especially pronounced for the EITC and CTC, are likely to have large effects whenever the imputations are used. The differences by income decile indicate how analyses of progressivity or the distributional impacts of taxes (more generally) are likely to be affected. On top of the overall errors and mean errors by group, the errors in family-level taxes will matter when

after-tax income is used as an explanatory variable for various analyses or as the baseline income when trying to identify who is poor or disadvantaged more broadly. In certain cases, such as when analyses are done at a level for which the imputations are close to correct (on average), there may be little bias in estimates that rely on these imputations.

In future work, we hope to extend the comparisons of tax calculators to linked samples using the Survey of Income and Program Participation (SIPP) and the American Community Survey (ACS). Differing levels of misreporting across Census surveys might affect the extent to which the magnitudes of errors found using the CPS extend to other surveys. We also hope to examine the distribution of errors for various tax components across additional demographic and socioeconomic characteristics, as these analyses may be relevant to recent studies (see, e.g., Goldin and Michelmore 2020, Thomson et al. 2020) that have analyzed the distribution of EITC and CTC receipt relying exclusively on survey data. We also plan to expand our analyses to other years, especially more recent ones. While the extensive tax data are currently available to us only for the 2010 processing year, we are able to access the limited tax data for a wide range of years. The analyses in this paper shed important light on the degree to which imputations relying on the limited tax data do a sufficient job of matching the values in the extensive tax data, although we plan on using the extensive tax data to fill in the remaining holes in the limited tax data imputations. We also plan on estimating other taxes that families and individuals pay, including sales and property taxes. A caveat with these taxes is that we are unlikely to obtain administrative values corresponding to them and must therefore make certain assumptions to simulate them.

The estimates of taxes calculated in this paper open the door for a number of distributional analyses that can be done using the family as the unit of analysis. Families are more natural for such an analysis than households or tax units, since family members share incomes in ways that unrelated roommates generally do not and multiple tax units within a family may also share resources and plan expenditures. These analyses include analyzing the redistributive value of taxes and transfers among families, focusing on how the progressivity of the U.S. tax and transfer system varies along the income distribution. Another analysis involves evaluating the poverty reduction of tax credits and government transfers. In all of these analyses, it is useful to examine the extent to which relying on survey data alone biases estimates not only of taxes and transfers but also of underlying income.

References

Auten, Gerald and David Splinter. 2019. Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends. Working Paper.

Bakija, Jon. 2014. Documentation for a Comprehensive Historical U.S. Federal and State Income Tax Calculator Program. Working Paper.

Brewer, Mike, Ben Etheridge, and Cormac O'Dea. 2017. Why are Households that Report the Lowest Incomes So Well-Off? *The Economic Journal*, 127(605): F24-F49.

Burkhauser, Richard V., Kevin Corinth, James Elwell, and Jeff Larrimore. 2019. Evaluating the Success of President Johnson's War on Poverty: Revisiting the Historical Record Using a Full-Income Poverty Measure. NBER Working Paper No. 26532.

Congressional Budget Office. 2018. The Distribution of Household Income, 2014. Washington, D.C.: Congressional Budget Office.

Dahl, Gordon B. and Lance Lochner. 2012. The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit. *American Economic Review,* 102(5): 1927-1956.

Duncan, Greg and Daniel H. Hill. 1985. An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data. *Journal of Labor Economics*, 3(4): 508-532.

Feenberg, Daniel and Elisabeth Coutts. 1993. An Introduction to the TAXSIM Model. *Journal of Policy Analysis and Management*, 12(1): 189-194.

Fox, Liana. 2019. The Supplemental Poverty Measure: 2018. Current Population Report P60-268. Washington, D.C.: U.S. Census Bureau.

Giertz, Seth H. 2007. The Elasticity of Taxable Income Over the 1980s and 1990s. *National Tax Journal*, 60(4): 743-768.

Goldin, Jacob and Katherine Michelmore. 2020. Who Benefits from the Child Tax Credit? NBER Working Paper No. 27940.

Gruber, Jon and Emmanuel Saez. 2002. The Elasticity of Taxable Income: Evidence and Implications. *Journal of Public Economics*, 84(1): 1-32.

Hoynes, Hilary W. and Erzo F.P. Luttmer. 2011. The Insurance Value of State Tax-and-Transfer Programs. *Journal of Public Economic,* 95(11): 1466-1484.

