Counting Elusive Nonfilers Using IRS Rather Than Census Data

Patrick Langetieg, Mark Payne, and Alan Plumley (IRS Research, Applied Analytics, and Statistics)[1]

Although not all individuals have a Federal income tax filing requirement, every year millions of required returns are not filed on time or at all. The Internal Revenue Service (IRS) would like to help taxpayers meet their filing obligations, so estimates of the extent of nonfiling[2] provide helpful insights. For many years, the IRS has used Census Bureau survey data when estimating the extent of individual income tax filing compliance. This paper describes an updated methodology for estimating the number of nonfilers and the corresponding Voluntary Filing Rate using IRS data alone—without relying on Census data. The reasons for this new approach are also explained.

The Voluntary Filing Rate (VFR)

The IRS has estimated the VFR since the mid-1990s to examine factors that influence individual income tax filing compliance.[3] It is defined for a given tax year as:

$$\text{VFR} = \frac{\text{Number of Required Returns Filed on Time}}{\text{Total Number of Returns Required To Be Filed}}$$

We derive the VFR numerator from IRS population data encompassing all filed returns, which we classify as *timely* or not based on a comparison of the filing date and the relevant filing deadline (accounting for all valid extensions) and as *required* or not based on a comparison of all relevant income (reported by the taxpayer) and the filing thresholds in place for the year in question. We have historically estimated the VFR denominator from the Census Bureau's Current Population Survey Annual Social and Economic Supplement (CPS ASEC, hereinafter referred to as CPS), grouping individuals into assumed tax units (e.g., marrieds, singles, or heads of households) and applying comparable logic to estimate whether a tax return was required.

Preliminary estimates for the denominator were first constructed in a necessarily approximate manner since the CPS lacks some of the information needed to confirm various tax-related concepts. Initially, both the numerator and denominator were estimated from samples each year. However, when the IRS began storing data on the whole population in a form that was accessible to IRS research staff, we began estimating the numerator from population data. That allowed us to examine in more detail what type(s) of taxpayers were driving fluctuations in the numerator.

Subsequent work addressed various measurement issues surrounding the numerator and denominator of the VFR ratio. This work included ensuring that the numerator and denominator more precisely represented the same population of taxpayers (U.S. residents over the age of 14), and that they reflected the same definitions (as much as the data would allow) for the requirement to file and what it means for a required return to be timely filed for VFR purposes.[4]

---

[1] The views expressed in this paper are those of the authors, and do not necessarily represent the positions of the Internal Revenue Service. This paper is an extension of our earlier papers: Erard, et al. (2012), and Erard, et al. (2014).

[2] We use the term "nonfiler" to include only those who are required to file a Form 1040 for income tax or employment tax purposes, but did not file a return on time. Therefore, nonfilers include taxpayers who filed a required return after the due date (plus extension). This compliance-oriented definition of nonfilers differs from a policy-oriented definition, which includes those who might be eligible for (current or proposed) benefits offered through the tax system without incurring a tax obligation. Timely filers exclude those who have no filing requirement, but who file solely to claim a refund of withholding or to claim a refundable credit.

[3] See Plumley (1996).

[4] We include in the numerator only those required returns that are filed on time by December 31 of the primary filing year. This excludes returns that are considered timely (e.g., due to combat extensions), but are filed much later than most, and also excludes returns filed by December 31 but after their filing deadline. Setting December 31 as the cut off date allows for a consistent measure to be produced each year.

Accounting for Missing Income

Even after the refinements mentioned above, it became necessary to augment the Census data used to construct the denominator of the VFR to account more fully for certain types of income that are reported to the IRS, but not fully reported on Census surveys (such as pensions, Social Security income, sole proprietor income, and unemployment compensation). An understatement of income in the denominator of the measure would contribute to an overstatement of the VFR. Erard, *et al.* (2014) documents the differences in the amounts of these types of income reported on the CPS survey versus what is reported on the third-party information returns sent to the IRS (in the case of pension and Social Security income, as well as unemployment compensation), or what is reported on filed income tax returns (in the case of self-employment income). To address these discrepancies, we developed econometric methodologies for imputing the missing income to the CPS records.

Insights from Tax Gap Research

In addition to estimating the number of nonfilers, the IRS also estimates the nonfiling tax gap—the estimated amount of tax not paid on time by those who do not file on time. The most recent estimate of the individual income tax nonfiling gap was calculated two ways: one based primarily on Census data matched to IRS data,[5] and another based mostly on IRS administrative data (such as income reported by third parties).[6] A curious observation from those estimates was that while they yielded very similar estimates of both the number of nonfilers and the corresponding dollar value of the gap, they suggested that the number of nonfilers was much larger than had been previously understood and the Census Method suggested that the number of required returns was much less than the number derived from the Administrative Data Method. That caused us to compare estimates of the VFR numerator and denominator derived in three ways: (1) using the latest VFR methodology; (2) using the Census Method for estimating the nonfiling gap; and (3) using the Administrative Data Method for estimating the gap. As Table 1 shows, both tax gap methods yielded an estimated 15 million nonfilers for Tax Year 2010, while the latest VFR methodology estimated only 6.3 million nonfilers for the same year. However, the Census Method for estimating the tax gap suggested that both numerator and denominator were around 11 million returns fewer than derived from the Administrative Data Method. This undercounting of required returns using Census data affected the denominator of the VFR, but not the numerator (which is tabulated from IRS data), resulting in an inflated VFR relative to either of the tax gap methods.[7] This observation exposes the inadequacy of the Census data to estimate the number of required returns—even after imputing pension, Social Security, sole proprietor, and unemployment compensation income to the Census data. Given the total counts observed in IRS data, it seems clear that many individuals who appear in the adjusted CPS data as *not* having enough income to be required to file actually *do* have a filing requirement. This affects both the number of required returns among those who file on time (the numerator of the VFR) and the estimated number of nonfilers. The common thread is the underreporting of income in the Census data.

TABLE 1. Thousands of Returns in VFR Components Estimated by Different Methods, Tax Year 2010

Item	VFR Method	Census Method	Administrative Data Method
Numerator (required returns filed on time)	115,900	105,001	115,900
Denominator (total required returns)	122,200	119,967	130,787
Difference (implied number of nonfilers)	6,300	14,966	14,937
Numerator/Denominator (Implied VFR)	94.8%	87.5%	88.6%

[5] The Census Bureau has developed the ability to assign an anonymous Protected Identification Key (PIK) to most respondents in the CPS and to all of the records Census routinely receives from the IRS for the population (including selected data from income tax returns and from third-party information documents). This allows Census to create a matched anonymized dataset containing both Census and IRS data for a representative sample of the population. However, there are some CPS records that cannot be matched to the IRS data because a PIK could not be assigned to them with adequate certainty. We therefore restricted our analysis to the records that could be matched, and re-weighted them to represent the entire population.

[6] See Langetieg, et al. (2016).

[7] The VFR method for estimating the denominator is very similar to the tax gap Census Method. The main difference is that the VFR method uses the entire CPS dataset, while the Census Method uses a subset of the CPS that is matched to IRS return data, which makes it possible to impute income to the Census records in a more sophisticated way.

Attempts To Impute More Income to the Census Data

Several attempts to impute still more income to the CPS yielded unsatisfactory results. One approach was to use the matched subsample, basing total income on the amount reported on the tax return (when available) instead of what was reported in the Census survey. Another approach was a backend imputation of gross income (without regard to source) that was calibrated to totals in IRS data. Neither produced a time series of required returns that was sufficiently similar to that produced from IRS data alone, particularly for key years when the number of required filers was known to have spiked.[8] Although further research may allow us to make adequate adjustments to the CPS data that would yield accurate estimates of the VFR, we concluded that it was necessary to estimate the VFR based solely on IRS administrative data.

Table 2A provides the population estimates based on the matched sub-sample. The numerator, denominator, and implied VFR for several approaches are provided, including the two attempts to correct (adjust) gross income mentioned above. It's important to note that the numerator and denominator estimates provided in Table 2A are from Census survey data so the numerator is not the actual IRS estimate.

TABLE 2A. VFRs and Millions of Returns in VFR Components Estimated by Census-Based Methods, Tax Years 2007 to 2014

Item	Tax Year							
	2007	2008	2009	2010	2011	2012	2013	2014
All Census Tax Units	239.0	240.6	243.9	245.9	247.8	249.8	251.4	253.9
All Census Tax Units that Filed a Return	146.0	135.4	133.7	133.8	137.0	138.2	138.2	140.6
	Previous Census-Based Methodology							
Numerator (Filed Required Returns)	113.5	111.5	108.7	108.9	110.9	112.2	112.7	114.9
Denominator (Total Required Returns)	124.0	124.2	121.0	120.9	123.2	124.8	126.0	129.8
VFR	91.5%	89.8%	89.8%	90.1%	90.0%	89.9%	89.4%	88.5%
	Updated Census-Based Methodology							
Numerator (Filed Required Returns)	114.3	112.5	109.8	110.2	112.1	113.4	114.0	116.4
Denominator (Total Required Returns)	125.9	127.0	124.3	124.7	126.6	128.5	130.0	133.7
VFR	90.8%	88.6%	88.3%	88.4%	88.5%	88.2%	87.7%	87.1%
	Updated Census Based-Methodology Form 1040 Income Amount Used When Available							
Numerator (Filed Required Returns)	125.1	118.5	116.7	117.8	120.0	120.8	120.5	123.6
Denominator (Total Required Returns)	136.7	133.0	131.2	132.3	134.5	135.9	136.4	140.9
VFR	91.5%	89.1%	88.9%	89.0%	89.2%	88.9%	88.3%	87.7%
	Updated Census-Based Methodology Backend Adjustment to Gross Income to All Tax Units							
Numerator (Filed Required Returns)	117.2	113.8	111.2	111.8	113.9	115.2	115.6	118.2
Denominator (Total Required Returns)	137.8	134.1	131.7	132.7	134.9	136.6	137.8	141.8
VFR	85.1%	84.9%	84.4%	84.3%	84.4%	84.3%	83.9%	83.4%
	Updated Census-Based Methodology Form 1040 Income Amount Used When Available Backend Adjustment to Gross Income to Tax Units Without a 1040							
Numerator (Filed Required Returns)	126.5	119.4	117.4	118.4	120.8	121.7	121.4	124.4
Denominator (Total Required Returns)	147.0	139.8	137.9	139.3	141.8	143.1	143.6	148.1
VFR	86.1%	85.4%	85.1%	85.0%	85.2%	85.0%	84.5%	84.0%

[8] For example, the number of required filers spiked in Tax Year 2007 because in order to receive the one-time Economic Stimulus Payment, people had to file a tax return for 2007. Many more returns were filed that year that had a requirement to file for regular tax reasons so the VFR spiked that year then fell somewhat in the subsequent years.

The Previous Methodology estimates for 2010 are slightly different from the estimates provided in Table 1 for two reasons. First, the Table 2A numerators include all required returns filed during the calendar year they were due so both timely and late required returns are included in the numerator. Second, the reweighting methodology to account for missing PIKs[9] was changed to an inverse probability method for this round of estimates.

The difference between the Previous and Updated Methodology estimates is due to a change to the sole proprietor income imputation. Previously, the sole proprietor imputation for the Census Method was calibrated to match the proportion of individuals with sole proprietor income observed on Form 1040. This approach was not used for the Administrative Data Method. In order to be consistent across methods, the Census Method was changed to match the Administrative Data Method. This change leads to a larger number of individuals receiving imputed sole proprietor income and, as a result, more required returns.

The Previous Method and Updated Method estimates show the same denominator deficiency discussed from Table 1. The last three sets of results show the effect of using Form 1040 income when available and/or computing a backend adjustment to gross income. These methods appear to resolve the denominator deficiency, some more reasonably than others.

Table 2B provides the same information shown in Table 2A, except the Census-based numerator estimates have been replaced with the actual IRS estimates.

TABLE 2B. VFRs and Millions of Returns in VFR Denominators Estimated by Census-Based Methods and Numerators Estimated From IRS Data, Tax Years 2007 to 2014

Item	Tax Year							
	2007	2008	2009	2010	2011	2012	2013	2014
All Tax Units	239.0	240.6	243.9	245.9	247.8	249.8	251.4	253.9
Tax Units that Filed a Return	146.0	135.4	133.7	133.8	137.0	138.2	138.2	140.6
	Previous Census Based-Methodology							
IRS Numerator (Filed Required Returns)	117.2	116.3	113.4	115.4	117.2	117.4	118.5	120.5
Denominator (Total Required Returns)	124.0	124.2	121.0	120.9	123.2	124.8	126.0	129.8
VFR	94.5%	93.6%	93.7%	95.5%	95.1%	94.1%	94.0%	92.8%
	Updated Census-Based Methodology							
IRS Numerator (Filed Required Returns)	117.2	116.3	113.4	115.4	117.2	117.4	118.5	120.5
Denominator (Total Required Returns)	125.9	127.0	124.3	124.7	126.6	128.5	130.0	133.7
VFR	93.1%	91.6%	91.2%	92.5%	92.6%	91.4%	91.2%	90.1%
	Updated Census-Based Methodology Form 1040 Income Amount Used When Available							
IRS Numerator (Filed Required Returns)	117.2	116.3	113.4	115.4	117.2	117.4	118.5	120.5
Denominator (Total Required Returns)	136.7	133.0	131.2	132.3	134.5	135.9	136.4	140.9
VFR	85.7%	87.4%	86.4%	87.2%	87.1%	86.4%	86.9%	85.5%
	Updated Census-Based Methodology Backend Adjustment to Gross Income to All Tax Units							
IRS Numerator (Filed Required Returns)	117.2	116.3	113.4	115.4	117.2	117.4	118.5	120.5
Denominator (Total Required Returns)	137.8	134.1	131.7	132.7	134.9	136.6	137.8	141.8
VFR	85.1%	86.7%	86.1%	87.0%	86.9%	85.9%	86.0%	85.0%
	Updated Census-Based Methodology Form 1040 Income Amount Used When Available Backend Adjustment to Gross Income to Tax Units Without a 1040							
IRS Numerator (Filed Required Returns)	117.2	116.3	113.4	115.4	117.2	117.4	118.5	120.5
Denominator (Total Required Returns)	147.0	139.8	137.9	139.3	141.8	143.1	143.6	148.1
VFR	79.7%	83.2%	82.2%	82.8%	82.7%	82.0%	82.5%	81.4%

[9] Refer to footnote 5.

