Imagine All the People:
Using Tax Data to Build a Representative Sample of the US Population

Peter J. Brady
Steven Bass
Investment Company Institute*
1401 H Street N.W.
Washington, DC 20005
pbrady@ici.org

Draft: April 24, 2023

Abstract

This paper documents our method of building a representative sample of the US population from tax data, the first step in larger research project measuring changes in the amount and composition of income over the life cycle. We supplement tax return data—which allows us to identify filers and dependent nonfilers—with information return data—which allows us to identify non-dependent nonfilers. The share of the population who are nonfilers increases with age, making their inclusion crucial for measuring elderly income. Our work builds on the existing literature, especially Cilke (2014) and Lurie and Pearce (2021). The data captures a larger share of the population than most previous studies using tax data because, following Lurie and Pearce (2021), we include in our sample individuals identified only on Form 1095, which reports health insurance coverage. Innovations in our method include sampling individuals rather than tax returns and using the individual rather than the tax return or family as the unit of observation. Our method allows us to examine individual income by single year of age—which is difficult, if not impossible, to do for tax returns or families. Inclusive of both filers and nonfilers, roughly 95 percent of the population older than 26 in 2016 has income. Among those with income, median income follows a hump shape over the lifecycle, peaking at $41,000 per individual at age 46. We find that tax data produce a population estimate similar to that of the U.S. Census Bureau (Census) and that the two estimates track closely by age. Our examination of 2010 and 2016 data show that, across years, differences between tax data and Census estimates are more correlated with birth year than they are by age.

* This research was conducted as part of the Statistics of Income Joint Research Program. Views presented are those of the authors and do not necessarily represent the views of the Internal Revenue Service or the views of the Investment Company Institute or its members. We thank Kevin Pierce for his assistance with this project.

1. Introduction

This paper explains our method of building a representative sample of the US population from tax data. The construction of this data is the first step in a larger project measuring changes in the amount and composition of income over the life cycle. We document our method both for readers of our broader research, so they understand how we derive our estimates, and for other researchers analyzing tax data, so that they can use or improve upon our methods.

Our method of estimating the population from tax data builds on the work of many at the US Department of the Treasury Office of Tax Analysis (OTA), the Joint Committee on Taxation (JCT), and the Internal Revenue Service Statistics of Income Division (SOI)—in particular Cilke (2014) and Lurie and Pearce (2021).[1] The information returns we use to identify nonfilers are largely the same as those used in Cilke (2014), with the addition of information returns created after 2010. The most important of these new information returns is the Form 1095 series (inclusive of Forms 1095-A, 1095-B, and 1095-C) which reports health insurance coverage and is the focus of the analysis in Lurie and Pearce (2021).[2]

The goal of this research is to use tax data to create independent estimates of the US population and their income, with both the population of interest and the measure of income based on the rules of the federal income tax. Our population of interest is generally individuals who, if required to file a tax return, would file a Form 1040, and those individuals' dependents. This would include US residents, but it would also include members of the US armed forces stationed overseas and US citizens or resident aliens living outside the US. Our measure of income generally follows the definition of income in the Internal Revenue Code. The only exceptions are that we include some items defined in the tax code but excluded from taxable

[1] Early work trying to measure the nonfiling population include Cilke (1998), Sailer and Weber (1998), Mortenson et al. (2009), and Lawrence et al. (2011). Examples of other recent studies using tax data to represent the US population include Saez (2016) and Larrimore et al. (2019).

[2] We wish to acknowledge and thank Ithai Lurie and James Pearce for providing us with the data they processed and analyzed in Lurie and Pearce (2021). Form 1095 data are complex and require careful processing to be usable. See Lurie and Pearce (2021) for a description of how they identified individuals with health insurance coverage using Form 1095.

income—such as tax-exempt interest and Roth-IRA distributions—and exclude others that are included in taxable income—such as Roth contributions.

Our method of estimating the population differs from the typical approach in that we create a sample of individuals rather than a sample of tax returns. Our full sample consists of three subsamples—one for filers (primary or secondary taxpayers listed on a return), one for dependent nonfilers, and one for non-dependent nonfilers. Multiple individuals from a single tax return would be included in the sample only if they independently meet our sampling criterion.

We can take this sampling approach because the individual—rather than the tax return, family, or household—is our unit of analysis. There is no need to link primary taxpayers to secondary taxpayers, and no need to link taxpayers to the dependents they claim. All income reported on tax returns is allocated to filers—the primary taxpayer in the case of non-joint returns and the primary and secondary taxpayers in the case of joint returns. Dependents have income only if they file their own return or have income reported on information returns. This approach allows us to tabulate individual income by single year of age.

Similar to the findings in the existing literature, filers and dependent nonfilers identified on tax returns represent the bulk (87 percent) of the population.[3] Like Lurie and Pearce (2021), however, the use of Form 1095 does allow us to identify more non-dependent nonfilers than is typical.

Although non-dependent nonfilers are not a large part of the overall population, their importance increases with age, making their inclusion crucial for measuring elderly income. From about five percent of the population through age 16, non-dependent nonfilers increase to 11 percent by age 24, 16 percent by age 60, and 30 percent by age 80.

Inclusive of both filers and nonfilers, the share of the adults with income remains fairly steady by age, while median income follows a hump shape over the life cycle. The share of the population with income averages 94 percent from age 27 through 61 and it averages 96 percent

[3] See, for example, Cilke (2014) and Larrimore et al. (2019).

from age 62 through 99. Among those with income, median income increases rapidly early in life, peaks at $41,000 at age 46, and then declines with age—to $38,000 at age 60 and $34,000 at age 70.

To get a sense of how well the tax data capture the US population, we compare our estimates to those of the U.S. Census Bureau (Census). Before we make the comparison, we the adjust our data to better conform to the Census population concept. Consistent with existing research, we find the two population estimates are similar and track closely by age. We also find some age ranges over which the two estimates diverge noticeably, but the ages differ from those noted in earlier studies. Replicating the results of the previous studies, we find that differences between tax data and Census estimates are more consistent over time by birth year than they are by age.

The paper is organized as follows. Section 2 describes the data we analyze. Section 3 defines our unit of analysis, measure of income, and population of interest. Section 4 describes our method of constructing our representative sample of the population. Section 5 examines the composition and income of the population by single year of age. Section 6 compares our estimates to Census population estimates. Section 7 summarizes our results and concludes the paper.

2. Description of the Data

This study uses Internal Revenue Service (IRS) administrative tax data from tax year 2016. These data include information from both federal individual income tax returns filed by taxpayers and information returns issued by third parties and sent to both taxpayers and the IRS. We also incorporate Social Security Administration (SSA) data on gender, date of birth, and date of death (if applicable).

In general, US citizens and resident aliens[4] are required to file a Form 1040 (inclusive of Form 1040, Form 1040-A, and Form 1040-EZ) if their gross income is above a certain threshold.[5] The filing requirement applies even if a US citizen or resident alien lives outside the US during the tax year.[6] An exception to this rule is US citizens or resident aliens living in US possessions, who generally file a return with their territory's tax authority rather than with the IRS.[7]

US citizens and resident aliens with gross income below the filing thresholds may also file. Regardless of gross income, filing is required for individuals who received any advance earned income tax credit (EITC), who owe payroll taxes that were not withheld, or who owe special taxes (such as the alternative minimum tax). In addition, although not required to file, individuals with gross income below the thresholds can only receive refunds of withheld taxes or refundable tax credit payments if they file a return.

Although they would not file a Form 1040, residents of US possessions and nonresident aliens may file a different type of tax return with the IRS. Residents of US possessions with self-employment earnings generally must file a Form 1040-SS or Form 1040-PR and, if necessary, pay self-employment tax.[8] Nonresident aliens who were engaged in a trade or business in the US or who had US-source income are required to file a Form 1040-NR.

---

[4] Resident aliens include individuals with a green card or who had a "substantial presence" in the US—inclusive of the 50 US states and the District of Columbia. For more information on US income tax treatment of both resident and nonresident aliens, see Internal Revenue Service (2017a).

[5] Gross income includes all income not exempt from tax. It includes both US-source and foreign-source income. For self-employed individuals providing services, gross income includes the gross receipts of the business. The gross income filing thresholds vary based on filing status, age, blindness, and whether a parent (or another taxpayer) can claim the individual as a dependent. For example, the 2016 filing thresholds for non-dependent single individuals younger than age 65 was $10,350 and the threshold for a non-dependent married couple filing a joint return with both spouses younger than 65 was $20,700. The threshold was $1,550 higher for single individuals aged 65 or older and $2,500 higher for joint filers with both spouses aged 65 or older, respectively. For more information on filing requirements, see Internal Revenue Service (2016b) and Internal Revenue Service (2016c).

[6] For filing requirement for US citizens and resident aliens living abroad, see Internal Revenue Service (2016a).

[7] US citizens and resident aliens who are bona fide residents of Guam, the US Virgin Islands, and the Northern Mariana Islands are not required to file a Form 1040 with the IRS. US citizens and resident aliens who are bona fide residents of American Samoa and Puerto Rico are only required to file a Form 1040 if they received income from a source outside of the territory. This would include US government employees, as their wages are considered US source income. For the definition of a bona fide resident and more information on filing requirements for individuals with income from US possessions, see Internal Revenue Service (2017c).

[8] Bona fide residents of Guam, American Samoa, the US Virgin Islands, the Northern Mariana Islands, and Puerto Rico must file a Form 1040-SS if they have $400 or more of net self-employment earnings and are not required to file

Tax return data include both information reported directly on Form 1040 and information reported on associated schedules (such as Schedule SE, which is used to calculate self-employment tax) and forms (such as Form 6251, which is used to calculate the alternative minimum tax) filed with the tax return.

In addition to tax return data, we incorporate data from various information returns, which are issued by third parties and sent to the taxpayer and to the IRS. Information returns allow us to identify nonfilers and measure their income. Information returns also allow us to allocate income reported by married couples on joint tax returns to the spouse who received it. And, in some cases, they provide information not reported on Form 1040—such as detailed codes for distributions from IRAs, pensions, and annuities.

To identify nonfilers, we use a large number of information returns. Many of the information returns report income—such as Form W-2, which reports wages, or Form 1099-INT, which reports interest income. Other information returns report expenses—such as Form 1098, which reports mortgage interest expense, or Form 1098-T, which reports tuition expense—or other tax-relevant information—such as Form 1095-B, which reports health insurance coverage. For a full listing of information returns used in this study and their description, see Appendix Table A.1.

3. Conceptual Issues

The goal of this research is to illustrate how the tax data can be used to create independent estimates of the size and income of the US population. As such, we are not attempting to match concepts used for existing income estimates derived from household surveys. More specifically, we do not alter the tax data to match the unit of analysis, the measure of income, or the population of interest used in the official income statistics published by the Census, which are derived from the Current Population Survey (CPS) Annual Social and Economic Supplement

Form 1040 with the IRS. Residents of Puerto Rico have the option to file a Form 1040-PR (which is in Spanish) instead. In addition to facilitating tax collection, the information reported on these forms is used by the SSA for calculating Social Security benefits. For a complete description of who must file Form 1040-SS/1040-PR, see Internal Revenue Service (2017c).

(ASEC). If the reader is primarily interested in comparing income measures derived from tax data with income measures derived from the CPS, see Brady and Pierce (2012), Bee and Mitchell (2017), Larrimore et al. (2019), and Brady and Bass (2021).

3.1 Unit of Analysis

We use the individual as the unit of analysis[9] rather than the tax return,[10] the family[11]—that is, the combination of all individuals reported on a tax return, inclusive of primary taxpayers, secondary taxpayers, and dependents (along with their income)—or the household[12]—that is, the combination of all families (along with their income) living in the same household. As such, we include in our sample non-dependent filers, dependent filers, dependent nonfilers, and non-dependent nonfilers.

We do not attempt to use the tax data to construct households, as is done in Larrimore et al. (2019); nor do we classify dependents using the income reported by the filers who claim them, as is done in Lurie and Pearce (2021). Non-dependent filers are classified by the income reported on their returns. Non-dependent nonfilers are classified by the income reported on information returns. Dependents are classified by their own income as reported on either a dependent return or on information returns.

The primary reason we use the individual as our unit of observation is that the focus of our research, using both the cross-sectional data described in this paper and panel data, is measuring changes in the amount and composition of income over the life cycle. The composition of an individual's tax return, family, or household rarely remains stable across the

[9] The individual is the unit of analysis used in Brady et al. (2017), Brady and Bass (2020a), Brady and Bass (2020b), and Brady and Bass (2021).

[10] The IRS Statistics of Income Division (SOI) uses the tax return (inclusive of both non-dependent and dependent tax returns) as the unit of analysis in their annual Complete Report. See, for example, Internal Revenue Service, Statistics of Income Division (2018).

[11] The Joint Committee on Taxation (JCT) and the US Department of Treasury Office of Tax Analysis (OTA) use the family as the unit of analysis in their distributional analyses (Joint Committee on Taxation 1993, Cronin 1999).

[12] The Congressional Budget Office (CBO) uses the household as the unit of analysis in its distributional analyses, as does Larrimore et al. (2019), which also includes nonfilers. CBO creates households for their microsimulation model using a combination of tax data and CPS data. Larrimore et al. (2019) uses the mailing address on tax returns and information returns to construct household measures of income from tax data.

life cycle. It is also difficult, if not impossible, to use any other of unit of analysis to study differences in income by age.

3.2 Measure of Income

Our measure of **total income** is the sum of six types of income: **labor** (wage and salary, self-employment earnings, unemployment compensation), **Social Security** (disability benefits and retirement benefits), **retirement** (IRA distributions and income from pensions and annuities), **investment** (taxable interest, tax-exempt interest, dividends, gains/losses), **business/farm/rents/royalties** (business and farm income in excess of self-employment earnings; income from rents, royalties, partnerships, S-corps, and trusts), and **other** (net alimony [alimony received less alimony paid] and other income).

In addition to total income and its components, we also measure **spendable income**—that is total income less federal income and payroll taxes.

While our total income measure is created from tax data, it differs from the tax code's definition of income. We include some types of income excluded from taxable income—such as tax-exempt interest and the nontaxable portion of Social Security benefits—and exclude some types of income included in taxable income—such as taxable state income tax refunds. In addition, we also treat, to the extent possible, all contributions and distributions from retirement plans the same--regardless of whether contributions were from an employer or an employee, and regardless of their tax treatment. This means we exclude from income not only tax-deferred employee contributions to employer plans and IRAs, but also Roth contributions and non-Roth after-tax contributions. It also means we include in income not only taxable non-Roth distributions, but also Roth distributions and the portion of non-Roth distributions that represents basis. Appendix Table A.2 describes how we calculate these income and tax measures for both filers and nonfilers.

