Missing Returns vs. Missing Income: Estimating the Extent of Individual Income Tax Filing Noncompliance From IRS and Census Data

Brian Erard (Brian Erard and Associates), and Pat Langetieg, Mark Payne, and Alan Plumley (IRS Research, Analysis, and Statistics: Office of Research)[1]

The Internal Revenue Service (IRS) and others have long used Census data to learn about the entire population—not just those who appear on income tax returns. However, many have observed that Census surveys sometimes do not fully reflect all relevant income. This paper uses administrative information available to the IRS to document the extent of such "missing income" for key income types, and applies this knowledge to estimate the number of "missing returns." 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. Many of these contribute to the nonfiling portion of the tax gap—the amount of tax liability imposed by law that is not paid on time. The IRS would like to help taxpayers meet their filing obligations, so research is under way to measure the extent of nonfiling[2] and to identify the major causes. This paper describes the current methodology for estimating the number of nonfilers and the corresponding Voluntary Filing Rate using both Census and IRS data.

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. 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 estimate the VFR denominator from the Census Bureau's Current Population Survey Annual Social and Economic Supplement (CPS ASEC), 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 fairly 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 is accessible

[1] The views expressed in this paper are those of the authors, and they do not necessarily represent the positions of the Internal Revenue Service.

[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 late filers of required returns, and timely filers exclude those who have no filing requirement, but file solely to claim a refund of withholding or to claim a refundable credit. 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.

to IRS research staff, we began estimating the numerator from population data. This required developing new systems to categorize each return as timely or late and as required or not required to be filed. After demonstrating that the new population data were able to replicate the results from the trusted samples used until then, we began using the population data each year. That allowed us to examine in more detail what type(s) of taxpayers were driving fluctuations in the numerator.

In general, the estimated trend in the VFR was fairly stable at just over 90 percent. The percentage increased significantly in 2007 and 2008, however, which we ascribed to the effects of the economic downturn and the economic stimulus.[3] When we estimated the VFR for 2009, however, we observed a noticeable decline, which we couldn't fully explain initially. So, we began analyzing what was causing the decline.

We soon realized that the estimated trend in the VFR was potentially misleading owing to the various measurement issues surrounding the numerator and denominator of the ratio. So, we set out to ensure 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[4]) for the requirement to file. In the process, we discovered that the instructions that the IRS provided taxpayers did not fully define the requirement to file; at issue was how losses were to be handled in the definition of gross income. Technically, the gross income concept disregards all losses; that is, losses do not offset positive income for the purpose of establishing a filing requirement. We also applied a consistent definition of what it means for a required return to be timely filed for VFR purposes; we include in the numerator only those required returns that are filed by December 31 of the primary filing year.[5]

Accounting for Missing Income

Our next significant task was to augment the Census data used to construct the denominator of the VFR to account more fully for certain types of income (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. Figures 1-4 illustrate the differences in the amounts of these types of

[3] The general observation is that as incomes fall, fewer people are required to file. If those who are no longer required to file were disproportionately less likely to have filed when they were required (as might be the case if their income was just over the filing threshold), then those who are still required to file would be disproportionately more likely to file, thus increasing the VFR—not because of a change in behavior, but because of a change in who is required to file. In addition, many additional returns were filed for Tax Year 2007 because in order to receive the one-time Economic Stimulus Payment, people had to file a tax return for 2007. This undoubtedly increased the number of required returns filed on time, and thus the VFR.

[4] We know whatever income is reported by third parties on information returns, but that doesn't include much income from self-employment and some other sources, and we do not generally know which filing status applies to those who do not file a return. We are undertaking other research to address these limitations, perhaps enabling us to estimate the VFR without using Census data, but this paper presents the Census-based alternative, which has its own advantages and disadvantages.

[5] This excludes returns that are considered timely (e.g., due to combat extensions), but are filed much later than most. Setting December 31 as the cut-off date allows for a consistent measure to be produced each year.

income reported on the CPS ASEC 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),[6] 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 ASEC records.

Figure 1. Social Security Income Reported on Form 1099 SSA-RRB vs. CPS ASEC, TY2012