The Table 2B estimates show that each of the attempts to account for the denominator deficiency led to a lower overall VFR in 2007. As discussed later, there is good reason to expect the 2007 VFR to be higher in 2007 than the following years, so this outcome suggests that the Census Method does not accurately estimate the number of required returns in the population (the VFR denominator). Further research is needed to understand and fully address this deficiency before moving forward with the Census Method.

The Administrative VFR Methodology

The VFR based on IRS administrative data relies on information from filed individual income tax returns and third-party information returns. The population is made up of filed returns, whether timely or late, and those who do not appear on a filed return—"not-filers." As in the CPS-based estimate, the numerator is the count of all required returns that were filed on time (including extensions). The denominator is the count of all required returns in the population.[10] For filed returns, the determination of whether the return was required is based on the income reported on the return. For most filers, this involves comparing the amount of gross income with thresholds linked to filing status and whether the taxpayer is over 65 years of age. Taxpayers with self-employment income are required to file if their net self-employment income is $433 or greater.

For individuals who do not appear on a filed return, third-party information documents are the main sources for estimating income. But, given the fact that most self-employment income is not reported on information documents, this type of income is imputed to not-filers using models developed from filed returns. In addition, filing status and family units are imputed. The method for determining the number of required returns among not-filers involves the following steps:

- Identify all individuals who appeared on a third-party information return for the tax year in question, but who did not appear on a filed return as a primary or secondary taxpayer for the given tax year by the end of the second year after the conclusion of the tax year.[11] This excludes (as potential primary taxpayers or spouses) those for whom no third-party information return was filed for the year in question.[12] The two-year cutoff for using information on late returns is put in place to limit the lag time after the conclusion of filing before the VFR estimate can be made while also ensuring consistency across the time series. The disadvantage is that the potentially more accurate information on income and tax unit structure available on the returns that are filed after this two-year cutoff is sacrificed.
- Identify the known income, prepayments, and State of residence for these not-filers from third-party information documents and other administrative tax data sources (e.g., Master File). In addition, the social security master file (DM-1) was used to obtain the age and gender of each individual. Finally, the individuals were matched to filed returns to determine which ones had been claimed as dependents.
- Impute self-employment income to the individual potential not-filers.
- Assign spouses and filing status and a number of dependents to the remaining not-filers using their age and gender, so that the combined age and filing status distribution of timely filers, late filers, and not-filers matches the corresponding distributions of singles, marrieds, heads of households, and dependents in CPS data, after the aggregate counts from those on filed returns is taken out.
- Convert net sole proprietor and farm income below the $433 threshold to gross income using multipliers based on mean ratios of gross to net observed on filed returns.
- Determine whether a tax return was required to be filed by comparing gross income for the tax unit to the filing thresholds for the relevant filing status and checking whether the net self-employment income threshold was exceeded.

The combined count of primary and secondary taxpayers on filed returns, dependents claimed on filed returns who do not file on their own, and not-filers who are not claimed as dependents is fairly close to

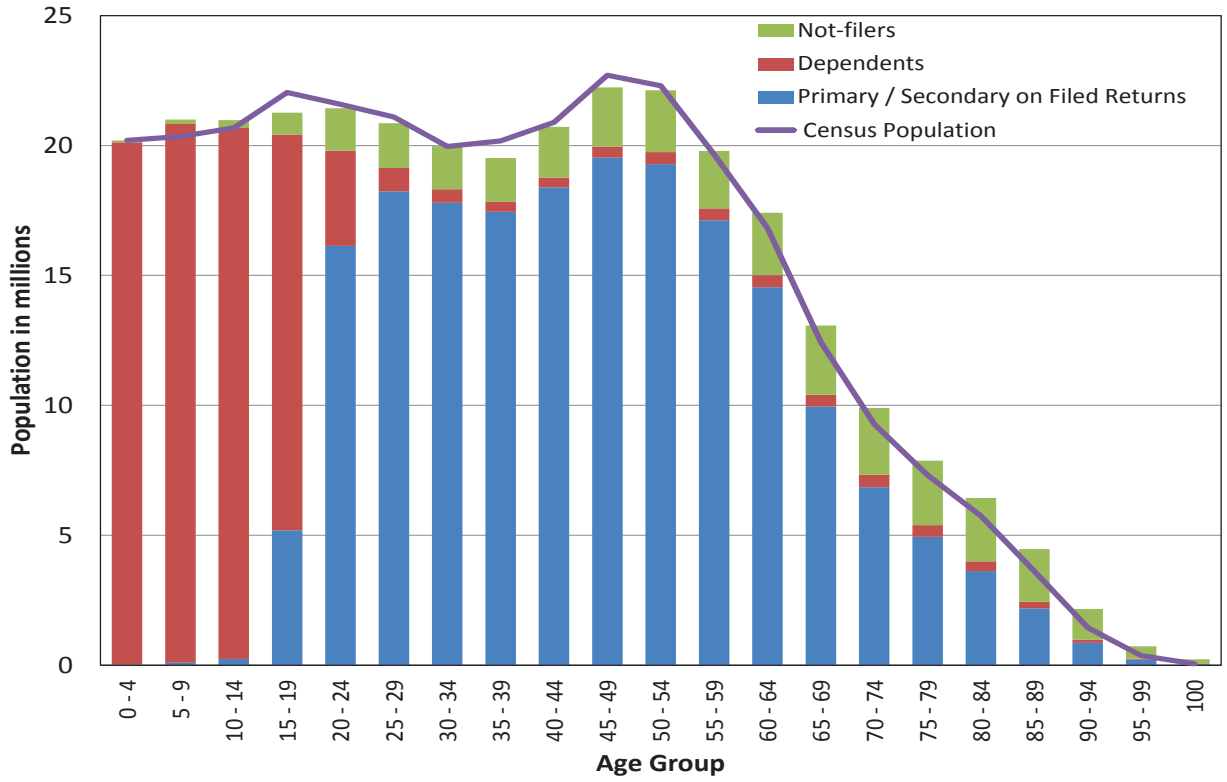
[10] Returns with foreign addresses are excluded from the counts of required filers in the numerator and denominator.

[11] Note that this treats all dependents as potentially required to file a return in their own right.

[12] This approach excludes people who had income only from sources not subject to third-party reporting (such as self-employment income).

Census-based estimates of U.S. population totals at most age levels (see Figure 1). Even though some not-filers are likely not present on information returns, their numbers appear to be offset by other individuals who are present on information returns but not counted in the U.S. Census, such as those who are living overseas. Hence, the population included in the VFR methodology is close to that of the whole U.S. population.

FIGURE 1. Population by Age: IRS Administrative Data vs. Decennial Census, TY2010



Imputation of Net Self-Employment Income

Most self-employment income is not reported on third-party information documents. About 12.9 million of the 25.7 million individuals reporting Schedule C income received a Form 1099 Miscellaneous Income in TY 2014 and an additional 800,000 were matched to a Form 1099 K. Approximately 400,000 were matched to both. However, these forms often provide just a partial picture of the income earned. And, in some cases the forms overestimate net income because they do not provide information about expenses and other deductions to gross income. Given the deficient coverage of information documents, we impute self-employment income to the not-filers based on models derived from filed tax returns.

The limitations of this approach include the fact that self-employment earnings are significantly under-reported on tax returns and that not-filers are likely different from filers. On the one hand, one reason why taxpayers may not file is because their net income is low enough to keep them below the filing threshold. At the same time, not-filers may be more likely to be self-employed and to have most of their income not subject to third-party reporting since this is the characteristic that provides the opportunity to avoid detection by IRS nonfiler enforcement programs. Thus, while our imputations aim to make the sole proprietor earnings of not-filers consistent with what is observed in tax return data, this measure may fall short of true earnings, meaning that we are not fully able to account for all returns that have a filing requirement solely as a result of the $433 self-employment earnings threshold.

The econometric framework for imputing self-employment income involves prediction equations for the presence and magnitude of this form of income as a function of individual characteristics, such as age, gender, and region, as well as sources and amounts of earnings, including wages, interest, Social Security and pension

income, unemployment compensation, and nonemployee compensation. In addition, the imputation of self-employment income is tailored to the ways in which this type of income relates to the filing requirement. If net sole proprietor income exceeds $433, that by itself creates a filing requirement because some self-employment tax is owed. Sole proprietors are also required to file if gross income from self-employment and other sources exceeds the filing threshold for their filing status. Thus, we also convert the imputed net amounts of sole proprietor income into gross income amounts by multiplying them by the average ratio of gross to net income in the population individual return data. Different gross-up factors are applied to different net self-employment earnings categories, with negative factors being applied when net self-employment income is negative (thereby resulting in a positive value for gross self-employment income) and a positive factor being applied when net self-employment income is between $0 and $433.[13] Thus, the econometric framework aims to estimate the likelihood that an individual has earnings falling into one of the following three categories: negative net self-employment earnings, net self-employment earnings between $1 and $433, and net self-employment earnings in excess of $433.

The econometric framework involves three separate models. The first is a probit specification for the likelihood that a filing unit has nonzero self-employment earnings:

$$SE^* = \gamma' x + \mu \quad (1)$$

where SE^* is a latent variable describing the propensity for net self-employment earnings to be present, x is a vector of explanatory variables, and γ is a vector of coefficients to be estimated. The error term u is assumed to follow the standard normal distribution. Estimation of this model permits us to develop a prediction equation for the unconditional likelihood that an individual has net income from self-employment. Results for this model applied to Tax Year 2014 data are shown in Table 3.

TABLE 3. Probit Model for the Presence of Self-Employment Income, Tax Year 2014

Variable	Parameter Estimate	Chi-square
Intercept	-1.475	(409.2)**
logage	0.152	(63.5)**
male	-1.101	(149.4)**
logage*male	0.341	(210.5)**
west	-0.066	(33.8)**
midwest	-0.108	(84.8)**
northeast	-0.042	(11.7)**
wagesind	0.585	(185.8)**
interestind	0.195	(68.1)**
socsecind	-0.471	(7.6)**
pensionind	0.215	(19.4)**
F1099miscind	0.717	(1847.1)**
unempcompind	0.042	(0.1)
logwages	-0.135	(1045.3)**
loginterest	-0.047	(90.4)**
logpension	-0.044	(70.6)**
logsocsec	-0.028	(2.5)
logunempcomp	-0.011	(0.3)
lognonempcomp	0.148	(4701.8)**
Number of observations	198,704	
R^2	0.217	
Max-rescaled R^2	0.408	

*significant at .05 level; ** significant at .01 level.
See the appendix for variable descriptions.

[13] Different multiplier factors are applied for net self-employment income less than -$5,000 and in the ranges of -$5,000 to -$3,000, -$3,000 to -$1,000, -$1,000 to 0, and $0 to $433.

Estimation of Net Self-Employment Income Category

Our second model is an ordered probit specification for the dollar amount category that net self-employment earnings fall into when they are present: negative, $1 to $433, or over $433:

$$I^{*}_{SE} = \delta' x + \nu \tag{2}$$

where I^{*}_{SE} is a latent variable for the propensity for net self-employment earnings to fall into one of these categories, x is the same set of explanatory variables used in our probit model, δ is a coefficient vector to be estimated, and v is a standard normal random disturbance. The model also includes a limit parameter l to be estimated.[14] The indicator I_{SE} for the net self-employment earnings category is assigned as follows:

$$I_{SE} = \begin{cases} 1 & \textit{net earnings} < \$0 \\ 2 & \$0 < \textit{net earnings} <= \$433 \\ 3 & \textit{net earnings} > \$433. \end{cases} \tag{3}$$

The estimation results for Tax Year 2014 are shown in Table 4.

TABLE 4. Ordered Probit Models for the Category of Self-Employment Income, Tax Year 2014

Variable	Parameter Estimate	t-statistic
Intercept	2.187	(12.5)**
logage	-0.346	(-7.6)**
male	-0.742	(-3.4)**
logage*male	0.205	(3.6)**
west	0.037	(1.6)**
midwest	0.122	(5.0)
northeast	0.223	(8.5)**
wagesind	1.384	(16.4)**
interestind	-0.019	(-0.4)
socsecind	-0.085	(-0.3)
pensionind	-0.036	(-0.4)
F1099miscind	-0.072	(-2.2)*
unempcompind	0.339	(1.1)
logwages	-0.212	(-25.8)**
loginterest	0.020	(2.0)*
logpension	-0.026	(-2.4)*
logsocsec	-0.027	(-0.8)
logunempcomp	-0.054	(-1.4)
lognonempcomp	0.082	(21.9)**
Limit	0.207	(38.9)**
Values of dependent variable I_{SE}	Number of observations	
1 = (SE Income < 0)	5,730	
2 = (0 < SE Income <= 433)	1,408	
3 = (SE Income > 433)	17,232	
Total number of observations	24,370	
Missing values	5	
Log Likelihood	-16,146	

* significant at .05 level; ** significant at .01 level.
See the appendix for variable descriptions.