The definition of income used in this paper also differs from previous studies focused on comparing tax data to household survey data. For example, Brady and Pierce (2012) and Brady and Bass (2021) measure four types of income which are measured in both data sources, and which have been found to be accurately reported on income tax returns: wage and salary

income, Social Security benefits, retirement income, and investment income excluding gains/losses (taxable interest, tax-exempt interest, and dividends). Bee and Mitchell (2017) use those four sources of income plus Supplemental Security Income (SSI). Larrimore et al. (2019) use a definition like the one used in this study but exclude realized capital gains/losses reported on Schedule D and bottom code income at zero to limit the effect of business losses.

The primary measure of income we use to analyze the incidence and amount of income by type is *per capita income*, which allocates the income reported on joint tax returns equally to each spouse. For primary and secondary taxpayers on joint returns, per capita income is the income derived from the tax return divided by two. For primary taxpayers on non-joint tax returns, per capita income is simply the income derived from the tax return. Similarly, for nonfilers, who are all assumed to be non-joint, per capita income is simply the amount derived from an individual's information returns.

In addition to per capita income, we also report *own income* for labor, Social Security, and retirement income. For joint filers, we use information returns to allocate the income reported on joint returns to the spouse who received the income. For non-joint filers and for nonfilers, there is no difference between own income and per capita income.

Importantly, neither per capita income nor own income adjusts for family size.[13] All income reported on a tax return is allocated to filers—the primary taxpayer in the case of non-joint returns and the primary and secondary taxpayers in the case of joint returns. The number of dependents claimed on a tax return has no impact on either measure, as no filer income is allocated to dependents. Dependents only have measured income if (1) they file their own tax return or (2) they do not file their own return, but they have income reported on an information return. For dependent filers, the calculation of income is the same as it is for non-dependent

[13] The discussion in the text is limited to family-size adjustments solely for brevity; the discussion also applies to household-size adjustments. Some, but not all, distributional analysis adjust income for family or household size. JCT does not adjust for family size before ranking families by income (Joint Committee on Taxation 1993). In contrast, OTA adjusts for family size and CBO adjusts for household size before ranking by income (Cronin 2022, Congressional Budget Office 2016).

filers. For dependent nonfilers, the calculation of income is the same as it is for non-dependent nonfilers.

We do not adjust our income measures for family size, in part, because we do not characterize our annual income measures as reflecting either well-being or the ability to pay taxes. Compared with the typical adjustment for family size—dividing family income by the square root of the number of individuals in the family[14]—we believe the transparency of our per capita income measure simplifies the interpretation of changes in the amount and composition of income by age, and that it simplifies the interpretation for both filers and dependents. We also note that both well-being and ability to pay depend on many factors other than annual income and the presence of children in the current year, and that the impact of children is unlikely to be properly captured by the typical family-size adjustment in any case.[15] That said, we encourage the reader to consider how the amount and composition of an individual's spending may change with age—including spending related to raising children—when interpreting changes in the amount and composition of per capita income over the lifecycle.

3.3 The Population of Interest

Our goal is to build a sample of US citizens and resident aliens who—provided their gross income exceeded the filing thresholds—would be required to file a 2016 Form 1040, excluding residents of US territories. This would include US citizens and resident aliens living in a state (inclusive of the 50 states and the District of Columbia), living outside the US, or living overseas

[14] Before ranking by income, both OTA and CBO divide income by the square root of the number of individuals in the family/household in their distributional analysis (Cronin 2022, Congressional Budget Office 2016), as does Larrimore et al. (2019).

[15] The rationale both for using the family or household (rather than the individual or tax return) as the unit of analysis, and for dividing family/household income by the square root of family/household size before ranking by income (rather than simply ranking by family income sans the size adjustment or ranking by family income divided by family size sans the square root), is that it produces a measure of income that better reflects well-being and/or the ability to pay taxes. This functional form is consistent with certain assumptions about economies of scale in family size. That is, it is consistent with the belief that two individuals living together cannot live as cheaply as one but can live more cheaply than two individuals living on their own. It is also consistent with the belief, common in movies, that children are "cheaper by the dozen"—that is, that the marginal cost of a second child is less than the marginal cost of the first, the marginal cost of a third child is less than the marginal cost of the second, and so on. While we believe it is reasonable to assume there are economies of scale in family expenses, we find it unlikely that (1) the square root of family/household size (conveniently) captures this effect or (2) that economies of scale are proportional to income.

as a member of the US armed forces. This would exclude nonresident aliens engaged in a trade or business in the US or who had US-source income—who, if required, would file a Form 1040-NR—and self-employed residents of US possessions—who, if required, would file a Form 1040-SS/1040-PR to pay self-employment taxes. We also exclude any residents of US territories who file a Form 1040 because bona fide residents of US territories generally do not file an income tax return with the IRS.[16]

Our population of interest differs from the populations represented in official measures of the US population and income. For example, Census releases annual estimates of the US resident population (inclusive of the 50 states and the District of Columbia).[17] Compared with our population, the US resident population excludes US citizens and resident aliens living abroad as well as members of the US armed forces stationed overseas. Census also produces an annual report on the income of the US resident civilian, noninstitutionalized population.[18] In addition to those excluded from the US resident population, this would exclude certain members of the US armed forces stationed in the US and individuals living in nursing homes, long-term care facilities, and other institutions.[19]

Our population of interest also differs from previous studies that use tax data to build a representative sample of the US population. For example, Cilke (2014), Lurie and Pearce (2021), and Larrimore et al. (2019) all attempt to represent the US resident population. They generally include only tax information associated with a US address (inclusive of the 50 states and the District of Columbia). This means, compared with our population, these studies exclude taxpayers and dependents listed on tax returns with an overseas US armed forces address or a foreign or missing address. They also exclude non-dependent nonfilers with an overseas US

[16] See note 7.

[17] See, for example, U.S. Census Bureau (2021).

[18] See, for example, Semega et al. (2019).

[19] Members of the US Armed Forces are only included in the CPS sample if they live in a household with at least one other adult who is a civilian.

armed forces address on their information returns, and generally have more stringent criteria than this study for excluding non-dependent nonfilers with foreign or missing addresses.[20]

4. Creating the Representative Sample

Our method of sampling individuals differs from the typical method of sampling tax returns. For example, the Individual and Sole Proprietor (INSOLE) file, which is used by the SOI to produce its annual *Individual Income Tax Returns Complete Report* publication and used by the JCT and OTA as the basis for their microsimulation models, is a sample of tax returns.[21] Similarly, the panel data used in analysis and as the basis of the OTA's panel model is also a sample of tax returns (Nunns et al. 2008). In these data, the sample consists of tax returns and information for all individuals on a selected tax return—primary taxpayers, secondary taxpayers, and dependents—is included in the data. In contrast, our sample consists of individuals, with multiple individuals from a given tax return included in our sample only if they independently meet our sampling criterion.

We can sample individuals rather than tax returns because individuals are the focus of our analysis. We do not use the tax data to construct families or households, so there is no need to link dependents to the taxpayers who claim them. All income reported on tax returns is allocated to the filers. In the case of joint filers, our per capita income measures allocate income equally to each spouse, while our own income measures (for labor, Social Security, and retirement income) use information returns to allocate the tax return income to the spouse who received it. The number of dependents listed on a tax return does not affect our measure of filer income, nor does filer income affect our measure of income for dependents listed on the return.

[20] We only want to include non-dependent nonfilers living abroad if they are US citizens or resident aliens. As discussed below, we include non-dependent nonfilers only if they have a US address (in one of the 50 states or the District of Columbia) or an overseas US armed forces address on at least one information return. In contrast, Cilke (2014) excludes individuals with only non-US addresses on their information returns (even individuals with only overseas US armed forces addresses) and excludes individuals with both US and non-US addresses on their information returns if taxable income reported on an information return sent to a non-US address exceeds a given threshold.

[21] For a description of the INSOLE sample, see Internal Revenue Service, Statistics of Income Division (2018).

We build a sample representative of the 2016 US population in three steps. In each step we generally only include individuals with a valid Social Security Number (SSN) or Taxpayer Identification Number (TIN) in the SSA data.

- First, we draw a *filer* sample: primary and secondary taxpayers who file a tax return.
- Second, we draw a *dependent nonfiler* sample: individuals who did not file a return but were claimed as dependents by a taxpayer who did file.
- Third, we draw a *non-dependent nonfiler* sample*:* individuals who did not file a return and who were not claimed as a dependent by a taxpayer who did file a return, but who receive at least one information return.

4.1 Filer Sample

To construct our *filer* sample, we first pull a sample of individuals appearing as primary or secondary taxpayers on tax-year 2016 tax returns, inclusive of Form 1040, Form 1040-SS/ 1040-PR, and Form 1040-NR. Primary taxpayers on non-joint returns are selected if they have an SSN or TIN that ends in one of 500 unique 4-digit numbers and the SSN/TIN is valid. Primary and secondary taxpayers on joint returns are selected if they have an SSN or TIN that ends in one of 500 unique 4-digit numbers and the SSN/TIN of either the primary or secondary taxpayer on their return is valid.[22] Although return and spousal information is retained for individuals selected from joint returns, spouses are not generally included in the sample (unless their SSNs/TINs independently meet our criteria for inclusion). This method results in a roughly 5.0 percent sampling rate of primary and secondary taxpayers.[23] The resulting sample represents 206 million primary and secondary taxpayers (Table 1).[24]

[22] As explained in Cilke (2014), the IRS uses a number of methods to validate an SSN/TIN reported on a tax form, with the primary method checking to see if the first four letters of the taxpayer's last name match the first four letters of the last name associated with the SSN/TIN. Married individuals who have recently changed their last name may be coded as invalid until their Social Security Administration records are updated.

[23] The last four digits of the SSN/TIN range in value from 0001 to 9999, resulting in a sampling rate slightly over 5.0 percent (=500/9,999).

[24] Sample weights are calculated as the inverse probability of selection. Taxpayers receive a weight of just under 20 (=9,999/500).

For analysis, we focus on primary and secondary taxpayers from 2016 Form 1040 who did not die prior to 2016, excluding residents of US possessions. This excludes individuals representing fewer than 2,000 taxpayers on 2016 tax returns who died before January 1, 2016. It also excludes individuals representing 748,000 non-resident aliens who filed a Form 1040-NR, 241,000 individuals who filed a Form 1040-SS/1040-PR, and 94,000 residents of US possessions who filed a Form 1040 (Table 1).

The final *filer* sample represents 204 million individuals, including 150 million primary taxpayers and 54 million secondary taxpayers (Table 1). Out of the 204 million filers, 687,000 were on returns with a foreign or missing address and 239,000 were on returns filed by US military personnel living overseas.

Among all filers, 98 percent had at least one information return. The most common information returns were Form 1095 (89 percent of taxpayers), Form W-2 (72 percent), and Form 1099-G (33 percent).

4.2 Dependent Nonfiler Sample

To construct our *dependent nonfiler* sample, we first pull a sample of individuals listed as a dependent on a 2016 Form 1040 or Form 1040-NR, or listed as a qualifying child on a Form 1040-SS/1040-PR.[25] The tax data include the SSN/TIN of up to four dependents/qualifying children listed on paper returns and the SSN/TIN of all dependents listed on electronic returns.[26] From this group we select all dependents with an SSN/TIN that ends in one of the same 500 unique 4-digit numbers used to select the filer sample. The resulting sample represents 94 million dependents (Table 2).

For analysis, we focus on dependents listed on 2016 Form 1040 who did not die prior to 2016 and who did not file a dependent return, excluding dependents claimed by residents of US

[25] Qualifying children are only listed on Form 1040-SS/1040-PR by bona fide residents of Puerto Rico who claim the additional child tax credit.

[26] There are two sources for dependent information. For all returns—including both paper returns and electronic returns—the SSN/TIN of the first four dependents listed on the return is reported and the validity of the SSN/TIN is checked. For electronic returns, the SSN/TIN of all listed dependents is reported, but the validity of the SSN/TIN is not checked. For the first four dependents listed on all returns, we include only those with a valid SSN/TIN. For additional dependents listed on electronic returns, we include all (because there is no validity indicator).

possessions. In addition to individuals representing 9,000 dependents who died before 2016, this excludes individuals representing 25,000 dependents claimed on Form 1040-NR, 171,000 qualifying children claimed on Form 1040-SS/1040-PR, and 41,000 dependents claimed on Form 1040 filed by residents of US possessions (Table 2). It also excludes 9.3 million individuals claimed as a dependent who filed their own 2016 Form 1040 (and who were already included in the taxpayer component of the sample).[27]

The final *dependent nonfiler* sample represents 84 million individuals (Table 2). Dependent nonfilers include 260,000 claimed on returns with a foreign or missing address and 121,000 claimed on returns filed by US military personnel living overseas.

Among all dependent nonfilers, 89 percent had at least one information return. The most common information returns were Form 1095 (88 percent of dependent nonfilers), Form SSA-1099 (6 percent), and Form W-2 (6 percent).

Note that the dependent nonfiler sample does not include dependents claimed on paper returns with more than four dependents if they were not among the first four listed. Provided they did not file their own dependent return, these individuals would be included in the non-dependent nonfiler sample if they had at least one tax-year 2016 information return with a valid SSN/TIN. If they neither filed their own return nor had an information return, they would be excluded from the final sample altogether.

4.3 Non-Dependent Nonfiler Sample

To construct our *non-dependent nonfiler* sample, we first pull a sample of individuals who had at least one 2016 information return with a valid SSN/TIN and who were neither a taxpayer who filed, nor a dependent/qualifying child claimed on, a 2016 Form 1040, Form 1040-NR, or

[27] The estimate of dependent filers in the *filer* component of the sample (9.4 million) is greater than the estimate of dependent filers excluded from the *dependent nonfiler* component of the sample (9.3 million). The difference in the estimates is presumably attributable to the fact that not all dependent filers included in the *filer* component of the sample were listed as dependents on a tax return. This is for two reasons. First, some dependent filers may have been claimed as a dependent on a paper return but not included in our dependent sample because they were not among the first four listed on the return. The paper Form 1040 only includes space for up to four dependents. Any additional dependents are reported to the IRS separately. Second, some dependent filers may not have been claimed as dependents, as individuals who could be claimed as a dependent must file a dependent return regardless of whether they were claimed as a dependent or not.

Form 1040-SS/1040-PR. From this group we select all individuals with an SSN/TIN that ends in one of the same 500 unique 4-digit numbers used to select the filer and dependent nonfiler samples. The resulting sample represents 50 million non-dependent nonfilers (Table 3).

For analysis, we focus on non-dependent nonfilers who did not die prior to 2016 and who were likely to be a US citizen or resident alien and who did not reside in a US territory. The largest group excluded from the initial sample were 4.3 million individuals who were sent 2016 information returns but died prior to 2016 (Table 3). We also exclude 3.4 million individuals for whom all information returns were sent to addresses in US territories or to foreign or missing addresses. Finally, we excluded 73,000 individuals who had one of three forms that indicated they were a foreign person.[28]

The final *non-dependent nonfiler* sample represents 42 million individuals (Table 3), including 17,000 who had at least one information return sent to an address for members of the US armed forces living overseas. Of the 42 million non-dependent nonfilers, 31 million (73 percent) had income. The most common information returns were Form 1095 (86 percent of non-dependent nonfilers), Form SSA-1099 (43 percent), and Form W-2 (25 percent).

