ELSEVIER

Journal of Public Economics 1 (2000) 000–000

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.nl/locate/econbase

Searching for ghosts: who are the nonfilers and how much tax do they owe?

Brian Erard[a,*], Chih-Chin Ho[b]

[a]*B. Erard and Associates, 2350 Swaps Court, Reston, Virginia 20191, USA*
[b]*Internal Revenue Service, Office of Research, Washington, D.C., USA*

Received 30 November 1999; received in revised form 30 July 2000; accepted 31 July 2000

Abstract

This paper is about 'ghosts' — individuals who fail to comply with their income tax filing requirements. As their name suggests, the identities and characteristics of these individuals are shrouded in mystery. In this paper we attempt to de-mystify the issues surrounding ghosts and examine their role in the compliance process. We begin by extending a standard model of tax evasion to account for the existence of ghosts. We then examine the empirical significance and policy relevance of our extension using a unique data set containing detailed tax and audit information for both filers and nonfilers of U.S. federal income tax returns.

Keywords: Tax evasion; Tax avoidance; Income tax; Ghost

JEL classification: H24; H26; H31

1. Introduction

Over the past three decades, researchers have devoted substantial attention to the decision concerning how much income to report on one's tax return and the tax authority's response to this report.[1] A group that has been largely neglected by this

*Corresponding author. Tel.: +01-703-390-9368.
E-mail address: brerard@aol.com (B. Erard).

[1]See Andreoni et al. (1998) for a recent survey of this literature.

0047-2727/00/$ – see front matter
PII: S0047-2727(00)00132-8

research is those individuals who simply choose not to file a return, a group sometimes referred to as 'ghosts' by academics and policy-makers.[2] Based on available evidence from the U.S. (Crane and Nourzad, 1993) and Jamaica (Alm et al., 1991), it appears that nonfiling poses a significant problem. However, very little is known about this form of evasion. In this paper we employ a unique data source to learn about the characteristics of ghosts, examine the factors driving their decision not to file a tax return, and measure their unpaid tax liability. We begin in Section 2 by developing an extended model of taxpayer reporting behavior that includes nonfiling as a strategic option. We then examine the empirical significance and policy relevance of our extension using detailed line-item tax and audit information for both filers and nonfilers of U.S. federal income tax returns. We lay out our econometric framework in Section 3, summarize our data in Section 4, and present the results of our analysis in Section 5. In Section 6, we employ our estimates to compare the profiles of the filer and ghost populations. Section 7 contains a discussion of the net tax liabilities of ghosts, and a brief conclusion is offered in Section 8.

2. Theoretical framework

In this section, a simple theoretical framework is presented for understanding the decision whether to file an income tax return. We begin by considering a standard model of taxpayer reporting behavior. We then extend the model to account for nonfiling as a strategic option. In the traditional economic model of evasion, a taxpayer approaches his reporting decision as he would a gamble, balancing the risk of audit and penalty against the benefits of a reduced tax payment. Formally, he chooses an amount of income to report X to maximize the following expression:

$$(1-p)U[Y-tX]+pU[Y-tX-(1+\theta)\,t(Y-X)], \tag{1}$$

where $U[\cdot]$ is his utility function, Y is his true income, p is the probability of audit, t is the proportional tax rate, and θ is the proportional penalty rate on undeclared taxes.[3] The optimal report depends on the taxpayer's preferences for risk as well as the values of the tax and enforcement parameters.

Although many elaborations of this model have been developed over the years,

[2] The term 'ghosts' is borrowed from Cowell (1990), who notes that it is commonly used by Inland Revenue in the U.K. to refer to individuals for whom no official record exists. Refer to Cowell and Gordon (1995) for a theoretical analysis of the role of ghosts in sales tax evasion.

[3] This model is the classic specification given by Allingham and Sandmo (1972), amended as in Yitzhaki (1974) to allow the penalty rate to depend on unreported taxes rather than unreported income. No penalty or reward is applied if reported income exceeds true income.

virtually all of them have followed the traditional specification in presupposing that an individual will choose to make a tax report. In fact, though, a nontrivial number of individuals elect each year to take the ultimate tax shortcut of not filing a return at all. To account for such 'ghosts', it is necessary to extend the above model to describe the incentives associated with not filing. In our extension, we focus on three fundamental choices facing a potential taxpayer. First, there is the decision whether to file a return at all. Second, if the individual should choose to file, he must decide (as in the standard model) how much income to report. Third, regardless of his filing decision, he must choose how much tax (if any) to prepay through withholding and estimated tax payments. This expanded set of compliance decisions raises some additional considerations for the individual to take into account when formulating his compliance strategy. In particular, his choices are likely to be shaped by the burden associated with preparing and filing a return, the risk of being identified as a nonfiler, and the penalties for not filing a return and for prepaying too little in taxes. As in the traditional model of evasion, we postulate that the individual approaches his compliance decisions by examining the expected utility associated with different alternatives. If the individual were to file a return, his utility would be determined by the following expression:

$$(1-p)U\left[Y-tX-\gamma(\overline{W}-W)-c\right]+pU\left[Y-tX-(1+\theta)\,t(Y-X)-\gamma(\overline{W}-W)-c\right]. \quad (2)$$

Although this expression is similar to Eq. (1), observe that the individual's net wealth has been reduced by a dollar measure of the burden of preparing and filing a return c.[4] In addition, the individual now chooses the amount of tax to prepay W as well as the amount of income to report on his return X. In the U.S., individuals are required to pay most of their tax liability over the course of the year, prior to filing their tax return. Employers normally withhold a portion of their salaried employees' paychecks for this purpose, submitting the amount withheld to the Internal Revenue Service (IRS). An employee can elect to have either more or less tax withheld than the standard amount to better address his personal tax situation. Self-employed individuals are required to make periodic tax installment payments based on their estimated tax liability for the year. Penalties are in place for those who fail to prepay a sufficient share of their taxes.[5] We capture the essence of the U.S. prepayment rules in Eq. (2) by assuming that if total prepayments W are

[4] See Blumenthal and Slemrod (1992) for evidence on the magnitude of the U.S. income tax compliance burden. Note that this model could be extended to allow c to be a function of the amount of effort that goes into legal and illegal tax avoidance schemes. See, for example, Cross and Shaw (1982) and Slemrod (1995).

[5] Normally, an individual must prepay the lesser of his tax obligation for the prior year or 90 percent of his current year's tax liability. The underpayment penalty is one-half of 1 percent of the shortfall per month, up to a maximum of 25 percent.

below the minimum prepayment threshold ($\overline{W}$), a penalty at the rate γ is applied to the shortfall.