Jones, Maggie R. and Amy B. O'Hara. 2016. Do Doubled-up Families Minimize Household-level

Tax Burden? *National Tax Journal*, 69(3): 613-640.

Jones, Maggie R. and James P. Ziliak. 2020. The Antipoverty Impact of the EITC: New Estimates from Survey and Administrative Tax Records. Center for Economic Studies Working Paper 19-14. Washington, D.C.: U.S. Census Bureau.

Langetieg, Patrick, Mark Payne, and Alan Plumley. 2017. Counting Elusive Nonfilers Using IRS Rather Than Census Data. In A. Plumley (ed.), *IRS Research Bulletin, papers Given at the 7th Annual Joint Research Conference on Tax Administration*, 197-222. Washington, D.C.: Internal Revenue Service.

Meyer, Bruce D. and James X. Sullivan. 2003. Measuring the Well-Being of the Poor Using Income and Consumption. *Journal of Human Resources*, 38(Supplement): 1180-1220.

Meyer, Bruce D. and James X. Sullivan. 2012. Winning the War: Poverty from the Great Society to the Great Recession. *Brookings Papers on Economic Activity*, Fall 2012: 133-200.

Meyer, Bruce D., Derek Wu, and Carla Medalia. 2020. Understanding Poverty by Linking Survey, Tax, and Program Data. Working Paper.

Meyer, Bruce D. Derek Wu, Victoria Mooers, and Carla Medalia. Forthcoming. The Use and Misuse of Income Data and Extreme Poverty in the United States. *Journal of Labor Economics.*

National Academies of Sciences. 1995. *Measuring Poverty: A New Approach*. Citro, Constance F. and Robert T. Michael, editors. Washington, D.C.: National Academy Press.

National Academies of Sciences, Engineering, and Medicine. 2019. *A Roadmap to Reducing Child Poverty*. Washington, DC: The National Academies Press.

O'Hara, Amy. 2004. New Methods for Simulating CPS Taxes. SESHD Working Paper 2004-08. Washington, D.C.: U.S. Census Bureau.

Piketty, Thomas and Emmanuel Saez. 2007. How Progressive is the U.S. Federal Tax System? A Historical and International Perspective. *Journal of Economic Perspectives*, 21(1): 3-24.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. Distributional National Accounts: Methods and Estimates for the United States. *Quarterly Journal of Economics*, 133(2): 553-609.

Saez, Emmanuel and Gabriel Zucman. 2019. Progressive Wealth Taxation. *Brookings Papers on Economic Activity*, Fall 2019.

Splinter, David. 2019. U.S. Taxes are Progressive: Comment on "Progressive Wealth Taxation."

Working Paper.

Splinter, David, Jeff Larrimore, and Jacob Mortenson. 2017. Whose Child is This? Shifting of Dependents Among EITC Claimants Within the Same Household. *National Tax Journal*, 70(4): 737-758.

Thomson, Dana, Lisa A. Gennetian, Yiyu Chen, Hannah Barnett, Medline Carter, and Santiago Deambrosi. 2020. State Policy and Practice Related to Earned Income Tax Credits May Affect Receipt Among Hispanic Families with Children. Child Trends Research Brief.

U.S. Census Bureau. 1986. Measuring the Effect of Benefits and Taxes on Income and Poverty: 1986. Current Population Reports, Series P60-164-RD-1. Washington, D.C.: U.S. Census Bureau.

U.S. Census Bureau. 1992. Measuring the Effect of Benefits and Taxes on Income and Poverty: 1992. Current Population Reports, Series P60-186RD. Washington, D.C.: U.S. Census Bureau.

Wagner, Deborah and Mary Layne. 2014. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research & Applications' Record Linkage Software. CARRA Working Paper 2014-01. Washington, D.C.: U.S. Census Bureau.

Webster, Bruce H. 2011. Evaluating the Use of the New Current Population Survey's Annual Social and Economic Supplement Questions in the Census Bureau Tax Model. Washington, D.C.: U.S. Census Bureau.

Wheaton, Laura and Kathryn Stevens. 2016. The Effect of Different Tax Calculators on the Supplemental Poverty Measure. Washington, D.C.: Urban Institute.

Zedlewski, Sheila and Linda Giannarelli. 2015. TRIM: A Tool for Social Policy Analysis. Washington, D.C.: Urban Institute.