4.4 Total Population Estimates

The final sample represents 331 million filers, dependent nonfilers, and non-dependent nonfilers (Table 4). Overall, 87 percent of the population are identified using Form 1040 (inclusive of both filers and dependent nonfilers), 10 percent are non-dependent nonfilers identified using at least one non-Form-1095 information return, and 3 percent are identified using Form 1095 only. Three-quarters of the population have income. Most without income are dependent nonfilers.

[28] The three forms are Form 8288-A (Statement of Withholding on Dispositions by Foreign Persons of U.S. Real Property Interests), Form 8805 (Foreign Partner's Information Statement of Section 1446 Withholding Tax), and Form 1042-S (Foreign Person's U.S. Source Income Subject to Withholding). About 85 percent of individuals with Form 8288-A or Form 8805 file a return, with nearly all filing Form 1040-NR. Just over one-third of individuals with Form 8288-A file a return, with over 60 percent filing a Form 1040-NR.

5. Population Composition and Median Income by Age

Dependent nonfilers (individuals identified as a dependent on Form 1040 and who do not file a dependent return) represent most of the population at younger ages, with their share falling rapidly after age 15, remaining low through middle age, and then increasing again at older ages (Figure 1, top panel). From over 95 percent at younger ages, the dependent nonfiler share of the population falls to 92 percent at age 15, 50 percent at age 18, and less than 5 percent at age 26. Their population share continues to fall with age, falling below 2 percent in the late 30s and remaining there through the early 50s. Dependent nonfilers then begin to increase as a share of the population, increasing to 4 percent by age 70 and 7 percent by age 80.

Largely mirroring the decline in dependent nonfilers, filers (individuals who are primary or secondary taxpayers on Form 1040) increase rapidly as a share of the population after age 15 and the share remains high for those from their mid-20s through early 60s before falling off at older ages (Figure 1, top panel). Filers increase from 3 percent of the population at age 15 to 43 percent at age 18 to 84 percent at age 26. Although filers are typically dependent filers at younger ages, non-dependent filers represent nearly one-quarter of filers by age 18 and nearly all filers after age 26 (not shown in Figure 1). The filer population share peaks at 86 percent from age 29 through age 46, remains above 84 percent through age 50, and is still 82 percent at age 60. After age 60, the filer share begins to decline more noticeably, falling to 75 percent by age 70 and 63 percent by age 80.

Non-dependent nonfilers (individuals not identified on a tax return but who receive at least one information return) increase as a share of the population with age (Figure 1, top panel). Non-dependent nonfilers average a bit less than 5 percent of the population from birth through age 15 and then increase to 11 percent of the population by age 26, with the decline in dependent nonfilers not completely offset by the increase in filers over this age range. After age 26, their population share increases slowly to 16 percent at age 60, an average increase of just over 0.1 percentage points per year of age. After age 60, their population share growth accelerates to 0.9 percentage points per year of age, on average, through age 99—increasing to 21 percent by age 70 and 30 percent by age 80.

Despite the filer share of the population falling at older ages, the share of the population with income remains consistently high throughout adulthood, and even increases slightly after age 61 (Figure 1, bottom panel).[29] The share of the population with income increases rapidly for teenagers and young adults, from 12 percent at age 12 to 71 percent at age 18 and 93 percent at age 26. The share generally remains steady for those older than 26—with 94 percent of the population, on average, from age 27 through age 61 and 96 percent of the population, on average, from age 62 through age 99 having income.

To inform comparisons with previous research using only data from tax returns or using only information on nonfilers prior to the existence of Form 1095, Figure 2 shows our population estimates broken down by how we identified those individuals.

As already noted, the bulk of the population is identified using Form 1040—inclusive of filers and dependent nonfilers—but this group represents a declining share of the population with age. Their population share is about 95 percent of the sample through age 16 but falls thereafter—to 89 percent by age 24, 84 percent by age 60, and 70 percent by age 80.

The next largest component of the population is non-dependent nonfilers identified on at least one information return other than Form 1095, who represent a small portion of population at younger ages but who increase in importance with age. This group's population share is less than 1.0 percent through age 13, but increases to 2.0 percent by age 16, and to 8.8 percent by age 24. Their population share growth slows after age 24, increasing to 12 percent by age 60. After age 60, share growth accelerates, exceeding 29 percent by age 80.

The final component of the population—which, to our knowledge, has not been included in analysis of tax data other than in Lurie and Pearce (2021)—is non-dependent nonfilers who are identified using Form 1095 alone. This group represents less than 5.0 percent of the population at all ages below 100, with their highest share among children and middle-aged adults and their lowest share among the elderly. Individuals identified only on Form 1095 represent 4.6 percent

[29] In this study, the presence of income is defined as having nonzero income in any of our broad income categories—*labor*, *Social Security*, *retirement*, *investment*, *business/farm/rents/royalties*, and *other* (see the definition of *total income* in Table A.2).

of the population younger than one year of age, with their share falling slowly to 3.1 percent by age 16, and then falling more quickly to 2.0 percent for those aged 20 through 22. After age 22, their population share begins to increase, peaking above 3.5 percent for individuals aged 52 through 61. After age 61, the share drops sharply, to less than 1.0 percent of the population from age 67 through age 90, before edging up at older ages.

As explained in section 3, our primary measure of income is *per capita*—that is, the income reported by married couples who file joint returns is allocated equally to each spouse. Again, no income is allocated to dependents listed on a tax return. All income reported on a tax return is allocated to filers—the primary and secondary taxpayers in the case of joint returns and the primary taxpayer in the case of non-joint returns. Nonfilers (inclusive of dependent nonfilers and non-dependent nonfilers) are allocated the income reported on their information returns (if any), with no effort made to impute marital status or to create family groups.

The composition of the population by filing-type group—joint filer, non-joint filer, and nonfiler (inclusive of both dependent nonfilers and non-dependent nonfilers)–varies with age, reflecting the fact that the group to which a given individual belongs changes over the life cycle (Figure 3, top two panels).

The nonfiler share of the population falls sharply with age early in life, as more and more have income above the filing threshold, and remains low through age 40. The nonfiler population share then begins to increase with age, particularly after age 61. Along with the elderly typically having lower total income, two tax code provisions reduce the share of the population who are required to file with age. First, Social Security benefits are either fully or

partially excluded from the gross income measure on which the filing requirement is based.[30] Second, the gross income filing threshold is slightly higher for individuals aged 65 or older.[31]

The joint-filer share of the population follows a hump-shaped pattern by age, increasing rapidly after age 18, hitting one-quarter of the population by age 27 and half of the population by age 37 before peaking at 56 percent at age 62. After age 62, the joint-filer share falls—at first slowly and then at an accelerated rate. In addition to spousal death, a portion of the decline in joint-filer share after age 62 may be attributable to a falling share of the population filing a return. Among filers, the share filing a joint return peaks at age 69, where it reaches 71 percent.

The relationship between median per capita total income and age differs by filing-type group (Figure 3, bottom panel).[32] For all but the elderly, joint filers have the highest per capita income. The median per capita income of joint filers increases rapidly at younger ages, peaks at age 46, then declines through about age 80. The median income of non-joint filers increases more slowly than that of joint filers after age 25, but continues to increase slowly through age 62 and then increases sharply between age 62 and age 67. As a result, the median income of non-joint taxpayers is higher than that of joint taxpayers for individuals aged 67 through 98. Nonfilers have the lowest median income at all ages, with median income increasing only modestly through age 40 and then remains roughly constant before drifting higher after age 62.

---

[30] The percentage of Social Security benefit payments included in gross income is based on a taxpayer's modified adjusted gross income (MAGI), which includes half of Social Security benefit payments plus other income included in gross income. For single, head of household, and qualifying widow(er) returns: if MAGI is $25,000 or less, no Social Security benefit payments are included in gross income; if MAGI is between $25,000 and $34,000, the lesser of 50 percent of Social Security benefit payments or 50 percent of MAGI in excess of $25,000 is included in gross income; if MAGI is in excess of $34,000, the lesser of 85 percent of Social Security benefit payments or 85 percent of MAGI in excess of $34,000 plus $4,500 [=50%*($34,000-$25,000)] is included in gross income. For joint returns, the MAGI thresholds are $32,000 and $44,000, respectively. For more information on the taxation of Social Security benefits, see Internal Revenue Service (2017b).

[31] See note 5.

[32] The medians presented in this study are approximate, as true medians could represent disclosure of an individual's tax data. To calculate approximate medians, we average the 48th, 49th, 50th, 51st, and 52nd percentile values and then round that average (to the nearest dollar for amounts less than $100, the nearest $10 for amounts from $100 to less than $10,000, the nearest $100 for amounts of $10,000 or more, and two decimal places for percentages). We then only report these approximate medians for groups with 100 or more observations. We use the same method to report other percentile measures. Others who wish to use this measure for their own research may cite this article for authority or simply refer to the measure as the Brady-Bass Adjusted Median (BBAM).

Within a filing-type group, changes in median income by age reflect not only the typical experience of individuals, but also changes in the composition of the group by age. The income of nonfilers grows more slowly with age than that of filers because of filing requirements, which effectively cap the amount of income an individual can have while remaining a nonfiler. At younger ages, the median income of joint filers increases more quickly with age than that of non-joint filers, at least in part, because single individuals with higher earning potential are more likely to get married.[33] At older ages, the tax treatment of Social Security benefits affects the relative income of all three groups. The upward drift in nonfiler income after age 62 is primarily because our total income measure includes Social Security benefits excluded from gross income,[34] and to a lesser extent because the gross income filing threshold increases for individuals aged 65 or older.[35] The income of joint filers declines relative to that of non-joint filers after age 62, at least in part, because the filing requirement effectively kicks in at lower levels of per capita total income for joint filers: although the gross income filing threshold is roughly twice as high for married couples,[36] the thresholds for determining the share of Social Security benefits included in gross income are not.[37]

Examining the entire population arguably provides a clearer picture of changes in income over the life cycle, as individuals shift between filing-type groups over the course of their lifetimes. For the entire population, conditional on having non-zero total income,[38] median per capita income follows a hump-shaped pattern with age (Figure 3, bottom panel, black line). Income increases rapidly early in life to $30,000 at age 30 and then continues to increase, but at a slower rate, peaking at $41,000 at age 46. After age 46, income declines with age—to $38,000 at age 60 and $34,000 at age 70.

[33] See, for example, Lundberg et al. (2016), which illustrates that, among individuals aged 30 through 44 in 2010, individuals with at least a bachelor's degree were more likely to be married than individuals with less education.

[34] See note 30.

[35] See note 5.

[36] See note 5.

[37] See note 30.

[38] See note 29.

6. Comparison with Census Population Estimates

This section compares our population estimates to those produced by Census to provide a general sense of how well the tax data captures the US population. Although the two populations that are estimated differ, there is enough overlap between the two to make the comparison informative.

The Census produces annual estimates of the US resident population as of July 1. The current 2016 Census estimate is *postcensal*—that is, it is based on the 2010 decennial census plus more recent data on births, deaths, and immigration. The *intercensal* 2016 estimate—that is, an estimate that also incorporate information from the 2020 decennial census—is scheduled to be released in 2023.[39]

As already discussed, our research focuses on a different population of interest than the Census. We are interested in individuals who would file a Form 1040 if required to file a return, excluding residents of US territories. This population includes individuals who reside outside the US. It also includes individuals alive at any point during the year. For example, it will include individuals who died during the year but had a return filed on their behalf. It will also include dependents born as late as December 31.

To make our population estimate from the tax data as comparable as possible to the Census estimate, we remove individuals we identify as living outside the US or who were not alive on July 1 (Table 5). This includes: 0.9 million filers and dependent nonfilers with a foreign or missing address; and 0.4 million filers, dependent nonfilers, and non-dependent nonfilers with an overseas US armed forces address.[40] It also includes 1.3 million individuals who died prior to July 1 and 1.9 million individuals born after June 30. The resulting Census-equivalent population estimate is 326.5 million individuals.

[39] For the release schedule, see https://www.census.gov/programs-surveys/popest/about/schedule.html.

[40] For filers and dependents, we remove them if they have an overseas armed forces, foreign, or missing address on their Form 1040. For non-dependent nonfilers, we remove them if at least one of their information returns has an overseas armed forces address.

The final adjustment we make to our data is to tabulate the data by age as of July 1 (rather than age at the end of the year).

6.1 Possible Explanations for Differences Between Tax Data and Census Estimates

Before we examine differences between the Census-equivalent and Census population estimates, we discuss why the two may differ. First, both are estimates and it is likely that neither measures the US resident population without error. Beyond measurement error, however, there are reasons that tax data may either overestimate or underestimate the US resident population.

One reason the tax data may overestimate the population is that, despite out best efforts, our estimate of the resident population may include US citizens or resident aliens who are civilians living outside the US or who are members of the US armed forces stationed overseas. As noted in both Cilke (2014) and Lurie and Pearce (2021), eliminating Form 1040 without a US address will not necessarily remove all non-resident filers and dependent nonfilers from the sample, as some may choose to list a US mailing address on their tax returns. Similarly, eliminating information returns without a US address will not necessarily remove all non-resident non-dependent nonfilers, as some may provide a US mailing address to the entities which generate information returns.

Another reason that the tax data may overestimate the US resident population is that, in addition to dependents who are US citizens and resident aliens, filers may claim children who are resident in Canada or Mexico as dependents. Both Cilke (2014) and Larrimore et al. (2019) note that some children in the tax data are attributable to filers legitimately claiming dependents who do not reside in the US.

The primary reason the tax data may underestimate the US resident population is that there is "… some portion of the population that is completely untouched by the Federal income tax system."[41] In particular, one group of individuals traditionally not captured by the tax data is individuals solely dependent on public assistance. This is because benefit payments from such

[41] Cilke (2014), p. 6.

programs as Temporary Assistance to Needy Families (TANF), Supplemental Security Income (SSI), and Veteran Affairs (VA) are not reported to the IRS—neither on tax returns nor on information returns.

In contrast to most previous studies using tax data, we believe we capture a large share of the population reliant solely on public assistance because—following Lurie and Pearce (2021), and using the data collected and processed by the authors of that study—we include individuals identified on Form 1095, which provides information on health insurance coverage. Although our income measure will not include benefit payments from public assistance programs, our population counts should include most, if not all, individuals who rely solely on public assistance because they typically are covered by government provided health insurance.

6.2 Comparison of 2016 Population Estimates

Overall and by age, the two population estimates are very close. The Census-equivalent estimate of 326.5 million is 3.4 million, or 1.1 percent, higher than the Census estimate of 323.1 million (Table 5). The two estimates also track very closely by single year of age (Figure 4).