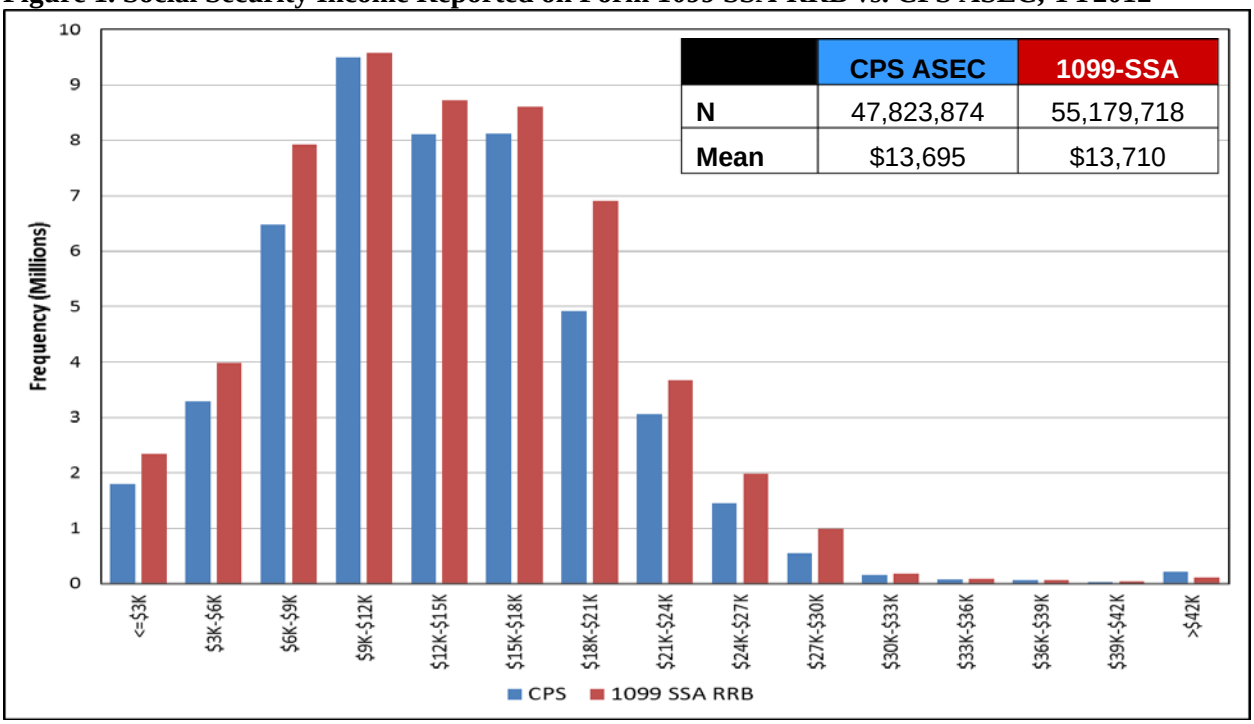


Figure 2. Pension Income Reported on Form 1099-R vs. CPS ASEC, TY2012

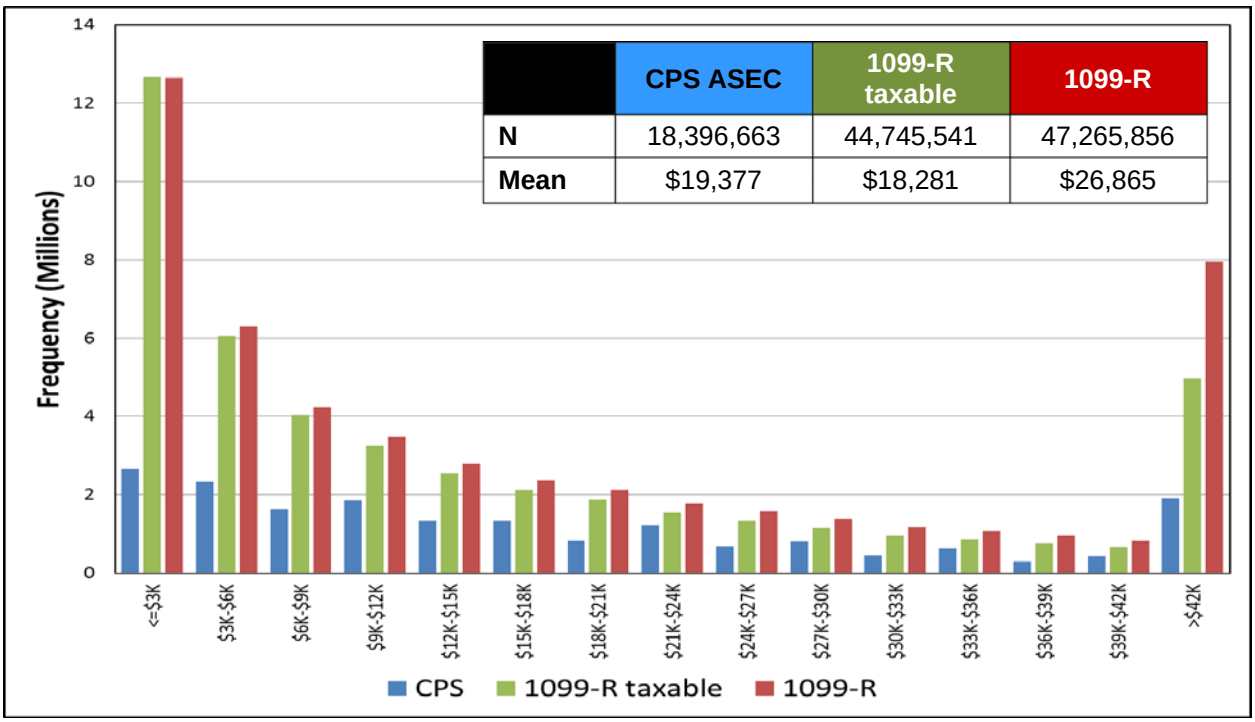
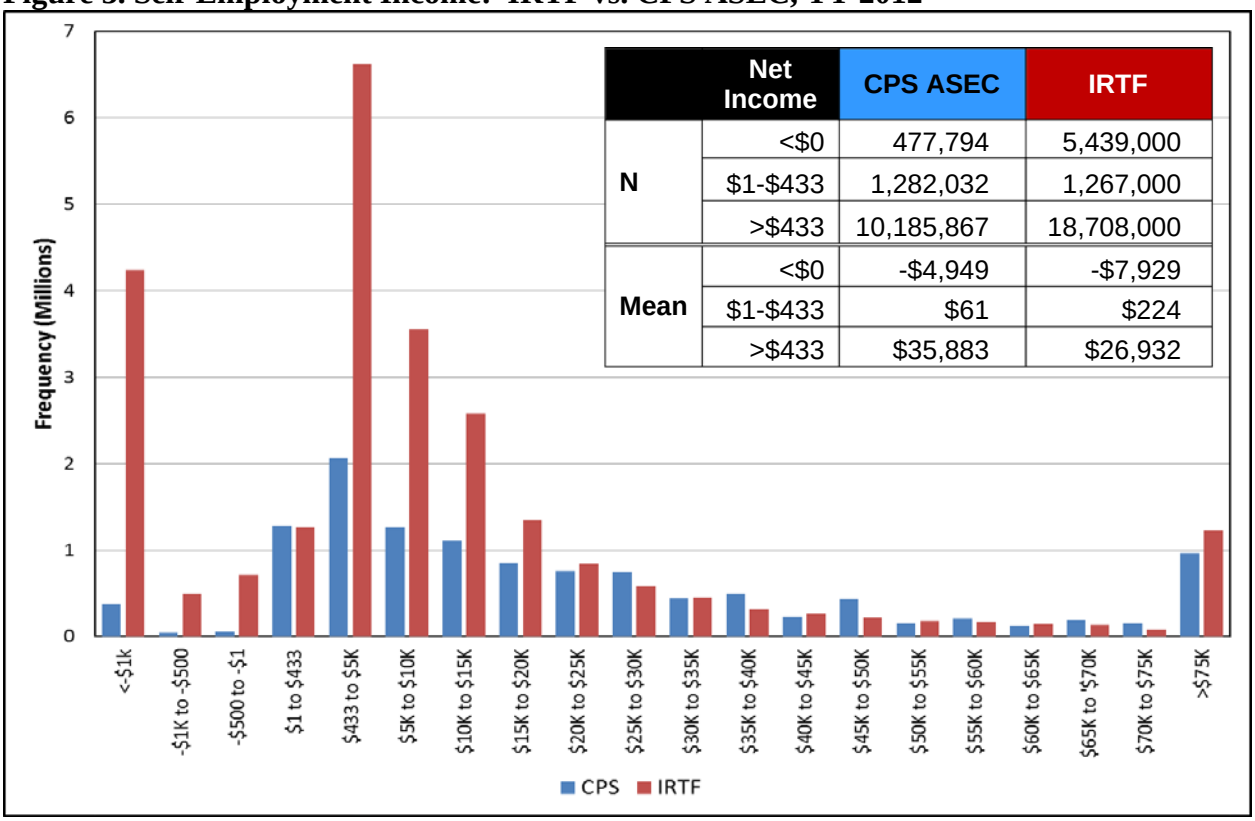
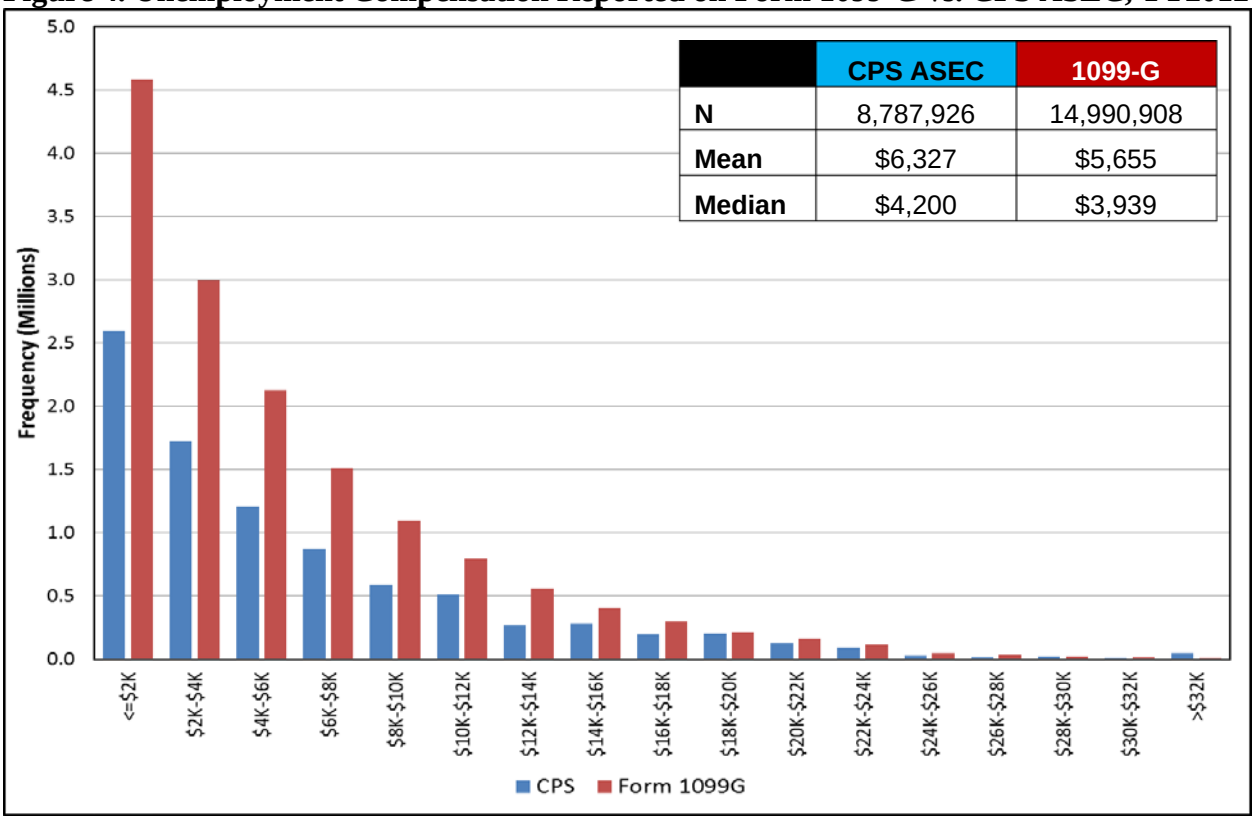