Tables

Table 1. Aggregate Income and Tax Components Using Various Tax Calculators, 2010 Tax Year, CPS Data

		1040 Universe	CPS ASEC Linked to Administrative Data							
	Independent Aggregates from IRS SOI or Other Sources	Extensive Tax Data Calculation	CPS Tax Imputation	CPS Data and TAXSIM Imputation	Limited Tax Data Imputation	Extensive Tax Data Calculation	Imputation for Limited Tax Data Filers Only	Imputation for Extensive Tax Data Filers Only	Imputation for Limited Tax Data Filers & Families with Missing PIKs	Imputation for Extensive Tax Data Filers & Families with Missing PIKs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Federal income tax liability	844,600	825,500	807,600	858,700	853,100	845,300	810,400	804,600	814,500	808,600
State income tax liability	243,400		216,800	235,500	246,500	231,900	235,000	219,900	237,000	222,000
Payroll tax liability	435,800		461,100	470,200	432,500	433,400	411,300	412,200	413,800	414,700
Payroll tax on wages	387,900			422,000	380,300	382,200	366,300	368,200	367,300	369,100
Payroll tax on self-emp.	47,900			48,260	52,230	51,250	44,930	44,000	46,520	45,590
Adjusted Gross Income	8,089,000	7,891,000	7,326,000	7,394,000	8,276,000	8,216,000	7,792,000	7,740,000	7,884,000	7,832,000
Taxable income	5,502,000	5,366,000	5,238,000	5,421,000	5,357,000	5,534,000	5,052,000	5,237,000	5,105,000	5,289,000
Fed. inc. tax before credits	1,065,000	1,042,000	917,100	979,100	1,014,000	1,052,000	960,300	999,400	969,100	1,008,000
Earned income tax credit	59,560	59,510	40,350	34,060	59,020	59,300	54,560	54,600	56,710	56,810
Child tax credit	56,260	54,580	45,860	44,460	52,910	54,230	50,320	51,560	51,500	52,740
Universe of returns		138,800,000								
Sample of individuals					170,000		140,000	139,000	147,000	146,000

Sources: IRS SOI line items; 2011 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC); IRS 1040, W-2, 1099-R extracts for tax year 2010; IRS extensive 2010 tax data; TAXSIM.
Approved for release by the U.S. Census Bureau, authorization numbers CBDRB-FY20-ERD002-014, CBDRB-FY20-ERD002-038.

Note: Amounts are in millions of dollars. We drop families with no PIKed members and families with any whole-imputed individuals, adjusting family survey weights using inverse probability weighting so that we approximately match population totals. Our IRS SOI aggregates come from publicly available line item totals. The CPS Tax Imputation sums income and tax components imputed by the Census Bureau over families. CPS Data and TAXSIM Imputation sums income and tax components imputed by TAXSIM using CPS inputs over families. Limited Tax Data refers to tax calculations using TAXSIM with inputs generated from the limited tax data linked to the CPS. The Extensive Tax Data calculation obtains federal income tax liabilities and its components directly from the extensive tax data and uses TAXSIM to generate tax liabilities and credits for CPS individuals not linked with an extensive tax data tax unit. The calculations for Columns (5) and (6) include imputed taxes for survey individuals (using survey information on demographics and income) for whom we cannot attach a 1040 return. Columns (7)-(10) calculate taxes for sub-samples of our overall sample; these estimates are compiled to better understand the estimates in Columns (5) and (6) and – on their own – do not match population totals. Specifically, Columns (7) and (8) calculate taxes only for those tax units to whom we can attach a 1040 return, and Columns (9) and (10) additionally calculate taxes for individuals that do not link to a PIK but are part of families where at least one person contains a PIK. Federal income tax liabilities are total tax (line 55) + additional tax on retirement (line 58) - refundable credits (line 63, 64a, 65, 66, 67, 70, 71).