Nonetheless, there some age ranges over which there are notable differences between the two population estimates, especially among children (with more in the tax data) and the elderly (with fewer in the tax data). The largest differences in numbers (239,000 more individuals per birth year in the tax data)—and also large as a percentage of the Census estimate (5.8 percent more in the tax data)—are among children aged 6 through 17, with the differences peaking at ages seven through nine (averaging 7.8 percent more in the tax data). Although much smaller in numbers, differences as a percentage of the Census estimate are large after age 75—with 3.3 percent fewer in the tax data from age 76 through 85, 6.5 percent fewer from age 86 through 95, and 11.6 percent fewer for those aged 96 or older.

Over other age ranges, the two estimates are much closer. By single year of age, differences between Census-equivalent and Census estimates averaged ±1.3 percent of the Census estimate for ages 0 through 5 and ages 18 through 75. Over these ages, about one-third of single-year-of-age population estimates were within 1.0 percent and about four-in-five were within 2.0 percent.

The Census-equivalent estimate from the tax data is higher than the Census estimate for older children and adults in their prime working years. The tax data are higher for individuals aged five through 20 (and markedly so, as already noted, for children aged 6 through 17). The tax data also identify more individuals aged 28 through 61, with the Census-equivalent estimate 1.3 percent higher, on average, than the Census estimate over this age range.

The Census-equivalent estimate is lower than the Census estimate for infants, young adults, and the elderly. There are 2.0 percent fewer newborns and 0.9 percent fewer one-year old children in the tax data than in the Census. On average, there are 1.6 percent fewer individuals in the tax data aged 21 through 27, with differences peaking at 2.8 percent fewer of the Census estimate at age 25. Tax data are consistently lower than the Census estimate for individuals older than aged 67, with (as already noted) the percentage differences increasing with age.

The two population estimates are closest for children aged two through four and adults aged 62 through 67. The Census-equivalent estimate averages about 10,400 more per birth year than the Census estimate for individuals aged two through four and about 16,500 more per birth year for individuals aged 62 through 67—both differences representing less than 0.5 percent of the Census population estimate over those ages.

We do note an anomalous pattern that occurs every five years of age from age 31 (the 1985 birth-year cohort) through age 61 (the 1955 birth-year cohort) where the Census-equivalent estimate from the tax data declines relative to the Census estimate, reducing the differences between the two estimates.[42] This pattern occurs because there are larger swings in the size of adjacent birth-year cohorts in the Census estimate than in the tax data. For example, differences in the number of individuals aged 37 (the 1979 birth-year cohort) and aged 35 (the 1981 birth-year cohort) are similar within the two data sources, with the age 37 population 4.8 percent lower than the age 35 population in the tax data and 4.5 percent lower in the Census estimate. In

[42] When comparing Census estimates to the Census-equivalent estimates from the tax data, the age in years is calculated as of July 1, 2016. For expositional ease, in the text we are effectively assigning the same birth year to all individuals born in the 12-month period ending on July 1 of a given year.

between, however, the change in cohort size differs. In the tax data, older age groups are progressively smaller, with 1.3 percent fewer individuals aged 36 (the 1980 birth-year cohort) than aged 35, and 3.5 percent fewer individuals aged 37 than aged 36. In the Census estimate, by contrast, the older age groups get larger before they get smaller, with 1.9 percent *more* individuals aged 36 than aged 35, and then 6.3 percent *fewer* individuals aged 37 than aged 36.

We also note why differences in population estimates peak for children aged seven through nine. Both the Census-equivalent and Census estimates show that the population aged six (the 2010 birth-year cohort) is roughly 5 percent lower than the population aged 16 (the 2000 birth-year cohort). In the tax data, cohort size drops sharply for children born after 2008, with the population aged eight and nine (the 2008 and 2007 birth-year cohorts) actually 2 percent higher, on average, than the population aged 16. In the Census estimate, there is a more consistent drop in cohort size between the 2000 and 2010 birth years, with the population aged eight and nine already 2 percent lower, on average, than the population aged 16. Further, cohort size continues to decline for children born after 2010, with 7 percent fewer newborns (the 2016 birth-year cohort) than children aged six in the tax data, but only 2 percent fewer in the Census estimate.

6.3 Comparison with Results of Previous Studies

This section compares our results to the results in Cilke (2014) and Larrimore et al. (2019)—two previous studies which use tax data to derive population estimates and which compare them to the Census estimate by single year of age. Our results are qualitatively similar to the previous studies, in that all three studies illustrate that population estimates derived from tax data are close to those of the Census and generally track Census estimates by age. That said, our results differ from the previous studies, particularly for children and middle-aged individuals.

Overall, we find 1.1 percent more individuals in Census-equivalent estimate from the tax data than the Census population estimate, whereas both Cilke (2014) and Larrimore et al. (2019) find about 0.5 percent fewer.[43]

[43] Cilke (2014) finds that, for individuals aged 16 and older, the tax data represent 99.5 percent of the Census estimate. Larrimore et al. (2019) finds that, for the total population (across all ages), the tax data represent 99.5 percent of the Census estimate.

We also find different patterns by age. Both previous studies find more children younger than age 15 in the tax data, with the tax data noticeably higher than the Census estimate for children as young as age one. We also find more in the tax data younger than age 15, but the overall difference is smaller than found in the previous studies and large differences do not emerge until children are aged five or older. Both previous studies find fewer in the tax data starting around age 15—Cilke (2014) discusses finding fewer individuals aged 16 through 24, Larrimore et al. (2019) discusses finding fewer individuals aged 15 to 20. In contrast, we find the Census-equivalent estimate exceeds the Census estimate through age 20. Both previous studies also find fewer middle-aged individuals in the tax data—Cilke (2014) discusses finding fewer aged 50 to 64, Larrimore et al. (2019) discusses finding fewer aged 40 to 55—whereas we generally find more individuals in tax data aged 27 through 67. Finally, we find increasingly large percentage differences beginning after age 67 with fewer individuals in the tax data than in the Census estimate, whereas the previous studies do not highlight similar differences.

There are two primary reasons our results differ from the previous studies.[44] The first is that we utilize tax data not available when the previous studies were done, allowing us to identify additional individuals. The second is that we compare Census-equivalent and and Census estimates for 2016, whereas the previous two studies compare estimates for 2011 and 2010, respectively.

To disentangle the effects of data availability from the effects of the time period analyzed, we first redo the 2016 comparison excluding the newly available tax data, and then redo that same comparison using 2010 data.

There are two sources of tax information that allow us to identify more individuals in 2016 than could be identified in 2010. The availability of Form 1095, which was first used in the 2014 tax year to report health insurance coverage (although it was not required until later), accounts for most of the additional individuals identified. We are also able to identify all dependents

[44] A third reason, which is less consequential for all but the elderly, is that the comparisons between the tax and Census data in the previous studies are slightly different than the comparison done in this paper. This issue is discussed in the appendix.

claimed on electronically filed returns rather than just the first four dependents listed, but this has a much smaller impact on our estimate.[45]

To account for differences in data availability, we exclude from our Census-equivalent estimate individuals who were (1) identified only using Form 1095 or (2) identified only as a dependent on an electronically filed return who was not among the first four dependents listed. Excluding these individuals reduces our estimate the most for children and middle-aged adults, with only very small effects for the elderly (Figure 5).

Overall, we no longer find more in the tax data than in the Census estimate after adjusting the Census-equivalent estimate down to account for data availability, but we still cannot replicate the results of the previous studies, as we now find too few (Table 5). Excluding individuals only identified using the newly available tax data, the 2016 Census-equivalent estimate represents 98.0 percent of the Census estimate—slightly below the 99.5 percent found in the earlier studies.

By age, adjusting the population estimate to account for data availability accentuates the discrepancy between our results and the previous studies for children (Figure 5). After removing individuals identified with the newly available tax data, there are now fewer children younger than age 15 in the tax data than in the Census estimate—considerably fewer aged six and younger and only a bit more aged seven through 14.

To investigate how the time period analyzed affects the results, we rerun the analysis using 2010 data. We then compare the 2010 results to the 2016 results adjusted for data availability.

When comparing the 2010 and 2016 population estimates by single year of age, we generally see similar changes in the age profile of both the Census-equivalent (Figure 6, top panel) and the Census (Figure 6, middle panel) estimates. In both comparisons, between 2010 and 2016 we see: more individuals in their early-20s through late-30s and in their early-50s or

[45] Cilke (2014) only includes information on up to four dependents for both paper returns and electronically filed returns. We believe Larrimore et al. (2019) does as well. Further, our investigation of the data indicates, that even if information on additional dependents from electronically filed returns were used in both 2010 and 2016, the results would not be directly comparable because the share of returns filed electronically increased considerably by 2016.

older; fewer older teenagers and individuals in their late-30s through early-50s; and about the same number aged eight through 16.

The exception to this rule is the change between 2010 and 2016 in the population aged seven or younger. For this age group, the Census estimate is about the same in both years, at just over 4.0 million per birth year, on average. In contrast, the size of this age group declines in Census-equivalent estimate, from 4.2 million per birth year, on average, in 2010 to 3.8 million per birth year, on average, in 2016.

The comparison of the 2010 and 2016 results illustrates that our results differ from those of the previous studies not only because of the availability of additional data, but also because we analyze a different tax year. Combined with the adjustments to account for the availability of data, switching to 2010 data allow us to generally replicate the results from the previous studies (Figure 6, bottom panel). In particular, we replicate the findings that, in 2010, there are more individuals younger than age 15 in the Census equivalent estimate than in the Census estimate, and that differences arise by age two. We also find fewer individuals in the tax data aged 15 through 24 in 2010. In addition, differences between the two population estimates for those aged 85 and older were smaller in 2010 than they were in 2016, possibly explaining why they were not highlighted in the previous studies. The differences between the 2010 and 2016 results, however, illustrate that the 2010 differences between the Census-equivalent and the Census estimates by age do not generalize to other years.

Rather than being consistent across time by age, we find differences between tax data and Census estimates are generally consistent across time by birth-year. When comparing the 2010 and 2016 population estimates by year of birth, we again see similar changes in the birth-year profile of both the Census-equivalent (Figure 7, top panel) and Census (Figure 7, middle panel) estimates. In both estimates between 2010 and 2016, the 1972 and later birth-year cohorts (aged 38 or younger in 2010) increase in size (implying net migration more than offsets deaths) and the 1971 or earlier birth-year cohorts (aged 39 or older in 2010) decrease in size (implying net migration does not fully offset deaths). As a result, the 2010 and 2016 differences between the tax and Census population estimates, expressed as a percentage of the Census population

estimate, are more highly correlated by birth-year (Figure 7, bottom panel) than they are by age (Figure 6, bottom panel).

Although birth-year differences in the two population estimates are correlated across time, there are three birth-year ranges over which the 2010 and 2016 results diverge consistently. Between 2010 and 2016, the Census-equivalent estimate from the tax data declines relative to the Census estimate for two birth-year groups: the 1989 through 2008 birth-year cohorts (aged two through 21 in 2010), and the 1944 and earlier cohorts (aged 66 or older in 2010). The Census-equivalent estimate increases relative to the Census estimate between 2010 and 2016 for only one birth-year group: the 1949 through 1954 birth-year cohort (aged 56 through 61 in 2010).

Of these changes, the easiest to explain is the increase relative to the Census estimate for the 1949 through 1954 birth-year cohorts. The age of these birth-year cohorts increased from 56 through 61 in 2010 to 62 through 67 in 2016. As discussed above and illustrated in Figure 2, in 2016 the share of the population identified using only Form 1095 declines sharply from 3.7 percent for individuals aged 61 to 0.9 percent for individuals aged 67. Thus, adjusted for data availability, the increase in the Census-equivalent estimate relative to the Census estimate for these birth cohorts between 2010 and 2016 presumably is because we identify a higher share of these individuals in 2016 using non-Form-1095 tax data.

The decline relative to the Census estimate for the 1944 and earlier birth-year cohorts is the result of relatively small differences between the tax data and the Census estimate in population changes (net migration minus deaths) between 2010 and 2016. For example, the age of the 1917 through 1944 birth-year cohorts increased from 66 through 93 in 2010 to 72 through 99 in 2016. Over this period, their population falls by 9.85 million in the tax data and by 9.80 million in the Census estimate—a difference of only 0.5 percent. In combination with existing 2010 differences, however, 2016 differences represent a much larger percentage of the shrinking population for these birth-year cohorts.

We do not have a ready explanation for the decline relative to the Census estimate for the 1989 through 2008 birth cohorts, and further investigation is beyond the scope of this research. The fact that the relative decline happened for children aged two through 21 in 2010 suggests

that the decline is associated with dependents. What is less clear is why the relative decline would occur for children young enough to be claimed as a dependent in both 2010 and 2016. We do note, however, that—despite the decline in the tax data relative to the Census estimate—2016 differences between the two estimates follow a similar pattern by birth year as the 2010 differences.

6.4 Discussion

To get a sense of how well the tax data represent our population of interest, we compare it to Census estimates of the US resident population. Although our sample of interest includes individuals living outside the US, US residents represent the bulk of our population. Therefore, we adjust our data as best we can to conform to the Census population concept and then make the comparison.

Consistent with previous research, the two total population estimates are similar. That said, most previous studies find slightly fewer in the tax data than in the Census estimate, whereas we find slightly more. This is because, consistent with Lurie and Pearce (2021), we include individuals who are only identified on Form 1095.

Also consistent with the previous research, we find the tax data generally track the Census estimate closely by age, but that there are differences over some age ranges. That said, we find differences over slightly different age ranges than the previous studies.

Some of the differences between the tax data and the Census estimate could be attributable to the tax data including non-residents even after our best efforts to remove them. For example, compared with the Census estimate, the 2016 Census-equivalent estimate from the tax data is higher for older children and prime working age adults. These differences would be consistent with the explanation that some individuals work and live outside the US but, for tax purposes, use a US mailing address for themselves and their dependents. It would also be consistent with the explanation that some filers who are US residents validly claim dependents who are Mexican or Canadian residents.

Our results, however, suggest some caution when suggesting explanations for the differences between the tax data and the Census estimate. This is because, across years,

differences between the tax data and the Census estimate appear to be more consistent by birth year than by age. For example, Larrimore et al. (2019) suggests there are fewer children older than 16 in the tax data because they no longer qualify for the child tax credit and are, thus, less likely to be claimed as a dependent on a return. Although that explanation is consistent with the 2010 data the study analyzes, it is not consistent with our 2016 results—which show more children in the tax data through age 20.

To the extent that differences between the tax data and the Census estimate are caused by the inclusion of non-residents in our Census-equivalent estimates, it is not of great concern to our broader research agenda. Although we attempt to remove non-residents to facilitate the comparison with the Census estimate, we include them in the data we analyze.

This exercise illustrates that the tax data appear to be representative of the population we wish to study. The Census-equivalent population estimate from the tax data is reasonably close to Census estimate of the US resident population. Differences between the two estimates are not unexpected given that the tax data are not particularly well suited to measuring the US resident population and given that both the tax data and the Census estimate likely measure the population with error.

