Advances in Nonfiling Measures

Brian Erard, B. Erard & Associates; and Mark Payne and Alan Plumley, IRS Office of Research[1]

The Internal Revenue Code places three basic obligations on individual income taxpayers: to file tax returns when required; to report accurately on those returns their income and tax; and to pay their tax on time. The Internal Revenue Service (IRS) therefore categorizes taxpayer noncompliance into three mutually exclusive and exhaustive types of behavior: nonfiling, underreporting, and underpayment. Underreporting accounts for the largest portion of the tax gap (the amount of tax imposed by law, but not paid on time) and receives most of the attention. Underpayment is generally observable, so the extent of underpayment can be tabulated and tracked from IRS systems. This paper focuses on two measures of individual nonfiling: the nonfiling gap (the amount of tax not paid on time by those who do not file on time) and the Voluntary Filing Rate (the number of required returns that are filed on time, expressed as a percentage of the total number of returns that are required to be filed, whether filed or not). In particular, we describe in this paper a number of methodological improvements we have made for estimating these two measures.

The Nonfiling Gap

The nonfiling gap focuses on those who are required to file tax returns, but do not file those returns on time. The nonfiling gap is defined as the difference between the total tax liability of these nonfilers and the amount of that tax that has been paid on time (such as through withholding and estimated tax payments). Thus, the nonfiling gap includes the amount of tax *not* paid on time by those who file late, but it excludes amounts of tax that *are* paid on time—both by late filers and by those who never file.[2]

The previous estimate of the individual income tax nonfiling gap (for Tax Year 2001) was based on aggregate tabulations derived from the Exact Match study conducted by the Census Bureau. This study "matched" records from the Current Population Survey with data that Census regularly receives from the IRS to identify potential nonfilers, whose tax liability could be estimated from the Census survey. This approach had several shortcomings, however. First, the IRS data did not clearly distinguish between timely and late returns, so Census treated all returns as timely. Second, the Census data do not include key tax-related information (e.g., eligibility for important tax benefits), and tend to understate some income types, so the calculation of tax liability is subject to much uncertainty. Finally, IRS tabulations of the tax paid on time by late filers and those who never filed were also subject to much uncertainty.[3]

Given our concerns about the Exact Match method for estimating the nonfiling gap, we decided to apply an alternative method pioneered by Treasury's Office of Tax Analysis and the congressional Joint Committee on Taxation. That method is based almost exclusively on IRS administrative data as opposed to Census data. The basic approach is as follows:

1. Select a large random sample of valid Social Security Numbers (SSNs) (e.g., excluding those of deceased persons).
2. Compare those SSNs with filed tax returns for a given tax year, putting them into three categories:
   a. Those that appeared on **timely** filed tax returns (as the primary taxpayer, the secondary taxpayer [i.e., spouse], or as a dependent);
   b. Those that appeared on **late** tax returns (as the primary taxpayer, the secondary taxpayer [i.e., spouse], or as a dependent); and
   c. Those that did not appear on any tax return at all (which we refer to as the "**no-return**" group).
3. Compile data for the no-return group of SSNs (primarily age, income, and income tax withholding) from Social Security records and third-party information returns.

4. Assemble the no-return group members into synthetic families, guided by the overall profile of the population from Census data, and taking into account the contribution to that profile by timely and late filers. For each synthetic family, identify one or more synthetic returns based on apparent filing statuses (e.g., married-joint, single, head of household) of the household members and assign income and exemptions to these returns based on the data collected in step 3.
5. Impute tax deductions, other income (such as capital gains), and likely tax credits to the synthetic returns, then compute their tax liability.
6. Subtract the amount of withholding from the estimated tax liability on each synthetic return to derive the hypothetical balance due (contribution to the nonfiling gap), then sum these amounts across all of the synthetic returns to derive the portion of the nonfiling gap attributable to the no-return group.

We applied this approach to Tax Year 2005 data, which resulted in a nonfiling gap estimate of $12.9 billion among the no-return population, as summarized in Table 1. That study was the subject of a paper presented at the 2011 IRS-Tax Policy Center Research Conference.[4]

TABLE 1. Estimates Related to Individuals Who Did Not Appear on Tax Returns for TY2005

	Total	Those Required to File	
		Before Credits	After Credits
Number of individuals	38.63 million	11.82 million	11.82 million
Number of synthetic tax returns	22.79 million	5.18 million	5.18 million
Income subject to tax	$233.7 billion	$196.7 billion	$196.7 billion
Tax liability	$20.2 billion	$21.1 billion	$20.2 billion
Tax balance due (not paid on time)	$12.9 billion	$13.8 billion	$12.9 billion

However, that represented just a portion of the overall nonfiling gap. The other major piece was the contribution from late filers. We found that the late returns identified above (in step 2. b.) reported an aggregate $8.7 billion balance due. However, this understates the true balance due on these returns, because the returns did not fully account for income that was independently reported on third-party information documents. Therefore, we accounted for additional income on those late returns using the logic summarized in Table 2 for each line item on the return.