In practice, of course, the individual may choose not to file a tax return. If he were to elect this option, his utility would instead be determined by the following expression:

$$(1-q)U[Y-W]+qU[Y-W-(1+f)(tY-W)-c], \tag{3}$$

where q represents the probability the individual will be apprehended and f is the nonfiler penalty rate that applies to the outstanding tax balance. In the U.S., the penalty for not filing is equal to five percent of the unpaid tax liability for each month the return is late, up to a maximum of 25 percent. In addition, the above-mentioned penalty for underpayment of estimated taxes may also be applied in some circumstances. If apprehended, a nonfiler would be required to submit a tax return. Eq. (3) therefore accounts both for the burden c associated with completing the return and any penalties for nonpayment of taxes.

We assume that the individual's actions proceed in the following sequence. At the beginning of the period, he makes a tax prepayment of W (which might be zero). For simplicity, we assume that the values of all parameters, including true income Y, are known to him at this point. At the end of the period, the individual either files a return or becomes a ghost. The individual is forward-looking and recognizes that the optimal choice of W depends on what behavior he will choose at the end of the period. He therefore compares the maximum expected utility he can achieve under the filing and nonfiling options, choosing the optimal value of W based on the more attractive option.

If the individual were to file a return at the end of the period, it would be optimal for him to make the minimum tax prepayment W^* that avoids a penalty; i.e., to choose $W^*=\overline{W}$ in Eq. (2).[6] Under this scenario, he would also want to report an income of X^* on his return, determined as the implicit solution to the following first-order condition:[7]

$$(1-p)tU'[Y-tX^*-c]=p\theta tU'[Y-tX^*-(1+\theta)t(Y-X^*)-c]. \tag{4}$$

The left-hand side of Eq. (4) represents the utility gain from successfully evading taxes by an additional dollar, weighted by the probability of not being audited. Analogously, the right-hand side represents the utility loss from having been caught evading taxes by an additional dollar, weighted by the probability of audit. At the optimal level of evasion, the marginal expected benefit of understating income just equals the marginal expected cost.

If the individual instead were to become a ghost, it would be optimal for him to

[6] In our model, we ignore any borrowing motive for making insufficient tax prepayments. We observe, though, that given the current penalty rate in the U.S., such a motive might drive some individuals to prepay less than $\overline{W}$.

[7] We are assuming here that $p<1/(1+\theta)$; otherwise, the optimal report would equal Y.

select the prepayment W^{**} that maximizes Eq. (3).[8] Specifically, he would want to choose W^{**} as the implicit solution to the following first-order condition:[9]

$$(1-q)tU'[Y-W^{**}] = qftU'[Y-W^{**}-(1+f)(tY-W^{**})-c]. \quad (5)$$

Similar to Eq. (4), this condition equates the marginal expected benefit from underpaying tax with the marginal expected cost.

If the value of Eq. (2), evaluated at X^* and W^*, exceeds that of Eq. (3), evaluated at W^{**}, the individual will recognize that he can achieve a higher expected utility by filing. He will therefore elect to make a tax prepayment of $W^* = \overline{W}$ at the beginning of the period. At the end of the period, he will file a return and report an income of X^*. On the other hand, if the above condition is not satisfied, the individual will prefer to become a ghost. In this case, he will make a tax prepayment of W^{**} at the beginning of the period and file no return at the end of the period.

Observe that in the absence of a filing burden c, the first-order conditions described by Eqs. (4) and (5) are isomorphic. Thus if $c = 0$, $p = q$, and $\theta = f$, the optimal choice of tax prepayments W^{**} under the nonfiling option will be precisely equal to t times the optimal choice of reported income X^* under the filing option, and the individual will be indifferent between filing and not filing. It follows that an individual will be relatively more likely to become a ghost the greater the filing burden c, the lower the perceived chances for successful underreporting $(1-p)$, the higher the penalty rate for underreporting θ, and the lower the probability q and rate of penalty f associated with not filing.

An issue not generally taken into account in studies of tax evasion is the dynamic nature of an individual's compliance decisions.[10] In practice, though, one would expect to observe a high degree of persistence in filing behavior. Consider, for example, an individual who failed to file in the previous tax year. If he were to file a return for the current year, he may perceive that this would increase the risk that his past filing violation would be uncovered. For similar reasons, a taxpayer who did file a return for previous year may fear that the tax authority would become suspicious if he elected not to file in the current year.[11] In our econometric

[8] In practice, a high value of W may provide a signal to the tax agency that the individual possesses sufficient income to have a tax filing requirement. A more general model would account for this possibility by allowing the probability of detection q to vary with W. Analogously, a low report X from a filer may serve as a signal to the tax agency of likely tax noncompliance, in which case p might tend to vary with X. However, the main factors influencing the choice between filing and not filing are adequately represented by the simpler fixed audit probability specification presented in this paper.

[9] We are assuming here that $q < 1/(1+f)$; otherwise, the optimal prepayment would equal tY.

[10] Two exceptions are Engel and Hines (1999) and Erard (1992).

[11] In fact, in the U.S. the IRS has what it calls a 'stop-filer' program designed to identify and investigate prior year taxpayers who have not filed a return for the current year.

207 analysis, we explicitly account for the recent filing history of the individuals in our
208 sample to address possible persistence in behavior.

209 3. Econometric framework

210 In this section we develop an econometric framework for analyzing the decision
211 whether to comply with one's income tax filing requirement. We restrict our
212 attention to individuals who were legally obliged to file a 1988 U.S. federal
213 individual income tax return. One was required to file a return in this year if
214 household gross income (excluding nontaxable sources of income) exceeded a
215 threshold, which varied according to one's age and marital status. For example, a
216 single individual under 65 years of age was required to file a return if his gross
217 income exceeded $4950. In contrast, the threshold for a married couple with both
218 spouses over 65 years of age was $10 100.[12]
219 The members of our sample are divided into two categories, *filers* and *ghosts*,
220 according to whether they have complied with their 1988 filing requirements. As
221 discussed in Section 4 our data includes detailed line-item tax and occupation
222 information for individuals from each category. The data on filers comes from a
223 stratified random sample of the overall filer population. The data on ghosts comes
224 from a stratified random sample of the 'locatable' nonfiler population. The latter
225 population includes all ghosts who could be located through an intensive search by
226 IRS agents. Sample weights are available that make the filers and ghosts in our
227 sample broadly representative of the overall filer and locatable nonfiler popula-
228 tions, respectively. The locatable nonfiler population is of considerable policy
229 interest, because it represents the portion of the overall ghost population that the
230 IRS would be able to uncover through an intensive search and audit process.
231 However, it is also desirable to learn about the number of unlocatable nonfilers,
232 the amount of taxes that these individuals owe, and the motivations behind their
233 decision not to file an income tax return. The econometric specification presented
234 below makes it possible to draw inferences about all ghosts, whether locatable or
235 not.