7. Conclusion

This paper was written to document our method of building a representative sample of the US population from tax data, which represents the first step in a larger project measuring changes in the amount and composition of income over the life cycle. We document our method both for readers of our broader research, so they understand how we derive our estimates, and for other researchers analyzing tax data, so that they can use or improve upon our methods.

To build a sample representative of the population, we supplement tax return data—which allows us to identify filers and dependent nonfilers—with information return data—which allows us to identify non-dependent nonfilers. Although non-dependent nonfilers are not a large part of the overall population, their importance increases with age, making their inclusion crucial for measuring elderly income.

Inclusive of both filers and nonfilers, we find the share of the population with income remains fairly steady after age 26 and actually increases slightly after age 61. For those with income, the median amount follows a hump shape over the lifecycle, peaking at age 46.

Many of our findings are similar to those in the existing literature. This is not terribly surprising given that our work builds on the work of many, especially Cilke (2014) and Lurie and Pearce (2021). For example, as with previous studies, we find that the tax data produce a population estimate similar to that of the Census, and that the two estimates track closely by age. We conclude from this comparison that the tax data adequately captures our population of interest.

Although our work builds on the existing literature, we note several innovations. First, we sample individuals rather than tax returns. Second, we use the individual—rather than the tax return, family, or household—as the unit of observation. All income reported on a tax return is allocated to filers—primary taxpayers in the case of non-joint returns and primary and secondary filers in the case of joint returns. We do not adjust filer income to account for the number of dependents they claim. Further, dependents have income only if they file their own tax return or have income reported on information returns. Our method allows us to examine individual income by single year of age—which is difficult, if not impossible, to do for tax returns or families.

Although we cannot claim credit for the innovation, we do note that—following Lurie and Pearce (2021)—we include in our sample individuals identified only on Form 1095, which reports health insurance coverage. There are some portions of the population that tax data traditionally does not identify, such as individuals whose only source of income is public assistance. We believe that the Form 1095 allows us to identify most individuals solely dependent on public assistance because they typically are covered by government provided health insurance.

Finally, our examination of 2010 and 2016 data show that, across years, differences between the tax data and the Census population estimate are more correlated with birth year than they are by age. This is important because many explanations for why the tax data would

overestimate or underestimate the true population are related to the age of individuals. Our work suggests that explanations for differences between the two estimates should also be consistent with differences by birth cohort that persist over time.

References

Bee, Adam C. and Joshua Mitchell. 2017. "Do Older Americans Have More Income Than We Think?" 2017. SEHSD Working Paper No. 2017-39. Washington, DC: U.S. Census Bureau Social, Economic, and Housing Statistics Division. Available at https://www.census.gov/content/dam/Census/library/working-papers/2017/demo/SEHSD-WP2017-39.pdf

Brady, Peter J. and Steven Bass. 2021. "Comparing the Current Population Survey to Income Tax Data." *SOI Working Paper*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/21rpcomparingcpstoincometaxdata.pdf.

Brady, Peter J. and Steven Bass. 2020a. "Reconciling Form 1040 and Form 1099-R Data." *SOI Working Paper*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/20rpreconciling10401099R.pdf.

Brady, Peter J. and Steven Bass. 2020b. "Decoding Retirement: A Detailed Look at Retirement Distributions Reported on Tax Returns." *SOI Working Paper*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/20rpdecodingretirement.pdf.

Brady, Peter, Steven Bass, Jessica Holland, and Kevin Pierce. 2017. "Using Panel Tax Data to Examine the Transition to Retirement." *SOI Working Paper*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/17rptransitionretirement.pdf.

Brady, Peter and Kevin Pierce. 2012. "The Promise and Potential Pitfalls of Using Administrative Tax Data: A Case Study." Unpublished manuscript, Investment Company Institute (April).

Cilke, James. 2014. The Case of Missing Strangers: What We Know and Don't Know About Non-Filers" *Proceedings of the 107th Annual Conference on Taxation.* Washington, DC: National Tax Association. Available at https://www.ntanet.org/wp-content/uploads/proceedings/2014/029-cilke-case-missing-strangers-know-don.pdf.

Cilke, James. 1998. "A Profile of Non-Filers." *OTA Working Paper 78*. Washington, DC: US Department of the Treasury. Available at https://home.treasury.gov/system/files/131/WP-78.pdf

Congressional Budget Office. 2016. *The Distribution of Household Income and Federal Taxes, 2013*. Washington, DC: Congressional Budget Office. Available at https://www.cbo.gov/publication/51361

Cronin, Julie-Anne. 2022. "U.S Treasury Distributional Analysis Methodology." *OTA Technical Paper 8*. Washington, DC: US Department of the Treasury. Available at https://home.treasury.gov/system/files/131/TP-8.pdf.

Cronin, Julie-Anne. 1999. "U.S Treasury Distributional Analysis Methodology." *OTA Working Paper 85*. Washington, DC: US Department of the Treasury. Available at https://home.treasury.gov/system/files/131/WP-85.pdf.

Internal Revenue Service, Statistics of Income Division. 2018. *Individual Income Tax Returns 2016*, Publication 1304. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/16inalcr.pdf.

Internal Revenue Service. 2017a. *U.S. Tax Guide for Aliens: For Use in Preparing 2016 Returns*, Publication 519. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/p519--2016.pdf.

Internal Revenue Service. 2017b. *Tax Guide for Seniors: For Use in Preparing 2016 Returns*, Publication 554. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/p554--2016.pdf.

Internal Revenue Service. 2017c. *Tax Guide for Individuals with Income from U.S. Possessions for Use in Preparing 2016 Returns*, Publication 570. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/p570--2016.pdf.

Internal Revenue Service. 2016a. *Tax Guide for U.S. Citizens and Resident Aliens Abroad: For Use in Preparing 2016 Returns*, Publication 54. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/p54--2016.pdf.

Internal Revenue Service. 2016b. *Exemptions, Standard Deduction, and Filing Information: For Use in Preparing 2016 Returns*, Publication 501. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/p501--2016.pdf.

Internal Revenue Service. 2016c. *1040 Instructions 2010*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-prior/i1040gi--2016.pdf.

Joint Committee on Taxation. 1993. "Methodology and Issues in Measuring Changes in The Distribution of Tax Burdens." JCS-7-93. Available at https://www.jct.gov/publications/1993/jcs-7-93/.

Larrimore, Jeff, Jacob Mortenson, and David Splinter. 2019. "Household Incomes in Tax data: Using Addresses to Move from Tax Unit to Household Income Distributions." *Journal of Human Resources*. Advance online publication. Available at http://jhr.uwpress.org/content/early/2019/07/02/jhr.56.2.0718-9647R1.abstract.

Lawrence, Joshua, Michael Udell, and Tiffany Young. 2011. "The Income Tax Position of Persons Not Filing Income Tax Returns for Tax Year 2005." In Alan Plumley ed. Recent Research on Tax Administration and Compliance. *IRS Research Bulletin*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/11rescontaxposition.pdf.

Lundberg, Shelly, Robert A. Pollak, and Jenna Stearns. 2016. "Family Inequality: Diverging Patterns in Marriage, Cohabitation, and Childbearing." *Journal of Economic Perspectives*, 30 (2): 79-102. Available at https://www.aeaweb.org/articles?id=10.1257/jep.30.2.79.

Lurie, Ithai Z. and James Pearce. 2021. "Health Insurance Coverage in Tax and Survey Data." *American Journal of Health Economics*, 7(2): 164-184. Available at https://www.journals.uchicago.edu/toc/ajhe/2021/7/2.

Mortenson, Jacob A., James Cilke, Michael Udell and Jonathan Zytnick. 2009. "Attaching the Left Tail: A New Profile of Income for Persons Who Do Not Appear on Federal Income Tax Returns." *Proceedings of the 102nd Annual Conference on Taxation.* Washington, DC: National Tax Association. Available at https://ntanet.org/wp-content/uploads/proceedings/2009/011-mortenson-attaching-left-tail-2009-nta-proceedings.pdf

Nunns, James R., Deena Ackerman, James Cilke, Julie-Anne Cronin, Janet H. Janet Holtzblatt, Gillian Hunter Emily Lin, and Janet McCubbin. 2008. "Treasury's Panel Model for Tax Analysis." *OTA Technical Papers 3*. Available at https://home.treasury.gov/system/files/131/TP-3.pdf.

Saez, Emmanuel. 2016. "Statistics of Income Tabulations: High Incomes, Gender, Age, Earnings Split, and Non-filers" *SOI Working Paper*. Washington, DC: Internal Revenue Service. Available at https://www.irs.gov/pub/irs-soi/16rpsaeztabulations.pdf.

Sailer, Peter and Michael Weber. 1998. "The IRS Population Count: An Update." Washington, DC: Statistics of Income Division, Internal Revenue Service. Available at: https://www.irs.gov/pub/irs-soi/indpopct.pdf.

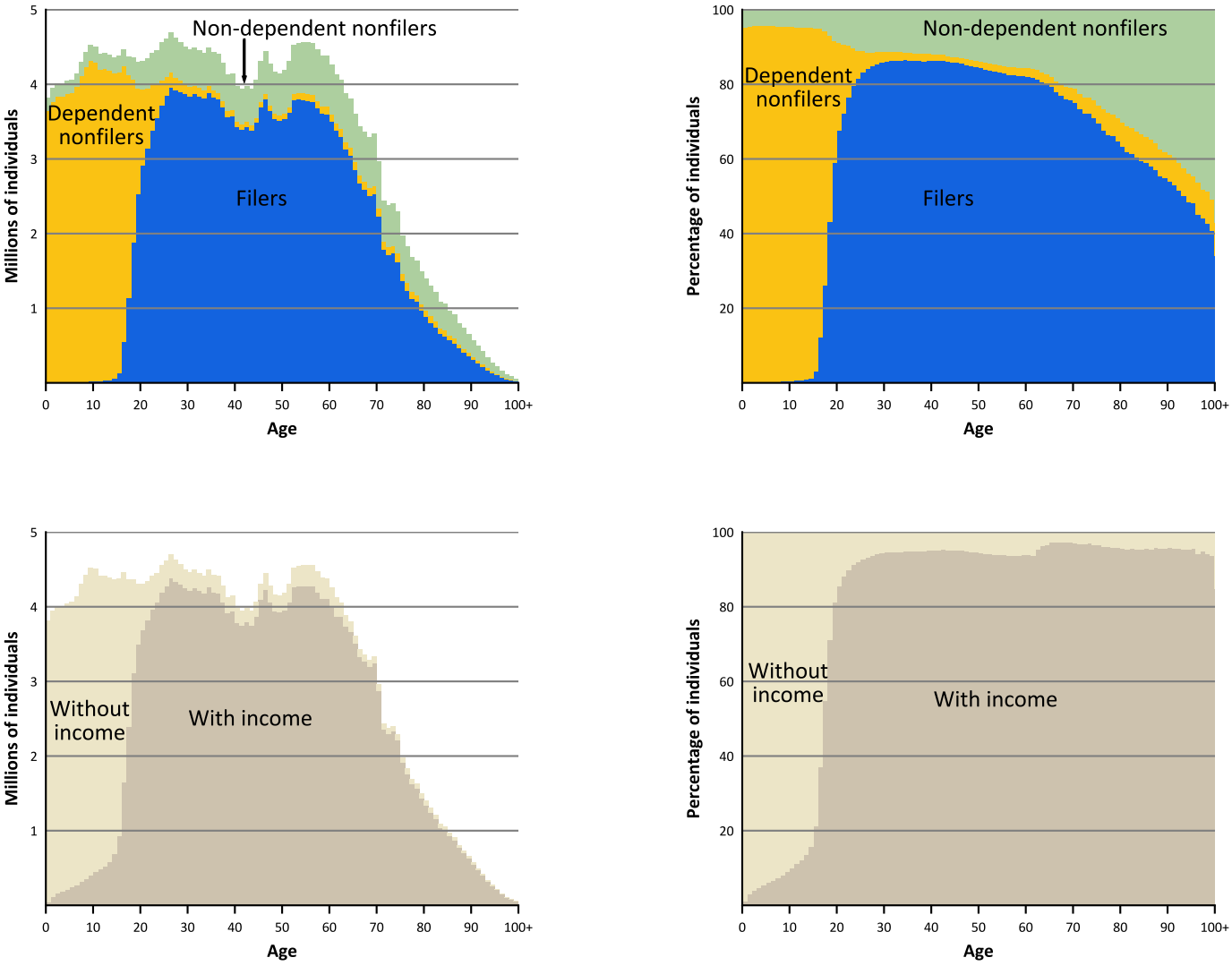
Semega, Jessica, Melissa Kollar, John Creamer, and Abinash Mohanty. 2019. "Income and Poverty in the United States: 2018." *Current Population Reports*, P60-266. Washington, DC: U.S. Census Bureau. Available at https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-266.pdf.

U.S. Census Bureau. 2021. Annual Estimates of the Resident Population by Single Year of Age and Sex. Available at: https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-national-detail.html.