[14] This parameter serves as a threshold for separating the various levels of the response variable.

Imputation of Net Self-Employment Income Amount

Our third model is a regression specification for the magnitude of net self-employment earnings when they exceed $433. Although a taxpayer is required to file a tax return as long as net earnings from self-employment exceed $433, it is desirable to predict their actual magnitude for the tax gap analysis and other research efforts. For instance, this will facilitate an econometric analysis of reporting compliance among self-employed taxpayers. Our specification is:

$$ln(SE) = \beta' x + \varepsilon, \tag{4}$$

where *ln(SE)* represents the natural log of net self-employment earnings, x is the same set of explanatory variables used in the preceding models, β is a vector of coefficients to be estimated, and ε is assumed to be a normal random error term with mean zero and standard deviation σ. Under this specification, the distribution of self-employment earnings is assumed to be log normal. The estimation results for this model for Tax Year 2014 are shown in Table 5.

TABLE 5. Regression Analysis for the Amount of Self-Employment Income > $433, Tax Year 2014

Variable	Parameter Estimate	t-statistic
Intercept	7.723	(44.1)**
logage	0.368	(8.0)**
male	-1.031	(-4.7)**
logage*male	0.340	(5.9)**
west	0.149	(6.3)*
midwest	-0.004	(-0.2)
northeast	0.083	(3.2)**
wagesind	0.303	(3.7)**
interestind	0.127	(2.6)*
socsecind	-2.700	(-7.4)**
pensionind	-0.151	(-1.3)
F1099miscind	-1.171	(-29.3)**
unempcompind	0.228	(0.7)
logwages	-0.110	(-13.2)**
loginterest	0.057	(5.6)**
logpension	0.007	(0.6)
logsocsec	0.201	(5.3)**
logunempcomp	-0.061	(-1.4)
lognonempcomp	0.138	(33.7)**
R^2	0.213	
Adjusted R^2	0.213	
Root MSE	1.213	
Coefficient of Variation	13.515	
N	17,856	

* significant at .05 level; ** significant at .01 level.
See the appendix for variable descriptions.

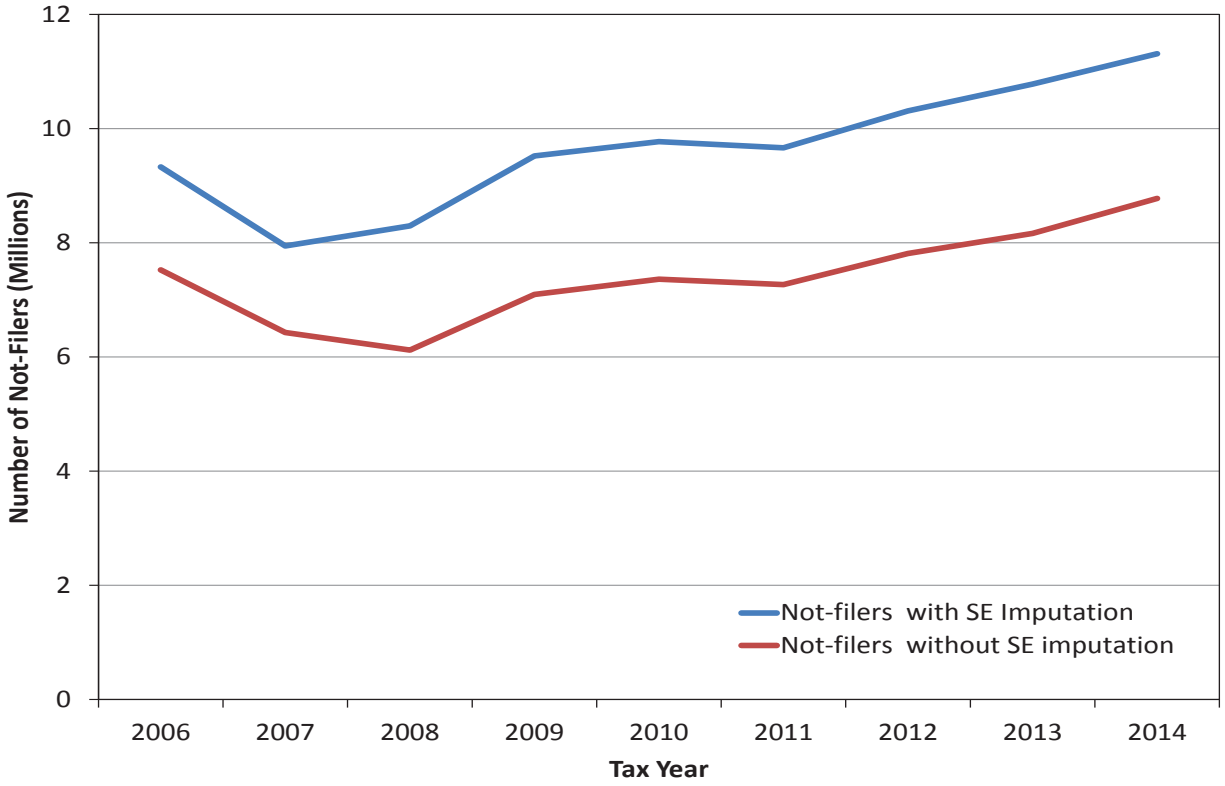
For not-filers flagged for imputation of net self-employment earnings in excess of $433, the parameter estimates are used to impute earnings (SE) as follows:

$$SE = exp\left(\tilde{\beta}' x + e\right), \tag{5}$$

where $\widetilde{\beta}'$ is the estimated coefficient vector and e is a random draw from a normal distribution with mean zero and standard deviation equal to the root mean-squared error of the regression.

For not-filers flagged for imputation of net self-employment losses, we assign a random draw from a lognormal distribution with parameters selected based on summary statistics for reported losses from administrative data. Finally, for not-filers flagged for imputation of net self-employment earnings between $1 and $433, we assign the mean reported earnings among taxpayers reporting earnings in that range. Imputing self-employment income causes the number of estimated required not-filer returns to increase by between 25 percent and 35 percent, or about 2.5 million returns, as illustrated in Figure 2.

FIGURE 2. Counts of Not-Filers With and Without Imputation of Self-Employment Income, TYs 2006–2014

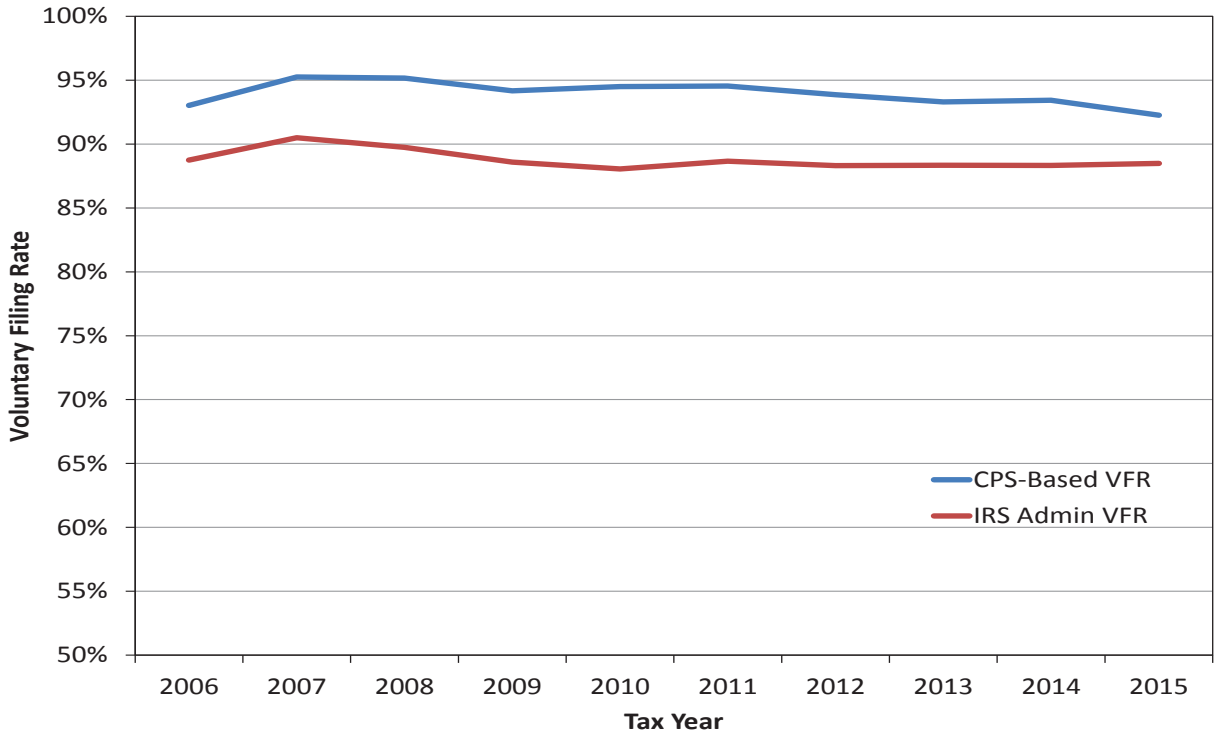


The Resulting VFR

Figure 3 shows the resulting VFR using the IRS Administrative Data Method and compares it with the prior estimates that used IRS filing data for the numerator and CPS as the starting point for the denominator. Given the higher estimate of required returns in the population, the VFR based on IRS data is about 5 percentage points lower than the VFR using CPS data. Both estimates show a significant increase in the VFR in 2007 due to the tax rebates associated with the Economic Stimulus Act of 2008.[15] The series based on IRS data shows a more rapid decline followed by considerable stability. By contrast, the series using CPS shows a slower decline from the peak in Tax Year 2007 with additional reductions more recently. But, by TY 2014, the average difference between the two estimates returned to within 0.3 percent of the difference in TY 2007.

[15] This law provided eligible taxpayers with a rebate of up to $300 per person ($600 for married couples filing jointly) and $300 per dependent child under the age of 17. In order to receive these payments, taxpayers had to file a 2007 tax return.

FIGURE 3. Individual Income Tax Voluntary Filing Rate, TYs 2006–2015



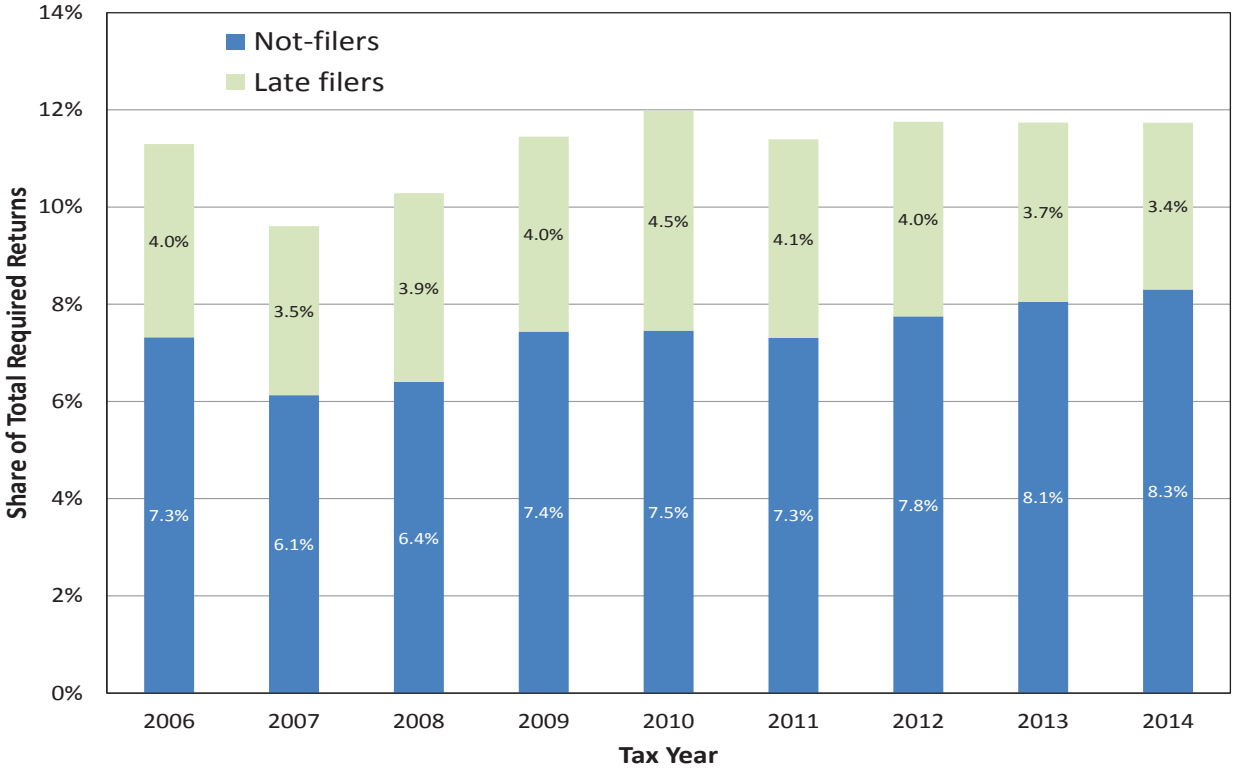
NOTE: The IRS Administrative VFR for Tax Year 2015 is provisional since it does not reflect a full two years of late-filer data.