236 *3.1. Model specification*

237 According to the theoretical framework presented in Section 2, an individual is
238 more likely to file a return when the likelihood of apprehension for not filing is

206

200 [12]An individual also was required to file a return if he owed certain special taxes (e.g., social security
201 tax for tips not reported to an employer); he had received advance Earned Income Credit payments
202 from an employer; he had net earnings from self-employment of at least $400; or if he had wages of
203 $100 or more from a church or qualified church-controlled organization that was exempt from
204 employer social security taxes. In addition, special rules applied for individuals who were claimed as a
205 dependent on another tax return.

high. One of the factors that will determine the likelihood of apprehension is the ease with which the tax agency can locate the individual. In our data sample, an intensive search by the IRS agents failed to locate a number of potential nonfilers. We therefore model the probability that an individual can be located jointly with the individual's filing decision. We begin by considering a specification in which the probability of being located only indirectly affects the filing decision. We then extend our specification to allow for a true simultaneous equations relationship.

Allow F^* to represent an index of the likelihood that an individual will file a return, and let L^* represent an index of the likelihood that the individual can be located. We specify the following model for these variables.

$$F^* = \beta_F' X_F + \epsilon_F \tag{6}$$

$$L^* = \beta_L' X_L + \epsilon_L, \tag{7}$$

where X_F and X_L are vectors of exogenous regressors and ϵ_F and ϵ_L are random disturbances. To complete the above model, it is necessary to specify the joint distribution of the error terms, or equivalently the joint distribution of the outcome variables. We define the binary outcomes of the filing decision as follows:

$$F = \begin{cases} 1 & \text{if the individual files a return;} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we define the marginal outcomes of the nonfiler search process as:

$$F = \begin{cases} 1 & \text{if the nonfiler is located;} \\ 0 & \text{otherwise.} \end{cases}$$

We specify a joint logistic distribution for F and L.[13] Let $P_{FL}(F = f, L = l)$ denote the joint probability that $F = f$ and $L = l$ (where $f, l \subset \{0, 1\}$). The joint probability distribution is summarized by the following equations:

$$P_{FL}(F = 1, L = 1) = \exp(\beta_F' X_F + \beta_L' X_L + K)/D \tag{8}$$

$$P_{FL}(F = 1, L = 0) = \exp(\beta_F' X_F)/D \tag{9}$$

$$P_{FL}(F = 0, L = 1) = \exp(\beta_L' X_L)/D \tag{10}$$

$$P_{FL}(F = 0, L = 0) = 1/D, \tag{11}$$

where

$$D = 1 + \exp(\beta_L' X_L) + \exp(\beta_F' X_F) + \exp(\beta_F' X_F + \beta_L' X_L + K).$$

The term K represents a measure of the strength of the correlation between the likelihood of filing and the probability of being located.

[13] See Nerlove and Press (1983), Mantel and Brown (1973), and Morimune (1979) for prior applications based on this distribution.

To understand the relationship between the above specification and an ordinary univariate logit framework, consider the implied conditional probability that F equals one given that L equals zero ($P_{F|L}(F=1|L=0)$):

$$P_{F|L}(F=1|L=0)=\frac{\exp(\beta_F'X_F)}{1+\exp(\beta_F'X_F)}.$$

This is clearly a univariate logit specification of the filing decision for those individuals who could not be located if they elected not to file. Similarly,

$$P_{F|L}(F=1|L=1)=\frac{\exp(\beta_F'X_F+K)}{1+\exp(\beta_F'X_F+K),}$$

which is a univariate logit specification of the filing decision for those individuals who could be located if they did not file. When $K=0$, we see that the above two conditional probabilities are the same, implying that F and L are independent events. When $K>0$, an individual who can be located is more likely to file than one who cannot be located, while the converse is true when $K<0$.

3.2. Allowing for simultaneity

In the above specification, the parameter K provides an indirect link between the filing decision and the probability of being located. However, it is plausible that an increase in the probability of being located would have a direct impact on one's filing choice. The following extended specification allows for this possibility:

$$F^*=\beta_F'X_F+\alpha L^*+\epsilon_F \tag{12}$$

$$L^*=\beta_L'X_L+\epsilon_L. \tag{13}$$

Observe that the propensity to be located now enters directly as a regressor for the filing decision. Since our extended model constitutes a simultaneous equations specification, it is necessary to consider model identification. The parameters of the filing equation will be identified if at least one of the regressors in X_L is excluded from the regressors in X_F.[14] We discuss our choice of exclusion restrictions below in Section 5.

To account for simultaneity within our logistic specification for F and L, we employ a limited information approach. In particular, we substitute for L^* in Eq. (12) to obtain:

$$F^*=\beta_F'X_F+\alpha\beta_L'X_L+u_F, \tag{14}$$

where $u_F=(\epsilon_F+\alpha\epsilon_L)$. From Eq. (14), it is apparent that we can account for the

[14]Note that Eq. (13) is identified even in the absence of any exclusion restrictions.

direct effect of L^* on the filing decision by including the term $\alpha\beta_L' X_L$ in our logistic specification of the joint probabilities. Our amended probability formulae are as follows:

$$P_{FL}(F=1, L=1) = \exp(\beta_F' X_F + (1+\alpha)\,\beta_L' X_L + K)/D \tag{15}$$

$$P_{FL}(F=1, L=0) = \exp(\beta_F' X_F + \alpha\beta_L' X_L)/D \tag{16}$$

$$P_{FL}(F=0, L=1) = \exp(\beta_L' X_L)/D \tag{17}$$

$$P_{FL}(F=0, L=0) = 1/D, \tag{18}$$

where D is now defined as:

$$D = 1 + \exp(\beta_L' X_L) + \exp(\beta_F' X_F + \alpha\beta_L' X_L) + \exp(\beta_F' X_F + (1+\alpha)\beta_L' X_L + K).$$

3.3. Conditional likelihood function

Our data contain detailed information pertaining to the filing decision for two groups of individuals: filers and located nonfilers. This information is not available, however, for the remaining group (unlocated nonfilers). Given the truncated nature of our sample, it is necessary to condition our analysis of the filing decision on the first two groups.

The conditional likelihood function involves separate expressions for filers and located nonfilers. For a member of the former group, our conditional likelihood expression (L_1) represents the probability that $F=1$ given that either $F=1$ or ($F=0$ and $L=1$).[15] In particular,

$$L_1 = \frac{\exp(\beta_F' X_F + \alpha\beta_L' X_L) + \exp(\beta_F' X_F + (1+\alpha)\,\beta_L' X_L + K)}{\exp(\beta_L' X_L) + \exp(\beta_F' X_F + \alpha\beta_L' X_L) + \exp(\beta_F' X_F + (1+\alpha)\,\beta_L' X_L + K).} \tag{19}$$

The conditional likelihood expression for a located nonfiler (L_2) represents the probability that ($F=0$ and $L=1$) given that either $F=1$ or ($F=0$ and $L=1$). In particular,

$$L_2 = \frac{\exp(\beta_L' X_L)}{\exp(\beta_L' X_L) + \exp(\beta_F' X_F + \alpha\beta_L' X_L) + \exp(\beta_F' X_F + (1+\alpha)\beta_L' X_L + K)}. \tag{20}$$

[15] Observe that this expression concerns the marginal probability that $F=1$, because we cannot deduce from the data whether a filer would have been located had he not filed.