TABLE 2. Logic for Using Information Return Data To Adjust Items Reported on Late Returns

	Form	Line	Item	Adjustment Logic
A	1040	7	Wages	Let GIC = Max[(D-E+G), (J+I+H+F), 0] • If A>0 and (B+C)>0 and GIC>0 and -150<(B+C+L-GIC)<150, then: ◦ Wages = (B+C) and ◦ Schedule C net income = Max[K-(B+C), 0] • Else, if A>0 and (B+C)>0 and GIC=0 and -150<(A-(B+C))<150, then: ◦ Wages = Max[A-L, (B+C), 0] and ◦ Schedule C net income = L • Else: ◦ Wages = Max[A, (B+C), 0] and ◦ Schedule C net income = Max[K, (L-GIC)+K]
B	W-2	1	Wages	
C	W-2	8	Allocated tips	
D	Schedule C	1	Gross receipts	
E	Schedule C	2	Returns & allowances	
F	Schedule C	4	Cost of goods sold	
G	Schedule C	6	Other income	
H	Schedule C	28	Total expenses	
I	Schedule C	30	Business use of home	
J	Schedule C	31	Net profit (loss)	
K	1040	12	Schedule C net income	
L	1099MISC	7	Nonempl compensation	

TABLE 2. Logic for Using Information Return Data To Adjust Items Reported on Late Returns—Continued

	Form	Line	Item	Adjustment Logic
M	1040	8a	Taxable interest	Interest income = Max[M, (N+O+P+Q+R)]
N	1099-INT	1	Interest income	
O	1099-INT	3	Interest on savings bonds	
P	K-1 (1041)	1	Interest income	
Q	K-1 (1120S)	4	Interest income	
R	K-1 (1065)	5	Interest income	
S	1040	9a	Ordinary dividends	Ordinary taxable dividends = Max[S, (T+U+V+W)]
T	1099-DIV	1a	Ordinary dividends	
U	K-1 (1041)	2a	Ordinary dividends	
V	K-1 (1120S)	5a	Ordinary dividends	
W	K-1 (1065)	6a	Ordinary dividends	
X	1040	9b	Qualified dividends	Qualified dividends = Min[X, Y] (The qualified dividends amounts from the Forms K-1 are not in our data.)
Y	1099-DIV	1b	Qualified dividends	
Z	1040	10	State tax refunds	State tax refund = Max[Z, Min[AA, AB]]
AA	1099-G	2	State tax refunds	
AB	Schedule A	5	Prior year deduction for S&L income taxes	
AC	1040	13	Capital gain (loss)	IRPCG = (AD+AE+AF+AG+AH+AI+AJ) Capital gain = Max[AC, IRPCG]
AD	1099-DIV	2a	Cap. gain distribution	
AE	K-1 (1041)	3	Net ST cap. gain (loss)	
AF	K-1 (1041)	4a	Net LT cap. gain (loss)	
AG	K-1 (1120S)	7	Net ST cap. gain (loss)	
AH	K-1 (1120S)	8a	Net LT cap. gain (loss)	
AI	K-1 (1065)	8	Net ST cap. gain (loss)	
AJ	K-1 (1065)	9a	Net LT cap. gain (loss)	
AK	1040	15a	IRA distributions	IRA and pension income combined to account for misclassification. If AK=0, then AK=AL If AM=0, then AM=AN IRA + Pension income = Max[(AL+AN), (AO-AK+AL), (AP-AM+AN)] AP=0 (to avoid double-counting pension income)
AL	1040	15b	Taxable IRA distrib'n	
AM	1040	16a	Pensions & annuities	
AN	1040	16b	Taxable pension, annuity	
AO	5498	3	Roth conversion amt	
AP	1099-R	2a	Taxable pension	
AQ	1040	18	Farm income or loss	Farm income = Max[AQ, (Max[AR,0] + Max[AS,0])]
AR	1099-G	7	Agricultural subsidy	
AS	1099-MISC	10	Crop insurance proceeds	
AT	1040	19	Unemployment comp.	Unemployment compensation = Max[AT, AU]
AU	1099-G	1	Unemployment comp.	
AV	1040	20a	Social security benefits	Social security benefits = Max[AV, AW]
AW	1099-SSA	3	SS benefits	
AX	1040	21	Other income	Line21Calc=AY+AZ+BA If (AX<0 and Line21Calc=0) or (Schedule C net income ≠ 0) or (Farm income ≠ 0) then: Other income = AX; Else: Other income = Max[AX, Line21Calc]
AY	W-2G	1	Gross winnings	
AZ	1099-C	2	Amt of debt cancelled	
BA	1099-G	5	ATAA payment	