Table 2. Mean Family Income and Tax Calculations or Imputations by Pre-tax Money Income Decile, 2010 Tax Year, CPS data

Income or Tax Component	Decile of Equivalized Survey Family Income	*Mean* Extensive Tax Data Calculation (1)	*Imputation Minus Extensive Tax Data Calculation* CPS Tax Imputation (2)	CPS Data and TAXSIM Imputation (3)	Limited Tax Data Imputation (4)
Federal income tax liability	All	6,824	-162	28	83
	1	-1,056	-524	-325	479
	2	-2,326	-915	-548	125
	3	-562	-1,475	-1,262	173
	4	-475	-45	24	256
	5	988	-5	-25	306
	6	3,903	-1,211	-1,240	331
	7	5,509	-578	-568	209
	8	7,846	287	567	116
	9	14,350	66	554	182
	10	40,080	2,780	3,103	-1,343
State income tax liability	All	2,095	-137	-62	113
	1	24	-50	-84	105
	2	47	-63	-105	103
	3	473	-275	-319	147
	4	611	-99	-150	113
	5	979	-116	-161	121
	6	1,638	-310	-333	105
	7	2,038	-196	-166	123
	8	2,592	-64	96	-102
	9	4,083	-387	-147	117
	10	8,470	189	747	96
Payroll tax liability	All	4,098	244	68	-16
	1	770	-439	-498	-37
	2	1,218	-171	-312	-47
	3	1,819	-149	-351	-23
	4	2,468	-32	-279	-24
	5	3,035	172	-99	-26
	6	3,915	207	-61	-21
	7	4,790	383	119	-24
	8	5,674	608	377	-8
	9	7,141	682	620	-32
	10	10,150	1,179	1,158	84
Total tax liability	All	13,020	-55	34	181
	1	-262	-1,013	-907	547
	2	-1,060	-1,149	-965	181
	3	1,729	-1,899	-1,932	298
	4	2,604	-176	-404	345
	5	5,001	51	-285	401
	6	9,456	-1,314	-1,634	415
	7	12,340	-391	-615	308
	8	16,110	831	1,039	209
	9	25,570	361	1,028	267
	10	58,710	4,149	5,008	-1,163

(continued on next page)

Table 2. Mean Family Income and Tax Calculations or Imputations by Pre-tax Money Income Decile, 2010 Tax Year, CPS data (Continued)

Income or Tax Component	Decile of Equivalized Survey Family Income	*Mean* Extensive Tax Data Calculation (1)	*Imputation Minus Extensive Tax Data Calculation* CPS Tax Imputation (2)	CPS Data and TAXSIM Imputation (3)	Limited Tax Data Imputation (4)
Adjusted Gross Income	All	73,680	-6,915	-9,653	403
	1	19,720	-15,410	-15,980	9
	2	22,670	-8,809	-10,390	-20
	3	34,350	-11,760	-14,030	120
	4	41,170	-8,606	-11,400	84
	5	50,470	-6,000	-9,266	199
	6	66,820	-9,509	-12,790	232
	7	76,480	-4,388	-7,606	138
	8	91,320	-1,711	-4,909	387
	9	121,800	-5,327	-8,251	416
	10	212,100	2,350	-1,922	2,466
Taxable income	All	48,010	-1,591	-2,169	-1,417
	1	9,028	-8,924	-8,997	-432
	2	6,682	-5,576	-6,010	-117
	3	14,350	-8,334	-9,244	-98
	4	17,910	-4,462	-5,775	-317
	5	25,790	-3,130	-4,537	-420
	6	39,890	-5,992	-7,100	-1,265
	7	49,310	-1,742	-2,171	-1,531
	8	61,180	2,655	3,280	-2,309
	9	87,200	2,201	3,813	-3,077
	10	168,800	17,370	15,030	-4,605
Federal income tax before credits	All	9,221	-1,260	-934	-393
	1	1,473	-1,462	-1,470	300
	2	996	-897	-927	-29
	3	2,523	-2,164	-1,993	-63
	4	2,462	-1,633	-1,110	-83
	5	3,663	-1,755	-1,075	-84
	6	6,288	-2,780	-2,048	-101
	7	7,766	-2,064	-1,318	-349
	8	9,884	-975	-163	-456
	9	15,930	-799	-240	-401
	10	41,230	1,920	1,000	-2,663
Earned Income Tax Credit	All	733	-193	-295	-5
	1	1,504	-375	-542	-17
	2	1,828	236	-83	-12
	3	1,390	-153	-386	-14
	4	1,013	-493	-653	-5
	5	679	-467	-536	0
	6	379	-271	-314	-2
	7	221	-168	-178	2
	8	170	-127	-142	2
	9	86	-70	-72	0
	10	52	-44	-43	-1