341 3.4. Two-stage estimation strategy

342 Since the conditional likelihood function excludes all unlocated nonfilers from
343 the analysis, it can be expected to generate poor estimates of the likelihood that a
344 given nonfiler can be located. This is a common problem in truncated regression
345 specifications. To get around this difficulty, we take advantage of the fact that
346 although details pertaining to the filing decision (X_F) are not available for
347 unlocated nonfilers, we do have details pertaining to the chances of being located
348 (X_L) for this group. From Eqs. (17) and (18), the conditional probability that an
349 individual will be located given that he does not file is of the logistic form:

350 $$P_{L|F}(L = 1|F = 0) = \frac{\exp(\beta_L' X_L)}{1 + \exp(\beta_L' X_L)}. \quad (21)$$

351 This observation leads us to estimate the parameters of our model in two stages.
352 First, we estimate β_L by performing a univariate logit analysis of Eq. (21) using
353 our sample of located and unlocated individuals who did not file. We then
354 substitute the estimated value of β_L into the conditional likelihood function defined
355 by Eqs. (19) and (20) and estimate the remaining parameters (β_F, K, and α). The
356 standard errors for the second stage parameter estimates are adjusted to account for
357 first-stage sampling error using the procedure described in Murphy and Topel
358 (1985).

359 3.5. Choice-based sampling

360 A minor complication for our analysis is that different sampling rates were used
361 to select the filers and nonfilers in our study, resulting in a choice-based sample.
362 Manski and Lerman (1977) have shown that weighting the likelihood function by
363 the inverse of the sampling rates will generate consistent estimates for choice-
364 based samples. We therefore apply this weighting strategy in both of the stages of
365 our analysis.[16]

366 4. Description of data

367 The data used for filers of 1988 federal income tax returns is based on a 25
368 percent random subsample of the IRS TCMP Phase III Survey. This survey
369 contains the results of intensive line-by-line audits of a stratified random sample of
370 approximately 54 000 individual income tax returns for tax year 1988. For most
371 line items both the amount that was reported by the taxpayer and the amount that

340

338 [16]We adjust the standard errors of our parameter estimates to account for the weighted estimation
339 procedure using the formula presented in Manski and Lerman (1977).

the examiner determined should have been reported are available. In addition, information is recorded about the prior filing history of the taxpayer, and a code is available for the taxpayer's occupational category.[17] A set of sample weights is included to make the data representative of the national return population.[18] Selection into the 25 percent subsample was restricted to taxpayers who were required to file a 1988 return.[19]

The data on potential nonfilers is from the collection-based segment of the IRS TCMP Phase IX Nonfiler Survey for tax year 1988. This survey includes information for a stratified random sample of approximately 23 000 cases from a population of 83 million individuals for whom there was no record of a 1988 individual income tax return. These individuals were identified through a social security number match of IRS tax records with the Social Security Administration Date of Birth/Date of Death Master File, which lists all individuals with valid social security numbers.[20] The potential nonfilers identified through this match include actual ghosts, late filers, and individuals who were not required to file a return.[21] An intensive effort was made by IRS agents to locate each of the individuals in the sample. Information that was known about each individual prior to the search is available, including the individual's age, whether a return had been filed for the previous tax year, and whether third-party information return documents were available for the 1988 tax year.

A total of 18 689 of the 23 286 potential nonfilers in the sample were successfully located through the search process. The sample weights for these 18 689 individuals sum to approximately 57 percent of the potential nonfiler population.[22] Revenue officers had access to information documents and past filing records. Armed with this information they conducted interviews or field visits to determine whether a successfully located individual's income was above the filing threshold. Tax returns were secured from 3549 individuals who were deemed to have been in violation of their tax filing requirements.

A separate segment of the nonfiler survey, the examination-based segment, is used to construct variables for analyzing the filing decision. A random subsample of 2195 of the 3549 secured delinquent returns from the collection-based segment

[17]This code is recorded by the IRS examiner based on his assessment of the taxpayer's occupation.

[18]These weights do not account for returns that were filed late or for the returns of nonresident taxpayers.

[19]According to our tabulations approximately 9.7 percent of the returns in the TCMP survey, representing 10.1 million households, were not legally required to file a return. In the majority of cases these individuals voluntarily filed a return to claim a refund or an Earned Income Credit.

[20]Nonresidents and individuals without valid social security numbers were excluded from the analysis.

[21]Recall that ghosts (i.e., nonfilers) are defined as individuals who fail to file a return in violation of federal filing requirements.

[22]Unlocated individuals in the sample tended to have much larger sample weights as a consequence of the way the sample was stratified.

were subjected to intensive line-by-line audits. The information recorded in the examination-based segment of the survey is comparable to that recorded in the TCMP Phase III Survey of filers discussed previously. We have adjusted the sample weights for the secured delinquent returns in this file so that they are broadly representative of all located nonfilers from the collection-based segment.[23] An additional adjustment to the sample weights was made to convert the individual-specific sample weights into return-specific weights. This adjustment was necessary to make the data on nonfilers comparable to the data on filers, which are recorded on a return-specific basis.[24]

5. Estimation results

In this section we present the results of our analysis of taxpayer filing behavior. We first present results for the probability that a nonfiler can be located, followed by results for the decision whether to file a return.

5.1. Locating potential nonfilers

The first stage of the two-stage analysis involves univariate logit estimation of odds of being located based on a large sample of individuals who did not file a 1988 tax return. We restrict the regressors for this portion of the model to information available to the IRS prior to conducting its search for these individuals. In addition to a constant term, the following variables are used as regressors (X_L) in this stage of the analysis:

1. **Prior Yr. Filer:** Dummy variable equal to one if the individual filed a 1987 income tax return; zero otherwise.
2. **IRP Income:** Dummy variable equal to one if there is an information returns program (IRP) record of any 1988 income; zero otherwise.
3. **Prior Yr. Filer*IRP Income:** Interaction of the above two dummy variables.

[23]The collection-based segment identifies a total of 4563 individuals who failed to comply with their filing requirement, including the 3549 from whom returns were secured. The collection-based segment divides returns into 23 sampling strata based on factors such as the presence or absence of information returns, the amount of income shown on those returns, the individual's filing history, and age. Within each stratum, all individuals have the same sample weight. For each of the 23 sampling strata employed for sample selection, we adjusted the sample weights for the returns in the examination-based segment upwards so that the sum equaled the stratum total for the nonfilers in the collection-based segment.