TABLE 2. Logic for Using Information Return Data To Adjust Items Reported on Late Returns—Continued

	Form	Line	Item	Adjustment Logic
BB	1040	17	Schedule E net income	GrossE = (BC+BD+Max[(BE+BF), BG]+BH+BJ+Max[BI, 0]) If BB > GrossE, Then GrossE = BB Note: any negative amount from any of the following components is set to zero: Line17Calc = BK+BL+BM+BN+BO+BP+BQ+BR+BS+BT+BU+BV+BW+BX+BY Schedule E net profit (loss) = Max[BB, BB + (Line17Calc – GrossE)]
BC	Schedule E	23c	Total rents received	
BD	Schedule E	23d	Total royalties received	
BE	Schedule E	29a (g)	Passive income from partnership or S corp	
BF	Schedule E	29a (j)	Nonpassive income from partnership or S corp	
BG	Schedule E	30	Passive + nonpassive inc. from partn or S corp	
BH	Schedule E	35	Estate & trust income	
BI	Schedule E	40	Farm rental net income	
BJ	Schedule E	41	REMIC net income	
BK	K-1 (1065)	1	Ordinary business inc.	
BL	K-1 (1065)	2	Net rental real estate inc.	
BM	K-1 (1065)	3	Other net rental income	
BN	K-1 (1065)	4	Guaranteed payments	
BO	K-1 (1065)	7	Royalties	
BP	K-1 (1041)	5	Other portfolio income	
BQ	K-1 (1041)	6	Ordinary business inc.	
BR	K-1 (1041)	7	Net rental real estate inc.	
BS	K-1 (1041)	8	Other rental income	
BT	K-1 (1120S)	1	Ordinary business inc.	
BU	K-1 (1120S)	2	Net rental real estate inc.	
BV	K-1 (1120S)	3	Other rental income	
BW	K-1 (1120S)	6	Royalties	
BX	1099-MISC	1	Rents	
BY	1099-MISC	2	Royalties	
BZ	1040	64	Tax withheld	Total withholding = CB+CC+CD+CE+CF+CG+CH+CI+CJ+CK+CL+CM+CN Total prepayments = Total withholding + CA
CA	1040	65	Estimated tax payments	
CB	W-2	2	Income tax withheld	
CC	W-2G	2	Income tax withheld	
CD	K-1 (1120S)	13(Q)	Backup withholding	
CE	1099-B	4	Income tax withheld	
CF	1099-SSA	6	Income tax withheld	
CG	1099-RRB	10	Income tax withheld	
CH	1099-G	4	Income tax withheld	
CI	1099-DIV	4	Income tax withheld	
CJ	1099-INT	4	Income tax withheld	
CK	1099-MISC	4	Income tax withheld	
CL	1099-OID	4	Income tax withheld	
CM	1099-PATR	4	Income tax withheld	
CN	1099-R	4	Income tax withheld	

After accounting for additional income using the logic presented in Table 2, we recalculated tax and the balance due for each return.[5] As indicated in Table 3, the sum of those balances due rose to $12.7 billion after

estimated credits. Combining our estimate ($12.7 billion) for late filers with our earlier estimate ($12.9 billion) for those who did not file at all, our overall estimate of the nonfiling gap for 2005 is $25.6 billion. This compares with an estimate of $25 billion for TY2001, which was derived using the Exact Match methodology.

TABLE 3. Estimates Related to Late Filers for Tax Year 2005

	All Late Returns		Those Required To File	
	Dollars (billions)	Returns (millions)	Dollars (billions)	Returns (millions)
Reported income subject to tax	$410.6	8.07	$401.7	5.99
Adjusted income subject to tax	$432.3	8.07	$423.4	6.15
Reported balance due	$8.8	2.27	$8.8	2.27
Adjusted balance due	$12.7	2.73	$12.7	2.73

The last step was to project the Tax Year 2005 estimate forward 1 year, to be consistent with the overall tax gap update for Tax Year 2006. The average *dollars* did not change appreciably from 2005 to 2006—due either to inflation or to tax law changes. However, our separate work to estimate the Voluntary Filing Rate (described below) indicated that the *number* of nonfilers declined from 11.3 million to 9.6 million during this interval, so the nonfiling gap undoubtedly declined as well. Although we could not determine a precise reduction in the nonfiling gap corresponding to that decline in the number of nonfilers, it was clear that the decline was due primarily to the infusion of low-income people who had an incentive to file in order to claim the one-time Telephone Excise Tax refund in 2006. This suggests that the reduction in the nonfiling gap was significantly smaller than the reduction in the number of nonfilers. Therefore, we reduced the overall nonfiling gap estimate from $25.6 billion in 2005 to $25.0 billion for 2006.