Figure 1

Filer Share Declines with Age, but Share with Income Remains Steady

Population by sample component and presence of income, tax-year 2016

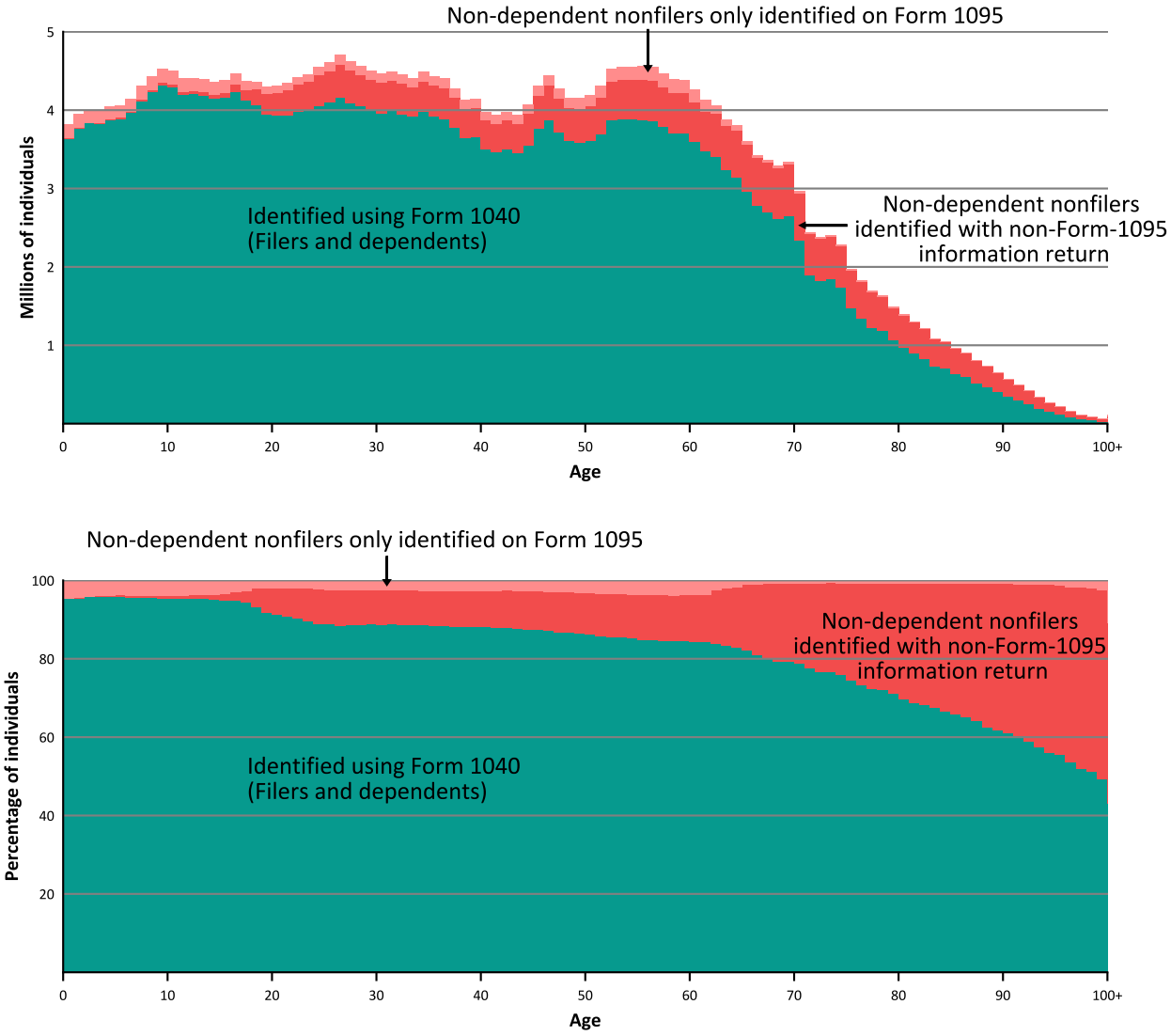


Source: Author's tabulation of IRS data

Figure 2

Younger Individuals More Likely to Be Identified with Form 1095 Only

Population by method of identification, tax-year 2016

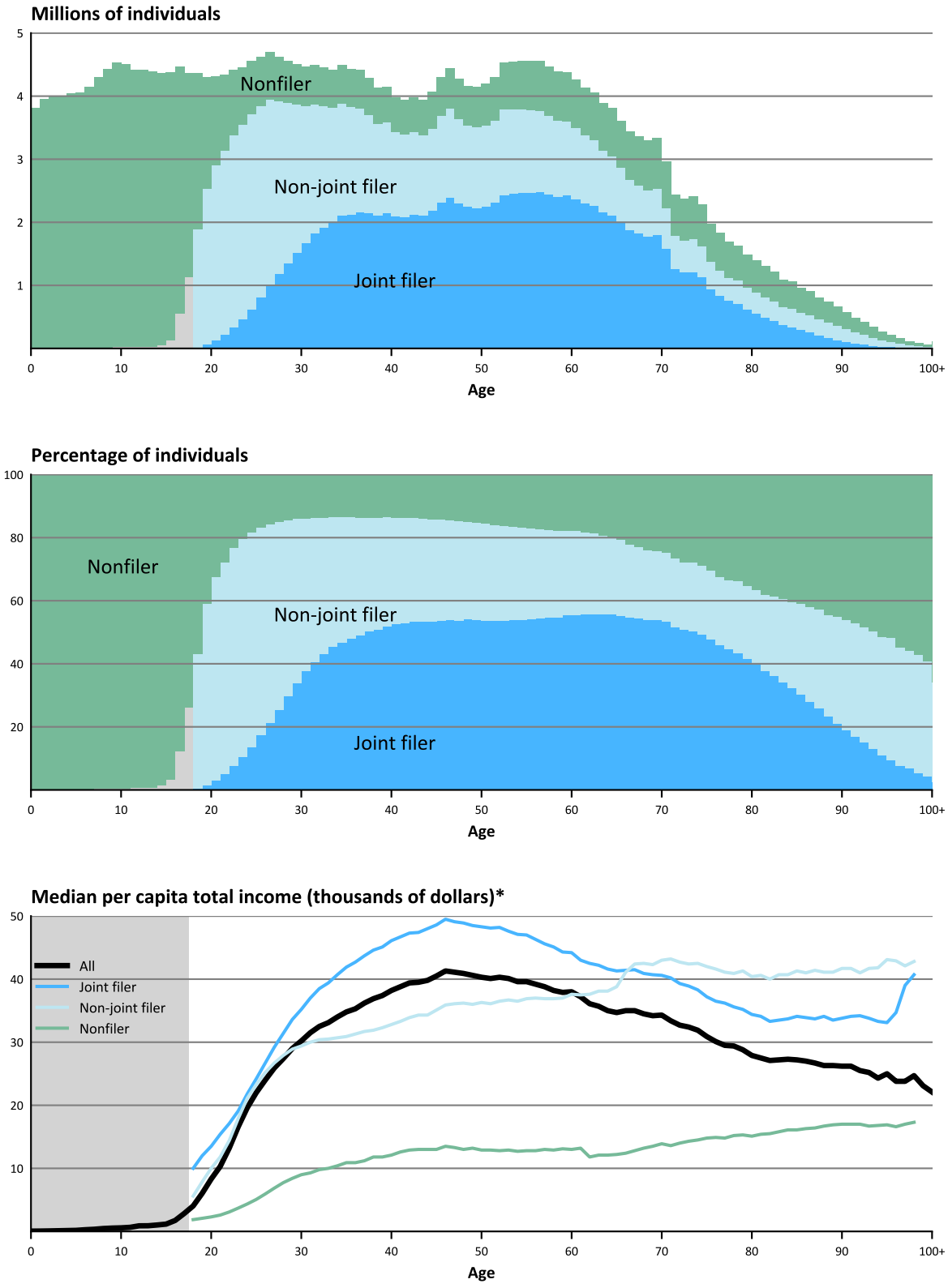


Source: Author's tabulation of IRS data

Figure 3

Composition of Filing-Type Groups Changes over the Life Cycle

Population and income by filing type, tax-year 2016



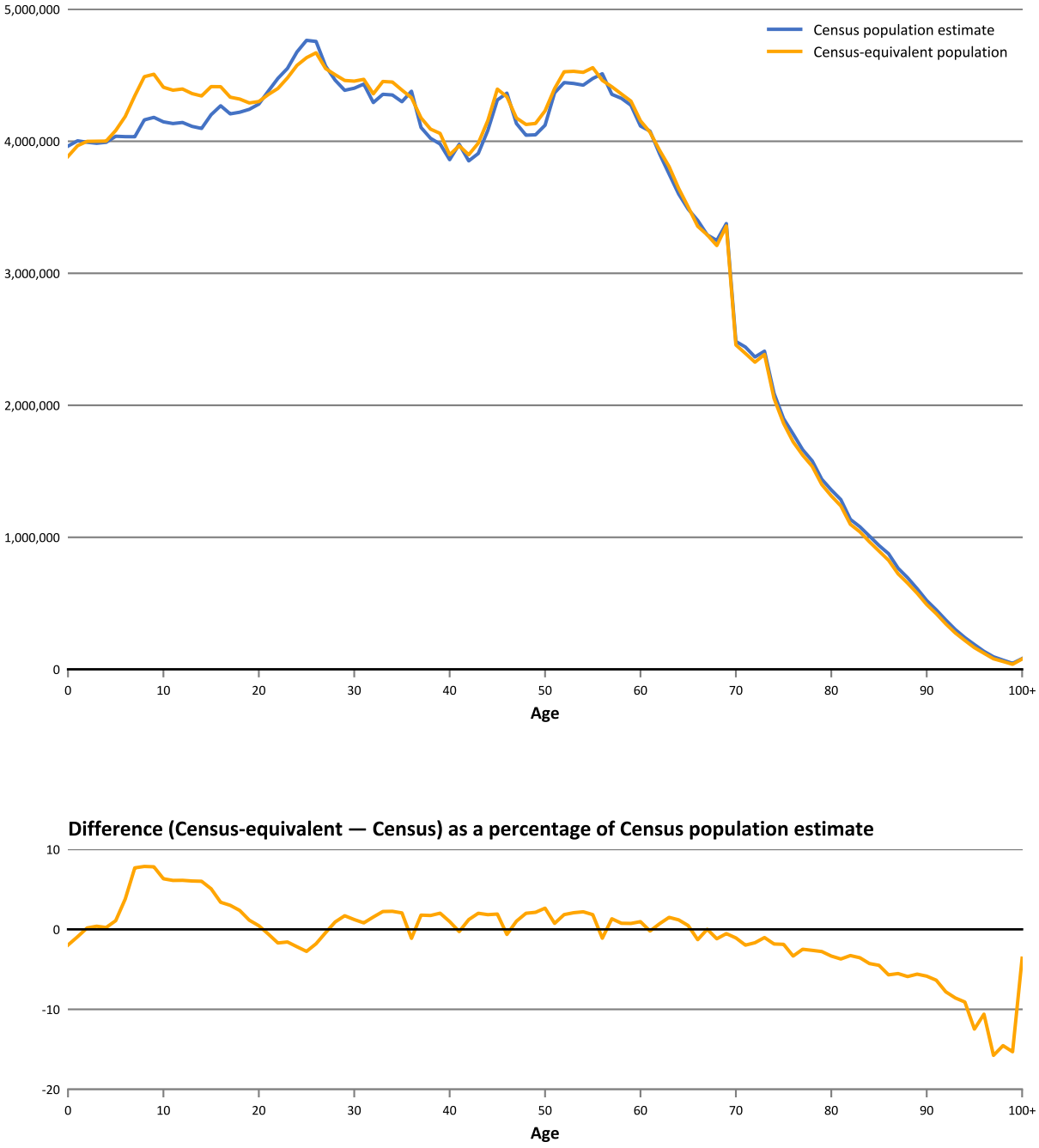
*Median per capita total income is calculated for individuals alive at year-end 2016.

Source: Author's tabulation of IRS data

Figure 4

Census-Equivalent Population Estimate from Tax Data Closely Tracks Census Estimate

Population estimate as of July 1, 2016

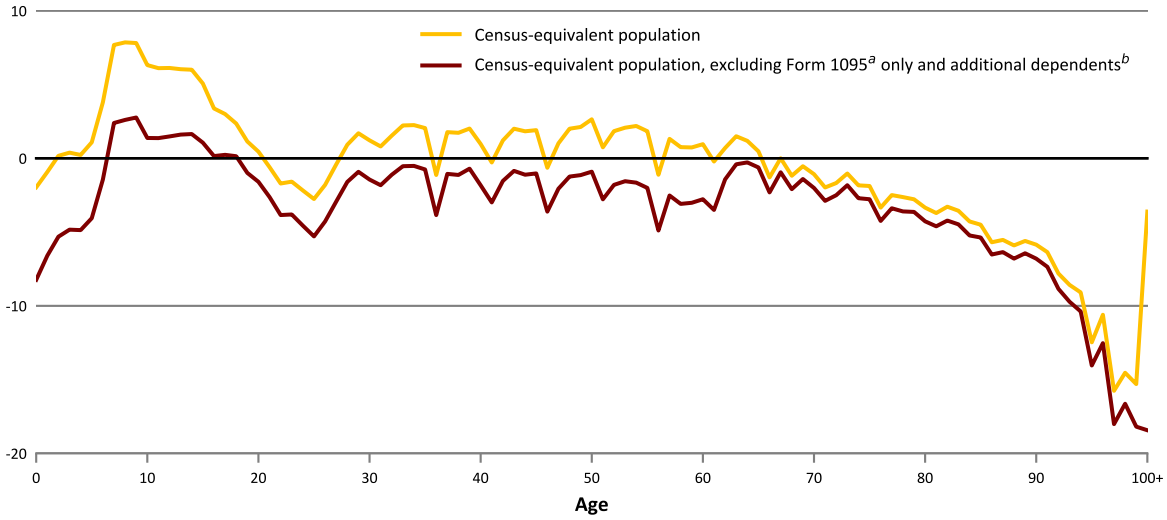


Sources: U.S. Census Bureau 2020 and author's tabulation of IRS data

Figure 5

Addition of Form 1095 and Extra Dependents Has Largest Effect on Non-Elderly Estimates

Difference between Census-equivalent and Census estimates as a percentage of Census, 2016



[a]Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

[b]Additional dependents are those not listed as one of the first four dependents on Form 1040.

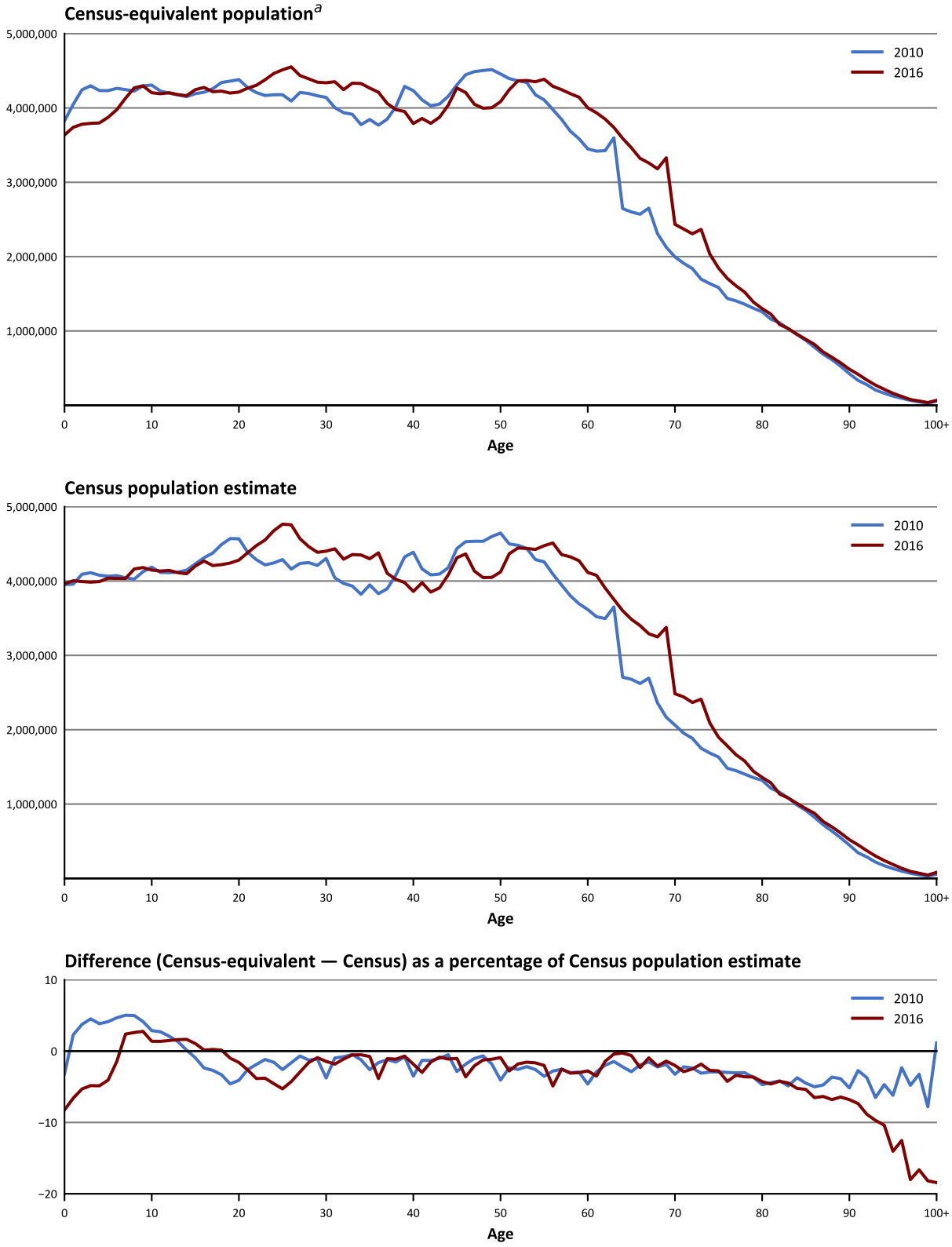
Note: Population estimates are as of July 1, 2016.