[24]To make the adjustment, we divided the sample weights for the secured delinquent returns of married joint nonfilers by a factor of two. All else equal, a delinquent married couple's return has approximately twice the chance of being included in our sample as a delinquent single individual's return.

Table 1
Mean values of first stage regressors

Variable	Weighted sample mean
Prior yr. filer	0.0762
IRP income	0.4835
Prior yr. filer*IRP income	0.0645
Spouse	0.0980
Age 65	0.3107

4. **Age 65:** Dummy variable equal to one if the individual's age is sixty-five or greater; zero otherwise.
5. **Spouse:** Dummy variable equal to one if available records indicate a spouse; zero otherwise.

Variables pertaining to the presence of prior year tax returns and third-party information reports are included, because these documents may contain relevant information about the individual's address, his place of work, or where he holds financial accounts. The age 65 and spousal dummies are included, because it is plausible that elderly individuals and married individuals are less mobile and therefore easier to locate than young and single individuals. The weighted mean values of the regressors in our sample are presented in Table 1.

The results of our logit analysis of the probability of being located are presented in Table 2.[25] Each of the parameter estimates is of the expected sign, and they all are statistically significant. The interaction between the prior year return and IRP income dummies is negative and rather large, indicating that having access to IRP information only modestly improves the odds of locating an individual when there is already a record of a prior year return.[26]

Table 2
Results of estimation — probability of being located[a]

Variable	Estimate	*t*-statistic
Constant	−1.1577	−77.46
Prior yr. filer	2.4027	3.83
IRP income	2.8288	75.19
Prior yr. filer*IRP income	−2.6725	−4.12
Spouse	1.9070	16.26
Age 65	0.2434	8.71

[a] Number of observations: 23 283; value of log-likelihood function: −11 124.8.

[25] The analysis incorporates the sampling weights, which make the observations representative of the overall population of individuals who did not file a return.

[26] For example, the probability of locating a single individual under 65 years of age rises from 77.6 percent to 80.2 percent when IRP information also becomes available.

Table 3
Observed and predicted outcomes of search for nonfilers[a]

Observed	Predicted		Total
	$L = 0$	$L = 1$	
$L = 0$	31.5 million	6.6 million	38.1 million
$L = 1$	10.0 million	40.4 million	50.4 million
Total	41.5 million	46.9 million	88.4 million

[a] Pseudo R^2: 0.3008.

Table 3 provides some measures of model fit. Overall, our logit specification performs well, correctly classifying over 80 percent of all located and unlocated individuals. The pseudo-R^2 for the specification is a respectable 30 percent.[27]

5.2. The decision whether to file

In the second stage of our analysis, we estimate the remaining parameters of our model using a data sample containing information on both filers and located nonfilers. These estimates are based on the conditional likelihood function presented in Eqs. (19) and (20). In addition to the constant term, the following variables are included as regressors (X_F) for the filing decision:

1. **Prior Yr. Filer** Dummy variable equal to one if the individual filed a 1987 income tax return; zero otherwise.
2. **Filing Burden:** An IRS estimate of the number of hours required to complete the tax return.
3. **Filing Threshold:** A dummy variable equal to one if the individual's gross income is within 5 percent of the filing threshold level for his age and filing status; zero otherwise.
4. **Burden*Threshold:** Interaction between the above two variables.
5. **State Tax:** Dummy variable equal to one for residence in a jurisdiction with a state-level income tax; zero otherwise.
6. **Business Income:** Dummy variable equal to one if Schedule C (business) income or loss is present; zero otherwise.
7. **Farm Income:** Dummy variable equal to one if the Schedule F (farm) income or loss is present; zero otherwise.
8. **Professional:** Dummy variable equal to one if the individual is a professional;

[27] This measure is computed as $1 - \ln L_\Omega / \ln L_\omega$, where $\ln L_\Omega$ is the value of the log-likelihood function for our model, and $\ln L_\omega$ is the value of the log-likelihood function when the model is restricted to have no regressors other than a constant term.

zero otherwise. (This dummy is excluded from the analysis, making this the omitted occupation category.)

9. **Supervisor:** Dummy variable equal to one if the individual is a supervisor or manager; zero otherwise.
10. **Service/Admin. Support:** Dummy variable equal to one if the individual works in a service occupation (including transportation) or provides administrative support; zero otherwise.
11. **Ag./For./Fishing** Dummy variable equal to one if the individual is employed in an agriculture, forestry, or fishing occupation; zero otherwise.
12. **Mechanic/Helper:** Dummy variable equal to one if the individual is a mechanic, helper, or handler; zero otherwise.
13. **Constr./Extrac./Prod.:** Dummy variable equal to one if the individual works in a construction, extraction, or production occupation; zero otherwise.
14. **Military:** Dummy variable equal to one if the individual works in the military; zero otherwise.
15. **Other:** Dummy variable equal to one if the individual doesn't work in any of the above occupations; zero otherwise.
16. **Age 65:** Dummy variable equal to one if the individual's age is 65 or greater; zero otherwise.
17. **Married:** Dummy variable equal to one if the individual's filing status is married joint return; zero otherwise.
18. # **Dependents:** Number of dependents.
19. **Unemployment Income:** Dummy variable equal to one if the individual received unemployment income; zero otherwise.
20. **AGI:** Adjusted gross income divided by \$100 000. (If AGI is negative, AGI is set equal to zero.)
21. **Locatability:** Index of the likelihood of being located (equal to $\beta_L' X_L$ in Eq. (14)).

The variables related to income, occupation, and filing status were based on the examiner-determined values rather than those originally reported by the taxpayer. Due to noncompliance, the former are likely to be more representative of the true values of these variables.

As discussed in Section 2, the decision whether to file a return should depend on an individual's past filing behavior, the burden associated with filing, the opportunities for successfully underreporting income, and the chances of being caught and penalized for not filing. The dummy variable for the presence of a 1987 tax return is included to account for the individual's past filing history. As a measure of the filing burden, we employ an IRS formula to estimate the number of hours it would take to complete a tax return given the sources of the individual's income and deductions. We also include a dummy variable for whether an individual's income is close to the filing threshold and an interaction between the

burden measure and the threshold dummy. Our intuition is that an individual may elect not to file if his income is only marginally above the threshold, particularly if his return is difficult to complete.[28]

The dummy variable for residence in a jurisdiction with a state income tax might be expected to have a positive association with filing a return. To the extent that such states also have nonfiler detection programs and share information with the federal government, an individual from a state with its own tax may perceive a greater risk of penalty for not filing. It is difficult to predict the sign on the business and farm income dummies a priori. An individual with these sources of income may have relatively good opportunities for underreporting income if he files. On the other hand, to the extent that his income from these sources is 'off-the-books', he may have relatively good opportunities for not filing as well.[29] We control for the influence of a variety of occupations on the filing decision. We also control for a number of demographic characteristics, including age (whether age 65 or over), marital status, number of dependents, receipt of unemployment insurance, and income. The final explanatory variable is an index of the likelihood that an individual could be located if he were to become a nonfiler. We anticipate that this variable will have a positive relationship with the filing decision.