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. In fact, when the IRS began developing a concerted nonfiler strategy, the VFR was selected as one of the strategic measures to be tracked. It is defined for a given tax year as:

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

The numerator is tabulated from IRS data, and the denominator is estimated from Census data (the Annual Social and Economic Supplement, ASEC, of the Current Population Survey, CPS). Both the numerator and the denominator were first estimated 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 to researchers, we began estimating the numerator from population data. This required developing new systems to categorize each return as timely or late and as required to be filed or not required. 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.[6] When we estimated the VFR for 2009, however, we observed a dramatic decline, which we could not 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 statistic. 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 15), and that they reflected the same definitions (as much as the data would allow) for the

requirement to file. In the process, we discovered that none of the existing instructions that the IRS provides to taxpayers fully defined 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. Therefore, we took steps to ensure that the instructions given to taxpayers reflect this nuance.

Our next significant finding was that the Census data used to construct the denominator of the VFR significantly understates certain types of income (such as pensions, Social Security income, and sole proprietor income). This understatement in the denominator of the measure contributes to an overstatement of the VFR. Figures 1-3 illustrate the differences in the amounts of these types of income reported on the 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), or what is reported on filed income tax returns (in the case of self-employment income). To address these discrepancies, we have developed an econometric methodology for imputing the missing income to the ASEC records. We employ age, gender, region, and citizenship, as well as indicators and amounts of wages, interest, and unemployment compensation to predict the amount of pension and Social Security income that should have been reported on the CPS.[7] For predicting self-employment income, we also employ filing status and the number of dependents, but we do not control for citizenship.[8] Imputing these types of income caused the number of estimated required returns to increase by over 7 million each year, as illustrated in Figure 4. Table 4 shows that the imputations to pension and Social Security income added roughly the same number of required returns as the imputations to self-employment income.

In addition, there are several types of income that are subject to little or no reporting in the 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 seen to have a filing requirement had these types of income been reflected in the CPS data, we determined from IRS data the number of required returns among both 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. This approach does not address 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.

After applying all of the adjustments to the numerator and denominator of our measure, we have re-estimated the VFR for each of the last 11 years (see Figure 5 and Table 5). Our updated results reveal that the decline in 2009 was not as pronounced as our preliminary measure had suggested and that the peak was in TY2007, not in TY2008. Our revised estimates further indicate that the decline in 2009 appears to represent a gradual return to historical filing behavior. Specifically, the temporary increase in the VFR was largely caused by the fact that many taxpayers who had traditionally filed late or not at all had a great incentive to file on time to be eligible for the Economic Stimulus Payment in 2007 (and to some extent the Telephone Excise Tax Refund in 2006). When that benefit lapsed, many of these taxpayers reverted to their old behavior. We suspect that this was especially true of those whose income put them above the filing threshold but below IRS enforcement thresholds.