(continued on next page)

Table 2. Mean Family Income and Tax Calculations or Imputations by Pre-tax Money Income Decile, 2010 Tax Year, CPS data (Continued)

Income or Tax Component	Decile of Equivalized Survey Family Income	*Mean*	*Imputation Minus Extensive Tax Data Calculation*		
		Extensive Tax Data Calculation	CPS Tax Imputation	CPS Data and TAXSIM Imputation	Limited Tax Data Imputation
		(1)	(2)	(3)	(4)
Child Tax Credit	All	740	-74	-108	-16
	1	590	-298	-332	-22
	2	964	-49	-136	-50
	3	999	-20	-95	-27
	4	1,026	-59	-108	-31
	5	971	-37	-74	-17
	6	928	-7	-31	-8
	7	812	14	-2	-12
	8	657	-20	-35	-1
	9	348	-166	-170	7
	10	109	-98	-96	5

Sources: 2011 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC); IRS 1040, W-2, 1099-R extracts for tax year 2010; IRS extensive 2010 tax data; TAXSIM.
Approved for release by the U.S. Census Bureau, authorization numbers CBDRB-FY20-ERD002-014, CBDRB-FY20-ERD002-038.

Note: We drop families with no PIKed members and families with any whole-imputed individuals, adjusting the individual survey weights using inverse probability weighting. The CPS Tax Imputation uses income and tax components imputed by the Census Bureau. CPS Data and TAXSIM Imputation uses income and tax components imputed by TAXSIM using CPS inputs. Limited Tax Data refers to tax calculations using TAXSIM with inputs generated from the limited tax data linked to the CPS. The Extensive Tax Data calculation obtains federal income tax liabilities and its components directly from the extensive tax data and uses TAXSIM to generate tax liabilities and credits for CPS individuals not linked with an extensive tax data tax unit. Federal income tax liabilities are total tax (line 55) + additional tax on retirement (line 58) - refundable credits (line 63, 64a, 65, 66, 67, 70, 71). Family pre-tax money income decile is based on total family income equivalized to represent a two adult, two child family.

Table 3a. Mean Absolute Difference Between Imputations and Extensive Tax Calculation, 2010 Tax Year, CPS Data

Income or Tax component	Quartile of Equivalized Survey Family Income	Extensive Tax Data Calculation	CPS Tax Imputation		CPS Data and TAXSIM Imputation		Limited Tax Data Imputation	
		Mean	*Mean absolute difference*	*% of Column (1)*	*Mean absolute difference*	*% of Column (1)*	*Mean absolute difference*	*% of Column (1)*
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Federal income tax liability	All	6,824	7,202	105.5	7,434	108.9	1,654	24.2
	1	-1,561	2,741	175.6	2,822	180.8	565	36.2
	2	184	3,258	1,768.0	3,259	1771.2	734	399.1
	3	5,159	5,285	102.4	5,429	105.2	1,349	26.1
	4	23,520	17,530	74.5	18,230	77.5	3,966	16.9
State income tax liability	All	2,095	1,367	65.3	1,415	67.5	342	16.3
	1	100	420	421.0	400	400.0	161	160.6
	2	752	691	91.9	693	92.2	219	29.1
	3	1,957	1,131	57.8	1,174	60.0	291	14.9
	4	5,571	3,226	57.9	3,394	60.9	699	12.5
Payroll tax liability	All	4,098	1,521	37.1	1,651	40.3	124	3.0
	1	1,131	669	59.1	747	66.0	66	5.9
	2	2,591	1,009	38.9	1,176	45.4	75	2.9
	3	4,564	1,543	33.8	1,700	37.2	102	2.2
	4	8,106	2,862	35.3	2,981	36.8	253	3.1
Total tax liability	All	13,020	9,385	72.1	9,870	75.8	1,913	14.7
	1	-330	3,225	976.4	3,232	979.4	712	215.7
	2	3,527	4,530	128.4	4,601	130.5	925	26.2
	3	11,680	7,401	63.4	7,832	67.1	1,565	13.4
	4	37,190	22,380	60.2	23,810	64.0	4,450	12.0
Adjusted Gross Income	All	73,680	32,050*	43.5	34,670	47.1	790	1.1
	1	23,150	15,100	65.2	16,190	69.9	525	2.3
	2	44,170	18,410	41.7	21,080	47.7	527	1.2
	3	74,800	26,800	35.8	29,760	39.8	520	0.7
	4	152,600	67,900	44.5	71,650	47.0	1,587	1.0
Taxable income	All	48,010	27,280	56.8	28,780	59.9	4,924	10.3
	1	8,724	8,655	99.2	8,580	98.3	1,104	12.7
	2	20,760	14,140	68.1	14,980	72.2	1,952	9.4
	3	47,170	23,950	50.8	25,830	54.8	4,399	9.3
	4	115,400	62,370	54.1	65,740	57.0	12,240	10.6