As discussed in Section 3, at least one regressor from the first stage of our analysis (for the probability of being located) must be excluded from our filing equation to identify the parameters of this equation. We have excluded the two terms from the first stage that involve the presence of income subject to third-party information reporting.[30] Our assumption is that third-party information reports influence the filing decision only indirectly, by raising the likelihood that the individual will be located and apprehended if he chooses not to file.[31] The weighted mean values of all regressors in our data sample for the second stage are presented in Table 4. The table includes both the means based on the overall sample and the means based on the subsample of located nonfilers.

Table 5 presents the results of our analysis of the decision whether to file an income tax return. In addition to providing the estimated parameter values and associated *t*-statistics, we have included estimates of the marginal effect for each variable on the unconditional probability of filing. These estimates reflect the

[28]Taxpayers may be able to reduce their filing burden by paying a tax practitioner to complete their returns. Refer to Erard (1997) for an analysis of the decision to use a tax preparer and its consequences for reporting compliance.

[29]As discussed by Simon and Witte (1982) it is commonly believed that individuals with substantial 'off the books' income are disproportionately represented among the nonfiler population.

[30]Specifically, these terms are IRP Income and Prior Yr. Filer*IRP Income.

[31]The Spouse dummy variable in the first stage equation also differs somewhat from the Married dummy variable in the filer equation, because the former variable is based on information from the previous year's records.

Table 4
Mean values of second stage regressors

Variable	Weighted mean overall sample	Weighted mean ghost subsample
PRIOR YR. FILER	0.9177	0.2474
IRP income	0.9837	0.8053
Pri. yr. filer*IRP Inc.	0.9116	0.2350
Spouse	0.4148	0.1614
Age 65	0.1022	0.0743
Filing burden	14.103	13.406
Filing threshold	0.0807	0.3167
Burden*threshold	0.7000	3.3129
State tax	0.8144	0.8123
Business income	0.1474	0.3040
Farm income	0.0250	0.0095
Supervisor	0.1092	0.0893
Service/admin. suppt.	0.2288	0.2035
Ag./for./fishing	0.0218	0.0152
Mechanic/helper	0.0958	0.2271
Constr./extrac./prod.	0.1307	0.0639
Military	0.0517	0.0065
Other	0.2425	0.2729
Married	0.4983	0.2951
#Dependents	0.6572	0.4731
Unempl. income	0.0749	0.0457
AGI	0.3184	0.1732
Locatability	2.2098	1.4124

marginal change in the probability of filing a return in response to a one unit increase in a given variable, holding all other variables fixed.[32]

The marginal effect for a given variable will tend to vary according to the values of the regressors being held fixed. For this reason, two separate sets of marginal effects are provided. The first set is computed using the weighted mean values of the variables over the entire sample. The second set is computed using the weighted mean values of the variables over the subsample of nonfilers. Thus, the first set will provide an indication of the marginal effect for an individual with the average characteristics of the overall population, while the second will provide

[32] The unconditional filing probability is:

$$\frac{\exp(\beta_F' X_F + (1+\alpha)\beta_L' X_L + K) + \exp(\beta_F' X_F + \alpha\beta_L' X_L)}{1 + \exp(\beta_L' X_L) + \exp(\beta_F' X_F + \alpha\beta_L' X_L) + \exp(\beta_F' X_F + (1+\alpha)\beta_L' X_L + K)}.$$

The value of $\beta_L' X_L$ is held constant in the computation of the marginal effects of all variables other than the index, itself.

678 Table 5
679 Results of estimation — probability of filing[a]

Variable	Parameter estimate	*t*-statistic	Marginal effect at full sample mean	*t*-statistic	Marginal effect at ghost subsample mean	*t*-statistic
Constant	−10.208	−15.858				
Prior yr. filer	4.036	22.133	0.3586	11.295	0.5986	15.719
Filing burden	0.005	0.385	0.0001	0.382	0.0012	0.385
Filing threshold	0.147	0.632	0.0019	0.247	0.0344	0.235
Burden*threshold	−0.094	−2.143	−0.0013	−2.146	−0.0223	−2.190
State tax	−0.168	−0.935	−0.0022	−0.992	−0.0391	−0.938
Business income	−1.424	−5.458	−0.0347	−3.178	−0.3388	−5.704
Farm income	−0.212	−0.458	0.0033	0.415	−0.0511	−0.449
Supervisor	−0.477	−2.634	−0.0161	−4.381	−0.1440	−3.916
Service/admin. suppt.	0.612	2.453	0.0058	2.099	0.1538	3.040
Ag./for./fishing	1.075	2.254	0.0082	2.754	0.2034	2.875
Mechanic/helper	−0.852	−3.576	−0.0291	−3.784	−0.2850	−5.686
Constr./extrac./prod	1.241	4.409	0.0111	6.375	0.2425	5.655
Military	−0.376	−0.977	−0.0123	−1.334	−0.1072	−1.185
Other	0.281	1.060	0.0005	0.140	0.0710	1.227
Age 65	−0.659	−2.513	−0.0121	−2.027	−0.1617	−2.490
Married	−0.030	−0.171	−0.0004	−0.171	−0.0072	−0.170
# Dependents	0.076	1.058	0.0011	1.039	0.0180	1.057
Unempl. income	−0.664	−3.925	−0.0124	−3.063	−0.1634	−3.892
AGI	−0.017	−0.456	−0.0002	−0.457	−0.0040	−0.456
Locatability	0.435	2.822	0.0061	2.923	0.1029	2.705
K	10.055	22.747				

[a] The marginal effect represents the change in filing probability for a 1 unit increase in an explanatory variable. In the case of a dummy variable, it represents the change in filing probability when the dummy value shifts from zero to one; for an occupation dummy, the effect is computed as the change in filing probability from when the dummy equals zero and the other occupation dummies are evaluated at the sample mean values to when the dummy equals one and all other occupation dummies set equal to zero. (The omitted occupation is Professional.) Number of observations: 15 489; value of log-likelihood: −1648.1.

an indication of the marginal effect for an individual with the average characteristics of the ghost population.[33]