Figure 3. Self-Employment Income: IRTF vs. CPS ASEC, TY 2012



	Net Income	CPS ASEC	IRTF
N	<$0	477,794	5,439,000
	$1-$433	1,282,032	1,267,000
	>$433	10,185,867	18,708,000
Mean	<$0	-$4,949	-$7,929
	$1-$433	$61	$224
	>$433	$35,883	$26,932

Figure 4. Unemployment Compensation Reported on Form 1099-G vs. CPS ASEC, TY2012



	CPS ASEC	1099-G
N	8,787,926	14,990,908
Mean	$6,327	$5,655
Median	$4,200	$3,939

Imputation of Pension and Social Security Income

In the case of pension and Social Security income, the imputation involves prediction equations for the presence and magnitude of these income sources as a function of individual characteristics such as age, gender, region, and citizenship, as well as indicators and amounts of wages, interest, and unemployment compensation.[7] Social Security and pension income need to be estimated jointly since, all else equal, individuals who have one of these income sources are relatively more likely to have the other source. The imputation methodology is supported by a large micro-level data sample of Social Security Administration (SSA) Master File records that is matched to data for Forms 1099 SSA-RRB, 1099 INT and 1099R and W2s. Adjustments are made to the sample weights to match population age targets derived from the CPS ASEC. Separate estimates are generated for each tax year from 1996 forwards.

To account for the presence of pension and Social Security income and their magnitudes, we use a two-part specification. A bivariate probit model is employed to predict the presence or absence of pension and Social Security income on records where they are not reported on the CPS ASEC survey. The propensity for pension (P*) and Social Security income (S*) to be present is estimated by the following pair of equations:

$$P^* = \gamma_P' x + \mu_P \qquad (1)$$

$$S^* = \gamma_S' x + \mu_S \qquad (2)$$

where x is a vector of explanatory variables, and γ_P and γ_S are vectors of coefficients to be estimated. To account for the joint association between these two sources of income, we assume that the error terms (μ_P and μ_S) are bivariate standard normal random deviates, with correlation coefficient ρ_{PS}. (Results for tax year 2012 are shown in Table 1).

[7] To be consistent with our Form 1099-R income measure, we impute IRA income along with pension income.