TABLE 6. Voluntary Filing Rate and Related Estimates, TYs 2006–2015

Tax Year	Millions of Required Returns			VFR (Ratio)
	Total Population (Denominator)	Timely Filed (Numerator)	Nonfilers (Difference)	
2006	127.3	113.0	14.3	88.8%
2007	129.5	117.2	12.3	90.5%
2008	129.4	116.3	13.2	89.7%
2009	128.0	113.4	14.6	88.6%
2010	131.0	115.4	15.6	88.1%
2011	132.1	117.2	15.0	88.7%
2012	132.9	117.4	15.5	88.3%
2013	134.1	118.5	15.6	88.3%
2014	136.4	120.5	15.9	88.3%
2015	138.4	122.5	15.9	88.5%

Given that the VFR has been quite stable over the past four years, this means that the percent of nonfilers has also been equally stable. But, in recent years, the share of nonfilers that remain as not-filers (rather than filing late returns) by two years after the end of the tax year has increased (Figure 4). This may be due to the reduction in IRS resources devoted to nonfiler enforcement. Since many of the returns that are secured from enforcement do not come in until the third year following the tax year, this effect would likely be even more evident if these returns were also included in the late-filing population.

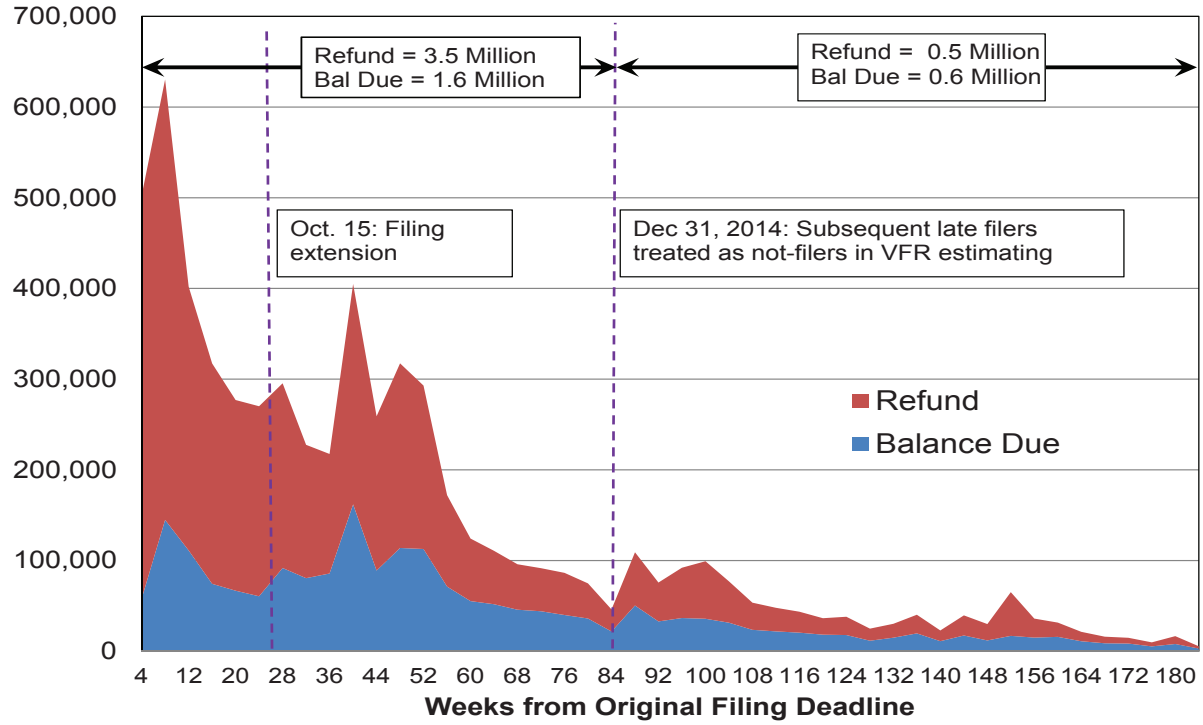
FIGURE 4. Nonfilers as a Percent of Total Required Population, TYs 2006–2014



Figures 5 and 6 show the counts of the late returns over the four-year period following Tax Year 2012. As you can see, in the VFR methodology about 5.1 million of the late returns (those filed within two years of the end of the tax year) are used in the estimation, while the remaining 1.1 million late returns filed three and four years after the end of the tax year are treated as not-filers. Figure 5 shows the breakdown of late returns between refund and balance-due conditions. Refunds are much more common among late returns filed within one year of the end of the tax year (2.2 million out of 2.9 million) than those filed in the second year (1.3 million out of 2.3 million). Refund returns make up only 60 percent of the returns filed in the third and fourth years following the end of the tax year. Figure 6 shows the breakdown between the late returns secured through enforcement efforts and those that come in without enforcement.[16] Because nonfiler notices do not start going out until the end of the filing year, very few returns that are filed within the first year can be attributed to enforcement. But, in the second year about 400,000 out of 2.3 million were filed after the taxpayers were sent nonfiler notices. And, in the third year, about 170,000 out of 700,000 returns had been sent nonfiler notices.

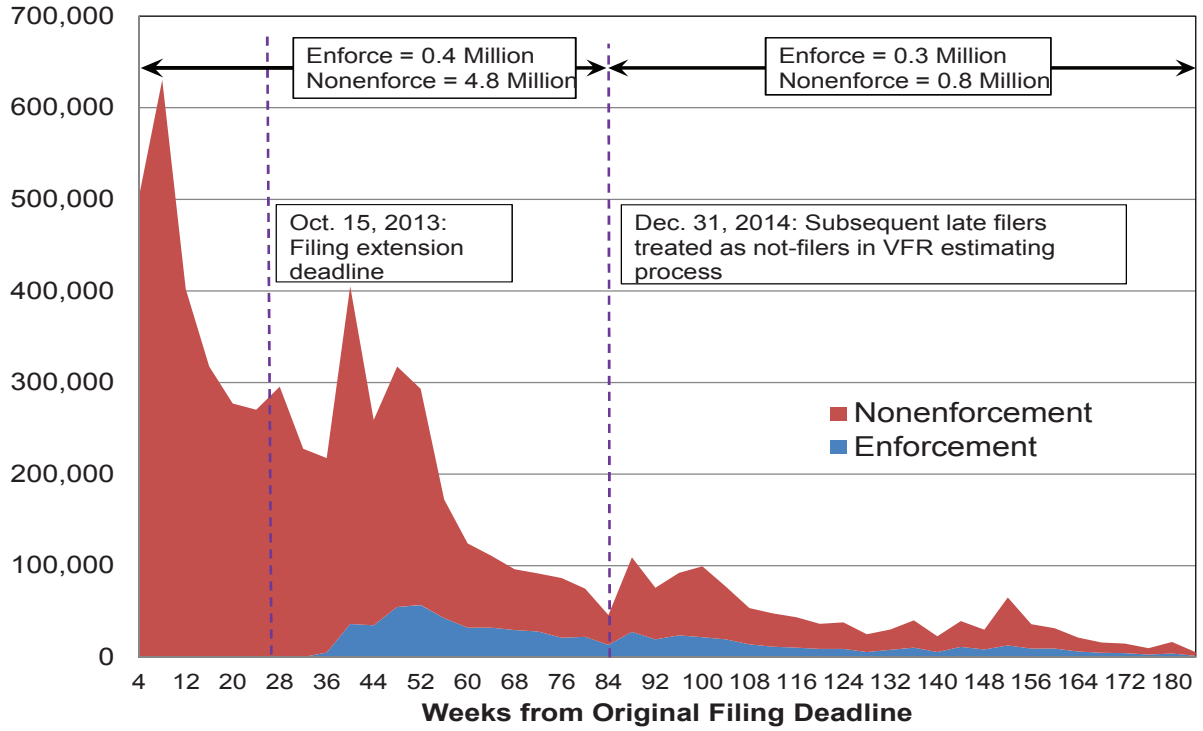
[16] The data shown here are an underestimate of the numbers of returns brought in by enforcement since only those returns that are processed and entered into the Individual Returns Transaction File (IRTF) are included. A fairly large number of returns secured through the Automatic Substitute for Return Program (ASFR) do not get processed in the same way and do not get posted in the IRTF. In addition, a significant number of returns might be filed because of enforcement related to other tax years, but no nonfiler notice was sent for the particular tax year in question.

FIGURE 5. Counts of Balance Due and Refund Returns by Weeks After the Filing Deadline, TY 2012



SOURCE: IRS Individual Returns Transaction File.

FIGURE 6. Counts of Enforcement and Nonenforcement Returns by Weeks After the Filing Deadline, TY 2012



SOURCE: IRS Enforcement Revenue Information System

Exploring Some Drivers of Filing Behavior

Aside from providing a more accurate estimate of the VFR and of the number of required returns in the population, the IRS Administrative Data Methodology offers the additional advantage of using the same data source for both the numerator and denominator of the ratio. This makes it easier to probe the potential causes of fluctuations in the VFR and to examine drivers of filing behavior. For example, if we want to explore the impact of the stimulus payments on the VFR increase in TY 2007, we can examine how many taxpayers who look like they were required to file in TY 2006 but did not file timely ended up timely filing for TY 2007 but then returned to not filing for TY 2008. Then we can compare the net fluctuations of taxpayers in and out of the numerator and denominator in TY 2007 with what we observe for prior and subsequent years.

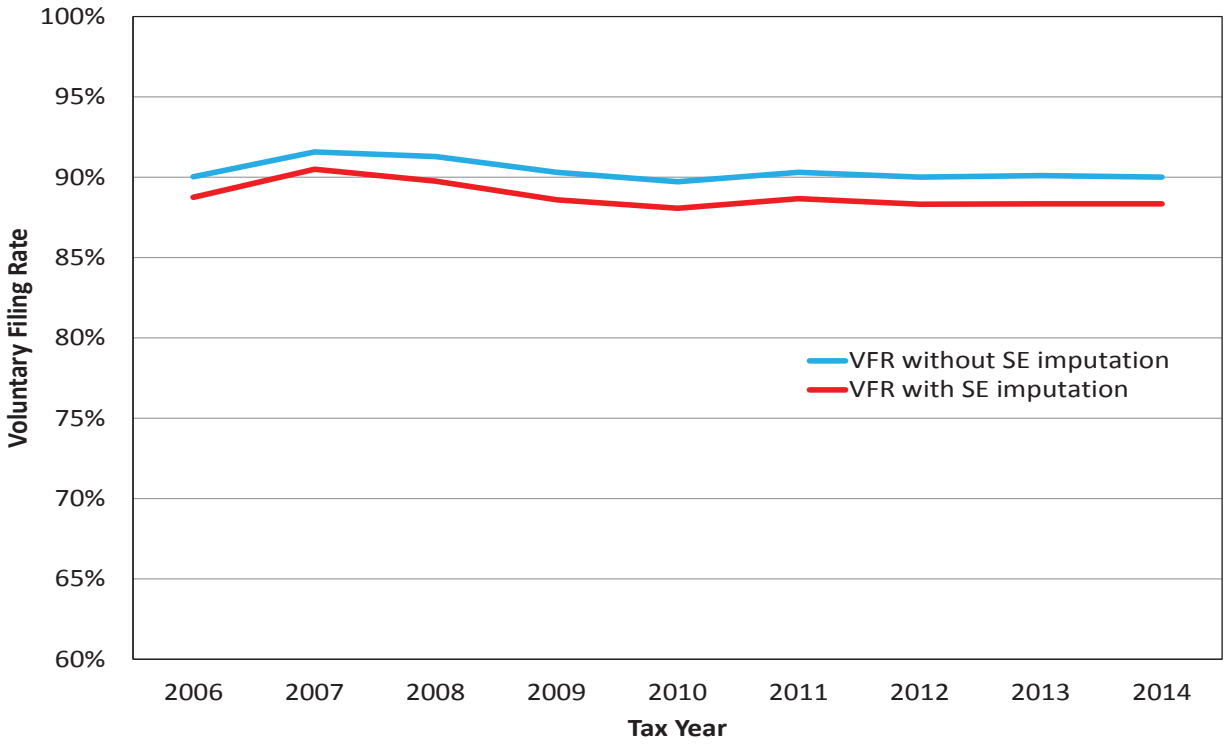
Nonetheless, the new methodology is not without some limitations. First, we do not know the filing status or family make-up—including the identity of the spouse and the number of dependents—of those who do not file. In the current estimates, we impute family units so that in the aggregate the profile of filers and nonfilers matches that evident in Census data. While this method should provide reasonable estimates of the number of required not-filers in the aggregate, it limits the reliability of analysis of filing behavior at the micro level.

Self-Employment Income

As with the Census-based method, we need to impute self-employment income to the not-filers since this income is only partially reported on third-party information returns. As stated above, the methodology assumes that not-filers look a lot like filers in terms of their propensity to have self-employment income given other characteristics, such as gender, age, region, and types and amounts of income present on third-party information returns. But, we also know from the analysis of the tax gap that self-employment income is significantly underreported on filed returns. Thus, if not-filers are generally like filers, our method underestimates the self-employment of not-filers as well. But, it could be that not-filers are less likely to have self-employment income than filers because, if they are being honest, they legitimately are not required to file. Or, given their income and other characteristics, they are more likely to have self-employment income since the relative invisibility of this form of income allows them to remain as not-filers. Whatever the conclusions about the effectiveness of the imputations in the aggregate, their effectiveness in accurately assigning self-employment income to particular taxpayers is certainly more limited. Thus, the imputation of self-employment income is another factor limiting the analysis of drivers of filing behavior at the micro level. These limitations might be partially overcome by considering the filing behavior and requirements of individuals instead of tax units and by analyzing filing drivers with and without the imputation of self-employment income.