As expected, there is substantial persistence in filing behavior. An individual

[33] For a given occupation dummy variable, this marginal effect is computed by taking the difference between the probability of filing when that occupation dummy is equal to one, the remaining occupation dummies are all zero, and the other variables are held at their mean values, and the probability of filing when that occupation dummy is zero and the other occupation dummies and all other variables are held at their mean values. The marginal effects for the non-occupation dummies are computed as the difference between the probability of filing when the dummy is equal to one and all other variables are held at their mean values and the probability of filing when the dummy is equal to zero and all other variables are held at their mean values.

who filed in the previous year is very likely to file in the current year. The first set of marginal results (based on the overall sample variable means) indicates that having filed last year increases the probability of filing this year by 36 percent. The second set of marginal results (based on the nonfiler subsample variable means) indicates that having filed previously raises the chances of filing in the current year by 60 percent! As discussed in Section 2, one explanation for the observed persistence in filing behavior is that a change in behavior might serve as a signal to the tax authority that enforcement action is warranted. For example, if an individual with no previous filing history completes a return, this may prompt the tax authority to investigate whether previous returns also should have been filed. Similarly, if an individual has routinely filed in previous years, the tax authority may find it suspicious if he should suddenly stop filing. An alternative interpretation of the observed persistence of filing behavior is that filing is a learned responsibility. Under this interpretation, some individuals fail to file simply because they are unaware of their filing obligation. It follows that if they should learn of their obligation, they will begin filing returns and continue doing so in future years.[34]

The estimated marginal effects for the burden and threshold variables are statistically insignificant. However, the marginal effect for the interaction between these variables is negative and significant. For an individual whose income is near the filing threshold, the estimated marginal effect of a 1 h increase in the time necessary to complete a return is about a two percent rise in the probability of filing (based on the sample mean characteristics of the ghost population).[35] One interpretation of this finding is that the burden of completing a return serves as a deterrent to filing for individuals with relatively low income (and hence, relatively low tax liability). An alternative interpretation is that individuals with low income are relatively less likely to be aware of their filing obligation or invest in learning about it. Under this interpretation, the measure of filing burden may be thought of as a proxy for the transparency of the individual's filing obligation. In other words, filing requirements may seem more obvious under simpler tax circumstances (i.e., when the filing burden is low). Consequently, low income individuals with low measures of tax burden may be relatively more likely to file than low income individuals with more complex tax circumstances.

Individuals with business income are relatively less likely to file a return. Among the different occupation categories, mechanics and helpers are the least likely to file (other factors equal). Presumably, their income is more easily concealed than that of workers in many other occupations (e.g., Professionals). Perhaps surprisingly, the results indicate that individuals employed in construction, extraction, and production are the most likely to file.

The elderly and the unemployed are relatively less likely to file. However, the

[34] We thank a Referee for pointing out this alternative interpretation.

[35] The estimated marginal effect remains at about two percent if one restricts the coefficients for the burden and threshold variables to zero.

other demographic controls (marital status, number of dependents, and adjusted gross income) are not significantly related to the filing decision.[36]

The estimated coefficient of the index for the likelihood that an individual can be located is positive and significant. A one unit increase in this index, evaluated at the weighted mean value of the index for the ghost population, results in an 11.4 percent increase in the likelihood of being located. The estimated marginal effect of 10.3 percent is therefore quite large, suggesting nearly a one-to-one relationship between the likelihood of being located and the probability of filing.

The estimated value of parameter K, which measures the strength of the correlation between the probability of filing and the probability of being located is also positive and significant. This indicates that unobserved factors which make an individual easier to locate also tend to make him likely to file.

Table 6 provides some measures of the fit of our specification for the likelihood that an individual will file a return. About 95 percent of the individuals in our weighted sample filed a 1988 federal income tax return. The model correctly classifies all but one percent of these individuals as filers. Not surprisingly, the model also classifies a number of the nonfilers in our sample as filers. However, the model does demonstrate a significant amount of discriminatory power. About 43 percent of the nonfilers are correctly classified, and the pseudo-R^2 for the specification is 45.2 percent.

The results from our structural model rely on the validity of our exclusion restrictions; specifically, the exclusion of the variables relating to the presence of third-party information reports from the filing equation. As discussed previously, we have assumed that these variables only indirectly affect the filing decision through their impact on the likelihood that one will be located if he chooses not to file. To examine the sensitivity of our results to this identifying assumption, we have estimated the reduced form version of our model. In this version, our index for the likelihood of being located is replaced as a regressor in the filing equation

Table 6
Observed and predicted filing outcomes[a]

Observed	Predicted		Total
	$F=0$	$F=1$	
$F=0$	2.0 million	2.7 million	4.7 million
$F=1$	0.8 million	91.7 million	92.5 million
Total	2.8 million	94.4 million	97.2 million

[a] Pseudo R^2: 0.4520.

[36] Observe that income does play an indirect role in the filing decision through the burden-filing threshold interaction term. As noted previously, filing by individuals with income near the threshold is sensitive to the level of burden they face in completing their returns.

with the third-party information report variables.[37] Not surprisingly, we find that the likelihood of filing increases when third-party information reports are available. The estimated marginal effects of the remaining regressors on the likelihood of filing are quite similar to the estimated effects of these variables in our structural specification. Thus, regardless whether the risk of being located is given a direct or an indirect role in the filing decision, our main findings seem to be robust.

6. Filer and nonfiler characteristics

In this section we employ the results of our econometric analysis to generate statistics on nonfiler income, adjustment, and deduction characteristics. We compare these statistics with the corresponding values from the filer population.

We provide separate estimates for the 'locatable' ghost and overall ghost populations. The former population is defined as the set of ghosts who would be located if an intensive search were performed by the IRS for all potential nonfilers. The latter is defined as the entire ghost population, including those ghosts who would not be located through an intensive search. To generalize our located nonfiler results to the overall ghost population, we adjust the sample weights for located nonfilers using the first-stage probability estimates from the two-stage analysis of Section 5. Specifically, the original sample weight for each located nonfiler is divided by the logit-based estimate of the probability that the individual would be located. Our statistics for the overall ghost population are then computed based on the adjusted weights. Our statistics for the filer population are based on a weighted analysis of the complete TCMP Phase III Survey data file, excluding those taxpayers who were not required to submit a return. Again, the statistics are computed using the examiner-determined values for the relevant variables.

Table 7 summarizes income and deductions for filers, locatable ghosts, and all ghosts. Relative to ghosts, filers tend to have substantially larger incomes. For example, their total income before adjustments is on average over two and one-half times larger than that of nonfilers. Taxable income for filers represents 68.8 percent of total income before adjustments. For ghosts, taxable income represents 71 percent of total income before adjustments, indicating that nonfilers have relatively fewer offsets to income. Intuitively, ghosts have little incentive to participate in tax planning. Similarly, nonfilers are relatively less likely to have itemized deductions in excess of the standard deduction threshold. Interestingly, though, among those ghosts whose deductions exceed the threshold, the average total deduction is actually larger than that of filers who itemize. Table 7 also

[37] In the reduced form specification, the spousal dummy variable also enters as a regressor in this equation.