(continued on next page)

Table 3a. Mean Absolute Difference Between Imputations and Extensive Tax Calculation, 2010 Tax Year, CPS Data (Continued)

Income or Tax component	Quartile of Equivalized Survey Family Income	Extensive Tax Data Calculation	CPS Tax Imputation		CPS Data and TAXSIM Imputation		Limited Tax Data Imputation	
		Mean	*Mean absolute difference*	*% of Column (1)*	*Mean absolute difference*	*% of Column (1)*	*Mean absolute difference*	*% of Column (1)*
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Federal income tax before credits	All	9,221	6,366	69.0	6,553	71.1	1,438	15.6
	1	1,427	1,447	101.4	1,413	99.0	327	22.9
	2	3,016	2,512	83.3	2,413	80.0	388	12.9
	3	7,436	4,637	62.4	4,903	65.9	1,018	13.7
	4	25,010	16,870	67.5	17,480	69.9	4,018	16.1
Earned Income Tax Credit	All	733	550	75.1	569	77.6	14	1.9
	1	1,642	1,094	66.6	1,157	70.5	31	1.9
	2	925	763	82.5	774	83.7	15	1.6
	3	283	266	93.9	268	94.8	5	1.9
	4	80	77	96.5	76	95.4	4	4.5
Child Tax Credit	All	740	275	37.2	306	41.3	43	5.8
	1	825	373	45.2	421	51.1	74	9.0
	2	994	293	29.5	341	34.3	55	5.5
	3	838	238	28.4	261	31.2	29	3.5
	4	304	197	64.6	200	65.9	12	4.1

* Alternatively, we calculated a winsorized mean absolute error for AGI (where we set mean absolute errors for greater than the 99th percentile equal to the 99th percentile), and this winsorized mean absolute error is equal to $26,420.

Sources: 2011 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC); IRS 1040, W-2, 1099-R extracts for tax year 2010; IRS extensive 2010 tax data; TAXSIM. Approved for release by the U.S. Census Bureau, authorization numbers CBDRB-FY20-ERD002-014, CBDRB-FY20-ERD002-038.

Note: We drop families with no PIKed members and families with any whole-imputed individuals, adjusting individual survey weights using inverse probability weighting. The CPS Tax Imputation uses income and tax components imputed by the Census Bureau. CPS Data and TAXSIM Imputation uses income and tax components imputed by TAXSIM using CPS inputs. Limited Tax Data refers to tax calculations using TAXSIM with inputs generated from the limited tax data linked to the CPS. The Extensive Tax Data calculation obtains federal income tax liabilities and its components directly from the extensive tax data and uses TAXSIM to generate tax liabilities and credits for CPS individuals not linked with an extensive tax data tax unit. Federal income tax liabilities are total tax (line 55) + additional tax on retirement (line 58) - refundable credits (line 63, 64a, 65, 66, 67, 70, 71). Family pre-tax money income decile is based on total family income equivalized to represent a two adult, two child family.

Table 3b. Median Absolute Difference Between Imputations and Extensive Tax Calculation, 2010 Tax Year, CPS Data