Sources: U.S. Census Bureau 2020 and author's tabulation of IRS data

Figure 6

Differences by Age Are Inconsistent Across Years

Population estimate as of July 1



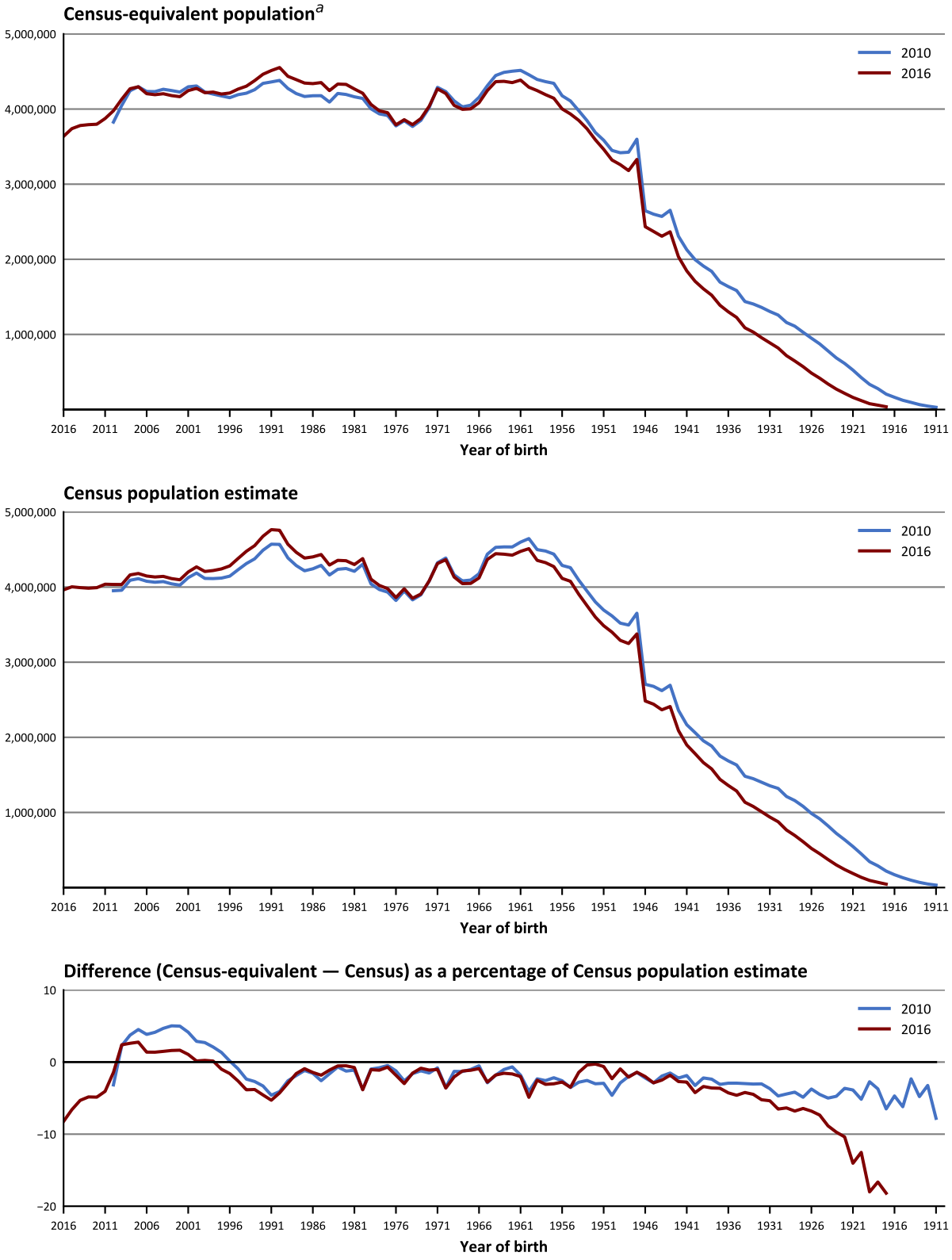
[a]In this figure, the Census-equivalent population excludes individuals identified only using Form 1095 (inclusive of Form 1095-A, Form 1095-B, and Form 1095-C) or only as the fifth or higher dependent on Form 1040.

Sources: U.S. Census Bureau 2020 and author's tabulation of IRS data

Figure 7

Pattern Is More Consistent by Year of Birth

Population estimate as of July 1



[a]In this figure, the Census-equivalent population excludes individuals identified only using Form 1095 (inclusive of Form 1095-A, Form 1095-B, and Form 1095-C) or only as the fifth or higher dependent on Form 1040.

Sources: U.S. Census Bureau 2020 and author's tabulation of IRS data

Table 1

Derivation of the Taxpayer Component of Sample

Population estimate of primary and secondary taxpayers filing any 2016 tax return (thousands)

	Primary filers	Secondary filers	Total taxpayers
Initial sample (Taxpayers who filed 2016 Form 1040,[a] Form 1040-NR, or Form 1040-SS/1040-PR)	**151,268**	**54,309**	**205,577**
– Died before January 1, 2016	1	1	2
– Filed Form 1040-NR	748	0	748
– Filed Form 1040-SS/1040-PR	178	63	241
– Filed Form 1040[a] with a US territory address	68	26	94
= **Final taxpayer sample**	**150,273**	**54,219**	**204,492**
Memo: Composition of final taxpayer sample			
Taxpayers on non-dependent returns			195,090
Taxpayers on dependent returns			9,402
Taxpayers with a state address			203,566
Taxpayers with an overseas US armed forces address			239
Taxpayers with a foreign or missing address			687
Memo: Most common types of information returns among taxpayers			
Any Information return			200,802
Form 1095[b]			182,447
Form W-2			147,828
Form 1099-G			67,417
Form 5498			60,275
Form 1099-INT			54,298
Form 1098			51,972
Form 1099-R			48,724
Form SSA-1099			39,403
Form 1099-DIV			36,234
Form 1099-MISC			24,509

[a] The term "Form 1040" is inclusive of Form 1040, Form 1040-A, or Form 1040-EZ.

[b] Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

Source: Author's tabulation of IRS data

Table 2

Derivation of the Dependent Nonfiler Component of Sample

Population estimate of dependents [a] claimed on any 2016 tax return (thousands)

	Dependents
Initial sample (Dependents[a] claimed on 2016 Form 1040,[b] Form 1040-NR, or Form 1040-SS/1040-PR)	**93,766**
– Died before January 1, 2016	9
– Claimed on Form 1040-NR	25
– Claimed on Form 1040-SS/1040-PR	171
– Claimed on Form 1040[b] with a US territory address	41
– Filed a Form 1040[b] dependent return	9,285
= Final dependent nonfiler sample	**84,235**
Memo: Composition of final dependent nonfiler sample	
With income	13,337
Without income	70,898
Form 1095[c] only	59,389
At least one form other than Form 1095[c]	2,504
No information return	9,005
Dependent nonfilers with a state address	83,854
Dependent nonfilers with an overseas US armed forces address	121
Dependent nonfilers with a foreign or missing address	260
Memo: Most common types of information returns among dependent nonfilers	
Any information return	75,230
Form 1095[c]	74,075
Form SSA-1099	4,860
Form W-2	4,765
Form 1098-T	3,621
Form 1099-INT	2,770
Form 1099-DIV	2,384
Form 1099-MISC	600
Form 1099-B	526
Form 1099-G	518
Form 1099-R	450

[a] Data include up to four dependents claimed on paper returns and all depdendents claimed on electronic returns.

[b] The term "Form 1040" is inclusive of Form 1040, Form 1040-A, or Form 1040-EZ.

[c] Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

Source: Author's tabulation of IRS data

Table 3

Derivation of the Non-Dependent Nonfiler Component of Sample

Number of individuals who are not identified on any 2016 tax return but who have at least one 2016 information return (thousands)

	Other nonfilers
Initial sample (Non-dependent nonfilers[a] with at least one 2016 information return)	**50,077**
– Died before January 1, 2016	4,285
– All information returns sent to US territory addresses	1,782
– All information returns sent to foreign, missing, or US territory addresses[b]	1,569
– Presence of information return indicating foreign person[c]	73
= Final non-dependent nonfiler sample	**42,368**
Memo: Composition of final non-dependent nonfiler sample	
With income	30,879
Without income	11,489
Form 1095[d] only	9,251
At least one form other than Form 1095[d]	2,238
Non-dependent nonfilers without an overseas US armed forces address	42,351
Non-dependent nonfilers with an overseas US armed forces address	17
Memo: The most common types of information returns among non-dependent nonfilers	
Any information return	42,368
Form 1095[c]	36,500
Form SSA-1099	18,374
Form W-2	10,786
Form 1099-R	6,457
Form 1099-INT	4,427
Form 1099-MISC	3,101
Form 5498	2,994
Form 1098	2,981
Form 1099-G	2,944
Form 1099-DIV	1,991

[a] Non-dependent nonfilers exclude all individuals identified on Form 1040 (inclusive of Form 1040-A and Form 1040-EZ), Form 1040-NR, or Form 1040-SS/1040-PR, including primary and secondary taxpayers and individuals claimed as a dependent.

[b] Individuals in this group have at least one information return sent to a foreign or missing address.

[c] The forms used to identify foreign persons are: Form 1042-S (Foreign Person's U.S. Source Income Subject to Withholding), Form 8288-A (Statement of Withholding on Dispositions by Foreign Persons of U.S. Real Property Interests), and Form 8805 (Foreign Partner's Information Statement of Section 1446 Withholding Tax).

[d] Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

Source: Author's tabulation of IRS data

Table 4

Composition of the Total Population

Number of unique individuals identified on 2016 Form 1040[a] or 2016 information return, excluding individuals with a US territory address and non-dependent nonfilers with a foreign or missing address

	Number of individuals (thousands)	Share of total population (percentage)
Total population estimate	**331,095**	**100**
Filers	***204,492***	***62***
Non-dependent filer	*195,090*	59
Dependent filer	*9,402*	3
Dependent nonfilers	***84,235***	***25***
With income	*13,337*	4
Without income	*70,898*	21
Non-dependent nonfilers	***42,368***	***13***
With income	*30,879*	9
Without income	*11,489*	3
Memo: Share of population by various characteristics (percentage)		
Identified on Form 1040 (filers plus dependent nonfilers)		87
Non-dependent nonfilers with at least one non-Form-1095[b] information return		10
Non-dependent nonfilers with Form 1095[b] only		3
Taxpayers on joint return		33
Taxpayers on non-joint return		29
Nonfilers (dependent plus non-dependent)		38
Taxpayers plus nonfilers with income		75
Dependent nonfilers without income		21
Non-dependent nonfilers without income		3

[a] The term "Form 1040" is inclusive of Form 1040, Form 1040-A, or Form 1040-EZ.

[b] Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

Source: Author's tabulation of IRS data

Table 5

Derivation and Composition of the Census-Equivalent Population

Number of unique individuals identified on 2016 Form 1040 [a] or 2016 information return, excluding individuals with a US territory address and non-dependent nonfilers with a foreign or missing address

	Number of individuals (thousands)
Total population estimate from tax data	**331,095**
– Taxpayers and dependent nonfilers with a foreign or missing address	947
– Taxpayers and dependent nonfilers with an overseas armed forces address	360
– Non-dependent nonfilers with an overseas US armed forces addresss	17
– Individuals who died before July 1, 2016	1,345
– Individuals born on or after July 1, 2016	1,918
= Census-equivalent population estimate from tax data	**326,508**
Identified only using Form 1095[b] or as an additional dependent[c]	9,873
Other	316,635
Memo:	
Census population estimate as of July 1, 2016	**323,072**
Census-equivalent population as a percentage of Census estimate	
Full Census-equivalent population	101.1
Excluding individuals identified only using Form 1095[b] or as an additional dependent[c]	98.0

[a] The term "Form 1040" is inclusive of Form 1040, Form 1040-A, or Form 1040-EZ.

[b] Form 1095 includes information about health insurance coverage through the Health Insurance Marketplace (Form 1095-A), health insurance coverage through other private or government sources (Form 1095-B), and an offer of health insurance coverage by certain employers (Form 1095-C).

[c] Additional dependents are those not listed as one of the first four dependents on Form 1040.

Sources: U.S. Census Bureau 2021 and author's tabulation of IRS data

Appendix

Comparison of Tax Data and Census Estimates in the Previous Literature

Beyond data availability and the year analyzed, the results of this study differ from those of Cilke (2014) and Larrimore et al. (2019) because the comparisons between the tax data and Census population estimates differ.

Without adjustment, population estimates from the tax data are not directly comparable to Census estimates. Annual Census population estimates provide a snapshot of the population alive on July 1. Tax data provide information on activities that could have occurred at any time during the year.

In this study, we create a Census-equivalent population estimate from the tax data and then compare that to the Census estimate. This is done by: (1) removing individuals who died before July 1 or who were born after June 30; and (2) tabulating the data by age as of July 1.

Cilke (2014) essentially creates a tax-equivalent population estimate from the Census data and then compares that to the tax data. This is done by (1) averaging, by single-year of age, the July 2011 and July 2012 Census population estimates to derive a population estimate as of December 31, 2011; and (2) adding back into the population an estimate of individuals who died during the year.

It is possible that the method used by Cilke (2014) could produce different results than our method. In theory, either approach produces valid comparisons. That said, the exact method used to add individuals who died during the year back into the Census estimates could affect the results, and we have not attempted to replicate that method here.

Unlike this study and Cilke (2014), Larrimore et al. (2019) does not appear to adjust either the tax data or the Census estimates. That is, it appears the study compares 2010 tax-year data—which represents individuals alive at any point during the year tabulated by age on December 31—to the decennial 2010 Census estimates.