Table 1. Bivariate Probit Models for the Presence of Social Security or Pension Income, Tax Year 2012

Variable	Presence of Social Security Benefits		Presence of Pension Income	
	Parameter Estimate	t-statistic	Parameter Estimate	t-statistic
Intercept	-8.875	(-133.3) **	-7.278	(-115.3) **
logage	2.266	(136.5) **	1.630	(105.9) **
logage*male	-0.262	(-11.4) **	0.128	(6.3) **
male	1.140	(12.5) **	-0.386	(-4.7) **
west	-0.194	(-18.2) **	-0.034	(-3.8) **
midwest	-0.002	(-0.2)	0.070	(8.0) **
northeast	-0.043	(-3.8) **	0.011	(1.2)
wagesind	1.806	(53.2) **	0.435	(14.2) **
interestind	-0.104	(-5.0) **	0.102	(6.2) **
logwages	-0.335	(-96.0) **	-0.054	(-18.3) **
loginterest	0.038	(9.5) **	0.063	(19.8) **
noncitizen			-1.462	(-14.7) **
ρ_{PS}	0.257			(42.7) **
Number of observations	231,631			
Missing values	24,888			
Log Likelihood	-155,482			

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

Next, these results are used to select respondents in the CPS ASEC survey who should be assigned Social Security and/or pension income. We assume that pension and Social Security income are truly present when they have been reported by a CPS ASEC respondent. However, when these income sources have not been reported, we require a means to predict whether they should have been reported.

By applying the parameter estimates from the first part of our model, we are able to estimate the unconditional joint probabilities associated with the presence or absence of pension and Social Security income for each respondent in the CPS ASEC survey. To estimate the likelihood that these income sources will be reported when they are present, we compute the ratio of the aggregate number of reporters of each income source in the weighted CPS ASEC sample to the aggregate number of recipients of that source based on the SSA and third-party information return records. By applying Bayes' rule, we are then able to predict the conditional probability that pension and/or Social Security income is present when one or both sources have not been reported.

Consider, for example, the case where neither income source has been reported by a CPS ASEC respondent. Based on our bivariate probit model, the unconditional probabilities associated with the presence of one or both income sources can be predicted. For instance, the probability that both income sources are present is estimated as:

$$\Pr(P = 1, S = 1) = BN(\tilde{\gamma}_P' x, \tilde{\gamma}_S' x, \tilde{\rho}_{PS}),$$

where *P* is a 0-1 indicator for the presence of pensions, *S* is a 0-1 indicator for the presence of Social Security benefits, and the terms with the tildes represent the parameter estimates.

Our above-described ratio estimates represent our predictions of the conditional probabilities:

$$\Pr(R_P = 1|P = 1)\ and\ \Pr(R_S = 1|S = 1),$$

where R_P and R_S respectively represent 0-1 indicators for whether pension and social security income have been reported. By subtracting each of these estimates from one, we can also predict:

$$\Pr(R_P = 0|P = 1)\ and\ \Pr(R_S = 0|S = 1).$$

Applying Bayes' rule, we can then predict the conditional probability that both income sources are present given that neither has been reported as:

$$\Pr(P = 1, S = 1|R_P = 0, R_S = 0) = \frac{\Pr(P = 1, S = 1)\Pr(R_P = 0|P = 1)\Pr(R_S = 0|S = 1)}{\mathrm{D}},$$

where

$$\begin{aligned} D = {} & \Pr(P = 1, S = 1)\ \Pr(R_P = 0|P = 1) Pr(R_S = 0|S = 1) \\ & + \Pr(P = 1, S = 0)\Pr(R_P = 0|P = 1) \\ & + \Pr(P = 0, S = 1)\ \Pr(R_S = 0 \mid S = 1) \\ & + Pr(P = 0, S = 0)\,.^{8} \end{aligned}$$

Similar calculations permit us to predict the probability that one or both income sources are present when neither source has been reported or when only one source has been reported. A comparison of the value from a pseudo-random draw against the probabilities associated with each possible outcome determines whether the respondent is flagged as having either or both types of income when they were not reported.

Once it is determined which CPS ASEC respondents will receive imputed values of pension and Social Security income, the next step is to describe the amounts to be imputed when they are present. This is done through equations of the form:

$$\mathrm{Ln}(A_P) = \beta_P' x + \varepsilon_P \qquad (3)$$

$$\mathrm{Ln}(A_S) = \beta_S' x + \varepsilon_S \qquad (4)$$

[8] When defining *D*, we impose the assumption that recipients do not report pension or social security income in cases where it has not been received.

where A_P and A_S are, respectively, the magnitudes of pension and Social Security income, x is the same vector of explanatory variables used in equations (1) and (2), and β_P and β_S are vectors of coefficients to be estimated. To account for the correlation between the magnitudes of these two income sources, we assume that the error terms (ε_P and ε_s) are bivariate normally distributed with zero means, standard deviations of σ_P and σ_S, and correlation term $\rho_{AP\,AS}$. To account for the skewed distribution of pension and Social Security earnings, we model their conditional distribution as lognormal. Based on the estimation results (see Table 2 for the tax year 2012 estimates) from the seemingly unrelated regression we then calculate the amounts to be imputed. A pair of pseudo-random deviates (e_p and e_s) to serve as proxies for the error terms (ε_P and ε_s) are drawn from a bivariate normal distribution with standard deviations equal to the root mean-squared error estimates and correlation coefficient equal to the estimated error term correlation from the seemingly unrelated regression analysis. For a CPS ASEC respondent who is flagged to receive pension income, we substitute the estimated regression coefficients into the following expression:

$$A_p = \exp(\beta_P' x + e_p) \qquad (5)$$

Similarly, for a CPS ASEC respondent who is flagged to receive Social Security income, we substitute the estimated regression coefficients into the following expression:

$$A_s = \exp(\beta_s' x + e_s) \qquad (6)$$

We then make an adjustment to the pension and Social Security income amounts reported by respondents so that in aggregate they equal the amounts predicted based on the administrative data. We apply equations (5) and (6) to each respondent who has reported pension and/or Social Security income to obtain a predicted value for the amounts of the income types received. We then aggregate this amount across the respondents and compute its ratio to the aggregate amounts actually reported. If the ratio is greater than one, we apply the ratio as an adjustment factor to inflate the amounts reported by the respondents to make them more commensurate with the IRS administrative data.