FIGURE 1. Social Security Income Reported on Form 1099 SSA-RRB vs. CPS, TY2010

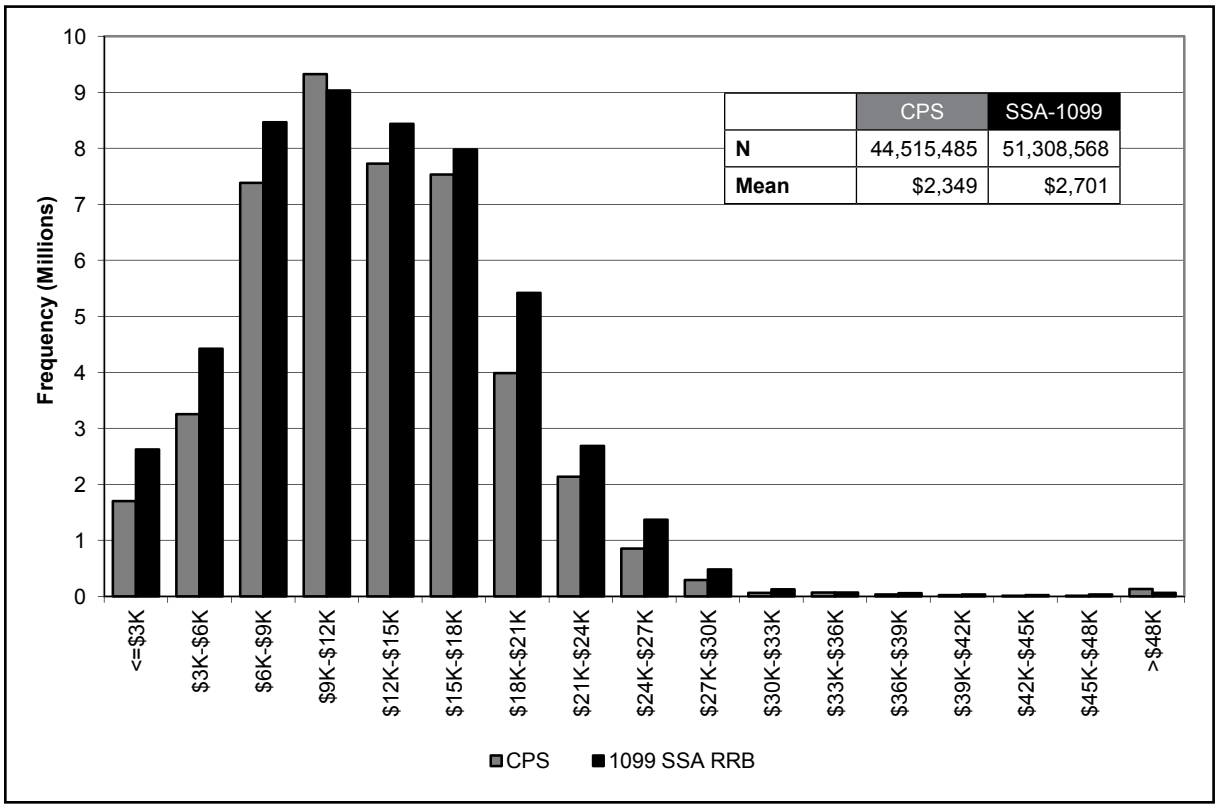


FIGURE 2. Pension Income Reported on Form 1099-R vs. CPS, TY2010

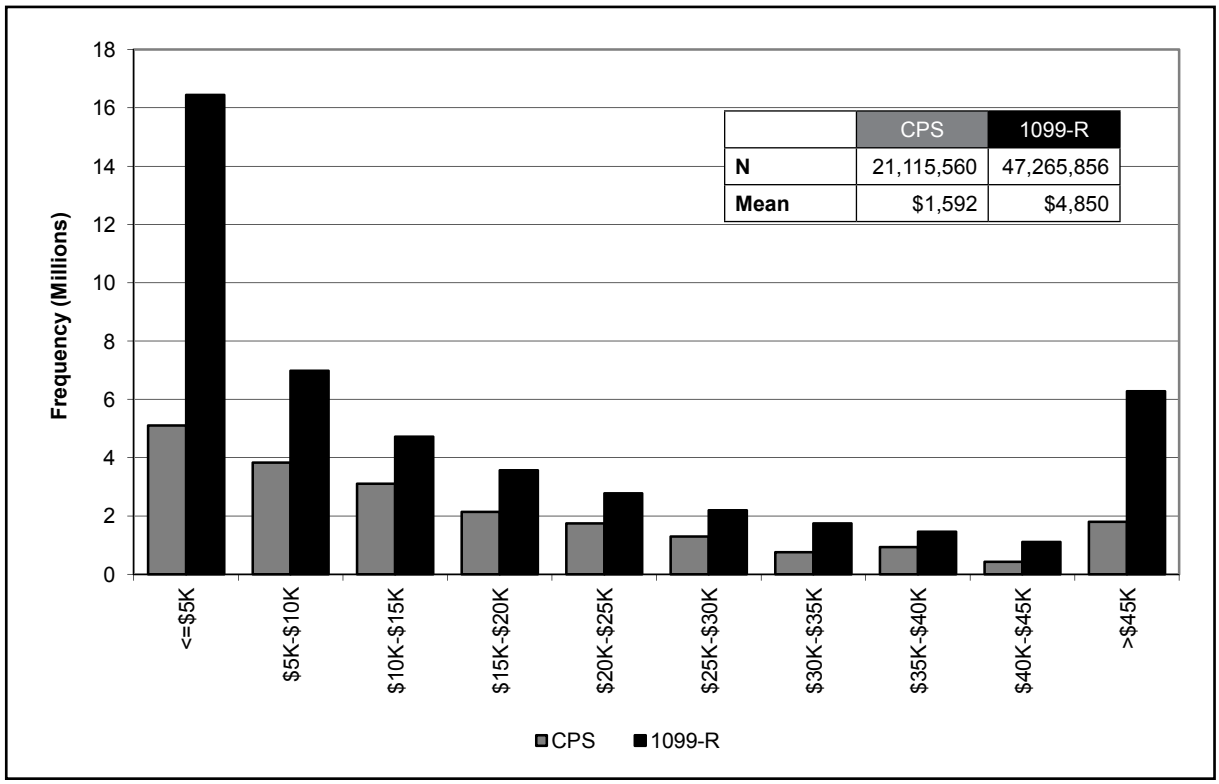