To illustrate how failing to adjust the data would impact the results, we make two adjustments to the tax data in addition to adjusting for data availability. First, we include individuals alive at any point in 2016. Second, we retabulate the data by age as of December 31

(while still comparing to Census estimates of the population on July 1 tabulated by age as of July 1).

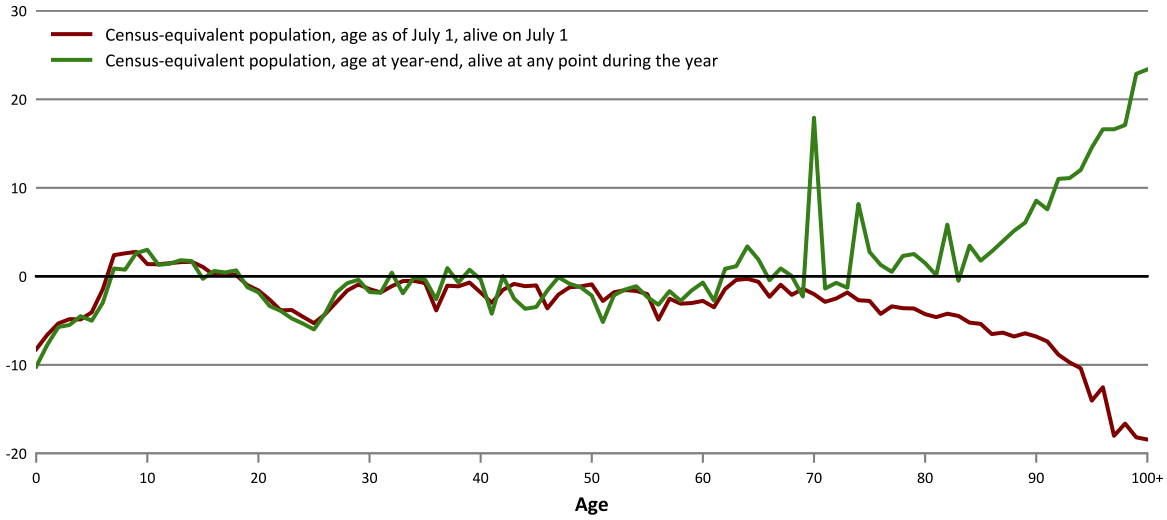
With these additional adjustments, the 2016 population estimate from tax data represent 99.0 percent of the Census estimate (up from 98.0 percent), which is closer to replicating the finding in Larrimore et al. (2019) that the tax data represent 99.5 percent of the Census population estimate.

By age, these additional changes affect the elderly comparisons the most (Figure A.1). Shifting the age groups by six months (from age as of July 1 to age as of December 31) generally has a small impact, except for individuals aged 70 or a bit younger, who were born as the baby boom accelerated in the late 1940s. The impact of including individuals who died before July 1 has a larger impact—at least among the elderly, where the tax data estimates switch from fewer individuals relative to the Census estimates to more individuals.

Figure A.1

Comparison to Census Requires Adjustments to Match Population Included in Census Estimate

Comparison of Census-equivalent population estimates, 2016



[a]In this figure, the Census-equivalent population excludes individuals identified only using Form 1095 (inclusive of Form 1095-A, Form 1095-B, and Form 1095-C) or only as the fifth or higher dependent on Form 1040.

Sources: U.S. Census Bureau 2020 and author's tabulation of IRS data

Table A.1

List of Information Returns Used to Identify Nonfilers

Form number	Form name
Forms used to identify nonfilers and measure nonfiler income	
Income reporting forms	
Form W-2	Wage and Tax Statement
Form W2-G	Certain Gambling Winnings
Form SSA-1099	Social Security Benefit Statement
Form 1065	Partner's Share of Income, Deductions, Credits, etc.
Form 1099-DIV	Dividends and Distributions
Form 1099-G	Certain Government Payments
Form 1099-INT	Interest Income
Form 1099-MISC	Miscellaneous Income
Form 1099-R	Distributions From Pensions, Annuities, Retirement or Profit-Sharing Plans, IRAs, Insurance Contracts, etc.
Form 1120-S	Shareholder's Share of Income, Deductions, Credits, etc.
Qualified account contribution reporting forms	
Form 5498	IRA Contribution Information
Forms used to identify nonfilers but not used to measure nonfiler income	
Income reporting forms	
Form 1041	Beneficiary's Share of Income, Deductions, Credits, etc.
Form 1042-S[a]	Foreign Person's U.S. Source Income Subject to Withholding
Form 1099-C	Cancellation of Debt
Form 1099-OID	Original Issue Discount
Form 1099-PATR	Taxable Distributions Received From Cooperatives
Form 8805[a]	Foreign Partner's Information Statement of Section 1446 Withholding Tax
Expenses and charitable contribution reporting forms	
Form 1098	Mortgage Interest Statement
Form 1098-T	Tuition Statement
Form 1098-E	Student Loan Interest Statement
Form 1098-C	Contributions of Motor Vehicles, Boats, and Airplanes
Transaction and financial account reporting forms	
Form 1099-A	Acquisition or Abandonment of Secured Property
Form 1099-B	Proceeds From Broker and Barter Exchange Transactions
Form 1099-CAP	Changes in Corporate Control and Capital Structure
Form 1099-K	Payment Card and Third Party Network Transactions
Form 1099-S	Proceeds From Real Estate Transactions
Form 8288-A[a]	Statement of Withholding on Dispositions by Foreign Persons of U.S. Real Property Interests
Form 8300	Report of Cash Payments Over $10,000 Received in a Trade or Business
FinCEN Form 103	Currency Transaction Report by Casinos
FinCEN Form 104	Currency Transaction Report
FinCEN Form 105	Report of International Transportation of Currency or Monetary Instruments
FinCEN Form 114	Report of Foreign Bank and Financial Accounts (FBAR)

Continued next page (notes at end of table)

Table A.1 (continued)

List of Information Returns Used to Identify Nonfilers

Form number	Form name
Forms used to identify nonfilers but not used to measure nonfiler income (continued)	
Qualified account distribution, transaction, and contribution reporting forms	
Form 1099-Q	Payments From Qualified Education Programs Under Sections 529 and 530
Form 1099-SA	Distributions From an HSA, Archer MSA, or Medicare Advantage MSA
Form 3921	Exercise of an Incentive Stock Option Under Section 422 (b)
Form 3922	Transfer of Stock Acquired Through an Employee Stock Purchase Plan Under Section 423 (c)
Form 5498-ESA	Coverdell ESA Contribution Information
Form 5498-SA	HSA, Archer MSA, or Medicare Advantage MSA Information
Other forms	
Form DS-11	Application for a US Passport
Form 1097-BTC	"Bond Tax Credit"
Form 1098-Q	"Qualifying Longevity Annuity Contract Information")
Form 1099-H	"Health Coverage Tax Credit (HCTC) Advance Payments"
Form 1099-LTC	"Long-Term Care and Accelerated Death Benefits"
Form 8596	"Information Return for Federal Contracts"
SCIR	State Corporate Information Return
SIIR	State Individual Information Return
SSSTIR	State, Sales, Service or Transaction Information Return
SWIP	State Withholding Information Return

[a] Non-dependent nonfilers with Form 1042-S, Form 8805, or Form 8288-A are assumed to be foreign citizens and excluded from the non-dependent nonfiler sample.

Table A.2

Definitions of Income and Tax Measures

Income/tax type	Definition for filers[a]	Definition for nonfilers[a]
Total income	*Labor* + *Social Security* + *retirement* + *business/farm/rents/royalties* + *investment* + *other*	
Positive total income	Calculated as total income above, with any income component that is negative set equal to zero	
Income originating from work	*Labor* + *Social Security* + *retirement*	
Total spendable income	**Total income – total taxes**	
Spendable income originating from work	**Income originating from work – total taxes on income originating from work**	
Components of Income		
Labor	Wage and salary + self-employment + unemployment compensation – health savings account deduction reported on Form 1040 (line 25) – self-employed health insurance deduction reported on Form 1040 (line 29) – self-employed SEP, SIMPLE, and qualified plans deduction reported on Form 1040 (line 28) – IRA deduction reported on Form 1040 (line 32) – Form 5498 Box 1 (IRA contributions) to traditional IRAs in excess of IRA deduction reported on Form 1040 – Form 5498 Box 1 (IRA contributions) to Roth IRAs	Wage and salary + self-employment + unemployment compensation – Form 5498 Box 1 (IRA contributions) to traditional IRAs – Form 5498 Box 1 (IRA contributions) to Roth IRAs
Wage and salary	Wages, salaries, tips, etc. reported on Form 1040 (line 7) – Roth contributions reported in Form W-2 Box 12	Form W-2 Box 1 (wages, tips, other compensation) + Form W-2 Box 8 (allocated tips) – Roth contributions reported in Form W-2 Box 12
Self-employment	Net self-employment earnings from Schedule SE (Section A line 4 or Section B Part I line 4a).	Business and farm * 92.35%
Unemployment compensation	Unemployment compensation reported on Form 1040 (line 19)	Form 1099-G Box 1 (unemployment compensation)

Continued next page (notes at end of table)

Table A.2 (continued)

Definitions of Income and Tax Measures

Income/tax type	Definition for filers[a]	Definition for nonfilers[a]
Social Security	The greater of: • Social Security benefits reported on Form 1040 (line 20a), or • Form SSA-1099 Box 5 (net benefits)	Form SSA-1099 Box 5 (net benefits)
Retirement	IRA distributions + pensions and annuities	
IRA distributions	Non-rollover IRA distributions reported on Form 1040 (lines 15a and 15b)[b]	Non-rollover IRA distributions reported on Form 1099-R[b]
Pensions and annuities	Non-rollover distributions from pensions and annuities reported on Form 1040 (lines 16a and 16b)[b]	Non-rollover distributions from pensions and annuities reported on Form 1099-R[b]
Investment	Taxable interest + tax-exempt interest + dividends + gains (or losses)	
Taxable interest	Taxable interest reported on Form 1040 (line 8a) – penalty on early withdrawal of savings reported on Form 1040 (line 30)	Form 1099-INT Box 1 (interest income) + Form 1099-INT Box 3 (interest on US Savings Bonds and Treas. obligations) + Form 1120-S Box 4 (interest income) + Form 1065 Box 5 (interest income) – Form 1099-INT Box 2 (early withdrawal penalty)
Tax-exempt interest	Tax-exempt interest reported on Form 1040 (line 8b)	Form 1099-INT Box 8 (tax-exempt interest) + Form 1099-DIV Box 10 (exempt-interest dividends)
Dividends	Ordinary dividends reported on Form 1040 (line 9a)	Form 1099-DIV Box 1a (total ordinary dividends) + Form 1120-S Box 5a (ordinary dividends) + Form 1065 Box 6a (ordinary dividends)
Gains (or losses)	Capital gain (or loss) reported on Form 1040 (line 13) + other gains reported on Form 1040 (line 14)	Form 1099-DIV Box 2a (total capital gain distr.)
Business/farm/ rents/royalties	Business and farm + rents, royalties, etc. – self-employment – deductible portion of self-employment tax reported on Form 1040 (line 27)	Business and farm + rents, royalties, etc. – self-employment – 0.5*self-employment tax

Continued next page (notes at end of table)

Table A.2 (continued)

Definitions of Income and Tax Measures

Income/tax type	Definition for filers[a]	Definition for nonfilers[a]
Business and farm	Business income on Form 1040 (line 12) + farm income on Form 1040 (line 18)	Form 1099-MISC [Boxes: 5 (fishing boat proceeds); 6 (medical and health care payments); 7 (nonemployee compensation); 10 (crop insurance proceeds); 14 (gross proceeds paid to an attorney)] + Form 1099-G [Boxes: 6 (taxable grants); 7 (agriculture payments); 9 (market gain)] + Form 1065 Box 4 (guaranteed payments)
Rents, royalties, etc.	Rental real estate, royalties, partnerships, S corporations, trusts, etc. on Form 1040 (line 17)	Form 1099-MISC Box 1 (rents) + Form 1099-MISC Box 2 (royalties)
Other	Other income reported on Form 1040 (line 21) + alimony received reported on Form 1040 (line 11) – alimony paid reported on Form 1040 (line 31a)	Form 1099-MISC Box 3 (other income) + Form 1099-MISC Box 8 (substitute payments in lieu of dividends or interest) + Form 1099-G Box 5 (RTAA payments) + Form W2-G Box 1 (gross winnings)
Taxes		
Total taxes	***Payroll taxes* + *federal income taxes***	
Total taxes on income originating from work	***Payroll taxes* + *federal income taxes on income originating from work***	
Payroll taxes	OASDI tax on wages + HI tax on wages + (0.5 * self-employment tax)	
OASDI tax on wages	Form W-2 Box 4 (Social Security tax withheld) + (6.2% * unreported tips reported on Form 4137 line 6) + (6.2% * wages reported on Form 8919 line 6), subject to maximum withholding ($7,347 in 2016)	Form W-2 Box 4 (Social Security tax withheld), subject to maximum withholding ($7,347 in 2016)
HI tax on wages	Form W-2 Box 6 (Medicare tax withheld) + (1.45% * unreported tips reported on Form 4137 line 6) + (1.45% * wages reported on Form 8919 line 6)	Form W-2 Box 6 (Medicare tax withheld)
Self-employment tax	Deductible portion of self-employment tax reported on Form 1040 (line 27) * 2	Minimum of 15.3% * self-employment or federal income tax withheld on Form 1099-MISC and Form 1099-G

Continued next page (notes at end of table)

Table A.2 (continued)

Definitions of Income and Tax Measures

Income/tax type	Definition for filers[a]	Definition for nonfilers[a]
Federal income taxes	Total tax reported on Form 1040 (line 63) – self-employment tax – payroll taxes on wages not reported on Form W-2 – credits treated as tax payments[c]	Sum of federal income tax withheld on Form W-2, Form SSA-1099, Form 1099-R, Form 1099-INT, and Form 1099-DIV + federal income tax withheld on Form 1099-MISC and Form 1099-G in excess of self-employment tax + Form W-2 Social Security tax withheld in excess of OASDI tax on wages
Federal income taxes on income originating from work	*Federal income taxes* * (**income originating from work** / **total income**)	

Notes

[a] The line numbers referenced in the definitions refer to form for tax year 2016.

[b] Non-rollover distributions exclude rollovers, Roth conversions, and Section 1035 exchanges of annuity contracts. In addition, we exclude from retirement income any amounts attributable to the recharacterization of IRA contributions, the return of excess contributions, or distributions related to prohibited transactions. For a detailed explanation of how we reconcile information reported by taxpayers on Form 1040 and associated Forms and Schedules with information reported by recordkeepers on Form 1099-R and Form 5498, see Brady and Bass (2020a).

[c] In 2016, credits treated as tax payments included the earned income credit (Form 1040 line 66a), additional child tax credit (line 67), American opportunity credit (line 68), and the net premium tax credit (line 69).