Income or Tax component	Quartile of Equivalized Survey Family Income	Extensive Tax Data Calculation	CPS Tax Imputation		CPS Data and TAXSIM Imputation		Limited Tax Data Imputation	
		Mean	*Median absolute difference*	*% of Column (1)*	*Median absolute difference*	*% of Column (1)*	*Median absolute difference*	*% of Column (1)*
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Federal income tax liability	All	6,824	2,265	33.2	2,442	35.8	238	3.5
	1	-1,561	700	44.8	723	46.3	2	0.1
	2	184	1,454	790.2	1,475	801.6	92	50.0
	3	5,159	2,430	47.1	2,607	50.5	471	9.1
	4	23,520	6,763	28.8	7,444	31.6	1,481	6.3
State income tax liability	All	2,095	260	12.4	261	12.5	0	0
	1	100	0	0	0	0	0	0
	2	752	182	24.2	172	22.8	0	0
	3	1,957	411	21.0	424	21.7	31	1.6
	4	5,571	1,065	19.1	1,170	21.0	163	2.9
Payroll tax liability	All	4,098	563	13.7	597	14.6	0	0
	1	1,131	207	18.3	225	19.9	0	0
	2	2,591	448	17.3	499	19.3	0	0
	3	4,564	723	15.8	759	16.6	0	0
	4	8,106	1,216	15.0	1,199	14.8	0	0
Total tax liability	All	13,020	3,116	23.9	3,345	25.7	446	3.4
	1	-330	1,000	303.0	1,032	312.7	30	9.2
	2	3,527	2,196	62.3	2,270	64.4	237	6.7
	3	11,680	3,679	31.5	3,961	33.9	652	5.6
	4	37,190	8,981	24.1	10,200	27.4	1,693	4.6
Adjusted Gross Income	All	73,680	10,090	13.7	11,210	15.2	0	0
	1	23,150	4,405	19.0	5,096	22.0	0	0
	2	44,170	7,801	17.7	8,997	20.4	0	0
	3	74,800	11,680	15.6	12,710	17.0	0	0
	4	152,600	23,160	15.2	24,470	16.0	0	0
Taxable income	All	48,010	8,660	18.0	9,621	20.0	0	0
	1	8,724	0	0	0	0	0	0
	2	20,760	6,593	31.8	7,345	35.4	0	0
	3	47,170	12,610	26.7	14,220	30.1	628	1.3
	4	115,400	27,280	23.6	30,130	26.1	6,368	5.5

(continued on next page)

Table 3b. Median Absolute Difference Between Imputations and Extensive Tax Calculation, 2010 Tax Year, CPS Data (Continued)

Income or Tax component	Quartile of Equivalized Survey Family Income	Extensive Tax Data Calculation	CPS Tax Imputation		CPS Data and TAXSIM Imputation		Limited Tax Data Imputation	
		Mean	*Median absolute difference*	*% of Column (1)*	*Median absolute difference*	*% of Column (1)*	*Median absolute difference*	*% of Column (1)*
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Federal income tax before credits	All	9,221	1,351	14.7	1,391	15.1	3	0.0
	1	1,427	0	0	0	0	0	0
	2	3,016	968	32.1	888	29.5	2	0.1
	3	7,436	1,916	25.8	2,194	29.5	136	1.8
	4	25,010	6,049	24.2	6,737	26.9	1,339	5.4
Earned Income Tax Credit	All	733	0	0	0	0	0	0
	1	1,642	146	8.9	125	7.6	0	0
	2	925	0	0	0	0	0	0
	3	283	0	0	0	0	0	0
	4	80	0	0	0	0	0	0
Child Tax Credit	All	740	0	0	0	0	0	0
	1	825	0	0	0	0	0	0
	2	994	0	0	0	0	0	0
	3	838	0	0	0	0	0	0
	4	304	0	0	0	0	0	0

Sources: 2011 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC); IRS 1040, W-2, 1099-R extracts for tax year 2010; IRS extensive 2010 tax data; TAXSIM. Approved for release by the U.S. Census Bureau, authorization numbers CBDRB-FY20-ERD002-014, CBDRB-FY20-ERD002-038.

Note: We drop families with no PIKed members and families with any whole-imputed individuals, adjusting individual survey weights using inverse probability weighting. The CPS Tax Imputation uses income and tax components imputed by the Census Bureau. CPS Data and TAXSIM Imputation uses income and tax components imputed by TAXSIM using CPS inputs. Limited Tax Data refers to tax calculations using TAXSIM with inputs generated from the limited tax data linked to the CPS. The Extensive Tax Data calculation obtains federal income tax liabilities and its components directly from the extensive tax data and uses TAXSIM to generate tax liabilities and credits for CPS individuals not linked with an extensive tax data tax unit. Federal income tax liabilities are total tax (line 55) + additional tax on retirement (line 58) - refundable credits (line 63, 64a, 65, 66, 67, 70, 71). Family pre-tax money income decile is based on total family income equivalized to represent a two adult, two child family.