How much of a difference do the self-employment imputations make in the VFR estimates? Figure 7 shows that when we consider only the self-employment income that is present on Forms 1099 Miscellaneous Income, the VFR is between 1.1 percent and 1.8 percent greater, with the difference being larger in recent years. Nonetheless, the trend over time in the estimates is fairly similar between the two approaches, which provides some confidence that the imputation of self-employment income mainly has the effect of lowering the level of the VFR but doesn't significantly alter the analysis of its year-to-year variation.

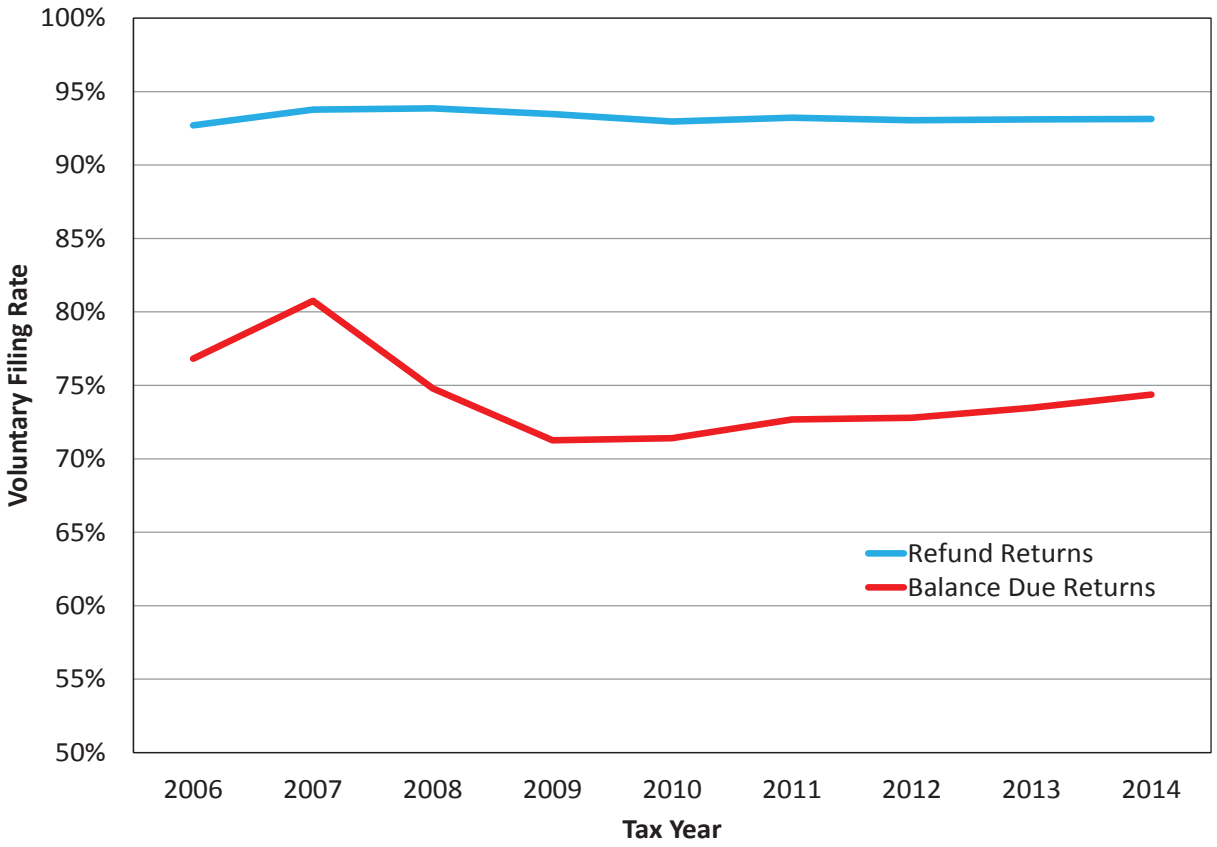
FIGURE 7. VFR With and Without Imputation of Self-Employment Income, TYs 2006–2014



Pre-Payment Position

Aside from the opportunity to obtain a tax rebate (as in the case of the 2008 Economic Stimulus Recovery Rebate Credit), another clear incentive for filing a tax return is to obtain a refund to recoup excessive tax withholding and/or overpayment of estimated taxes, or to claim refundable tax credits, such as the Earned Income Tax Credit or Additional Child Tax Credit. As Figure 8 shows, the filing rate for those owed a refund is very stable and substantially higher (about 20 percentage points greater) than the filing rate for those with a balance due. In addition, this graph shows that the effect of the economic stimulus payments in increasing the filing rate in TY 2007 was largely due to the fact that it encouraged a larger share of those with a balance due to file a tax return. Partly due to the penalties for underpayment of taxes during the year, the percent of timely filers owed a refund typically exceeds by a large magnitude those with a balance due. In TY 2014, for instance, about 78 percent of timely filers were owed a refund while only 22 percent had a balance due. As a result, the VFR is much closer to the rate of the refund group rather than the balance due group. This finding corroborates the value of the IRS' Withholding Compliance Program in ensuring both timely payment of taxes and encouraging timely filing of tax returns. Clearly, when taxpayers have sufficient taxes withheld from their paychecks to meet their tax obligations, they are also much more likely to file a tax return.

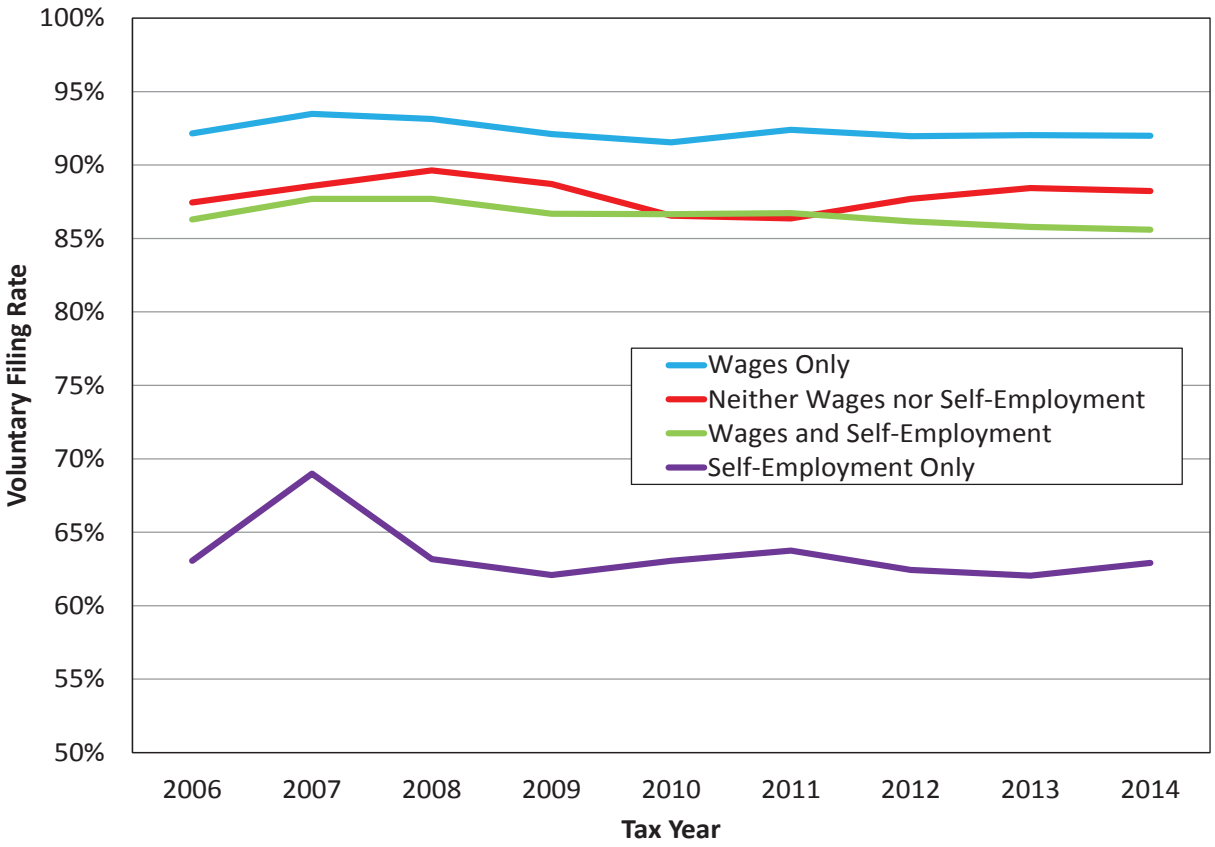
FIGURE 8. VFR by Whether Taxpayer Has a Balance Due or Owed a Refund, TYs 2006–2014



Income Visibility

Another factor that likely influences taxpayers' filing decisions is the visibility of their income to enforcement authorities. Thus, taxpayers with mostly wage income reported on Form W2 are more likely to file than those who earn most of their income from self-employment, especially when their self-employment income is not reported on third-party information returns. In addition, some taxpayers may not be aware that they are obligated to file and pay self-employment tax when their income is as low as $433. Figure 9 shows that the VFR among those whose earned income is comprised of only wage income is significantly higher (30 percentage points more) than among those with only self-employment income. The effect of the stimulus rebates in 2007 seems to have had a larger effect in encouraging filing among those with self-employment income than among those with only wage income, who were already filing at a high percentage.

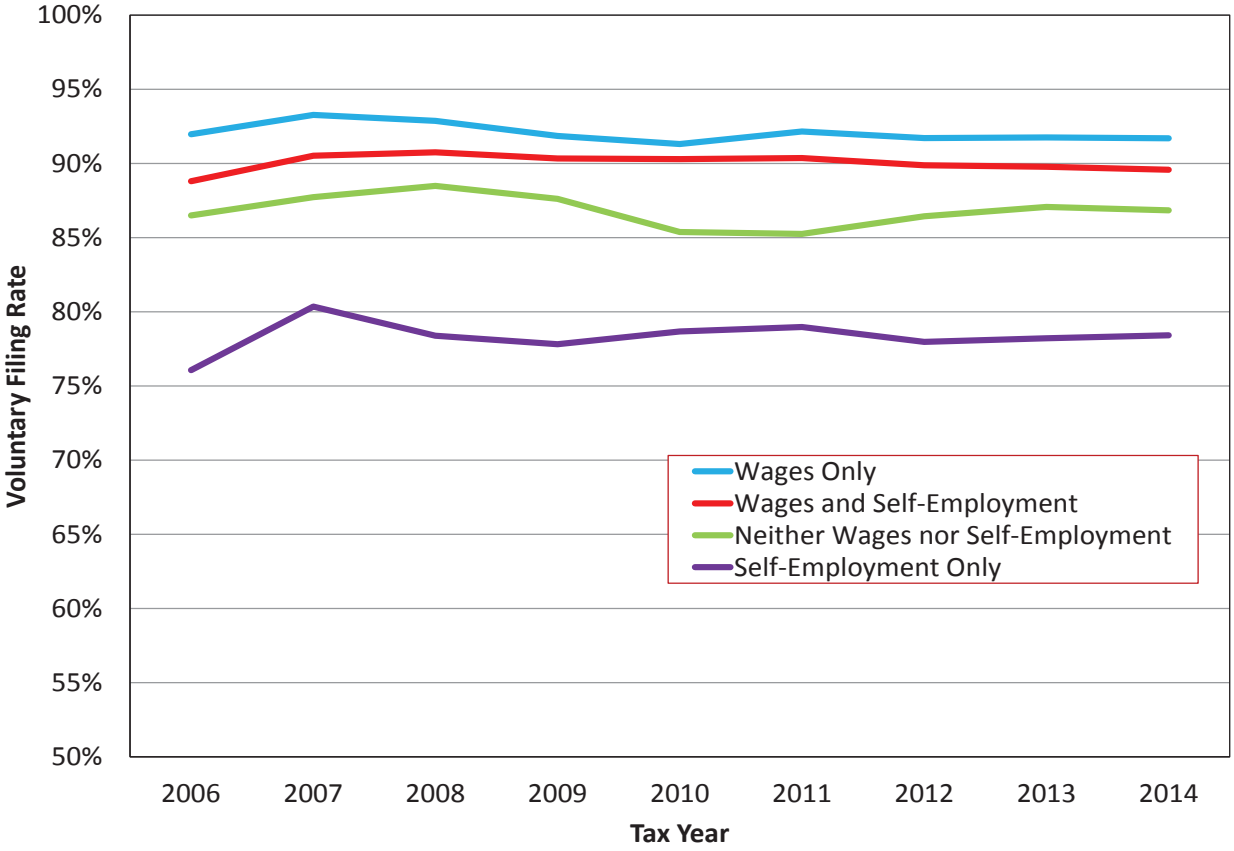
FIGURE 9. VFR by Primary Income Types (with SE Imputation), TYs 2006–2014



After imputation, more potential not-filers are assigned self-employment income. As stated above, taxpayers with self-employment income greater than $433 are required to file. However, the imputed net Schedule C amount could cause the taxpayer's gross income to exceed the relevant filing threshold even if they are below the $433 net self-employment income threshold. But, when only nonemployee compensation on Forms 1099-Misc Miscellaneous Income is considered (that is, self-employment income is not imputed), the VFR for those with just self-employment income remains significantly lower than for those with just wage income (about 78 percent instead of about 92 percent for TY 2014) (Figure 10).[17]

[17] Clearly some of those with nonemployee compensation of greater than $433 on a Form 1099-MISC do not have a filing requirement because they have sufficient expenses to reduce their net income to $433 or less. However, these same individuals, as well as other potential not-filers, likely receive self-employment income that is not reported on third-party documents. So the estimates without the self-employment imputations are likely a conservative estimate of the extent of nonfiling for this population.