Table 2. Seemingly Unrelated Regressions on the Amounts of Social Security Income and Pension Income, Tax Year 2012

Variable	logsocsec		logpension	
	Parameter Estimate	t-statistic	Parameter Estimate	t-statistic
Intercept	8.125	(46.9) **	12.771	(29.5) **
logage	0.287	(7.1) **	-0.955	(-9.5) **
logage*male	-0.325	(-5.8) **	0.668	(4.7) **
male	1.645	(6.7) **	-2.322	(-3.8) **
west	-0.008	(-0.8)	0.132	(5.2) **
midwest	0.010	(1.1)	-0.052	(-2.2) *
northeast	0.060	(5.9) **	-0.015	(-0.6)
wagesind	0.108	(3.0) **	0.411	(4.7) **
interestind	0.036	(2.3) *	0.231	(6.0) **
logwages	-0.011	(-2.7) **	-0.028	(-2.9) **
loginterest	0.008	(3.2) **	0.046	(7.1) **
R^2	0.052		0.070	
Adjusted R^2	0.051		0.070	
Root MSE	0.553		1.382	
Correlation between error terms	0.123		0.123	
N	28,742		28,742	

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

Imputation of Net Self-Employment Income

A third type of income that we impute to CPS ASEC records is self-employment income. We had initially predicted self-employment income at the return (rather than person) level, making it possible to use filing status as a predictor. However, self-employment income is an important predictor of the presence and amount of unemployment compensation, so we developed a way to impute self-employment income at the person level as we do for each of the other income sources. After Tax Year 2006, our tax return data include the Taxpayer Identification Number (TIN) of the Schedule C proprietor, which allows us to know which spouse on a joint return was the proprietor. Then, using data from Tax Year 2007 and later, we developed a predictive model to allocate self-employment income between spouses in the case of joint returns. We applied this model to earlier tax years.

Taxpayers with net earnings from self-employment of more than $433 are required to file an individual income tax return and report self-employment tax. Since most of such earnings are not subject to third-party information reporting, we rely instead on information reported on individual tax returns to impute additional self-employment earnings to CPS ASEC respondents. Our imputations are restricted to net sole proprietorship income reported on Schedule C and Schedule F, because most flow-through sources of self-employment earnings (such as partnership and S-corporation income) are not subject to the $433 threshold. A challenge with our approach is that self-employment earnings are sometimes understated on tax returns. Thus,

while our imputations adjust the CPS ASEC-based measure of proprietor earnings to be more in line with tax return data, this measure will still 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.[9]

Our overall measure of gross income will also tend to be understated, both because sole proprietorship earnings are not fully accounted for in our imputations and because partnership and S-corporation income also may be understated in the CPS ASEC. The CPS ASEC questionnaire inquires about income from an unincorporated business or farm that the respondent owns, which presumably includes partnership and S-corporation income as well as earnings from a sole proprietorship. However, it seems likely that these income sources, much like sole proprietor earnings, tend to be underreported by respondents. We currently do not impute additional partnership or S-corporation income when estimating gross income from the CPS ASEC. Consequently, some taxpayers with gross income above the filing threshold may go uncounted as a result of the incomplete measurement of business income.

The econometric framework for imputing self-employment income also 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 of earnings, including wages, interest, and Social Security and pension income. But, 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 since that amount results in some self-employment tax being 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. Since the CPS ASEC survey only asks about net self-employment earnings, we convert these amounts into gross income amounts by multiplying them by the average ratio between gross and 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. [10] Thus, the econometric framework aims to estimate the likelihood that an individual that reports no net self-employment earnings actually 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:

$$SE^* = \gamma' x + \mu \qquad (7)$$

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

[9] This undercounting may be partially offset by the inclusion of certain flow-through income reported as self-employment earnings by CPS ASEC respondents.

[10] Different gross-up 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.

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 2012 data are shown in Table 3.

Table 3. Probit Model for the Presence of Self-Employment Income, Tax Year 2012

Variable	Parameter Estimate	Chi-square
Intercept	-2.486	(1988.5) **
logage	0.358	(579.5) **
male	-1.175	(265.8) **
logage*male	0.364	(374.4) **
west	-0.036	(15.1) **
midwest	-0.085	(77.9) **
northeast	-0.073	(51.1) **
wagesind	0.673	(194.5) **
interestind	0.024	(1.7)
socsecind	-1.490	(114.2) **
pensionind	-0.184	(19.9) **
logwages	-0.083	(341.3) **
loginterest	0.019	(27.2) **
logpension	0.007	(2.7)
logsocsec	0.080	(29.1) **
Number of observations	233,849	
R^2	0.094	
Max-rescaled R^2	0.150	

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

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:

$$I^*_{SE} = \delta' x + \nu \qquad (8)$$

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.[11] The indicator I_{SE} for the net self-employment earnings category is assigned as follows:

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

$$I_{SE} = \begin{cases} 1 & net \;\; earnings < \$0 \\ 2 & \$0 < net \;\; earnings <= \$433 \\ 3 & net \;\; earnings > \$433. \end{cases} \qquad (9)$$

The estimation results for tax year 2012 are shown in Table 4.

Table 4. Ordered Probit Models for the Category of Self-Employment Income, Tax Year 2012

Variable	Parameter Estimate	t-statistic
Intercept	1.769	(10.6) **
logage	-0.244	(-5.6) **
male	-0.789	(-3.7) **
logage*male	0.228	(4.2) **
west	0.093	(4.3) **
midwest	0.121	(5.2) **
northeast	0.218	(8.8) **
wagesind	0.188	(1.9)
interestind	-0.074	(-1.8)
socsecind	-0.476	(-1.6)
pensionind	-0.174	(-1.9)
logwages	-0.053	(-5.8) **
loginterest	0.030	(3.6) **
logpension	-0.010	(-1.0)
logsocsec	0.032	(1.0)
Limit	0.166	(36.7)
Values of dependent variable I_{SE}	Number of observations	
1 = (SE Income < 0)	5,440	
2 = (0 < SE Income <= 433)	1,265	
3 = (SE Income > 433)	18,573	
Total number of observations	25,278	
Missing values	10	
Log Likelihood	-17,444	

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

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 other research projects that rely on the CPS ASEC data base. For instance, this will facilitate an econometric analysis of reporting compliance among self-employed taxpayers. Our specification is:

$$\ln(SE) = \beta' x + \varepsilon, \qquad (10)$$

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 2012 are shown in Table 5.

Table 5. Regression Analysis for the Amount of Self-Employment Income > $433, Tax Year 2012

Variable	Parameter Estimate	t-statistic
Intercept	6.777	(35.9) **
logage	0.565	(11.2) **
male	-0.975	(-4.0) **
logage*male	0.356	(5.6) **
west	0.060	(2.3) *
midwest	-0.075	(-2.7) **
northeast	0.071	(2.5) *
wagesind	-0.021	(-.2)
interestind	-0.009	(-.2)
socsecind	-1.529	(-4.0) **
pensionind	-0.665	(-5.7) **
logwages	-0.060	(-5.4) **
loginterest	0.102	(10.4) **
logpension	0.050	(4.0) **
logsocsec	0.074	(1.8)
R^2	0.113	
Adjusted R^2	0.112	
Root MSE	1.356	
Coefficient of Variation	15.0337	
N	18,573	

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

For CPS ASEC respondents flagged for imputation of net self-employment earnings in excess of $433, the parameter estimates are used to impute earnings (SE) as follows:

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

where β 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 CPS ASEC respondents 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 CPS ASEC respondents 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.

