# The Individual Income Tax and Self-Employment Tax Nonfiling Gaps for Tax Years 2008–2010

*Pat Langetieg, Mark Payne, and Alan Plumley*
*IRS, Research, Applied Analytics, and Statistics*

## Introduction

Taxpayers are required by the Internal Revenue Code to file income tax returns with the IRS by the established due date, on which they are to report all of their tax liability; it also requires them to pay that tax liability on time. However, not all taxpayers file required tax returns on time (or at all), and some of their tax liability is therefore not paid on time. The nonfiling gap is the amount of true tax liability not paid on time by those who do not file on time. Since some nonfilers pay some or all of their true tax liability on time (e.g., through withholding), not all nonfilers actually contribute to the tax gap. Nonetheless, the nonfiling gap is comprised of two major components: the portion associated with those who file late ("late filers"), and the portion associated with those who never file at all ("not-filers"). Thus, from a tax gap perspective, nonfilers include both late filers and not-filers. With the passage of time, some not-filers file late returns, so the distinction between these two groups is merely a pragmatic one for estimating the gap. It is easier to estimate the contribution that late filers make to the nonfiling gap since we have their tax return; it is much harder to estimate the gap associated with those who have not filed any return by the time the estimate is made.

We estimate that the average annual individual income tax nonfiling gap over the TY 2008 through TY 2010 period was $26 billion, and the corresponding self-employment tax nonfiling gap was $4 billion. We have estimated the individual income tax nonfiling gap and the self-employment tax nonfiling gap together (since self-employment tax is reported and reconciled on the Form 1040 individual income tax return), but we report them separately herein. Our estimates of the not-filer portion of the gap are based on two different methodologies—one based primarily on IRS data (the Administrative Data Method), and the other based primarily on Census data matched with a limited amount of IRS data (the Census Method).

The Administrative Data Method is based on IRS administrative data (providing income from third parties paid to individuals who were not accounted for on filed returns), but uses aggregate information from Census data about the grouping of individuals into families and tax units. The Census Method is based on annual Census demographic surveys that have been linked to the limited IRS administrative data that the Census Bureau receives by law and regulation. Both methods have strengths and weaknesses, so it is helpful to derive estimates of the gap (and related tax return amounts) both ways. This paper provides details about these estimates and the methodologies used to produce them.

Section 1 below explains the two methods used to estimate the gap associated with not-filers; Section 2 explains the steps for estimating the gap associated with late filers; Section 3 summarizes the resulting nonfiling gap estimates; Section 4 compares the two sets of estimates introduced in Section 1, accounting for late filers; and Section 5 examines some of the distributional characteristics of nonfilers.

## 1. Not-Filers

Since not-filers do not file income tax returns, this is the most difficult portion of the nonfiling gap to estimate. Several methods have been employed in the past to estimate this portion of the tax gap. For example, in the early 1990s, the IRS estimated the nonfiling gap using a special study of Tax Year 1988 nonfilers under the Taxpayer Compliance Measurement Program (TCMP). This study selected a random sample of nonfilers, and attempted to contact them and secure delinquent returns from them when possible; those secured returns were then subjected to line-by-line examinations to determine the true tax liability.[1] This approach is not only very costly but still requires estimating the gap associated with any not-filers for whom the IRS could not secure a delinquent return.

For the Tax Year 2001 tax gap estimates, the IRS turned to a different method: an "Exact Match" between Census and IRS data. This approach involved identifying respondents in the annual Current Population Survey who appeared not to have filed an income tax return, then estimating their income tax liability. This approach is much simpler, but yields uncertain results because the Census data do not capture all income, it is not always possible to determine whether a Census respondent filed on time, and there was not a good method available to estimate the extent to which nonfilers paid at least some of their tax liability on time.

To estimate the Tax Year 2006 tax gap associated with not-filers, the IRS assembled a sample of individuals not appearing on filed tax returns, identified the income reported to the IRS for them by third parties, grouped them into family (tax) units (guided by Census data), imputed some additional income, deductions, and credits to them, then estimated their tax liability less credits and withholding. However, this approach (which we call the Administrative Sample Method) lacked information on income not reported to the IRS by third parties, and starting with a sample of individuals created challenges for grouping people together into presumed tax units.[2]

The current estimates are based on improvements to the last two of those three methods. We call these the Administrative Data Method and the Census Method. We estimate the not-filer portion of the gap using each of these two distinct methods, then add in the appropriate amounts for late filers, and compare the estimates. We average our estimates over Tax Years 2008 through 2010 to correspond with the individual income tax underreporting gap estimates provided in the combined tax gap report.

### *1.1 Administrative Data Method*

The main improvement we made to the methodology used for Tax Year 2006 was to apply the approach to population data, rather than to a sample, thus avoiding the disadvantages inherent in the sample approach. The method involved 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 that year;[3]
- Exclude (as potential primary taxpayers or spouses) those for whom no third-party information return was filed for the year in question;[4]
- 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 not-filer records;
- Assign a marital status (and filing status and spouse in the case of married-joint returns) and a number of dependents to the remaining not-filers using their age and gender, so that the total 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 Census data;
- Impute adjustments, deductions, and credits to the tax units;
- Compute the tax liability of the not-filers; and

[1] See Internal Revenue Service (1996) and Erard and Ho (2001).

[2] See Internal Revenue Service (2012).

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

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

- Deduct allowable credits and all prepayments made by these taxpayers, as reported on third-party information documents or found in IRS transactional history data, to derive their contribution to the tax gap.

The methods we developed for imputing self-employment income, deductions, and credits to the administrative dataset, and our methods for grouping individuals into tax units, are described in greater detail below.

### 1.1.1 Imputing Income

The basic method described above takes into account nonemployee compensation from Forms 1099 *Miscellaneous,* but this is only a small share of self-employment income reported on tax returns. Since other self-employment income is not reported to the IRS on third-party information documents, we used regression models to impute some of this income to not-filers based on the characteristics of taxpayers who reported self-employment income on filed returns. We developed the imputation model using a large sample of individuals from the Social Security master file and then matching these individuals to tax returns and information returns. Our imputations were restricted to net sole proprietorship income reported on Schedule C, and followed the method we employed for imputing self-employment income to Census records. Thus, we estimated the likelihood that a not-filer has self-employment earnings falling into one of the following three categories: (a) negative net self-employment earnings; (b) net self-employment earnings between $1 and $433; and (c) net self-employment earnings in excess of $433 (since taxpayers with more than $433 in net self-employment earnings are required to file a tax return).

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

$$SE^{*} = \gamma' x + \mu \tag{1}$$

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 $\gamma$ is a vector of coefficients to be estimated. The explanatory variables include gender, the log of age, region, indicators for the presence of key income types (wages, interest, Social Security, pension, unemployment compensation, and nonemployee compensation), and the log of each of these income amounts. The error term $\mu$ 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. Each individual was assigned a random number from a uniform distribution, and, if the value of this number was below the predicted probability, then the person was determined to have some net self-employment income.

Our second model was 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 the probit model, $\delta$ 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.[5] The indicator $I_{SE}$ for the net self-employment earnings category is assigned as follows:

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

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

Our third model is a regression specification for the magnitude of net self-employment earnings when they exceed $433. 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, $\beta'$ is a vector of coefficients to be estimated, and $\varepsilon$ is assumed to be a normal random error term with mean zero and standard deviation $\sigma$. Under this specification, the distribution of self-employment earnings is assumed to be log normal. Furthermore, we have imposed the constraint that the imputed self-employment income amount cannot exceed the amount corresponding to the 99.99 percentile of self-employment income on filed returns.

There are other income line items that are incomplete or missing from our income calculations for not-filers, such as capital gains, rental income, etc. However, these are either absent from information returns or it is difficult to estimate the net income that should be reported if a tax return were required.

### *1.1.2 Forming Tax Units*

The next task was to allocate these individual not-filers into tax units (families), which we did using the overall demographic profile of the CPS-ASEC. Specifically, this took the following steps:

1. Tabulate the population counts of persons in the CPS-ASEC in cells defined by gender, age group, marital status, and the number of dependents;
2. Tabulate the corresponding counts of persons present on filed returns for each cell (from IRS administrative data);
3. Subtract the IRS counts from the Census counts in each cell. This generates the target count of not-filers in each cell;[6]
4. Randomly allocate the not-filer individuals in each age group to the cells identifying married vs. single people and the number of dependents (0, 1, or 2 dependents). Persons assigned to a married status were randomly matched to other persons assigned to a married status. Persons assigned a dependent were randomly assigned characteristics of dependents from the CPS-ASEC population; and
5. Add the individual income and prepayment data among spouses to derive total amounts for each tax unit.

### *1.1.3 Estimating Elective Benefits and Prepayments*

The estimation of tax liability for each tax unit involved calculating and/or imputing adjustments, exemptions, deductions, refundable credits, and nonrefundable credits.

**Statutory Adjustments**

The adjustment for self-employment was calculated as one half of the self-employment tax associated with the imputed self-employment income amount for the taxpayer(s) on the return. Then the rest of the adjustments were imputed based on formulas estimated from the National Research Program (NRP) sample of returns for the specific tax year.

Once again, a probit model, with the same form and assumptions as for self-employment income, was used to predict the likelihood that a given return would have some positive amount of total adjustments (other than the self-employment tax deduction). For this model, the explanatory variables included the age category and gender of the primary taxpayer, filing status (restricted to married-joint or single), region, number of dependents, indicators for the presence of various types of income (wages, interest, dividends, business, farm, Social Security, pension and IRA distributions, and unemployment compensation) and the log of these same

[6] In practice, it was necessary to collapse the age and dependent categories for the CPS-ASEC counts in order to avoid having too few observations. Also, no not-filer Head of Household returns resulted from this procedure.

income amounts. In all cases, the values after correction during the NRP audit were the ones used in the estimation.

This model provided a prediction equation for the unconditional likelihood that an individual could have claimed some adjustments. Once again, each tax return was assigned a random number from a uniform distribution and, if the value of this number was below the predicted probability, then the person was determined to have been able to claim some amount of adjustments.

For those returns predicted to have adjustments, we predicted the amount by regressing the log of total adjustments on the same independent variables as in the probit model.

### Deductions

The standard deduction was calculated for all returns based on whether their filing status was single or married-joint, whether either or both taxpayers were 65 years of age or older, and whether either or both taxpayers were or could have been claimed as a dependent on a filed return.[7]

In addition to the standard deduction, we once again used the NRP sample for the same year to estimate a probit model to choose which of the not-filers would be most likely to file a Schedule A to itemize deductions. We used the same independent variables in this probit specification as in the case of adjustments. We then used the predicted probability for each return from this estimated equation and a random number assignment to select those who might choose to itemize. Then we used a regression model to predict the magnitude of itemized deductions for those who were predicted to itemize. Like the specification for adjustments, this model included the variables age category, gender, filing status, region, number of dependents, and indicators for the presence of the different income types, but in this case the log of total income is used rather than the log of each income type separately. The larger of the standard deduction amount and the itemized deduction amount is the one used in the tax calculation.

### Exemptions and Credits

The total exemption amount was calculated in a straightforward manner based on the number of taxpayers on the return, whether either was claimed on another return or under 21 years of age, and the number of dependents.

Nonrefundable credits were calculated in two parts: (1) the child tax credit; and (2) all other credits. The child tax credit was calculated following the worksheets on the Form 1040 instructions using the information available in the not-filer file on the number of dependents of eligible age, filing status, adjusted gross income and earned income. All other types of credits were imputed based on probit and OLS regression equations estimated from the NRP sample for the same year.

Once again the imputation followed a two-step process. First, the incidence of tax returns claiming these other types of credits was estimated using a probit model. Second, for those predicted to claim these credits, a regression model was used to develop prediction equations for the amount of the credits that could be claimed. The same independent variables were used in both of these models as in the probit model for itemized deductions.

The Earned Income Credit and Additional Child Tax Credit were also calculated from the worksheets on the Form 1040 instructions using the information on the not-filer file for the age and number of dependents, filing status, and income. These estimated amounts were used solely to offset unpaid tax balances. Since the portions of these credits that could have been refunded have no bearing on the tax gap, they were ignored.

[7] An indicator for being claimed as a dependent was created for those who actually were claimed on a filed return in the tax year in question or who were 21 or under.

**Prepayments**

Withholding amounts were summed for each taxpayer from all of the withholding reported in third-party information documents and then combined to a return-level aggregate. Any estimated payments reported in the IRS administrative transaction history file are also added to the prepayments made by the tax unit.

### *1.1.4 Calculating the Tax Gap*

Based on these imputed and calculated amounts of income, credits, and deductions, we estimated the total balance due for each tax unit. Our estimate of the not-filer tax gap using this IRS administrative data approach is the sum of all non-zero balances due. For this initial component, our final estimates are based on the average of five replicates, each of which used different family unit files and imputation processes for sole proprietor income and for adjustments, deductions, and credits, including a new set of random number draws to determine incidence. In order to control for the presence of unusually large values for income amounts on a small number of information documents in each tax year,[8] two steps are taken in the calculation of the estimate for each of the five replicates. First, observations are removed where total income exceeds the 99.99 percentile of the amount of total income on filed returns for the given tax year. Second, each estimate is calculated by drawing 25 one-percent samples of the population of not-filers having a tax liability, and sorting these by the estimated balance due. Then the average of the middle seven of these samples (weighted by a factor of 100) is the basis for the estimates.

The resulting average estimates for not-filers for Tax Years 2008 through 2010 are summarized in Table 1, with their contribution to the combined income tax and self-employment tax gap amounting to just under $18 billion. This estimate is not completely comparable to the estimate from the Census Method since the matched Census-IRS data include late filers who file by December 31 of the normal filing year. Once we account for late filers (see Section 2), we will be able to compare the two sets of estimates.

**TABLE 1. Administrative Data Method Estimates of the Not-Filer Gap ($ in Billions), Tax Years 2008–2010†**

| Key Items | Amount |
|---|---|
| Total income | $216.3 |
| Total adjustments that offset income* | $7.1 |
| Total personal and dependent exemptions that offset income* | $27.3 |
| Total deductions that offset income* | $58.5 |
| Total taxable income | $123.2 |
| Tentative tax | $20.1 |
| Tax offset by nonrefundable credits* | $0.4 |
| Self-employment tax | $6.9 |
| Net tax due | $26.6 |
| Tax offset by prepayments* | $8.5 |
| Tax offset by refundable credits* | $0.5 |
| Total payments of tax | $9.0 |
| Total contribution to the nonfiling gap | $17.6 |

† Estimates averaged over Tax Years 2008 through 2010.

* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

## 1.2 Census Method

We improved on the old "Exact Match" method in several important ways:

- Census has improved their ability to assign an anonymous Protected Identification Key (PIK) to most respondents in the Current Population Survey Annual Social and Economic Supplement (CPS-ASEC) and to all of the records Census receives from the IRS for the population (including selected data from income tax returns and from third-party information documents). This allows them to create a better

[8] This is often the case in raw population data due both to data errors and genuine outliers.

matched dataset, allowing us to identify not-filers more accurately.[9] However, there are some ASEC records that could not 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 of not-filers.

- We used the third-party information associated with the not-filers, together with demographic information about them contained in the ASEC, to make better imputations of certain income, deduction, and credit amounts.
- We estimated the tax liability of the not-filers using a more detailed tax calculator than had been used in prior Exact Match studies.
- We estimated the aggregate amount of withholding and other prepayments made by the not-filers in the matched dataset based on rates of withholding derived from tabulations of late filers and not-filers identified in the Administrative Data Method.
- We supplemented this estimate of the gap associated with not-filers with a separate estimate for late filers using IRS administrative data (see Section 3).

The Census Method therefore involves a five-step process to estimate the not-filer portion of the nonfiling gap: (1) impute income that either does not exist in the CPS data or was grossly underreported; (2) create tax units according to CPS household relationships; (3) re-weight the CPS data in order to account for survey respondents who could not be assigned a unique PIK identifier and for records whose income was completely imputed; and (4) impute tax return level line items not observed in the CPS; and (5) calculate tax and balance due. These five steps are described in greater detail below.

### *1.2.1 Imputing Income at the Individual Level*

**Impute Retirement, Pension, and Social Security Income**

Individuals in the matched dataset who had a Form 1099-R but did not report any Form 1099-R income were assigned the gross distribution amount from Form 1099-R. In order to determine what portion of the gross distribution is taxable, a series of models were estimated on IRS data. The following models were estimated and applied in the following order: (a) an incidence model to determine if any of the distribution is taxable; (b) an incidence model to determine if all of the distribution is taxable; and (c) a regression model to determine the taxable portion of the distribution.

Additional Pension and Social Security income were imputed using a set of models developed on IRS data. The models were estimated and applied in the following order: (a) a multinodal model to determine if an individual had no retirement income, only pension income, only Social Security income, or pension and Social Security income; and (b) regression models to impute the amount of pension and/or Social Security income given the outcome of the multinodal model. IRS data were used to determine the total number of individuals who should have reported receiving pension and Social Security income. The multinodal model probabilities were adjusted so that the CPS totals were proportional to the IRS totals in application. Variation was applied to the regression estimates using the mean squared error and a random normal draw. In order to ensure the imputations stayed within a realistic range, a cap (upper limit) was established using observed IRS data. Any imputation that exceeded the cap was re-imputed.

**Impute Schedule C Income**

Schedule C (nonfarm sole proprietor) income was imputed using a set of models developed on IRS data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had Schedule C income; (b) a multinodal model to determine if Schedule C income was negative, between $1 and $433, or greater than $433; and (c) regression models to impute the amount of Schedule C income given the outcome of the multinodal model. IRS data were used to determine the total number of individuals that

[9] See Jones and O'Hara (2014), and Wagner and Layne (2012).

should have reported receiving Schedule C income. The incidence and multi-nodal model probabilities were adjusted so that the CPS totals were proportional to the IRS totals in application. Variation was applied to the regression estimates using the mean squared error and a random normal draw. In order to ensure the imputations stayed within a realistic range, a cap was established using IRS data. Any imputation that exceeded the cap was re-imputed.

#### Impute Additional Self-Employment (SE) Income

Additional SE income is the net SE income remaining after accounting for net Schedule C income (described above) and Schedule F income (from the Census data).[10] Additional SE income was imputed using two models developed on IRS data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had additional SE income; and (b) a regression model to impute the amount. Variation was applied to the regression estimate using the mean squared error and a random normal draw. In order to ensure the imputations stayed within a realistic range, a cap was established using IRS data. Any imputation that exceeded the cap was re-imputed.

#### Impute Unemployment Compensation

Unemployment compensation was imputed using two models developed on IRS data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had unemployment compensation; and (b) a regression model to impute the amount. IRS data were used to determine the total number of individuals that should have reported receiving unemployment compensation. The incidence model probabilities were adjusted so that the CPS total matched the IRS total in application. Variation was applied to the regression estimate using the mean squared error and a random normal draw. In order to ensure that the imputations stayed within a realistic range, a cap was established using IRS data. Any imputation that exceeded the cap was re-imputed.

#### Impute Dividends and Qualified Dividends

Total dividend income was imputed using two models developed on IRS data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had dividend income; and (b) a regression model to impute the amount. IRS data were used to determine the total number of individuals that should have reported receiving dividend income. The incidence model probabilities were adjusted so that the CPS total matched the IRS total in application. Variation was applied to the regression estimate using the mean squared error and a random normal draw. In order to ensure that the imputations stayed within a realistic range, a cap was established using IRS data. Any imputation that exceeded the cap was re-imputed.

The portion of total dividends that are qualified was imputed using three models developed on IRS data. These models were estimated and applied in the following order: (a) an incidence model to determine if any portion of the dividend is qualified; (b) an incidence model to determine if the entire dividend is qualified; and (c) a regression model to determine the qualified portion of the dividend.

### *1.2.2 Forming Tax Units*

The following steps were used to create tax units using CPS data:

**Step 1:** Combine married individuals into one record.

**Step 2:** Assign dependents to taxpayers.

Check 1—Individuals under age 19 (or age 24 and a full-time student) who are children of someone in the household are assigned as dependents of their parents.

Check 2—Remaining individuals under age 19 (or age 24 and a full-time student) who are related to an eligible person in the household are randomly assigned as a dependent to an eligible related person.[11]

[10] So, the additional SE income reflects any income reported on: (1) Schedule K-1 (Form 1065), box 14, code A (other than farming); and (2) Schedule K-1 (Form 1065-B), box 9, code J1.

[11] Note: An eligible person is someone with income who is not claimed as a dependent.

Check 3—All remaining individuals under the age of 19 are randomly assigned to an eligible person in the household.

Check 4—All remaining individuals with no income, but who have an eligible relative in the household, are randomly assigned to an eligible relative.

### *1.2.3 Re-weighting the CPS Data*

Roughly 10 percent of the CPS records cannot be matched to a PIK and/or contain income amounts that were all imputed. The records without a PIK or with completely imputed income were dropped from the sample. The remaining records were re-weighted at the strata level to account for the dropped records. The strata weights were re-weighted using a multiplicative adjustment factor that makes the sum of the adjusted strata weights equal the sum of the original strata weights.

### *1.2.4 Imputing Offsets at the Tax Return Level*

**Adjustments Other Than the Self-Employment Tax Deductions**

Our tax calculator derived the adjustment for one-half of the self-employment tax paid. The sum of other adjustments was imputed using two models developed on the NRP sample of returns for the relevant tax year. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had other adjustments; and (b) a regression model to impute the amount. Variation was applied to the regression estimate using the mean squared error and a random normal draw.

**Itemized Deductions**

Itemized deductions were imputed using two models developed on NRP data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual itemized; and (b) a regression model to impute the amount. Variation was applied to the regression estimate using the mean squared error and a random normal draw. Only tax units who would have had taxable income after using the standard deduction were eligible for the itemizer imputation.

**NonRefundable Credits Other Than Child Tax Credit**

Our tax calculator derived the Child Tax Credit applicable to each return. Other nonrefundable credits were imputed using two models developed on NRP data. The models were estimated and applied in the following order: (a) an incidence model to determine if an individual had other nonrefundable credits; and (b) a regression model to impute the amount. Variation was applied to the regression estimate using the mean squared error and a random normal draw. Only tax units who had tentative tax were eligible for the other nonrefundable credit imputation.

### *1.2.5 Calculating Tax and Balance Due*

Based on these imputed and calculated amounts of income, credits, and deductions, we estimated the net tax due for each tax unit, and subtracted from this any prepayments (such as withholding) and estimated refundable credits.

**Prepayments and Refundable Credits**

We estimated prepayments and refundable credits in the aggregate using the ratio of each aggregate amount to the aggregate total amount of tax among the population of later late filers and the not-filers in the Administrative Data Method. This population corresponds to the population of not-filers in the matched Census-IRS dataset. This is especially important because IRS administrative data indicate that the ratio of these amounts to tax liability appears to be determined jointly with filing behavior[12] and because we have no way of knowing which

[12] For example, the ratio of prepayments is much larger among early late filers than it is among the later late filers, and it is much larger among the later late filers than it is among not-filers.

not-filer "returns" in the matched Census-IRS dataset were actually filed after December 31 of the primary filing year.

### Contribution to the Tax Gap

The resulting average estimates for not-filers for Tax Years 2008 through 2010 are summarized in Table 2, with their contribution to the combined income tax and self-employment tax gap amounting to just over $26 billion.

**TABLE 2. Census Method Estimates of the Not-Filer Gap ($ in Billions), Tax Years 2008–2010†**

| Key Items | Amount |
|---|---|
| Total income | $421.4 |
| Total adjustments that offset income* | $9.8 |
| Total personal and dependent exemptions that offset income* | $45.2 |
| Total deductions that offset income* | $88.5 |
| Total taxable income | $278.0 |
| Tentative tax | $48.7 |
| Tax offset by nonrefundable credits* | $2.1 |
| Self-employment tax | $8.6 |
| Net tax due | $55.2 |
| Tax offset by prepayments* | $26.2 |
| Tax offset by refundable credits* | $2.6 |
| Total payments of tax | $28.8 |
| Total contribution to the nonfiling gap | $26.4 |

† Estimates averaged over Tax Years 2008 through 2010.

* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

## 1.3 Imputing Tax Benefits to Not-Filers

The tax gap is a fairly simple concept to understand, but it's not so simple to define operationally. The complication arises from the fact that the Internal Revenue Code doesn't define "true tax liability." One reason for this may be that the tax liability imposed by the Code depends on taxpayer choices; taxpayers are *required* by the law to report their income, but they are *not required* to claim all the tax benefits[13] for which they are eligible.[14] Perhaps the simplest example of this is the election to claim the standard deduction in lieu of itemizing. Some taxpayers avoid itemizing for rational reasons (e.g., to avoid the recordkeeping and filing burden, or to avoid disclosing information to the government or to an estranged spouse)—even if they could lower their tax liability by itemizing. So, what is "true tax liability" in that case? Should we estimate unclaimed itemized deductions (and the corresponding reduction in tax) when estimating the tax gap? We have never done so in the past (except to the extent that the TCMP or NRP data include cases in which some taxpayers become itemizers as a result of the audit). Other taxpayers undoubtedly choose not to claim other benefits (such as credits) because of the effort (and/or out-of-pocket cost) it would take to determine how much, if any, they could legitimately claim, or due to fear that claiming such items would subject them to a higher probability of audit.

This is an even bigger issue when estimating the nonfiling gap, since not-filers—by definition—didn't claim *any* tax benefits. Should we define their "true tax liability" by assigning them assumed offsets to income that they haven't claimed? Should we determine their timely payments as including credits that they haven't claimed?

The nonfiling gap estimates presented in Tables 1 and 2 reflect the imputation of straightforward exemptions, adjustments, deductions, and credits. These imputations are based on the auditor-corrected amounts

[13] We use the term "tax benefits" in a fairly broad sense to include all opportunities allowed by the Tax Code to reduce one's taxable income or tax. Perhaps we can think of them as tax expenditures that are subject to taxpayer choice.

[14] It uses terms like "may claim" and "allowed."

among the random sample of filers included in the National Research Program for the same year.[15] However, if we didn't impute *any* benefits that involve a choice on the part of the taxpayer, we estimate that the gap would be on the order of $5 billion (which is 17 percent) larger.[16]

## 2. Late Filers

In addition to not-filers, who don't file a tax return at all, late filers also make a significant contribution to the nonfiling gap since they have a lot of unpaid tax, but did not meet the filing deadline. As in the case of not-filers, however, they do not contribute to the tax gap to the extent that they pay their tax liability on time, such as through withholding and tax credits. Unlike not-filers, of course, we have tax returns for the late filers, so estimating their contribution to the gap is much more straightforward. On the surface, the gap is their aggregate balance due. However, we adjust this amount to take into account income and payments that are not reported on the late returns, but are reported to the IRS on third-party information documents.[17] All of the data needed to estimate the nonfiling gap due to late filers is present in IRS administrative data, and we estimate it from multiple large samples drawn from population data (to mitigate the effects of data errors). Our estimates are provided in Table 3. The reason there are separate estimates for the Census Method and the Administrative Data Method is that the matched Census-IRS data include late returns filed by December 31 of the ordinary filing year, but they do not allow us to identify which returns were filed on time. We could have estimated this, but it's easier to treat all filed returns in the matched data as timely. However, that means that the *later* late filers (those who file after the IRS extract provided to Census[18]) appear as "not-filers" in the matched dataset, causing us to overstate the true not-filer portion of the gap. To avoid double-counting, we need to add only the *early* late filers to the Census Method estimate of not-filers. So, the total nonfiling gap is still the sum of the not-filer and late filer portions. See Figure 1. As shown in Table 3, the gap due to early late filers is less than half of the full amount associated with late filers. The difference between the two sets of estimates is due to those who file *after* December 31, and most of the dollars on those are identified through enforcement.

**TABLE 3. Tax Gap Due to Late Filers ($ in Billions), Tax Years 2008–2010§**

| Key Items | Census Method (Early Late Filers) | Administrative Data Method (All Late Filers) |
|---|---|---|
| Total income | $183.2 | $375.4 |
| Total adjustments that offset income* | $3.3 | $6.5 |
| Total exemptions that offset income* | $17.9 | $40.0 |
| Total deductions that offset income* | $42.1 | $86.8 |
| Total taxable income | $119.9 | $242.2 |
| Tentative tax | $25.3 | $48.8 |
| Tax offset by nonrefundable credits* | $1.4 | $2.6 |
| Self-employment tax | $1.5 | $4.0 |
| Net tax due † | $25.6 | $50.7 |
| Tax offset by prepayments* | $18.9 | $35.1 |
| Tax offset by refundable credits* | $2.3 | $4.3 |
| Total payments of tax | $21.3 | $39.4 |
| Total contribution to the nonfiling gap | $4.4 | $11.3 |

§ As of November 2015; estimates averaged over Tax Years 2008 through 2010.

* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

† Includes other miscellaneous taxes reported.

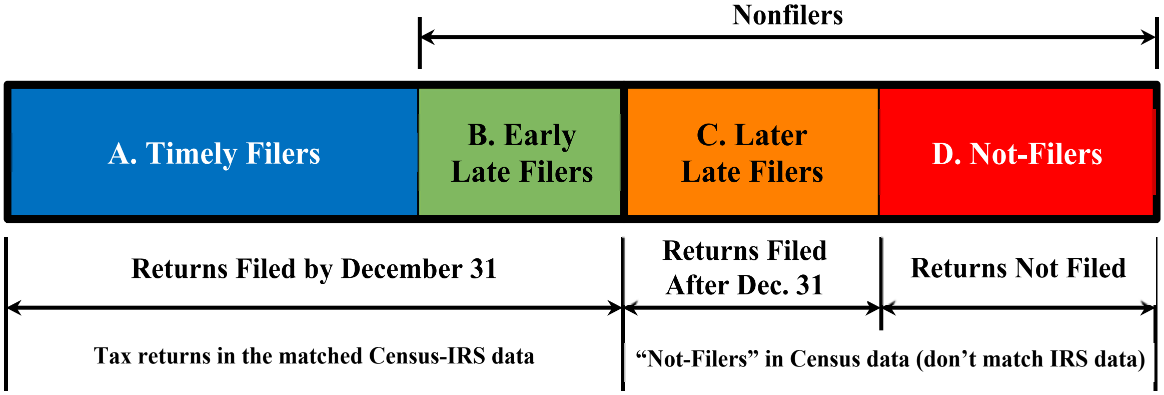
[15] Since many taxpayers overstate the amount of elective benefits to which they are entitled, basing these imputations on the amounts claimed by filers would project to not-filers the same degree of overstatements. The NRP audits correct both overstatements and understatements, but the overall net effect is to reduce most elective benefits.

[16] This assumes that "true tax liability" for not-filers is always based only on the standard deduction, personal (not dependent) exemptions, and the adjustment for self-employment tax (but no other statutory adjustments), and that payments against that liability exclude credits since none were claimed. It also makes the same filing status assumptions reflected in Tables 1 and 2 since the Tax Code determines filing status based on a person's characteristics rather than his choice of status.

[17] We do not impute other kinds of income to them (such as from self-employment). However, late filers already report a significant amount of these kinds of income.

[18] That extract is created during the last posting cycle of the calendar year for filing, generally around December 31.

**FIGURE 1. The Role of Late Filers in the Census Method**


Nonfilers
A. Timely Filers
B. Early Late Filers
C. Later Late Filers
D. Not-Filers
Returns Filed by December 31
Returns Filed After Dec. 31
Returns Not Filed
Tax returns in the matched Census-IRS data
"Not-Filers" in Census data (don't match IRS data)

Since the later late filers (C) are already accounted for among the Census "not-filers," only the early late filers (B) need to be added to the Census "not-filers" to estimate the nonfiling gap.

Since we observe fairly complete information about late filers, we can analyze how their portion of the nonfiling gap is distributed across various income categories as well as by the extent of lateness. As documented in Table 4, in Tax Years 2008 through 2010, low-income late filers (those with total incomes of less than $25,000) account for $0.7 billion (or 5.8 percent) of the late filer portion of the nonfiler gap. In contrast, high-income late filers (those with total incomes of $200,000 or more) account for $4.8 billion (or 42.3 percent) of the late filer portion of the nonfiler gap. Table 4 also shows that 49 percent of the late filer gap among lower-to-middle-income taxpayers (those with less than $100,000 of total income) is associated with returns that are filed over 12 months late, while for high-income late filers (those with $100,000 or more of total income), this is just 28 percent. Indeed, a large share of the late filing gap among high-income late filers is associated with returns that are filed less than 3 months late.

While it is true that we observe substantially complete information about late filers, this is true only after sufficient time has passed for most of the late filers to file. Some of them file in response to IRS enforcement actions, which can take place years after the filing deadline has passed. This means that we have less complete information about late filers for more recent tax years than we have for much older tax years. As Table 4 indicates, 36 percent of the late filing gap is associated with returns that are filed over 12 months late. This lag means that any estimate of the nonfiling gap based solely on data compiled shortly after the close of the filing year will need to include an estimate of the gap due to later filers, while much later estimates can take advantage of their actual late returns.

**TABLE 4. Average Nonfiling Gap and Share of Returns Among Late Filers by Total Income and Extent of Lateness, Tax Years 2008 to 2010§**

| Total Income | Extent of Lateness (months)* | | | | | Row Total | Share of Total |
|---|---|---|---|---|---|---|---|
| | < 3 | 3 to < 6 | 6 to < 9 | 9 to < 12 | 12 + | | |
| ***Contribution to the Nonfiling Gap ($B)*** | | | | | | | |
| < $25,000 | 0.09 | 0.08 | 0.07 | 0.07 | 0.35 | 0.66 | 5.8% |
| $25,000 to < $50,000 | 0.17 | 0.17 | 0.15 | 0.14 | 0.67 | 1.30 | 11.5% |
| $50,000 to < $100,000 | 0.34 | 0.32 | 0.29 | 0.25 | 1.02 | 2.23 | 19.7% |
| $100,000 to < $200,000 | 0.43 | 0.38 | 0.35 | 0.27 | 0.91 | 2.34 | 20.7% |
| $200,000 or more | 1.69 | 0.84 | 0.72 | 0.45 | 1.09 | 4.79 | 42.3% |
| Column Total | 2.72 | 1.78 | 1.59 | 1.17 | 4.05 | 11.31 | 100.0% |
| ***Share of Row Total*** | | | | | | | |
| < $25,000 | 13.0% | 12.5% | 10.7% | 10.0% | 53.8% | 100.0% | Heavily 12+ months |
| $25,000 to < $50,000 | 13.5% | 12.9% | 11.4% | 10.5% | 51.7% | 100.0% | |
| $50,000 to < $100,000 | 15.4% | 14.4% | 13.2% | 11.2% | 45.8% | 100.0% | |
| $100,000 to < $200,000 | 18.3% | 16.1% | 15.0% | 11.5% | 39.1% | 100.0% | |
| $200,000 or more | 35.3% | 17.5% | 15.1% | 9.4% | 22.7% | 100.0% | Earlier |
| Column Total | 24.1% | 15.8% | 14.0% | 10.4% | 35.8% | 100.0% | |
| ***Average Contribution to the Gap Per Late Return ($)*** | | | | | | | |
| < $25,000 | 683 | 800 | 863 | 799 | 839 | 808 | |
| $25,000 to < $50,000 | 1,604 | 1,932 | 2,109 | 1,896 | 1,984 | 1,920 | |
| $50,000 to < $100,000 | 3,213 | 3,806 | 4,005 | 3,893 | 4,065 | 3,844 | |
| $100,000 to < $200,000 | 7,390 | 8,621 | 8,664 | 8,492 | 9,579 | 8,687 | |
| $200,000 or more | 59,430 | 43,311 | 41,556 | 39,727 | 42,621 | 46,941 | |
| Column Total | 6,374 | 5,302 | 5,604 | 4,477 | 3,576 | 4,635 | |

§ As of November 2015. Cells with bold borders are the largest ones in each row.

## *2.1 Method for Incorporating Third-Party Information for Late Filers*

Like filers, some late filers do not report amounts consistent with the information reported on their behalf by third parties. We accounted for this for each late filer using the logic summarized in Table 5 for each line item on the return.

After accounting for additional income using the logic presented in Table 5, we recalculated tax and the balance due for each return. We assumed that the total of all withholding for a given taxpayer that was documented by third parties on information returns was not more accurate than the amount reported by the taxpayer on his or her Form 1040.

## *2.2 Method for Handling Outliers in Population Data*

A sampling method was applied to the late filer estimates in order to minimize the impact of administrative transcription errors and other outlier data issues that exist in the raw administrative data. The sampling method consisted of tabulating results for 100 to 125 one percent samples. The samples were ordered by aggregate balance due, and the middle 10 were selected and averaged to create our final estimates.

**TABLE 5. 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]<br>• If A>0 and (B+C)>0 and GIC>0 and -150<(B+C+L-GIC)<150, then:<br>Wages = (B+C) and<br>Schedule C net income = Max[K-(B+C), 0]<br>• Else, if A>0 and (B+C)>0 and GIC=0 and -150<(A-(B+C))<150, then:<br>Wages = Max[A-L, (B+C), 0] and<br>Schedule C net income = L<br>• Else:<br>Wages = Max[A, (B+C), 0] and<br>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 | 1099 MISC | 7 | Non-employee compensation | |
| 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]<br>(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)<br>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.<br>If AK=0, then AK=AL<br>If AM=0, then AM=AN<br>IRA + Pension income = Max[(AL+AN), (AO-AK+AL), (AP-AM+AN)]<br>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 amount | |
| 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. | |

## TABLE 5. Logic for Using Information Return Data To Adjust Items Reported on Late Returns—Continued

| | Form | Line | Item | Adjustment Logic |
|---|---|---|---|---|
| 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<br>If (AX<0 and Line21Calc=0) or (Schedule C net income ≠ 0) or (Farm income ≠ 0), then: Other income = AX;<br>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 | |
| BB | 1040 | 17 | Schedule E net income | GrossE = (BC+BD+Max[(BE+BF), BG]+BH+BJ+Max[BI, 0])<br>If BB > GrossE, then GrossE = BB<br><br>Note: Any negative amount from any of the following components is set to zero:<br>Line17Calc = BK+BL+BM+BN+BO+BP+BQ+BR+BS+BT+BU+BV+BW+BX+BY<br><br>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) | Non-passive inc. from partnership or S corp. | |
| BG | Schedule E | 30 | Passive + non-passive 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<br><br>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 | |

## 3. Nonfiling Gap Estimates

Our overall estimates of the individual income tax nonfiling gap, averaged over Tax Years 2008 through 2010 are provided in Table 6—adding the gap associated with late filers and not-filers. We average the estimates over the TY2008–2010 period to arrive at an estimate that is comparable to the underreporting gap estimates. We also average the estimates derived from the Census Method and the Administrative Data Method since each method has its own strengths and weaknesses. In particular, the Census Method benefits from much richer demographic microdata on family composition, but it is based on a sample of individuals and families, and has weaker information on incomes, which forced us to rely on imperfect imputations of many income types based on IRS data. In contrast, the Administrative Data Method takes advantage of population data and much stronger data on income from third-party information returns, but it relies on a very rough assignment of not-filers into tax units as primary taxpayers, spouses, and dependents, guided by Census tabulations. Since there are multiple—but different—approximations employed in the two methods, averaging the estimates produced by them may result in a better estimate than either of the two methods produces by itself.

**TABLE 6. Individual Income Tax and Self-Employment Tax Nonfiling Gap Estimates ($ in Billions)**

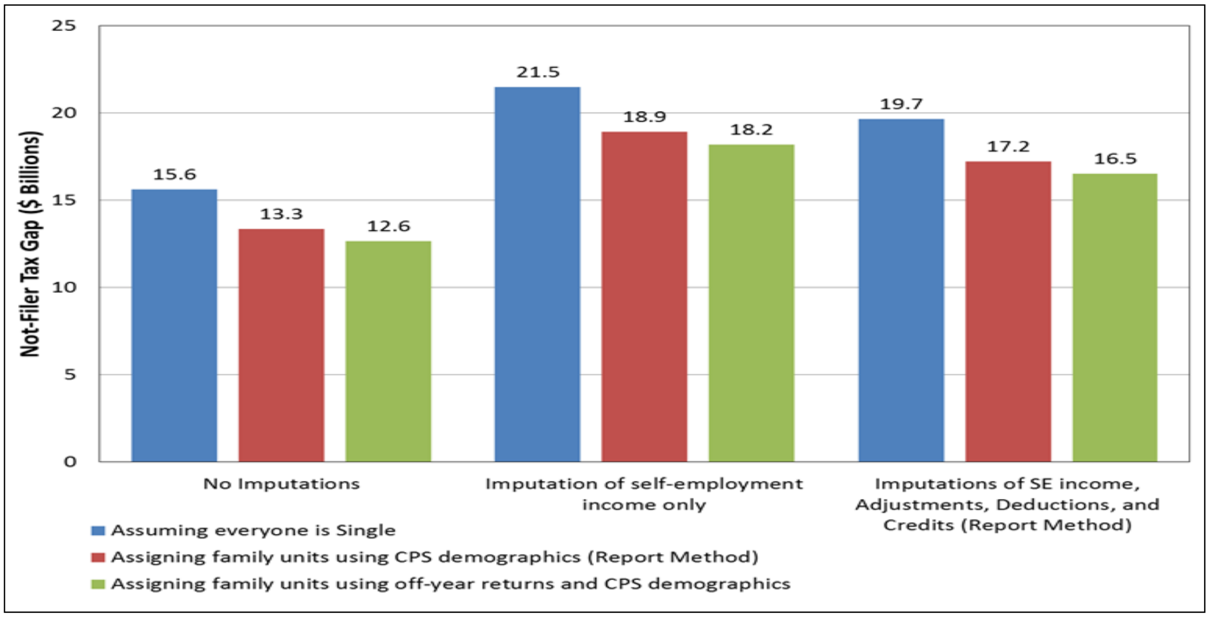
| | TY 2005 | Average 2008–2010 |
|---|---|---|
| Final Nonfiling Gap Estimate* | | 29.8 |
| Census Method | | 30.8 |
| Not-Filers | | 26.4 |
| Late Filers | | 4.4 |
| Administrative Data Method | 24.7 | 28.9 |
| Not-Filers | 13.6 | 17.6 |
| Late Filers | 11.1 | 11.3 |
| Administrative Sample Method | 25.0 | |
| Not-Filers | 13.0 | |
| Late Filers | 12.0 | |

* The portion of this attributable to self-employment tax is $3.8 billion, assuming that payments are allocated to income tax and self-employment tax proportional to the magnitude of tax liability.

Table 6 also compares our estimates for Tax Year 2005 using the Administrative Data Method (which is based on population microdata) vs. the Administrative Sample Method, which was the basis for the Tax Year 2006 nonfiling gap estimate. Although one might expect the Administrative Data Method to produce a larger estimate (since, in contrast with the Administrative Sample Method, it includes imputations of self-employment income), the estimate for late filers is somewhat smaller and the estimate for not-filers is only slightly larger than were estimated using the Administrative Sample Method. There are two primary reasons for this: (1) the Administrative Sample Method was based on one simple sample of individuals from the overall population, taking as correct even the outlier values in the information document data, while the Administrative Data Method is based on a multi-step process to mitigate against the records with implausibly high dollar amounts; and (2) the Administrative Data Method excludes individuals with overseas addresses (in part to get closer to the population represented in the Census data), while the Administrative Sample Method did not. The difference between not-filers and late filers is larger under the Administrative Data Method because we didn't impute self-employment income to late filers. In any case, the growth in the nonfiling gap, as estimated by both methods, appears to be due to growth in the population and the economy (and possibly changes in taxpayer behavior), but not to the change in methodology.

In Figure 2 we compare Administrative Data Method estimates of the not-filer portion of the nonfiler tax gap for Tax Year 2010 using three different sets of assumptions about family unit construction and three different levels of imputations. Assuming all not-filers are single results in larger estimates of this gap than the approach followed in making the tax gap estimate—a random allocation guided by CPS demographics—as well as an approach that uses prior year and subsequent year tax return information and CPS demographics to construct tax units. At each level of imputation, the more informed method for building family units leads to a slightly lower estimate. In addition, as would be expected, for each method of building tax units the largest tax gap estimates results when self-employment income is imputed but adjustments, deductions and credits are not. The lowest estimates occur when neither self-employment income nor adjustments, deductions, and credits are imputed.

**FIGURE 2. Not-Filer Portion of Administrative Data Method Nonfiler Gap Under Different Assumptions, Tax Year 2010**



## 4. Comparing the Census Method and the Administrative Data Method

It is helpful to have two methods to estimate the same concept. Each has its own strengths and weaknesses, which we discuss below. Our decision to average the two nonfiling gap estimates produced by these methods is admittedly a judgment call, recognizing the strengths and weaknesses of each. In the absence of an objective way to quantify (or weight) those strengths and weaknesses, we believe that averaging them is warranted—particularly since the estimates are quite close anyway.

Table 7 compares the two sets of estimates at the line-item level. The Administrative Data Method identifies slightly more nonfilers, but slightly less Total Income, Total Tax, and nonfiling gap among them. Wages in the Census are slightly greater than in the administrative data, which suggests that Census respondents may be characterizing certain other income types as wages. More dividends and pension income is imputed to the Census records than appears in the administrative data for the nonfilers. In the case of pensions, the inability to distinguish taxable IRAs from pensions in the Census data may contribute to the difference. The Census Method results in close to $13 billion more income than the Administrative Data Method, an amount that was very similar to the difference in wage income. We also estimate a much larger aggregate deduction amount in the Administrative Data Method, possibly reflecting the presence of standard deductions among its slightly larger population of nonfilers. Nonetheless, the Total Tax and nonfiling gap estimates are quite close between the two methods, illustrating the benefit of estimating the gap both ways.

**TABLE 7. Comparison of Estimates From the Two Methods: Late Filers and Not-Filers Combined ($ in Billions), Tax Years 2008–2010§**

| Key Items | Administrative Data Method | Census Method | $ Difference | % Difference |
|---|---|---|---|---|
| Number of required returns (millions) | 14.2 | 14.0 | 0.2 | 1.7% |
| Wages | $356.2 | $368.8 | -$12.6 | -3.5% |
| Interest | $11.5 | $13.2 | -$1.7 | -14.9% |
| Dividends | $10.9 | $29.8 | -$18.8 | -171.9% |
| Taxable refunds | $2.9 | $1.2 | $1.7 | 58.2% |
| Alimony received | $0.5 | $0.7 | -$0.2 | -54.9% |
| Schedule C net income | $70.2 | $60.7 | $9.5 | 13.5% |
| Form 4797 income | -$2.3 | -$0.6 | -$1.7 | 74.1% |
| Schedule D net income | $18.0 | $9.4 | $8.6 | 47.9% |
| Taxable IRA and pension income | $61.3 | $63.6 | -$2.3 | -3.7% |
| Schedule E net income | $21.3 | $18.8 | $2.4 | 11.4% |
| Schedule F net income | -$1.0 | $0.4 | -$1.5 | 142.2% |
| Unemployment compensation | $14.1 | $11.9 | $2.2 | 15.4% |
| Taxable SSI income | $12.7 | $12.2 | $0.5 | 4.2% |
| Other income | $3.9 | -$1.1 | $5.0 | 128.0% |
| Additional self-employment income ** |  | $4.9 | -$4.9 |  |
| Total income† | $591.8 | $604.7 | -$12.9 | -2.2% |
| Adjustments that offset income * | $13.6 | $13.1 | $0.5 | 3.7% |
| Deductions that offset income * | $145.5 | $130.6 | $15.0 | 10.3% |
| Exemptions that offset income * | $67.3 | $63.1 | $4.2 | 6.2% |
| Taxable income | $365.4 | $397.9 | -$32.5 | -8.9% |
| Tentative tax | $68.8 | $73.9 | -$5.1 | -7.4% |
| Tax offset by nonrefundable credits* | $3.0 | $3.4 | -$0.4 | -12.8% |
| Self-employment tax | $10.9 | $10.1 | $0.8 | 7.5% |
| **Net tax due** | **$77.3** | **$80.8** | **-$3.6** | **-4.6%** |
| Tax offset by prepayments* | $43.6 | $45.1 | -$1.5 | -3.4% |
| Tax offset by refundable credits* | $4.8 | $5.0 | -$0.2 | -4.0% |
| Total payments of tax* | $48.4 | $50.1 | -$1.7 | -3.5% |
| **Total nonfiling gap** | **$28.9** | **$30.8** | **-$1.9** | **-6.5%** |

§ Estimates averaged over Tax Years 2008 through 2010.

† The Total Income amount is slightly larger than the sum of the components because Total Income cannot be less than zero on any given return.

* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

** The net self-employment income remaining after accounting for net Schedule C and F income in the Census imputations.

## 5. Characteristics of Nonfilers

In this section we use Tax Year 2010 nonfiler data to examine the types of taxpayers who make up the nonfiler population. In this year, we estimate that not-filers accounted for about 51 percent of all nonfiler required returns and late filers accounted for 49 percent. Of those returns submitted late, about half were filed during the 2011 filing year, while the other half were filed after December 31, 2011. In addition, about 18.8 percent of all late required returns were submitted after the IRS sent a notice reminding the taxpayer of the requirement to file a tax return.

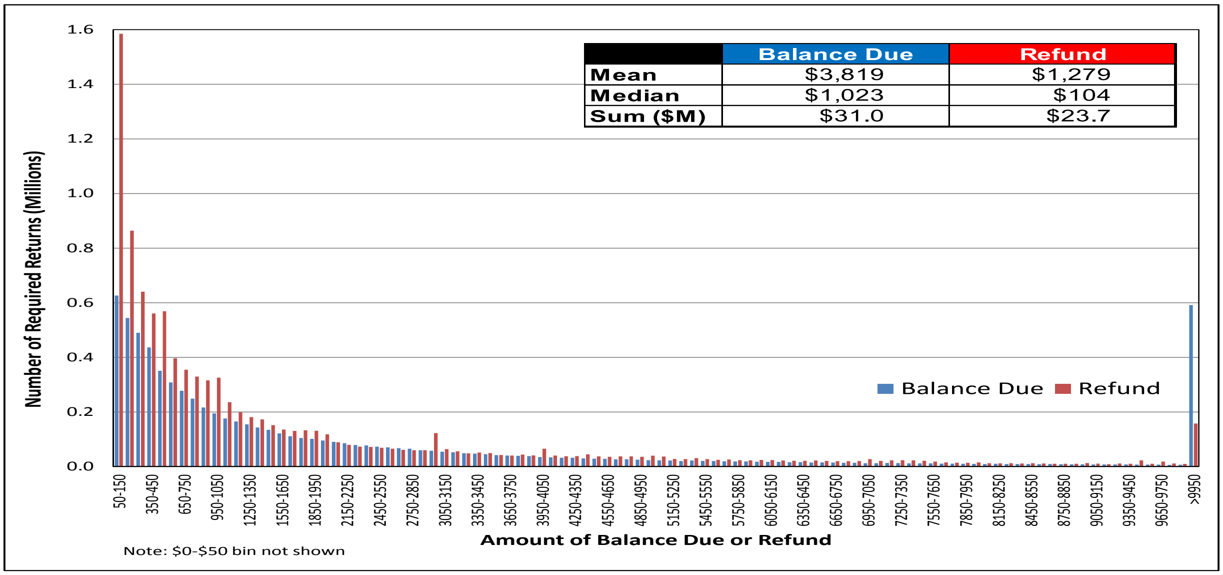
The nonfiling tax gap is concentrated in a relatively small share of the total population of nonfilers. Focusing on Tax Year 2010 data, the top decile of nonfilers is responsible for about 64 percent of the nonfiling tax gap, and the top 20 percent, 78 percent (Table 8). The Tax Year 2010 nonfiling tax gap is estimated by this method to be about $31 billion spread among taxpayers with a balance due. The mean balance due amount is $3,819, and the median amount is $1,023 (Figure 3).

**TABLE 8. Nonfiling Tax Gap by Decile, Tax Year 2010**

| Percentile | Balance Due by Decile, All Nonfilers ($Billions) | % Share of Balance Due by Decile, All Nonfilers |
|---|---|---|
| 10 | 0.0 | 0.1% |
| 20 | 0.2 | 0.5% |
| 30 | 0.2 | 0.8% |
| 40 | 0.4 | 1.4% |
| 50 | 0.7 | 2.2% |
| 60 | 1.0 | 3.4% |
| 70 | 1.5 | 5.3% |
| 80 | 2.4 | 8.2% |
| 90 | 4.3 | 14.4% |
| 100 | 18.7 | 63.7% |

The nonfiling tax gap is computed as the sum of the balance due for all taxpayers with a positive balance due amount. A large number of nonfilers are owed refunds since the sum of their prepayments and refundable credits exceeds their total tax liability. Following the IRS administrative method for Tax Year 2010, we estimate that if all not-filers and late-filers (i.e., taxpayers with a filing requirement) who had a balance due or refund filed a tax return on time, the net increase in revenues would be $7.4 billion, since $23.7 billion would be paid out in refunds (Figure 3). The mean refund amount is $1,279, and the median amount is $104. Not-filers had a net ***balance due*** of $13.8 billion, while late filers were due a net ***refund*** of $6.4 billion.

**FIGURE 3. Distribution of Balance Due and Refund Amounts Among Nonfilers, TY2010**

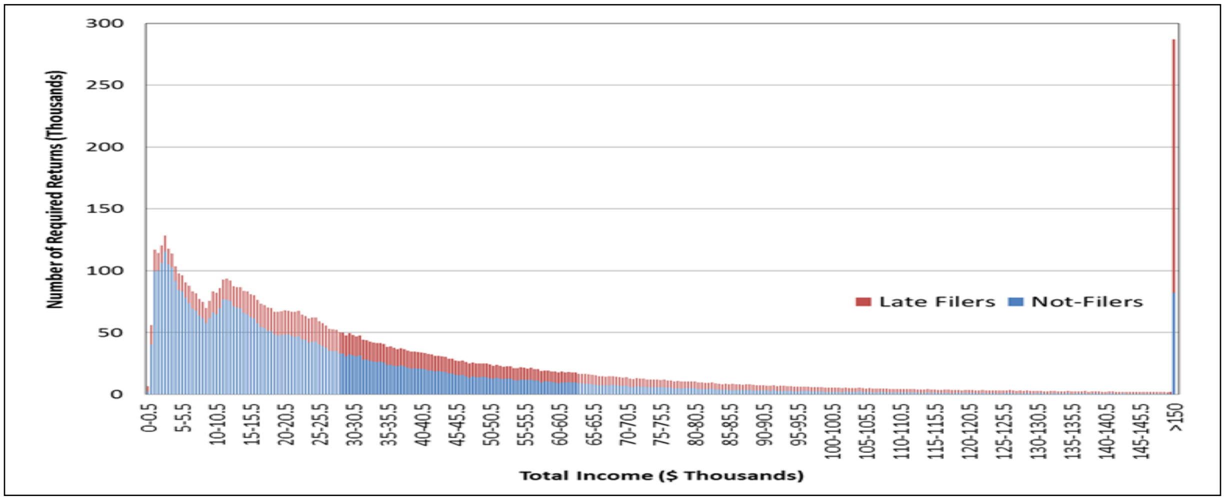


| | Balance Due | Refund |
|---|---|---|
| Mean | $3,819 | $1,279 |
| Median | $1,023 | $104 |
| Sum ($M) | $31.0 | $23.7 |

Note: $0-$50 bin not shown

Consistent with the relative concentration of balance due, the distribution of total income is also highly skewed (Figure 4). Of those with a balance due, the 20 percent with the lowest total income have less than 2.5 percent of the income of all nonfilers. By contrast, the top 20 percent have almost half of total income. Roughly 600,000 nonfilers have more than $100,000 in total income, while almost 300,000 nonfilers have more than $150,000 in total income.

**FIGURE 4. Distribution of Total Income Among Nonfilers, TY2010**



Using the IRS Administrative Data Method nonfiler data, we can examine how the nonfiler tax gap estimate is distributed geographically. Figure 5 shows the relationship between the nonfiler tax gap and the number of required returns contributing to that gap for the nine Census regions. The slope of the line represents the overall average tax gap per required return—about $2,000. The regions above the line have a larger-than-average nonfiling gap per required return, while those below it have a smaller than average nonfiling gap per return. In particular, the tax gap per required return is estimated to be fairly large in the South Atlantic and Mid-Atlantic regions, while it is estimated to be relatively small in the West South Central, Mountain, and West North Central regions.

Figure 6 shows the relationship between the total balance due and the total tax of filers and nonfilers. The slope of the line represents the overall average of the nonfiling tax gap as a percent of total tax—about 2.7 percent. Regions above the line, such as the South Atlantic, Pacific, and West South Central regions, are estimated to have larger nonfiling tax gaps as a share of total tax. By contrast, the Mid-Atlantic, East North Central, New England, and West North Central regions are estimated to have smaller nonfiling tax gaps.

**FIGURE 5. Nonfiler Tax Gap vs. Nonfiler Returns, TY2010**

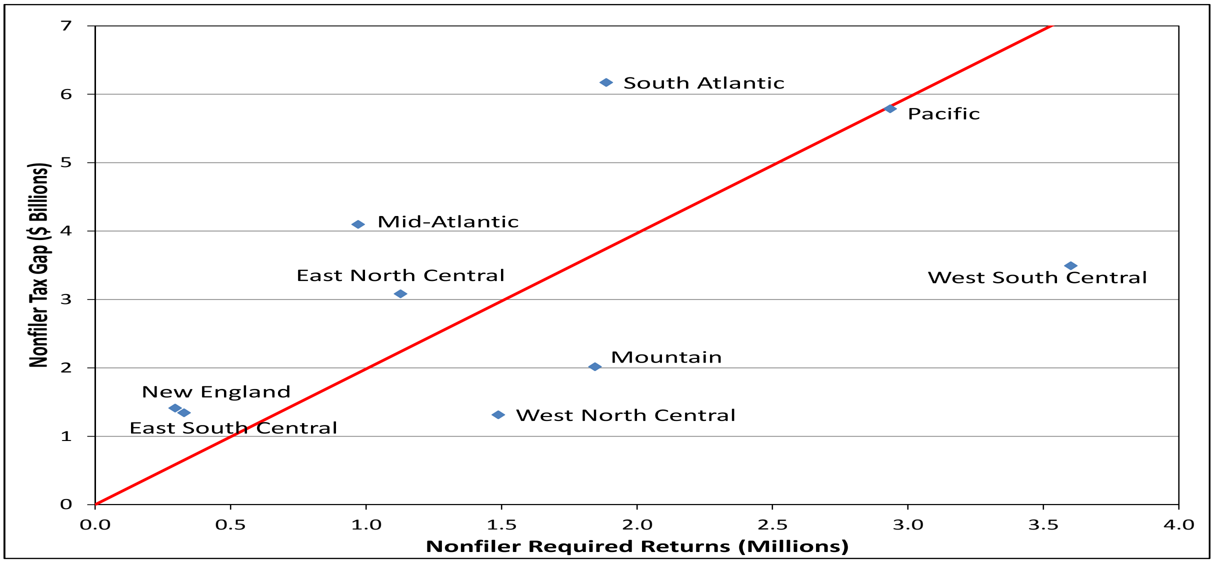
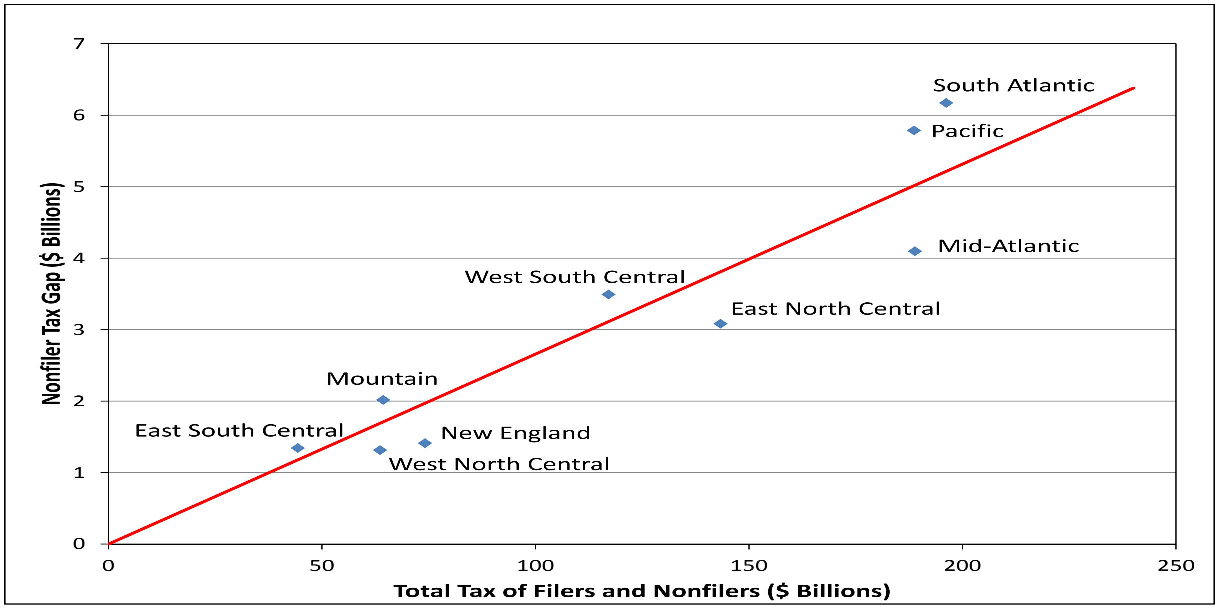


**FIGURE 6. Nonfiler Tax Gap vs. Total Tax of Filers and Nonfilers, TY2010**



As discussed above, a significant share of nonfiler returns, both late and not-filed, would be due a refund. Figure 7 shows that the later a return is filed past the filing deadline, the greater is the likelihood that it will have a balance due. Returns that are submitted after receipt of an IRS nonfiler notice are also more likely to have a balance due. The median balance due amount for such enforcement returns is also higher than for returns that are submitted without an enforcement treatment (Figure 8).

**FIGURE 7. Percent of Late Required Returns with a Balance Due by Posting Date, Enforcement vs. Non-Enforcement, TY2010**

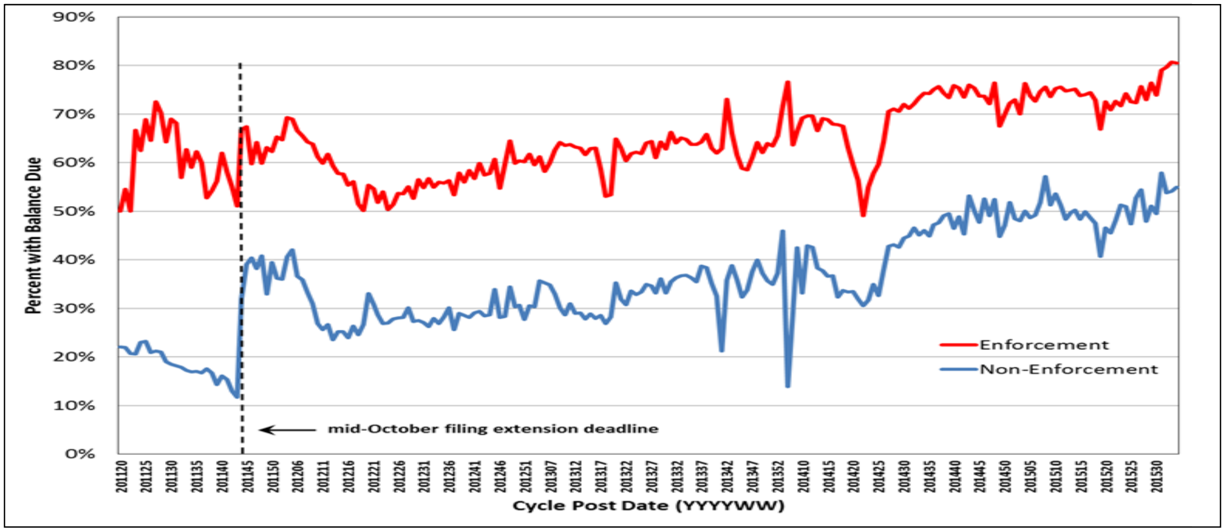
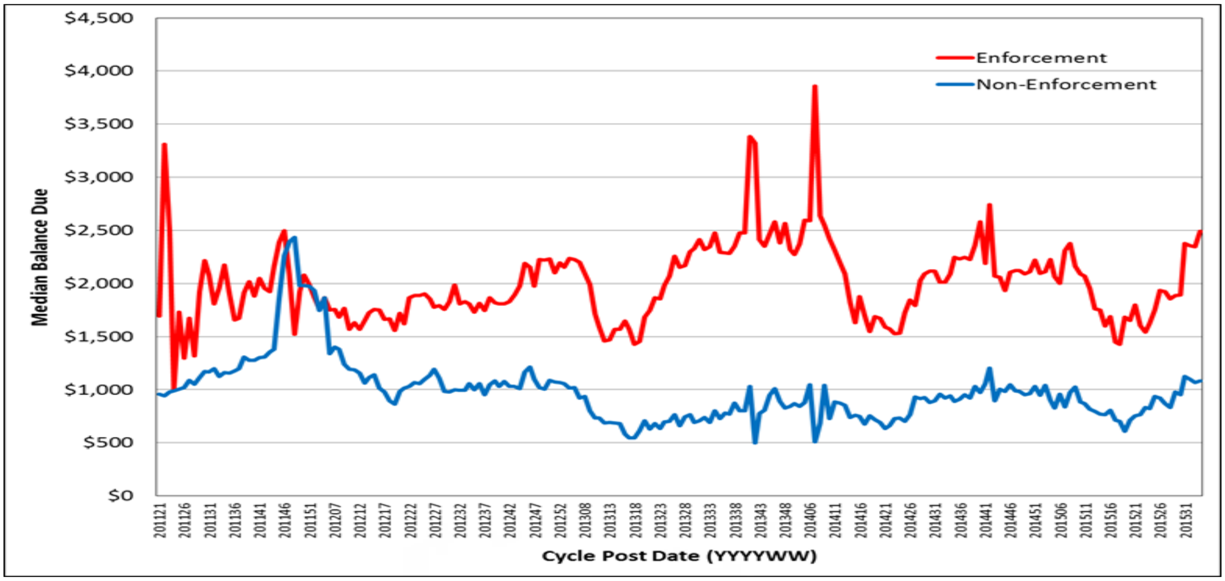


**FIGURE 8. Median Balance Due of Late Returns by Posting Date**



## References

Erard, B. and Ho, C. (2001). "Searching for ghosts: who are the nonfilers and how much tax do they owe?," *Journal of Public Economics* 81, p. 25–50.

Internal Revenue Service (1996). *Federal Tax Compliance Research: Individual Income Tax Gap Estimates for 1985, 1988, and 1992*, Publication 1415 (Rev. 4–96).

____________ (2012). *Federal Tax Compliance Research: Tax Year 2006 Tax Gap Estimation,* at http://www.irs.gov/pub/irs-soi/06rastg12workppr.pdf.

Jones, Maggie R. and O'Hara, Amy (2014). "Do Doubled-Up Families Minimize Household-Level Tax Burden?," *2014 IRS Research Bulletin*, Publication 1500, p. 181–203.

Wagner, D. and Layne, M. (2012). "Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' Record Linkage Software," Washington, DC: Center for Administrative Records Research and Applications Internal Document, U.S. Census Bureau.