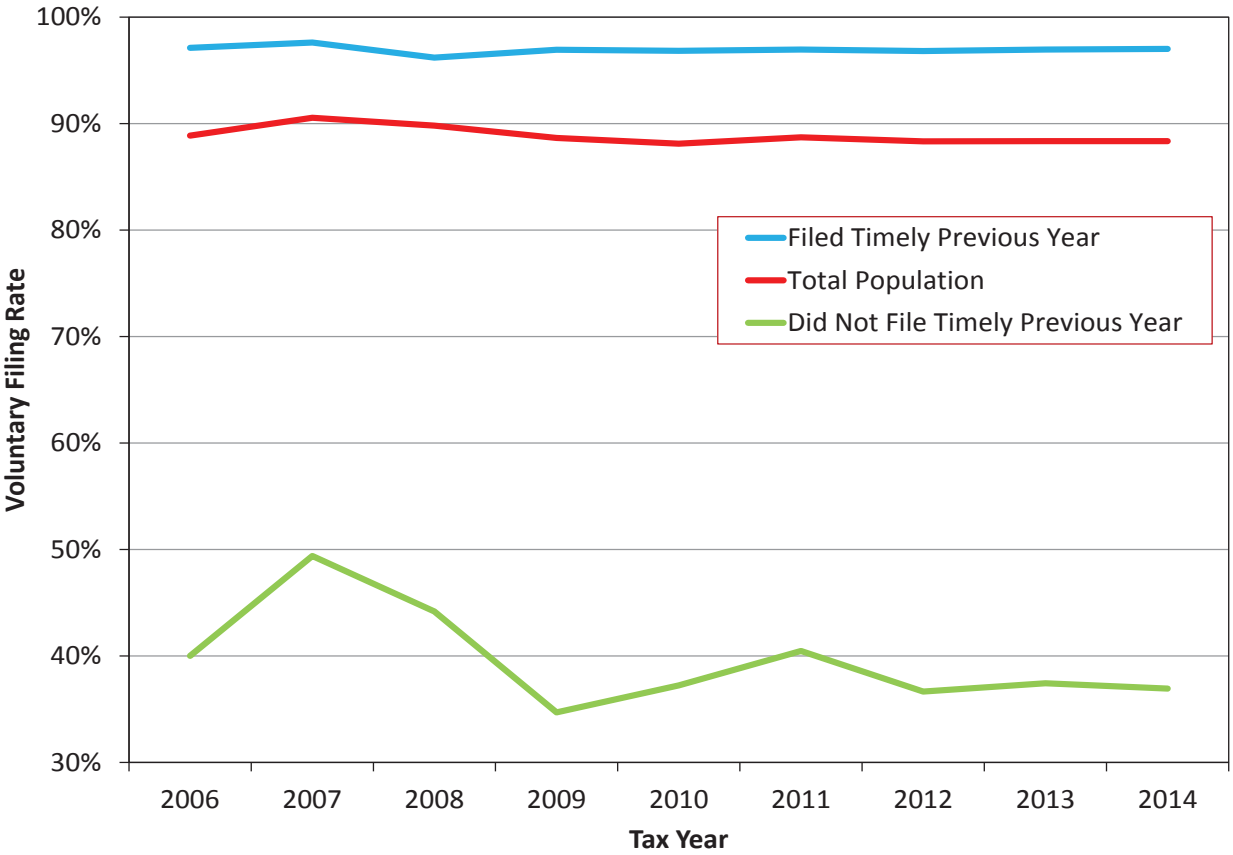
FIGURE 10. VFR by Primary Income Types (Without SE Imputation), TYs 2006–2014



Prior-Year Filing

Another presumption in the tax compliance literature is that there is significant persistence in filing behavior (Erard and Ho, 2001). On the one hand, a taxpayer who failed to file in the previous tax year has less incentive to file in the current year because he likely perceives that doing so increases the probability that his past filing noncompliance will be discovered. On the other hand, a taxpayer who did file a return in the previous year risks raising the suspicion of the tax authority if he does not file in the current year. In addition, once a taxpayer starts filing a tax return, it can become a matter of habit. Thus, it is not very surprising that, in fact, the VFR is much higher among those who filed timely in the previous tax year. For those who filed timely the prior year, the VFR hovers between 96 percent and 97 percent, while for those who did not file timely in the prior year, the VFR ranges between 35 percent and 50 percent (Figure 11).

FIGURE 11. VFR for Taxpayers Who Filed Timely in the Previous Year vs. Those Who Did Not, TYs 2006–2014

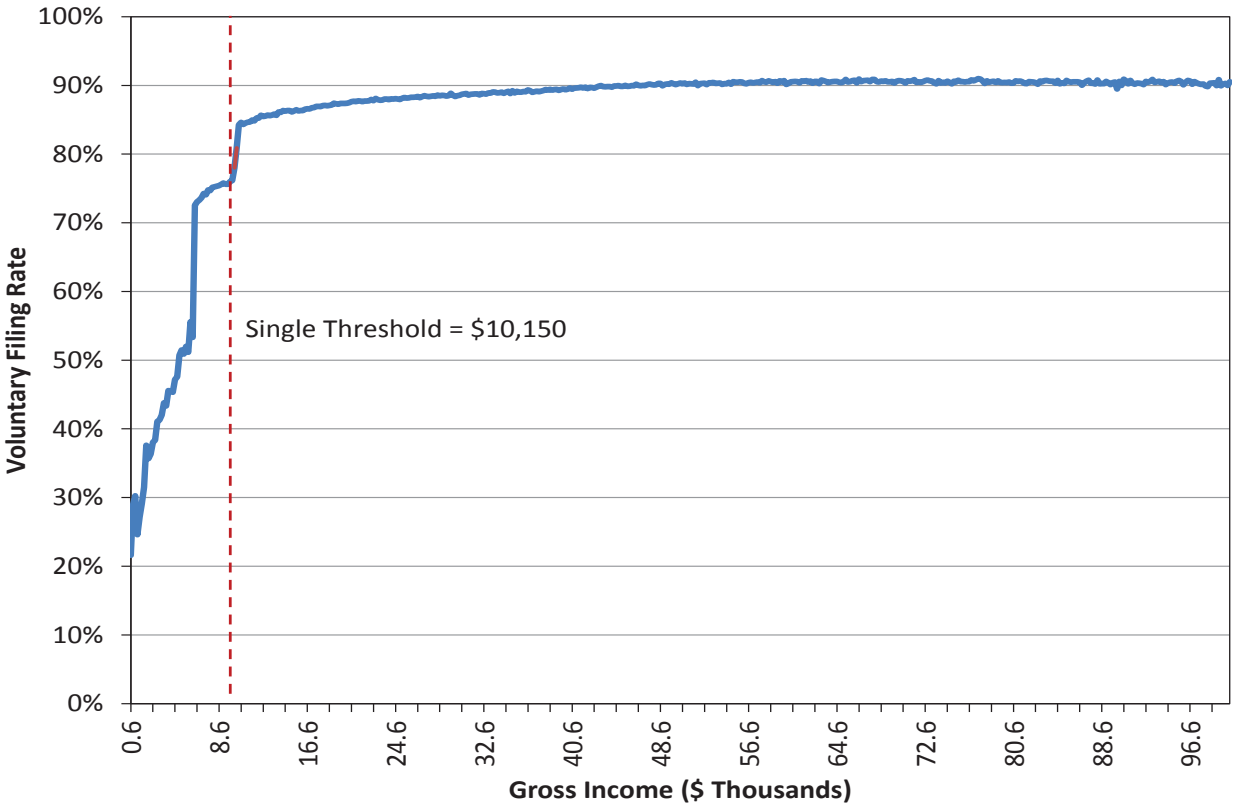


Amount of Income

Another factor that could be expected to make a taxpayer more likely to file is the level of income relative to the pertinent filing threshold. Taxpayers with gross income below the filing threshold, but with enough Schedule C income that they owe self-employment tax, might not be aware that they are obligated to file a tax return. When gross income barely exceeds the threshold, taxpayers may believe that they are below tolerance for IRS enforcement, especially if they would be owed a refund if they filed a return. A taxpayer whose income is only slightly above the filing threshold may owe very little tax or be due a refund that is too small to offset the financial costs and burden of filing a return.

Figures 12 and 13 tend to support these expectations about how filing relates to income. First, the VFR is very low for those well below the gross income filing threshold, but required to file because of self-employment tax or because they are claimed as dependents on another return. Second, the VFR jumps sharply as gross income nears the threshold and then gradually increases logarithmically as gross income increases beyond the threshold. This pattern holds true for both single and married taxpayers, but the income effect seems more prolonged and stronger in the case of married taxpayers.

FIGURE 12. VFR of Single Taxpayers (Under 65 Years Old) by Gross Income, TY 2014



Age

Filing rates may also vary by age, but the nature of this influence is not clear a priori. Changes in cultural norms might affect age groups differentially and the composition and level of income, filing status, and numbers of dependents may change as taxpayers age, which could affect taxpayers' propensity to file. Given the interrelationship of age with many filer characteristics,teasing out its effects would likely require a multivariate analysis. But, the plots in Figures 14, 15 and 16 show the relationship between the VFR and age for all taxpayers as well as for single and married taxpayers separately, which at least raise questions for further research. Figure 14 suggests that the VFR is relatively high for young taxpayers, declines as taxpayers reach middle age, increases again as taxpayers near retirement, and then falls off in later years. This pattern is accentuated for single taxpayers (Figure 15) and modulated for married taxpayers (Figure 16). Without controlling for other factors, married taxpayers generally have a higher filing rate than single taxpayers. It would be interesting to explore the factors other than age that are driving the apparent decline in the voluntary filing rate in the middle age ranges and leading to lower apparent filing rates among elderly taxpayers.

One factor that we thought might contribute to lower apparent filing rates for the elderly is the fact that amounts reported on 1099Rs as taxable pensions are not necessarily subject to taxation.[18] This could have resulted in an overestimation of the numbers of nonfiling elderly taxpayers with a filing requirement and an underestimate of the VFR. However, using the random audit data from the National Research Program to correct for the misidentification of pension income as taxable on 1099Rs, we found that this problem was a not a major factor in the apparent VFR decline for these taxpayers.[19]

[18] About 18 percent of taxpayers with a taxable pension amount on a 1099R were found in NRP audits to have no taxable pension income. For those who did have verified taxable pension amounts on their Form 1040s, the amounts on average were about 10 percent lower than the amounts reported on the 1099Rs.

[19] Other than measurement limitations, several other factors, including a greater prevalence of health problems, lower awareness of filing rules, and lower computer literacy could be driving the lower filing rates among the elderly.

FIGURE 13. VFR of Married Taxpayers (Under 65 Years Old) by Gross Income, TY 2014