Imputation of Unemployment Compensation

The fourth form of income imputed is unemployment compensation. In this case we just need to apply two models. The first is a probit specification for the likelihood that an individual has nonzero earnings from unemployment compensation:

$$U^* = \gamma' x + \mu \qquad (11)$$

where U^* is a latent variable describing the propensity for unemployment compensation to be present, x is a vector of explanatory variables, and γ is 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.

We estimate the conditional likelihood of reporting unemployment compensation given that it is present as the ratio of the aggregate number of reports in the weighted CPS ASEC sample to the aggregate number of recipients based on Form 1099-G data.

Next, we apply Bayes' rule to predict the probability that unemployment compensation is present given that it has not been reported. A comparison of a pseudo-random draw against this probability is made to flag CPS ASEC non-reporters for imputation of unemployment compensation.

The magnitude of unemployment compensation is estimated through a regression of the form:

$$\ln(U) = \beta' x + \varepsilon \qquad (12)$$

where $\ln(U)$ represents the natural log of unemployment compensation, 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 earnings from unemployment compensation is assumed to be log normal. Results for the probit and regression models for tax year 2012 are shown in Tables 6 and 7.

Table 6. Probit Model for the Presence of Unemployment Compensation, Tax Year 2012

Variable	Parameter Estimate	Chi-square
Intercept	-2.334	(1292.1) **
logage	0.211	(146.4) **
male	-0.077	(0.9)
logage*male	0.080	(13.4) **
west	0.181	(265.7) **
midwest	0.086	(54.8) **
northeast	0.240	(419.2) **
wagesind	1.608	(962.0) **
interestind	-0.023	(0.8)
socsecind	-0.213	(1.8)
pensionind	0.587	(189.0) **
seind	0.536	(68.2) **
logwages	-0.165	(1118.9) **
loginterest	-0.052	(84.9) **
logsocsec	-0.065	(14.3) **
logpension	-0.046	(89.4) **
logabsseinc	-0.088	(132.6) **
Number of observations	233,849	
R^2	0.026	
Max-rescaled R^2	0.068	

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

We impute income sources in the following order: retirement income, then self-employment income, then unemployment compensation. However, since the presence and amount of unemployment compensation depends on self-employment income, we implement our imputations 25 times (independently) and average the results.

Table 7. Regression Analysis for the Amount of Unemployment Compensation, Tax Year 2012

Variable	Parameter Estimate	t-statistic
Intercept	6.229	(32.8) **
logage	0.425	(8.4) **
male	-0.506	(-2.1) *
logagemale	0.176	(2.8) **
west	0.213	(8.1) **
midwest	0.050	(1.8)
northeast	0.296	(10.8) **
wagesind	1.880	(13.8) **
interestind	0.191	(2.8) **
socsecind	0.307	(0.7)
pensionind	-0.042	(-0.5)
seind	-0.187	(-1.0)
logwages	-0.202	(-14.4) **
loginterest	-0.004	(-0.3)
logpension	0.046	(4.4) **
logsocsec	-0.062	(-1.4)
logabsseinc	0.039	(1.7)
R^2	0.062	
Adjusted R^2	0.061	
Root MSE	1.197	
Coefficient of Variation	14.8135	
N	15,036	

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

Aggregate Adjustment for Other Missing Income

In addition to these underreported income types, there are several types of income that are subject to little or no reporting in the CPS ASEC survey. These include capital gains and losses, other gains and losses, state and local tax refunds, royalties, and miscellaneous other incomes reported on Form 1040 Schedule E. To account for returns that would be observed to have a filing requirement had these types of income been reflected in the CPS ASEC data, we determined from IRS data the number of required returns among the pool of timely and late filers first with and then without these types of income, then added the difference (i.e., the number of filed returns that were required due to the presence of these types of income) to the denominator of the VFR. This approach, which adds fewer than 900,000 returns to the denominator each year, does not address a possible undercounting of required returns (due to the absence of these income sources) among those who never file, but we anticipate that any such undercounting is likely to be small.

Impact of Imputations and Adjustment

Imputing these types of income caused the number of estimated required returns to increase by 7 to 10 million each year, as illustrated in Figure 5. Table 8 shows the increase in the total number of required returns resulting from each imputation.

Figure 5. CPS ASEC Estimates of Required Returns in the Population
Before and After Imputations and Adjustments, Tax Years 2005-2012