Table 7
Mean income and deductions for filers and ghosts, tax year 1988[a]

	Filers	Ghosts	
		Locatable	All
Mean total income (before adjustments)	$32 376	$15 974	$12 448
Mean taxable income	$22 276	$11 349	$8838
Percentage of itemizers	32.12%	9.66%	6.63%
Mean total deductions among itemizers	$11 832	$13 061	$12 911

[a] Statistics weighted to be representative of all filers who are required to file, all locatable ghosts, and all ghosts, respectively.

indicates that income is on average larger for locatable ghosts than for the overall ghost population. However, their mean income is still only about half that of filers.

Table 8 displays income, adjustment, and itemized deduction amounts as a percentage of total income before adjustments for filers, locatable ghosts, and all ghosts. Wages and salaries, interest, dividends, and pension income make up a much more substantial share of total income for filers than nonfilers, while business income and net capital gains receipts are relatively more important for nonfilers. The findings for wages and salaries and business income reflect the fact that the ghost population includes a disproportionate share of self-employed individuals. The findings for interest, dividends, and pension income may reflect an aversion by nonfilers to leaving a paper trail. A possible explanation for the

Table 8
Income and offsets as a percentage of total income for filers and ghosts, tax year 1988[a]

	Filers	Ghosts	
		Locatable	All
Income items			
Wages and salaries	72.73%	61.56%	69.89%
Taxable interest	5.78%	4.87%	4.34%
Dividends	2.29%	0.64%	0.58%
Taxable pensions	4.24%	2.92%	2.55%
Taxable soc. sec.	0.48%	0.17%	0.15%
Unemployment comp.	0.37%	0.54%	0.49%
Net business (Sch. C)	5.17%	20.85%	14.27%
Net farm (Sch. F)	0.11%	0.54%	0.51%
Net cap. gains (Sch. D)	4.77%	10.75%	10.05%
Net. supplemental (Sch. E)	2.24%	0.08%	0.07%
All other	1.82%	−2.92%	−2.90%
Total adjustments	0.82%	0.40%	0.34%
Total itemized deductions	11.74%	7.90%	6.85%

[a] Statistics weighted to be representative of all filers who are required to file, all locatable ghosts, and all ghosts, respectively.

capital gains finding is that nonfilers have relatively less incentive to offset taxable capital gains with capital losses. Perhaps for similar reasons, discretionary adjustments and itemized deductions tend to be relatively less important as a share of total income for nonfilers than they are for filers.

7. Net tax liability

We have used our adjusted sample weights for located nonfilers to generate an estimate of the net tax liability of the overall ghost population.[38] The results indicate that ghosts were responsible for approximately $5 billion in unpaid income taxes for tax year 1988, after accounting for tax prepayments such as taxes withheld and estimated tax payments they had made. Approximately 43 percent of all nonfilers made at least some form of prepayment, compared to 93 percent of filers.[39] Overall, prepayments by nonfilers covered about half of their aggregate income tax liability.

Not all individuals who are required to file a return owe taxes. In fact, our estimates indicate that 29 percent of all ghosts had no tax liability for tax year 1988. Moreover, we estimate that 22.2 percent of the overall nonfiler population for this year would have been entitled to a refund if they had filed a return. The median size of this refund would have been $407, a figure which presumably exceeded the burden of filing in many cases. It therefore seems plausible that some of these nonfilers were unaware of the magnitude of the refund to which they were entitled.

In addition to the $5 billion in aggregate unpaid income taxes, our estimates indicate that nonfilers owed approximately $2.8 billion in self-employment taxes. Our estimates tend to understate the true unpaid tax liability of ghosts, because even experienced examiners are unable to uncover all income that has gone unreported. In its most recent tax gap report (U.S. Internal Revenue Service, 1996), the IRS has used an approach similar to ours to estimate the nonfiler tax gap.[40] However, its estimate includes a sizeable adjustment that attempts to account for any income that might not have been detected during the audits. The official IRS estimate of nonfiler net income tax liability (excluding self-employ-

[38] The estimate accounts both for ghosts who would be located if an intensive search and audit process were carried out and ghosts who would not be located.

[39] Approximately 41 percent of nonfilers had at least some income taxes withheld, while 4.3 percent made at least one installment payment of estimated taxes. The comparable figures for filers are 86.8 percent and 12.2 percent, respectively.

[40] In the preliminary stage of our research, we employed a probit analysis of the probability an individual could be located rather than a logit analysis. The results were quite similar. The IRS employed our probit analysis in generating its tax gap estimates using a somewhat different weighting scheme than that employed in this study.

954 ment taxes) for tax year 1988 amounts to $11 billion after adjusting for undetected
955 noncompliance. No official estimate is available for understated self-employment
956 taxes.
957 The estimated size of the ghost population based on our approach is 7.9
958 million.[41] The IRS estimate of the tax gap for the 110 million filers of tax year
959 1988 returns is $73 billion. Thus, while we find that the number of ghosts is only
960 about 7 percent (i.e., 7.9/110) as large as the number of filers, the nonfiler tax gap
961 is approximately 15 percent (i.e., 11/73) as large as the filer tax gap.
962 As discussed previously, even an intensive search by the IRS was unable to
963 locate all potential nonfilers. However, as shown in Table 7, locatable nonfilers
964 tend to have higher incomes (and hence, higher tax liabilities) than ghosts who
965 cannot be located. In fact, our results (based on detected net tax liabilities) indicate
966 that approximately 82 percent of the overall nonfiler tax gap is attributable to
967 locatable nonfilers.

968 ## 8. Conclusion

969 Nonfilers have been a neglected group in theoretical and empirical research on
970 tax compliance. Much of this neglect has been due to the lack of reliable
971 information about their characteristics, a problem so severe that nonfilers are
972 sometimes referred to as 'ghosts' by academics and policy-makers. This study
973 provides important evidence on the characteristics of nonfilers and the taxes for
974 which they are liable. We find that nonfiling is more prevalent among self-
975 employed individuals and within occupations where income may be more easily
976 concealed from the tax authority, such as mechanics and helpers. In addition, for
977 taxpayers with incomes near the filing threshold, the burden associated with
978 completing a return appears to serve as a deterrent to filing. Thus, initiatives that
979 reduce the burden of filing (such as existing taxpayer assistance programs and
980 simplified tax returns) may encourage individuals with relatively low incomes to
981 file. Moreover, to the extent that the failure to file is due to an ignorance of the tax
982 laws (and even of potential tax refund opportunities), programs to educate
983 individuals about filing requirements may be useful. