What Drives Income Tax Filing Compliance?[1]

Brian Erard (B. Erard & Associates) and John Guyton, Patrick Langetieg, Mark Payne, and Alan Plumley (IRS, Research, Applied Analytics, and Statistics: Office of Research)

This paper summarizes the findings from a recent effort to better understand what drives households to comply with their Federal income tax filing obligations.[2] The decision whether to file a tax return is essentially a participation decision, and there is a long history in empirical research of applying qualitative choice models, such as logit and probit specifications, to model the determinants of participation behavior. However, standard qualitative choice models assume that one has access to a representative data sample of participants and nonparticipants, including an indicator for the participation status of each subject in the sample.

Although IRS has detailed tax return information for filers, it lacks comparable information on nonfilers. To fill this void, we supplement IRS information on filers with survey information from the general population of filers and nonfilers. However, this latter data source does not identify which respondents are filers and which are nonfilers. We therefore apply a novel econometric methodology (calibrated probit analysis) to estimate the drivers of filing compliance.[3]

Calibrated Probit Analysis

We apply a calibrated probit analysis (Erard, 2017) to our combined data sample to estimate the drivers of filing compliance. Intuitively, whereas a standard probit analysis relies on differences between the characteristics of participants (filers) and nonparticipants (nonfilers) to infer what drives behavior, the calibrated probit methodology relies on differences between the characteristics of a representative sample of participants and the characteristics of a supplementary sample that (when weighted) is representative of the overall population of participants and nonparticipants. We estimate our calibrated probit model by solving the following constrained optimization problem:

$$\max_{\beta} \sum_{i=1}^{n_f} ln[\Phi(\beta' X_i)]$$

$$s.t. \quad \sum_{j=1}^{n_o} w_j \Phi(\beta' X_j) = N_f$$

where n_f represents the number of filers in our filer-only data sample, n_o is the size of our supplementary sample of filers and nonfilers, X is a vector of explanatory variables, β is a vector of coefficients to be estimated, $\Phi(z)$ represents the value of the standard normal cumulative distribution function evaluated at z, w refers to sample weights in the supplementary sample, and N_f is the population number of returns filed. The solution to this problem is the value of β that maximizes the predicted likelihood of filing among households in the filer-only data sample subject to the constraint that the weighted aggregate predicted number of filers in the supplementary sample based on this solution is consistent with the actual number of filers in the population. In other words, the estimated value of β is calibrated to be consistent with the population count of filed required returns (hence the name "calibrated probit analysis").We estimate our model using the constrained maximum likelihood application (CML) in GAUSS@.

[1] The views expressed in this paper reflect those of the authors, and they do not necessarily represent the position of the Internal Revenue Service. This research was funded under IRS contracts TIRNO-10-D-00021-D0004, TIRNO-14-P-00157, and TIRNO-15-P-00172.

[2] For readers interested in more details, a full presentation of our methodology and findings is provided in Erard, *et al.* (2016).

[3] Prior research on this topic includes the aggregate longitudinal analysis of filing compliance across states by Plumley (1996) and the micro-level analysis of filing behavior by Erard and Ho (2001).

To account for variations in behavior both across households and over time, we have extended the above estimation framework to permit an analysis of a time series of cross-sections. This extended framework imposes a separate constraint equation for each time period in the data sample. To ensure that all restrictions can be satisfied, a set of tax-year dummies is included among the explanatory variables in this specification.

Data Description

Our IRS data source on filers is the Individual Returns Transaction File (IRTF). Many households have no legal filing obligation because their income is below the filing threshold, and they do not meet certain other filing criteria, such as a need to report self-employment tax or taxes on tip income. Some of these households do file, however, to claim refunds of withheld earnings or to claim a refundable tax credit, such as the Earned Income Credit. Since our focus is on filing compliance, we restrict our IRTF sample to households with a legal filing obligation for income tax or self-employment tax purposes. This is achieved by applying an algorithm to check whether a given return satisfies any of the various conditions (such as gross income above the relevant filing threshold or net self-employment earnings in excess of $400) that trigger a filing requirement. Our supplementary sample of filers and nonfilers is drawn from the Current Population Survey Annual Social and Economic Supplement (CPS-ASEC), which is compiled annually by the Census Bureau. In past research, we have found that certain income sources are understated in this survey. Therefore, in order to more accurately identify households with a legal filing obligation, we follow Erard, Langetieg, Payne, and Plumley (2014) in imputing additional income across the sample (in many cases, so that the Census data become consistent with third-party information return data reported to the IRS). To assign household members to tax returns, we also impute tax filing status. The CPS-ASEC is a stratified random sample; however, the stratification criteria are not publicly available. A desirable feature of our econometric methodology is that we are able to effectively control for the stratified nature of the sample simply by applying the sample weights.[4]

For both data sources, we have large cross-sectional samples for each tax year over the period from 2000 through 2012. On average, the data include a simple random sample of approximately 113,000 filed required returns per year from the IRTF and a stratified random sample of approximately 76,000 required returns per year from the CPS-ASEC.

Estimation Results

In this section, we present the results of our calibrated probit analysis of the drivers of filing compliance based on our time series of cross sections over the period from TY2000 through TY2012. Because our estimation methodology relies on a comparison of variables from two separate data sources (IRTF and CPS-ASEC), it is important to restrict the set of regressors to those variables that are comparably measured in the two sources. So, for instance, while the IRTF provides information on whether a taxpayer is owed a refund or has a balance due (which is likely to be relevant to the filing decision), comparable information is not available in the CPS-ASEC. It also would be desirable to include some indicators of filing status as explanatory variables. However, a nontrivial number of taxpayers claim the incorrect status on their return. For instance, the percentage of filers claiming head of household status greatly exceeds the estimated percentage of required returns with this status based on the CPS-ASEC. Instead of filing status indicators, we include an indicator for marital status in our specifications. Similarly, we would like to account for Earned Income Credit eligibility in our analysis. However, a claims-based measure from the IRTF would be misleading, owing to a nontrivial number of Earned Income Credit claimants who are not truly eligible. Ultimately, we have selected the following explanatory variables for our analysis, which we believe are measured reasonably comparably (and generally accurately) across our two data sources:

CONST: Constant term.

AGE 65: Dummy for primary or secondary filer age 65 or over.

[4] Other existing models for use with supplementary samples (Lancaster and Imbens (1996); Manski and McFadden (1981)) require knowledge of the stratification criteria, which are not provided in public use samples of Census surveys. An exception is the Steinberg-Cardell (1992) approach, which can be adapted for use with sample weights. However, this estimator produces relatively inefficient estimates and is subject to convergence issues.

MARRIED: Dummy for married taxpayer.

CHILD3UP: Dummy for three or more children.

TY2009UP*CHILD3UP: Dummy for three or more children and TY2009 or later.

LN(GROSSINC): Natural log of gross income, where gross income is computed as the sum of the positive amounts of wages and salaries, interest, taxable dividends, pensions, rents, unemployment compensation, taxable social security benefits, alimony, and gross self-employment earnings.

NO TAX STATE: Dummy variable for residence in a State with no individual income tax (AK, FL, NH, NV, SD, TN, TX, WA, WY).

SEFILREQ: Dummy variable for a filing requirement triggered by having net profit from farm and nonfarm self-employment in excess of $433.

NEARTHRESH: Dummy variable for gross income less than 1.10 times the filing threshold, where the filing threshold for nonependent joint filers is applied for married joint filing status and the filing threshold for single filers is applied to all other nondependent filers. The lower statutory thresholds are applied to single and married dependent filers.

LN(BURDEN): Natural log of taxpayer burden.[5]

LN(BURDEN)*NEAR THRESHOLD: Interaction between LN(BURDEN) and NEARTHESH.

MIDATL: Dummy variable for residence in the Middle Atlantic division.

EASTNC: Dummy variable for residence in the East North Central division.

WESTNC: Dummy variable for residence in the West North Central division.

SOUTHATL: Dummy variable for residence in the South Atlantic division.

EASTSC: Dummy variable for residence in the East South Central division.

WESTSC: Dummy variable for residence in the West South Central division.

MOUNTAIN: Dummy variable for residence in the Mountain division.

PACIFIC: Dummy variable for residence in the Pacific division.

TY2001–TY2012: Tax year dummies.

The omitted Census division is New England and the omitted tax year is 2000.[6] Our measures of taxpayer burden and gross income have been converted to real 2010 amounts based on the CPI-U price index.

Table 1 breaks down the average of the weighted mean values of our explanatory variables (excluding the year dummies) over the 13 year estimation period separately for the IRTF and CPS-ASEC samples. Over this period, filers were relatively more likely to be of age 65 or older, reside in States without an income tax as well as in the Middle Atlantic or East North Central divisions, and have gross income near the filing threshold. They were relatively less likely to be married, have three or more children, receive nontrivial self-employment earnings, or reside in the Mountain or Pacific divisions. On average, their gross earnings were somewhat lower than the overall population of households with a filing requirement, and they faced a somewhat lower filing burden.[7]

[5] Monetized value of time and out-of-pocket expenses incurred to meet one's filing obligation, estimated using IRS methodology applied to the limited set of explanatory variables available in this analysis.

[6] Census divides the country into four "regions" (Northeast, Midwest, South, and West), each of which is further broken down into two or more "divisions." In this study, we control for possible variations in behavior across the eight Census divisions (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific).

[7] Technically, our estimate of burden (i.e., compliance cost) reflects the time and money expense that filers actually expend, not the amount that they would have to expend to be fully compliant.

TABLE 1. Average Values of Explanatory Variables Over the TY2000–TY2012 Period, by Data Source

Variable	Data Source	
	IRTF	CPS-ASEC
AGE 65	0.6352	0.6059
MARRIED	0.4396	0.4646
CHILD3UP	0.0653	0.0658
LN(GROSSINC)	10.7026	10.7395
NO TAX STATE	0.1984	0.1972
SEFILREQ	0.1200	0.1266
NEARTHRESH	0.0565	0.0537
LN(BURDEN)	5.9386	5.9689
LN(BURDEN)*NEARTHRESHOLD	0.2922	0.2798
MIDATL	0.1406	0.1383
EASTNC	0.1582	0.1550
WESTNC	0.0695	0.0689
SOUTHATL	0.1908	0.1910
EASTSC	0.0559	0.0557
WESTSC	0.1065	0.1073
MOUNTAIN	0.0670	0.0694
PACIFIC	0.1594	0.1636
NEW ENGLAND	0.0520	0.0508

Our specification includes a dummy variable (TY2009UP*CHILD3UP) equal to one if both the household has three or more children and the tax year is 2009 or later to explore whether the introduction of a larger earned income credit amount in Tax Year 2009 for families with three or more children induced more households to file their (required) tax returns. Table 2 compares the average shares of households in the IRTF and CPS-ASEC samples before and after Tax Year 2009. Whereas filers were relatively less likely to claim three or more children prior to Tax Year 2009, the opposite was true from Tax Year 2009 on, even though such families appear to have become less common in the general population. Our econometric analysis explores whether this pattern continues to hold after controlling for other factors related to the filing decision.

TABLE 2. Average Share of Households With Three or More Children, by Tax Year Period and Data Source

Tax Year Period	Data Source	
	IRTF	CPS-ASEC
TY2000–TY2008	6.42%	6.76%
TY2009–TY2012	6.79%	6.18%

Formulas for estimating taxpayer burden were developed through regression analyses involving four different IRS surveys performed for Tax Years 2007, 2010, 2011, and 2012. A separate burden estimation formula was developed from each survey based on a set of explanatory variables that was common to our IRTF and CPS-ASEC samples. In principle, the burden estimate for a given tax year could be based on any of the four alternative formulae. In practice, the alternative formulae yield burden estimates that are extremely highly correlated, so the choice of which formula to use is not of much importance. In the results presented below, we rely on the Tax Year 2007 survey formula for Tax Years 2000-2008, the Tax Year 2010 survey formula for Tax Years 2009 and 2010, the Tax Year 2011 survey formula for Tax Year 2011, and the Tax Year 2012 survey formula for Tax Year 2012. We have confirmed that our results are very similar when the Tax Year 2007 survey formula is employed for all tax years.

Table 3 presents the estimated coefficients of our calibrated probit analysis along with their associated t-statistics.[8] Remember that the data were restricted to those having a filing requirement. A positive estimated coefficient indicates that, after controlling for other factors, the variable is positively associated with filing compliance. The results indicate that filing compliance tends to be positively associated with gross income, being age 65 or older, and being married. On the other hand, filing compliance tends to be negatively associated with self-employment. The results also indicate significant variations in filing compliance across Census divisions with high compliance (relative to those who live in New England) among residents of the Middle Atlantic and East North Central states and relatively low compliance among those living in the West South Central, Mountain, and Pacific States. However, residence in a State that does not administer an income tax does not appear to have any significant impact on Federal income tax filing compliance.

The positive estimated coefficient of the interaction term TY2009UP*CHILD3UP supports notion that the introduction of a larger Earned Income Credit for households with three or more children in Tax Year 2009 had a positive impact on filing compliance. The tax-year dummies indicate that filing compliance declined between Tax Year 2000 and Tax Year 2006, rebounded in Tax Year 2007 to around the Tax Year 2000 level, and then declined again in Tax Year 2009.

TABLE 3. Calibrated Probit Estimation Results

Variable	Coefficient Estimate	t-statistic
CONST	-0.8135	-6.81
AGE 65	1.3090	40.25
MARRIED	0.2025	4.72
CHILD3UP	-0.1453	-5.50
TY2009UP*CHILD3UP	0.5032	8.09
LN(GROSSINC)	0.5124	36.67
NO TAX STATE	0.0034	0.15
SEFILREQ	-0.2735	-13.74
NEARTHRESH	-1.0188	-8.96
LN(BURDEN)	-0.5721	-26.35
LN(BURDEN)*NEARTHRESH	0.1843	9.01
MIDATL	0.1048	2.02
EASTNC	0.0746	1.69
WESTNC	-0.0463	-0.19
SOUTHATL	-0.0706	-0.29
EASTSC	0.0812	0.24
WESTSC	-0.0852	-1.83
MOUNTAIN	-0.2766	-6.47
PACIFIC	-0.2869	-7.27
TY2001	-0.0224	-0.46
TY2002	-0.1629	-3.54
TY2003	-0.1518	-3.31
TY2004	-0.1637	-3.56
TY2005	-0.2103	-4.71
TY2006	-0.1537	-3.37
TY2007	0.0618	1.26
TY2008	0.0372	0.76
TY2009	-0.3490	-7.19
TY2010	-0.3188	-6.46
TY2011	-0.3991	-7.98
TY2012	-0.5022	-10.10

[8] The standard errors of the estimated coefficients and marginal effects were computed using the generalized moment conditions implied by the first-order conditions of the constrained maximization problem.

This rise in filing compliance in Tax Year 2007 is presumably at least partially attributable to the Economic Stimulus Act of 2008. In order to claim a stimulus payment in 2008, households were required to file a Tax Year 2007 income tax return. The stimulus payment was rather substantial, ranging from $300 to $600 for eligible individuals and from $600 to $1,200 for joint filers, plus an additional $300 for each qualifying child.[9] Households that did not claim the stimulus for Tax Year 2007 were permitted to file a return and claim it the following tax year. The estimated decline in filing compliance beginning in Tax Year 2009 indicates that the stimulus-induced boost in compliance was temporary in nature. This is a noteworthy finding as it contrasts with the commonly held view that once a taxpayer enters the system, the taxpayer tends to remain in the system.

Taxpayers who have income near the filing threshold are relatively less likely to file a return. As well, burden is negatively associated with filing compliance. However, this burden effect is partially mitigated for households with gross income near the filing threshold. This may reflect a relatively high degree of filing compliance among households near the threshold who are eligible for benefits such as the Earned Income Credit. Although claiming such benefits is associated with an additional filing burden, this represents a "voluntary burden" that many households would find acceptable in relation to the benefits to be received.

Concluding Remarks

We have applied a novel econometric methodology to examine the drivers of filing compliance across households and over time. This methodology allows us to draw inferences based on a primary sample that includes only return filers and a supplementary survey sample from the overall population of filers and nonfilers. We have found evidence that filing compliance is linked to a variety of factors, including taxpayer burden, age, and self-employment. The results also indicate a positive role for various tax benefits, such as the Economic Stimulus Act of 2008 and the Earned Income Credit. We are currently working to extend the estimation framework to permit an analysis of how one's filing history impacts subsequent filing choices. In addition, we have recently discovered that our estimates of the number of required returns in the population based on the CPS-ASEC tend to be understated, despite our imputations of additional income to the data base. Preliminary estimation results for our model based on improved measures of the required return population are qualitatively similar to those presented here. However, further work is needed to refine these measures and extend them over a longer time span.

References

Erard, B. (2017) "Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A Calibrated Qualitative Choice Estimation Approach," Working Paper.

Erard, B., J. Guyton, P. Langetieg, M. Payne, and A. Plumley (2016) "What Drives Filing Compliance?," Working Paper.

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

Erard, B., P. Langetieg, M. Payne, and A. Plumley (2014) "Missing Returns vs. Missing Income: Estimating the Extent of Individual Income Tax Filing Compliance from IRS and Census Data," paper presented at the National Tax Association Annual Conference, Santa Fe, New Mexico, November 13–15.

Lancaster, T., and G. Imbens (1996) "Case Controlled Studies with Contaminated Controls," *Journal of Econometrics* (71) 145–160.

Manski, C.F., and D. McFadden (1981) "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden, Cambridge: MIT Press, 2–49.

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

Steinberg, D., and N.S. Cardell (1992) "Estimating Logistic Regression Models When the Dependent Variable Has No Variance," *Communication in Statistics—Theory and Methods* (21) 42–50.

[9] The value of this payment was phased out for households with high levels of AGI.