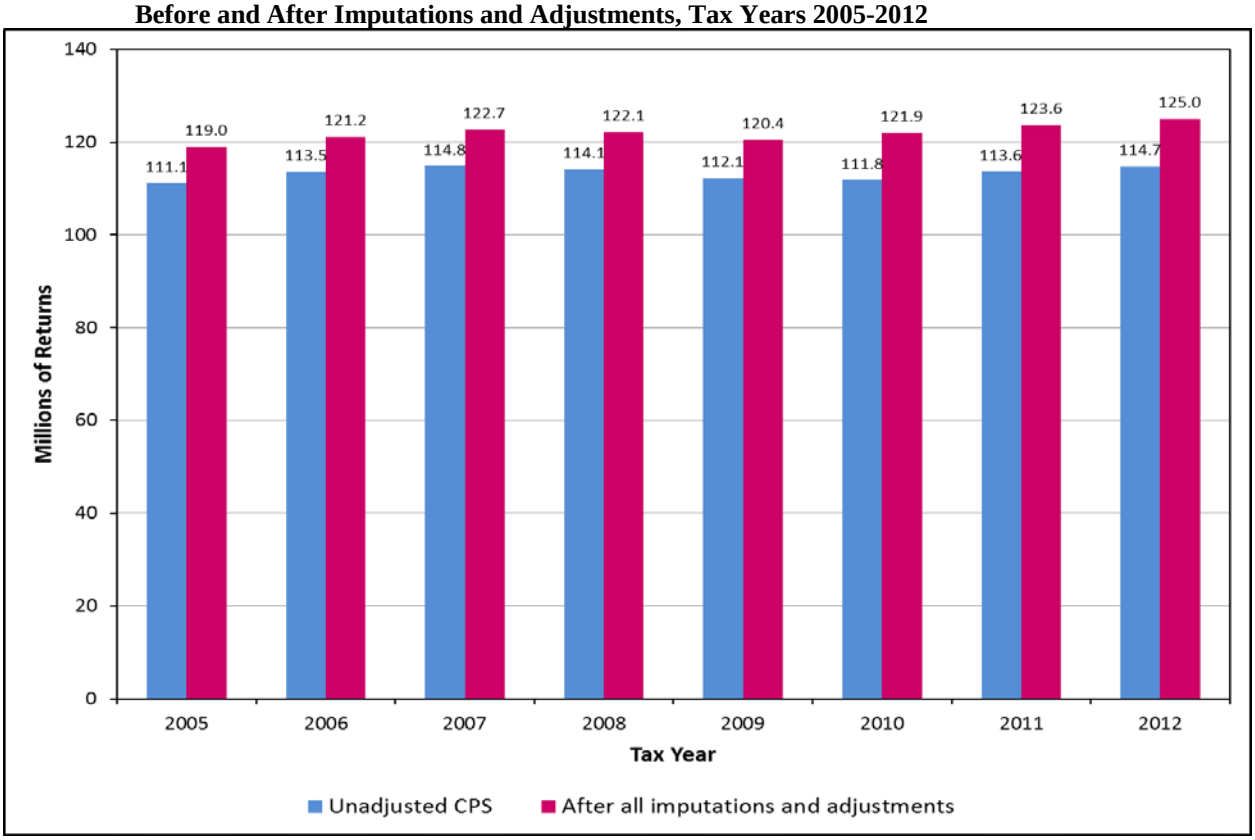


Table 8. CPS ASEC Estimates of the Number of Required Returns in the Population (in Millions) Impacts of imputing or adjusting for each type of income, Tax Years 2005-2012

Imputation/Adjustment Stage	2005	2006	2007	2008	2009	2010	2011	2012
Without imputing any income to the CPS ASEC	**111.1**	**113.5**	**114.8**	**114.1**	**112.6**	**111.8**	**113.6**	**114.7**
Increment from imputing only **Social Security** and **pension** income	4.6	4.6	4.8	4.9	5.8	6.1	6.0	6.5
Increment from also imputing **self-employment** income	1.9	1.9	1.8	2.2	1.3	2.4	2.5	2.4
Increment from also imputing **unemployment compensation**	0.4	0.4	0.4	0.3	0.1	1.2	0.8	0.7
Total after all imputations	**118.1**	**120.4**	**121.8**	**121.6**	**119.9**	**121.4**	**122.9**	**124.3**
Increment from accounting for income types not reflected in CPS ASEC	0.8	0.8	0.9	0.6	0.5	0.5	0.7	0.7
Total after all adjustments	**119.0**	**121.2**	**122.7**	**122.1**	**120.4**	**121.9**	**123.6**	**125.0**

The Resulting VFR

After applying all of the adjustments to the numerator and denominator of our measure, the resulting VFR for each of the last 13 years is given in Figure 6 and Table 9. Our estimates suggest that the VFR is somewhat higher in recent years after increasing markedly in 2007 due to the Economic Stimulus payments. When that benefit lapsed, some of the new filers remained. We suspect that some may have done so because of the incentives created by higher refundable tax credits after 2007.

Figure 6. Individual Income Tax Voluntary Filing Rate, Tax Years 2000-2012

Number of Required Returns Filed on Time / Total Number of Returns Required to be Filed

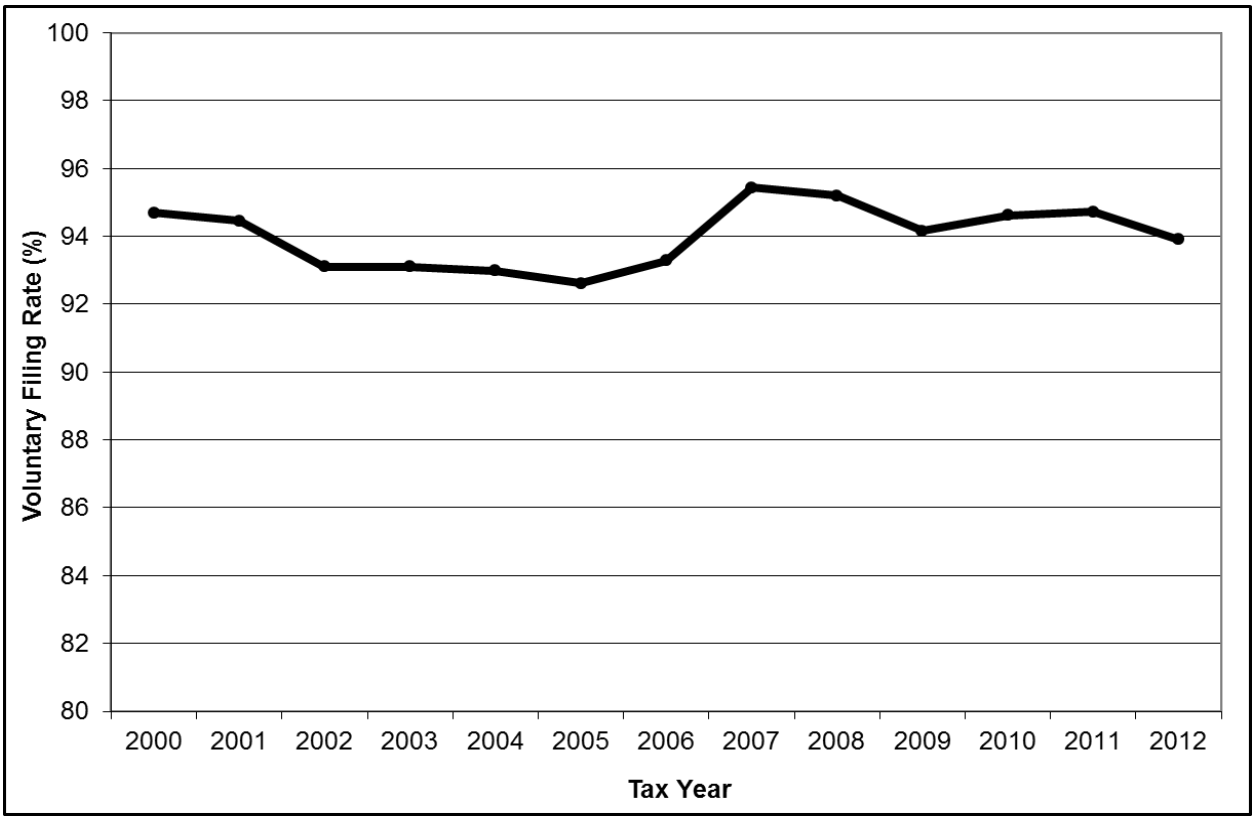


Table 9. Voluntary Filing Rate and Related Estimates, Tax Years 2000-2012

Tax Year	Millions of Required Returns			VFR (Ratio)
	Total Population (Denominator)	Timely Filed (Numerator)	Nonfilers (Difference)	
2000	113.5	107.5	6.0	94.7
2001	114.2	107.9	6.3	94.5
2002	115.0	107.0	7.9	93.1
2003	114.7	106.8	7.9	93.1
2004	116.5	108.3	8.2	93.0
2005	119.0	110.2	8.8	92.6
2006	121.2	113.0	8.1	93.3
2007	122.7	117.1	5.6	95.4
2008	122.1	116.2	5.9	95.2
2009	120.4	113.3	7.0	94.2
2010	121.9	115.3	6.6	94.6
2011	123.6	117.1	6.5	94.7
2012	125.0	117.4	7.6	93.9