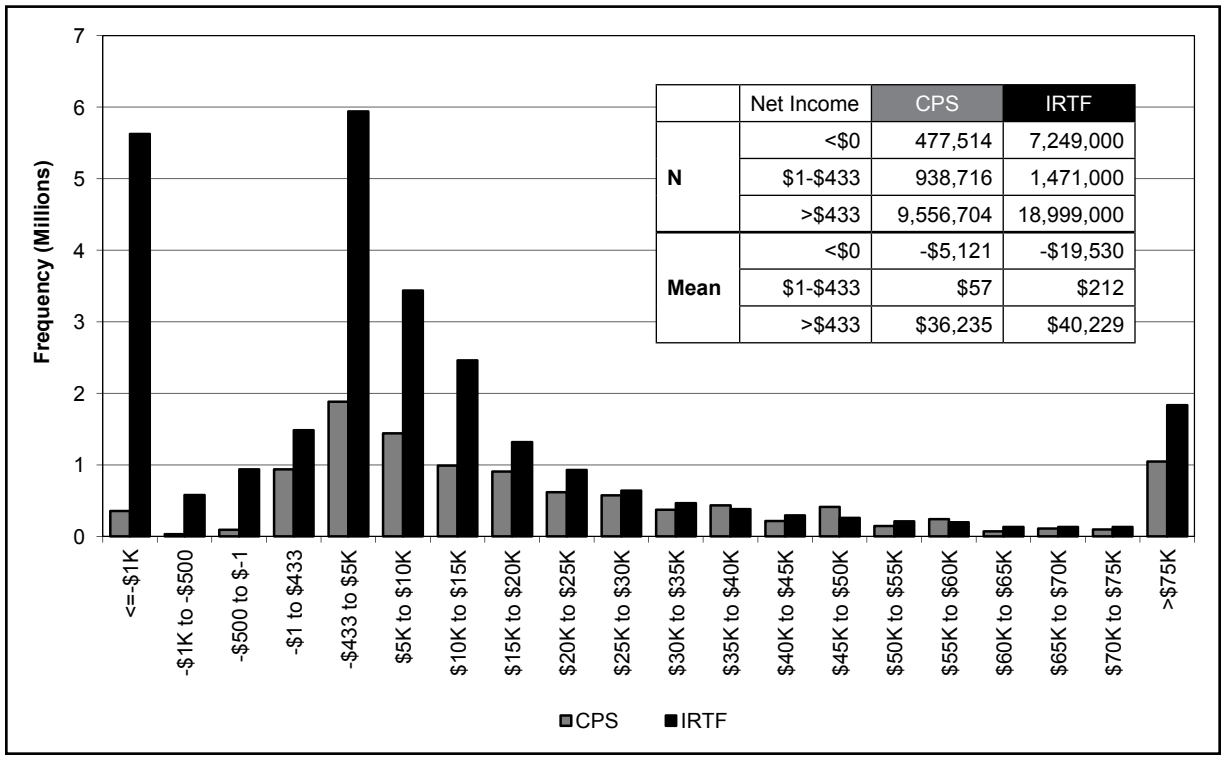


FIGURE 3. Self-Employment Income: IRTF vs. CPS, TY 2010



	Net Income	CPS	IRTF
N	<$0	477,514	7,249,000
	$1-$433	938,716	1,471,000
	>$433	9,556,704	18,999,000
Mean	<$0	-$5,121	-$19,530
	$1-$433	$57	$212
	>$433	$36,235	$40,229