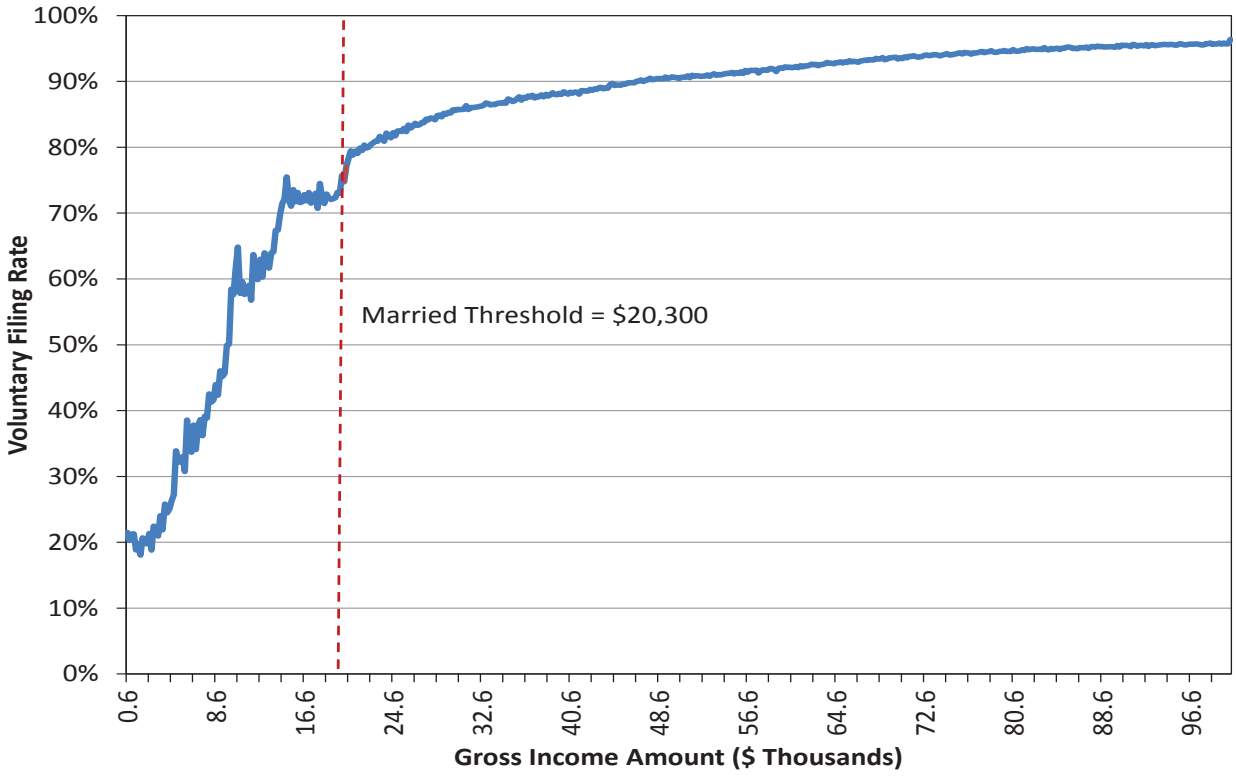


FIGURE 14. VFR by Age, TY 2014

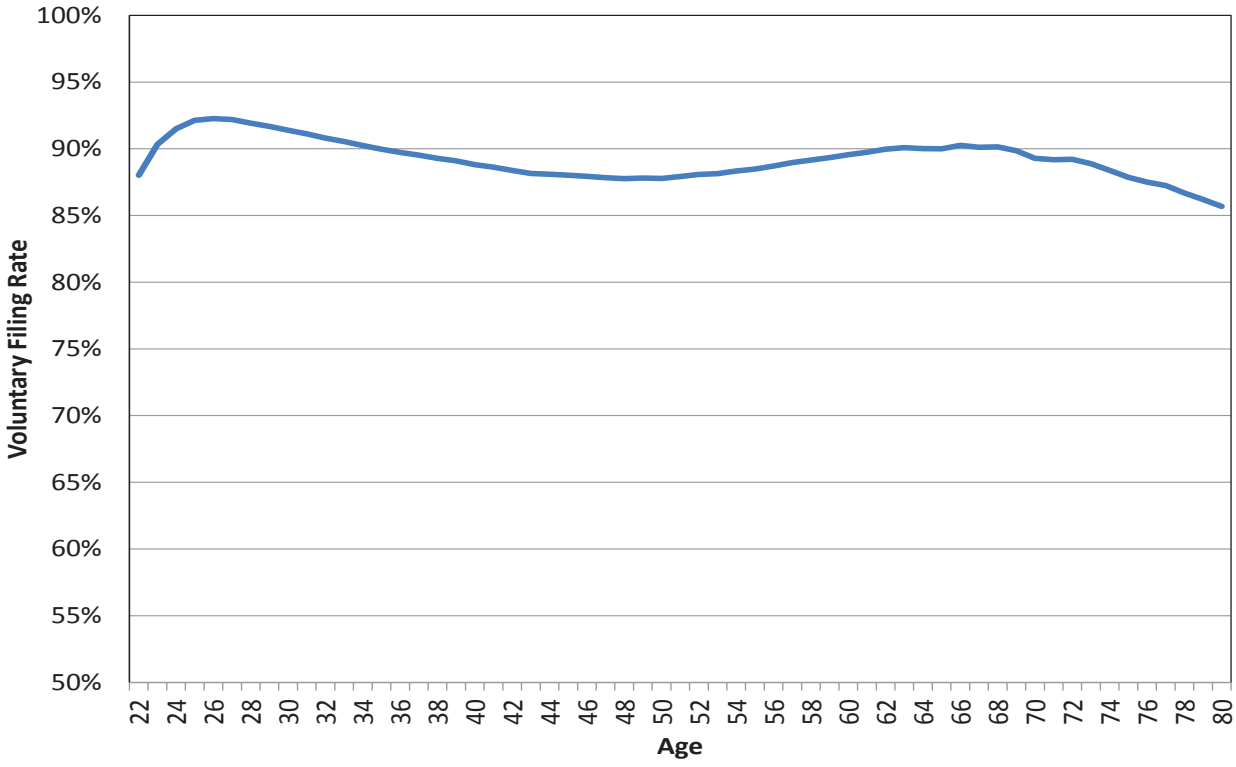


FIGURE 15. VFR by Age, Single Taxpayers, TY 2014

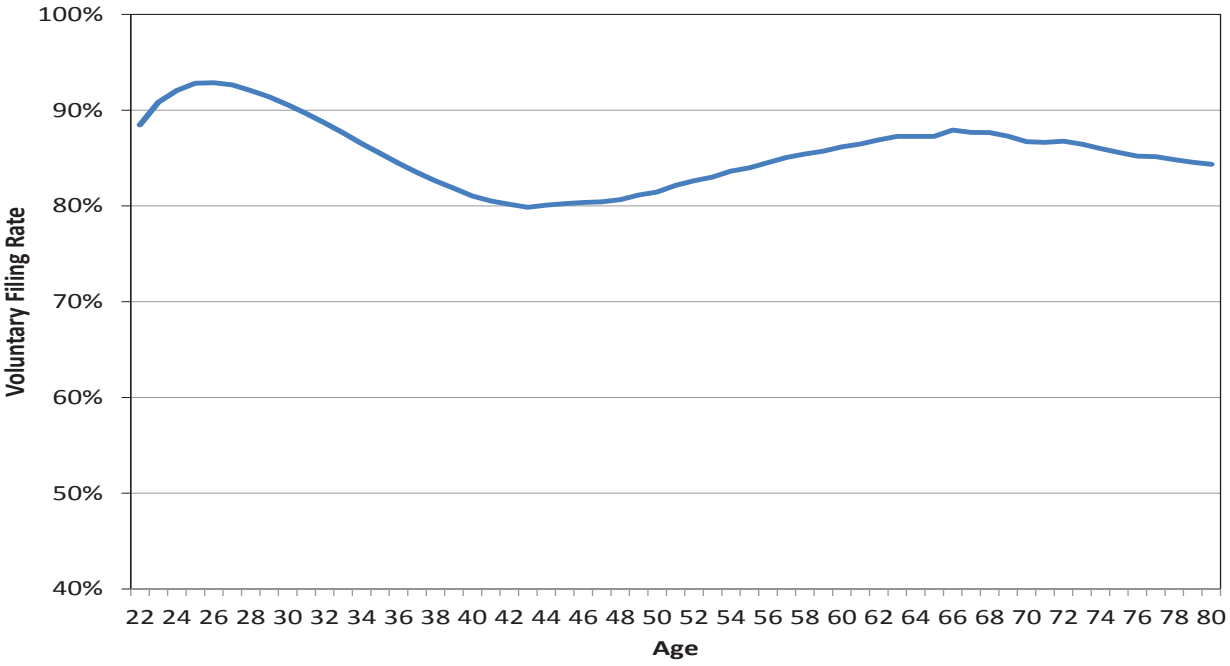
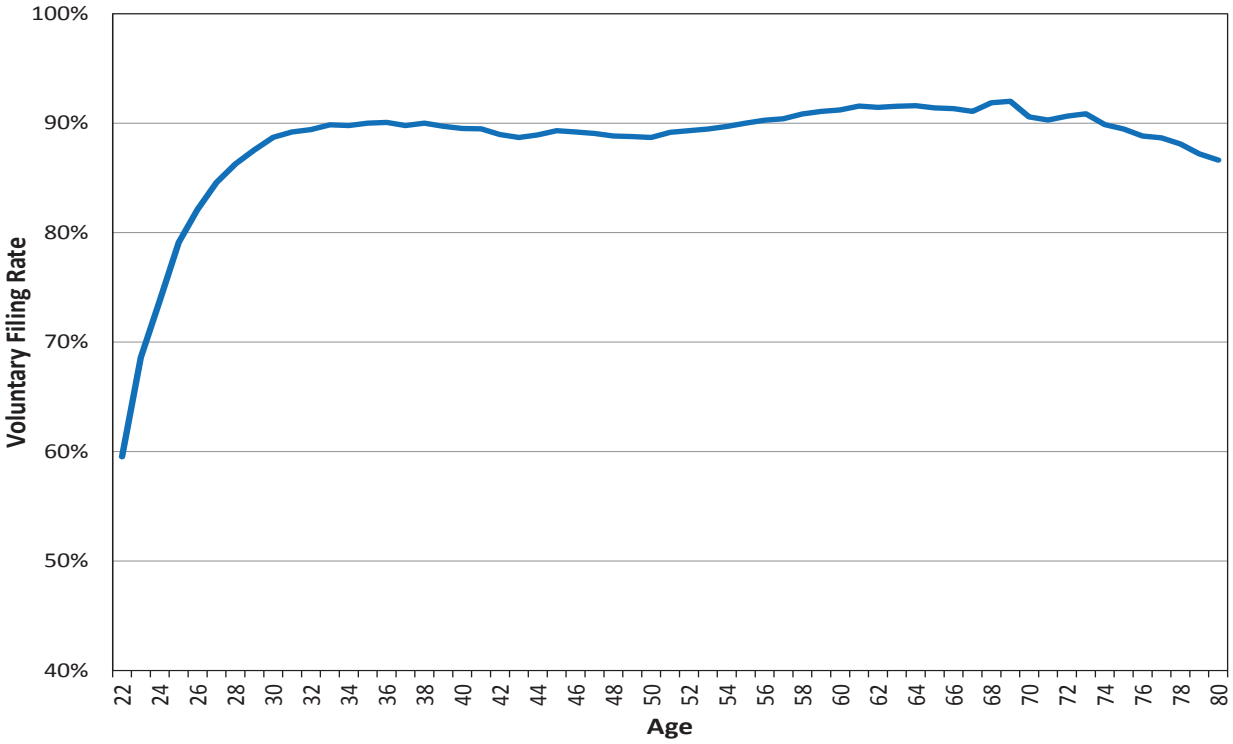


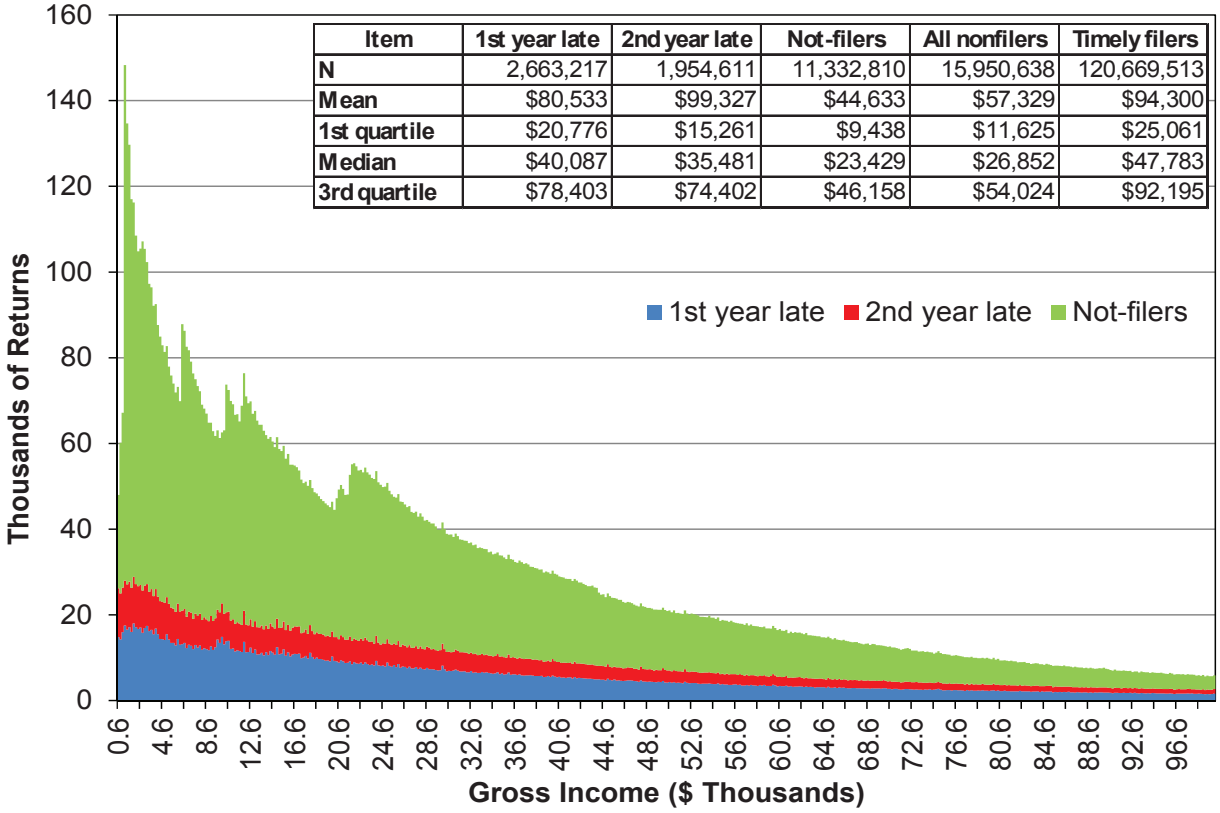
FIGURE 16. VFR by Age, Married Taxpayers, TY 2014



Income and Tax Comparisons With Timely Filers

The next several figures examine some of the distributional characteristics of nonfilers compared with timely filers. Given the fact that the filing rate is lower for taxpayers close to the filing threshold, it is not surprising that the average gross income of required nonfilers is lower than that of required timely filers (Figures 17 and 18). For nonfilers, mean gross income in TY 2014 was $57,329, while for timely filers it was $94,300. The mean gross income for late filers is very close to that of timely filers, while that of not-filers is much smaller ($44,633). The returns that come in during the first year after the tax year tend to report less income and to be less unequally distributed than returns that are filed in the second year after the tax year. The distribution of gross income for not-filers is skewed just as much as for timely filers, with the upper one-quarter of the distribution responsible for about one-half of total gross income.

FIGURE 17. Distribution of Gross Income, Late and Not-Filers, TY 2014



Because not-filers tend to make less income than timely and late filers, they are also estimated to owe less tax. While timely filers report an average tax liability of $11,530, not-filers are estimated to owe an average of $7,957. The distribution of total tax is considerably more skewed for not-filers than it is for timely filers (Figure 19).