Benefits of the VFR Analysis

Our efforts to improve the VFR measure have produced several important benefits. Perhaps one of the most significant of these has been to document the extent to which the CPS ASEC data do not include all of the income that is reported to the IRS, and to develop reasonable approaches to imputing these income sources to the CPS ASEC each year so that they mirror the known distributions in IRS administrative data. Our efforts have also resulted in a more accurate definition of the criteria underlying the filing requirement, which we now apply as closely as possible to both the numerator and denominator of the VFR measure. Our improved understanding of these criteria has even prompted a revised description of the gross income concept in the filing requirement section of the Form 1040 instruction booklet. Ultimately, these improvements enhance the quality of the measure, and allow us to develop a deeper understanding of the drivers of fluctuations in the VFR over time.

Future Work

Work is under way to evaluate other potential sources of misreporting on both tax returns and Census samples. In the case of tax return data, two key issues are at the forefront of this research. The first has to do with filers who misreport their filing status, and therefore their filing threshold. In particular, a very large number of singles incorrectly claim Head of Household status on their tax returns, and by so doing, many of them appear to have no filing requirement when in fact they do. We have therefore begun research using audit data from the IRS National Research Program to impute corrected filing statuses to those who claim Head of Household status; this will increase the number of required returns filed on time, and therefore the VFR.

The second tax return data issue is that some returns that appear not to be required on the basis of the income information reported (because the reported income is under the filing threshold) would be designated as required (and, therefore, part of the count in the numerator of the VFR) if non-reported income from information returns were taken into account. Thus, another refinement of the VFR will involve adding in this additional non-reported income to the returns that appear to not be required when only self-reported income is considered. These returns (which are timely filed but underreport income) would then be properly considered as compliant with the obligation to file a required return, though likely not compliant with the obligation to fully report all income.

In the case of the CPS ASEC data, we found that a large share of respondents report amounts for various income sources that are rounded to multiples of $1,000, $5,000 or $10,000. This doesn't have much impact on many uses of the data, but it has a pronounced effect when comparing someone's income against a threshold. For example, the filing threshold for Singles in Tax Year 2012 was $9,750—just under $10,000, which was an extremely frequent amount of wages reported in the CPS ASEC. It seems apparent that many individuals had income under the filing threshold (and were therefore not required to file a return), but made themselves appear to have a filing requirement by rounding their wages up to $10,000. This inflates the denominator and therefore decreases the (unadjusted) VFR. Likewise, when the filing threshold climbs above $10,000 due to indexing, we would expect the opposite to be true. We are therefore exploring

ways of "unrounding" the CPS ASEC data. In addition, we hope to address the fact that many individuals who earn a small amount of wages fail to report it to the CPS ASEC.

We also plan to explore ways to estimate the denominator of the VFR solely from administrative data (i.e., without Census data). This would present both advantages and disadvantages. A key advantage would be having greater ability to explore the role of the numerator and denominator together—rather than just the numerator—in affecting fluctuations in the VFR. The biggest challenge is determining how best to combine no-return individuals into tax units (combining spouses and allocating children); we will also need to find a way to account for income sources, such as self-employment income, which are not subject to third-party reporting. Eventually, we hope to compare the Census-based VFR estimates with the IRS-only VFR estimates to determine which one is the best approach to estimating the extent of missing returns and missing income.

In addition to working to refine our measure of the VFR, we are undertaking research to better understand why some taxpayers do not meet their filing obligations. Our approach involves the application of microeconometric methods to compare the characteristics of filed required returns (based on tax return data) to the characteristics of required returns in the overall population (based on CPS ASEC data). This research may provide insights into ways to better assist taxpayers in meeting their filing obligations.

References

Brown, Robert E. and Mark J. Mazur, "IRS's Comprehensive Approach to Compliance Measurement," *National Tax Journal*, 2003, vol. 56, issue 3, pp. 689-700.

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

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

Plumley, A.H., "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.

Appendix
Variable Descriptions

Variable	Description
interestind	Indicator of presence of interest income on Form 1099INT
pensionind	Indicator for positive gross distribution amount on Form 1099R
seind	Indicator for nonzero self-employment income of the individual taxpayer as reported on Schedule C
socsecind	Indicator for positive amount of net social security benefits reported on Form SSA-1099
wagesind	Indicator for positive amount of wages, tips and other compensation reported on Form W2
logabsseinc†	Natural logarithm of the absolute value of self-employment income of the individual taxpayer as reported on Schedule C
loginterest†	Natural logarithm of interest income reported on Form 1099INT
logpension†	Natural logarithm of the gross distribution amount from Form 1099R
logsocsec†	Natural logarithm of amount of net social security benefits reported on Form SSA-1099
logwages†	Natural logarithm of amount of wages, tips, and other compensation from Form W2
logage†	Natural logarithm of the age of the taxpayer at the end of the tax year
male	Indicator for male from SSA Master File
noncitizen	Taxpayer is not a citizen according to indicator in SSA Master File
midwest	Taxpayer resides in one of the following states: WI, IL, IN,MI, OH, ND, MN, SD, IA, NE, KS, or MO. Residency based on address provided on tax return, if filed, or on information returns
northeast	Taxpayer resides in one of the following states: NH, VT, ME, MA, RI, CT, NY, PA, or NJ. Residency based on address provided on tax return, if filed, or on information returns
west	Taxpayer resides in one of the following states: WA, OR, CA, AK, HI, MT, ID, WY, NV, UT, CO, AZ, or NM. Residency based on address provided on tax return, if filed, or on information returns

† Logarithms are taken of the value plus one to avoid taking the log of zero.