FIGURE 4. CPS Estimates of Required Returns in Population

With and Without Imputed Income, Tax Years 2007–2009

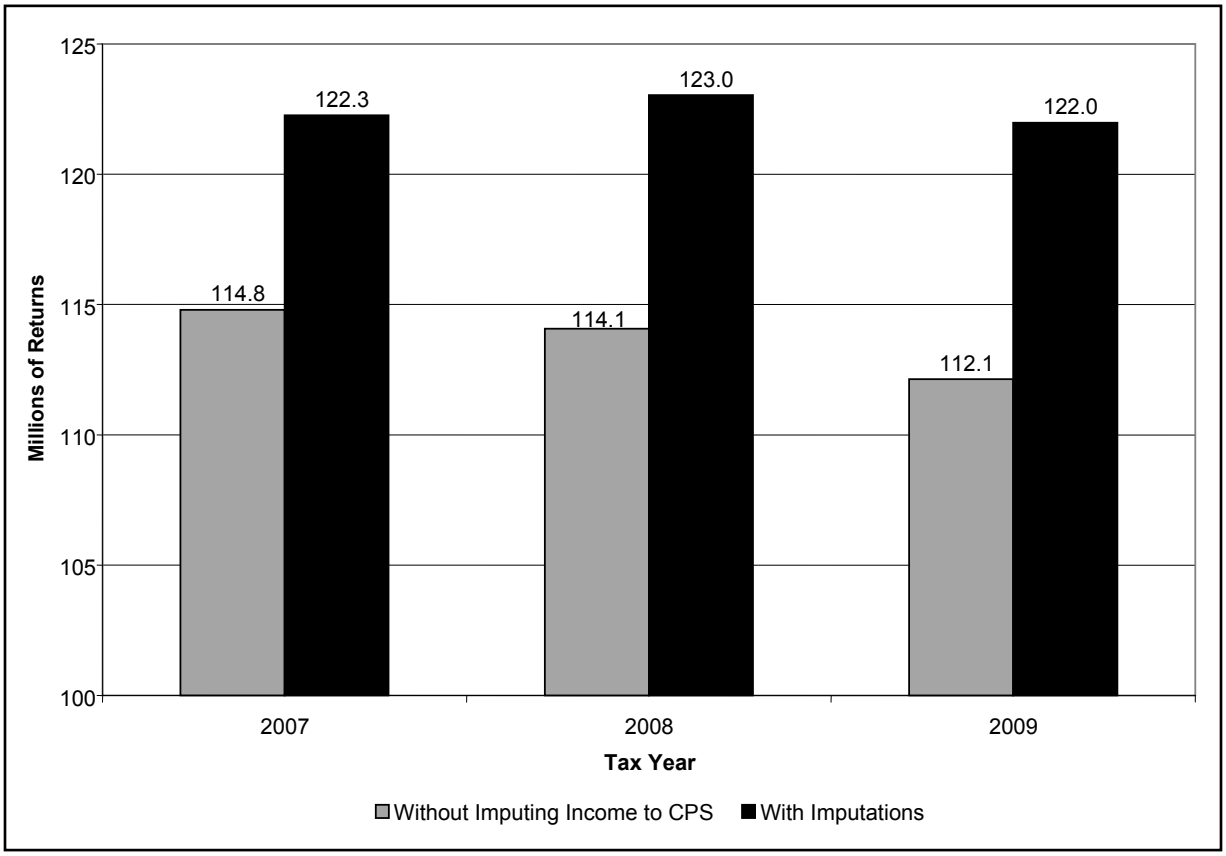


TABLE 4. CPS Estimates of Required Returns in Population

Impacts of imputing pension/Social Security income vs. self-employment income, Tax Years 2007–2009

Estimated Number of Required Returns (millions)	2007	2008	2009
Without imputing any income to CPS	114.8	114.1	112.1
Increment from imputing only Social Security and pension income	4.4	4.2	5.1
Increment from imputing only self-employment income	3.6	4.7	5.2
Number double-counted in the two increments	-0.5	0.0	-0.5
Total after all imputations	122.3	123.0	122.0

FIGURE 5. Individual Income Tax Voluntary Filing Rate

VFR = Number of Required Returns Filed on Time / Total Number of Returns Required To Be Filed