FIGURE 18. Distribution of Gross Income, Timely Filers (TY 2014)

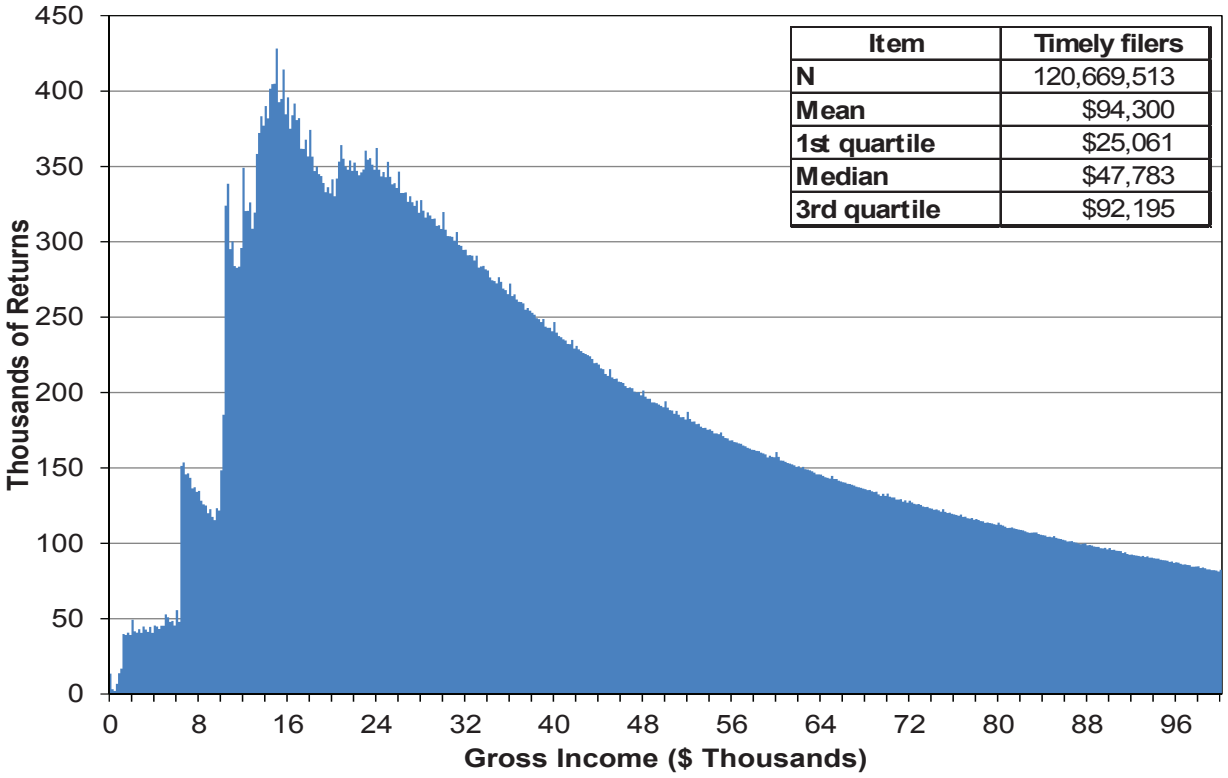
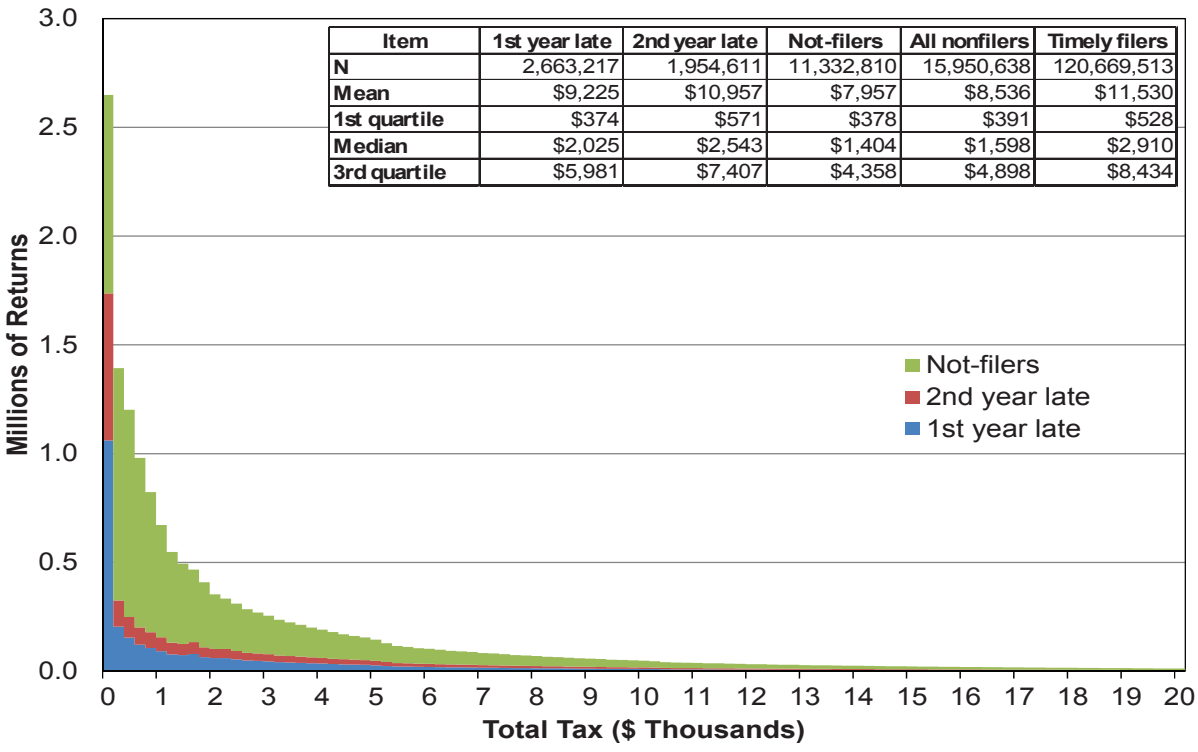
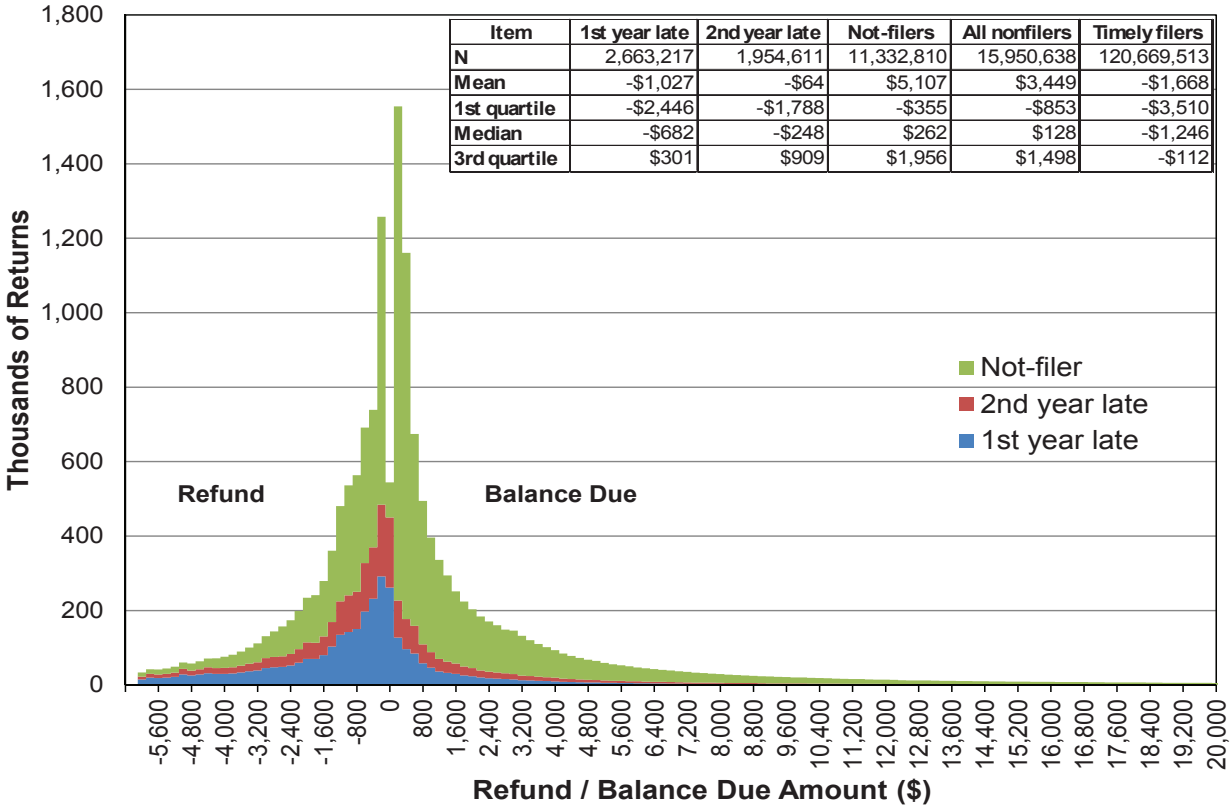


FIGURE 19. Distribution of Total Tax, Late Filers and Not-Filers, TY 2014



Given that about 78 percent of taxpayers filing a timely return are owed a refund, it is not surprising that on average the tax returns of timely filers have a negative balance due (Figure 20). This is also true for late filers, although in the case of second year late filers, refund and balance due returns are closer in number than for returns filed late in the first year following the tax year. However, not-filers are estimated on average to owe over $5,000. As discussed above, the presence of a tax debt is likely a key reason why such taxpayers choose not to file a return.

FIGURE 20. Distribution of Refund/Balance Due Amount, 1st Year Late, 2nd Year Late, and Not-Filers, TY 2014



Benefits of the VFR Research

Our efforts to develop a VFR measure using the Census-IRS matched data and by using just IRS data have resulted in important benefits. In the attempt to use the Census-IRS matched data, we found that even after imputing additional income to the Census records following the previous methodology, the estimated number of required taxpayers was still dramatically lower than the estimate from IRS data. This indicates that there are likely other forms of income that are insufficiently reported in the CPS survey that contribute to this undercount. Thus, the work on the VFR has revealed additional gaps in the CPS income data relative to the income reported in IRS data. This led to some efforts using the matched IRS-CPS data to try to fill these gaps, but more work along these lines is needed to ensure that the resulting distribution of income in the CPS is comparable to that found in the IRS population data.

The work on the IRS administrative VFR has resulted in a more accurate VFR that allows us to more carefully examine the drivers of filing behavior and to explain the fluctuations in the VFR over time. There are benefits to being able to explore the role of the numerator and denominator together—rather than just the numerator—in affecting fluctuations in the VFR. The time series data resulting from this work could, with some limitations, potentially be used in a multivariate analysis of the drivers of filing behavior.

Future Work

For the IRS-based VFR presented in this paper, we assume that the income amounts and filing status as reported and processed on the return are correct. However, we know that it is likely that a significant number of single filers incorrectly claim Head of Household status on their tax returns, and by so doing, many of them appear to not have a filing requirement, when in fact they do. Thus, one potential modification of the measure would be to use the audit data from the IRS National Research Program to impute corrected filing statuses to those who claim Head of Household status. This would tend to increase the number of required returns filed on time, and therefore increase the VFR. Another potential refinement of the IRS-based VFR would be to add to the filed returns the income amounts reported on third-party information documents but not on the returns. This change might also add to the count of required returns that were filed timely, while likely adding proportionately fewer to the category of late returns.

Greater leverage in understanding the drivers of filing might be obtained by using the IRS administrative data approach in a longitudinal probit analysis of the decision to file. Given the inaccuracies introduced at the micro-level in the imputation of tax units for the not-filers, one approach might be to analyze filing at the individual level, rather than the tax unit level. Another approach would be to improve the imputation of tax units by drawing on information from prior year returns and other sources, such as Social Security Administration data. Since the self-employment imputations also introduce inaccuracies at the micro-level, the size and significance of the parameter estimates could be compared with and without both sets of imputations.

It would be valuable to explore the factors driving variation in filing rates across different demographic groups. A better understanding of these drivers might suggest outreach programs or targeted enforcement efforts that could encourage filing for groups having lower filing rates.

Another area of future work is to further explore the potential of the expanded IRS-Census matched dataset in supporting the development of a reliable alternative VFR measure, as well as to further examine the drivers of filing behavior. With complete IRS data in the Census environment, it should be possible to explore more directly the deficiencies in the matching process as well as in the reporting of income on CPS records. The matched data also provides an opportunity to undertake an analysis of the drivers of filing without the same limitations of knowledge about the structure of the nonfiling tax units.

References

Erard, Brian and Chih-Chin Ho (2001). "Searching for Ghosts: Who Are the Nonfilers and How Much Tax Do They Owe?," Journal of Public Economics. Vol. 81 pp. 25–50.

Erard, Brian, Pat Langetieg, Mark Payne, and Alan Plumley (2014), "Missing Returns vs. Missing Income: Estimating the Extent of Individual Income Tax Filing Noncompliance From IRS and Census Data," Proceedings of the National Tax Association Annual Conference, November 2014.

Erard, Brian, Mark Payne, and Alan Plumley (2012), "Advances in Nonfiling Measures," 2012 IRS Research Bulletin, Internal Revenue Service, Publication 1500, Washington, DC, pp. 79–89.

Langetieg, Pat, Mark Payne, and Alan Plumley (2016), "The Individual Income Tax and Self-Employment Tax Nonfiling Gaps for Tax Years 2008–2010," 2016 IRS Research Bulletin, Internal Revenue Service, Publication 1500, Washington, DC, pp. 38–59.

Plumley, Alan (1996), "The Determinants of Individual Income Tax Compliance: Estimating the Impact of Tax Policy, Enforcement and IRS Responsiveness," Internal Revenue Service, Publication 1916 (Rev. 11-96), Washington, DC, 1996.