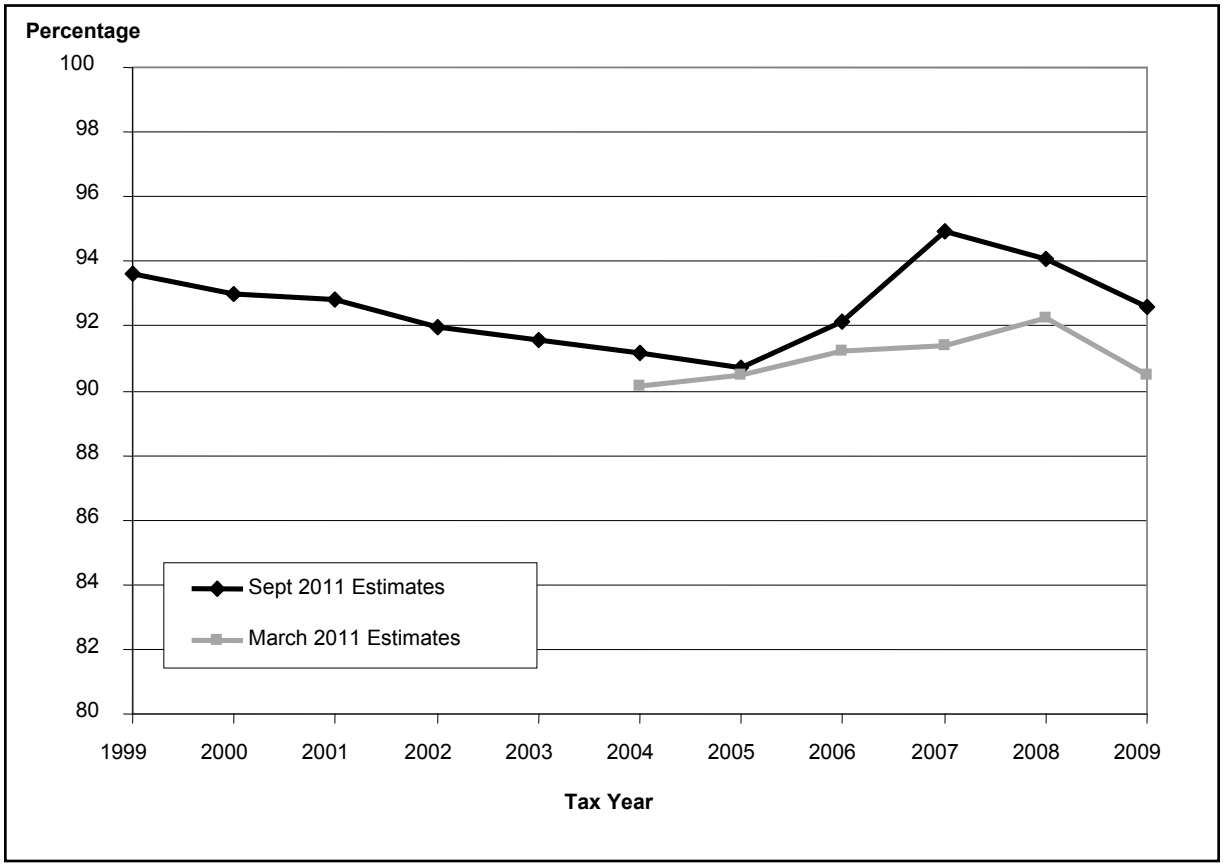


TABLE 5. Updated Voluntary Filing Rate and Related Estimates, Tax Years 1999–2009

Tax Year	Required Returns (Millions)					VFR
	Denominator (Total Required)	Numerator (Timely Filed)	Change in Denominator	Change in Numerator	Number of Nonfilers	
1999	112.6	105.4			7.2	93.6
2000	115.4	107.4	2.8	2.0	8.1	93.0
2001	116.1	107.8	0.7	0.4	8.3	92.8
2002	116.3	106.9	0.2	-0.9	9.4	91.9
2003	116.5	106.7	0.2	-0.2	9.8	91.6
2004	118.7	108.2	2.2	1.5	10.5	91.2
2005	121.4	110.1	2.7	1.9	11.3	90.7
2006	122.6	112.9	1.2	2.8	9.6	92.1
2007	123.3	117.0	0.7	4.1	6.3	94.9
2008	123.5	116.1	0.2	-0.9	7.3	94.1
2009	122.3	113.2	-1.2	-2.9	9.0	92.6

Influenced by the Telephone Excise Tax Refund

Influenced by the Economic Stimulus Payment

Benefits of the VFR Analysis

Our efforts to enhance 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 understate certain types of income, and to develop a reasonable approach to imputing these income sources to the CPS each year. 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. Under our revised methodology, we now apply a consistent definition of what it means for a required return to be timely filed for VFR purposes; we now include in the numerator only those required returns that are filed by December 31 of the primary filing year.[9] 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 if other types of income are significantly understated in the Census samples, and to use the VFR to analyze the factors that influence filing compliance. 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 in affecting fluctuations in the VFR, rather than just the numerator.

Endnotes

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

[2] This means that some nonfilers who were required to file did not contribute to the tax gap—either because they had no tax liability, or because they fully paid their tax liability on time. Thus, for tax gap purposes, the key issue is the amount of tax not paid on time—not the technical requirement to file a return. Note, however, that overpayments by some nonfilers do not offset underpayments by others (just as refunds paid to filers at the time of filing do not offset the underreporting gap). Also, returns filed before an officially extended due date are considered timely.

[3] Although it is possible to identify in the IRS Master File the timely payments made by late filers and nonfilers, the Master File does not indicate how much of those payments was in excess of true tax liability—particularly for those who never filed a return. Thus, the Master File tabulations are bound to overstate the timely payments of true tax liability, resulting in an underestimate of the nonfiling gap.

[4] Josh Lawrence, Michael Udell, and Tiffany Young, "The Income Tax Position of Persons Not Filing Returns for Tax Year 2005," *The IRS Research Bulletin*, Publication 1500 (Rev. 4-2012), pp. 143–155.

[5] We assumed that the total of all withholding for a given taxpayer that was documented by third parties on information returns was more accurate than the amount reported by the taxpayer on his or her Form 1040.

[6] 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. Furthermore, 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.

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

[8] To be consistent with the CPS measure of self-employment income, we impute partnership income along with nonfarm sole proprietorship income.

[9] 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.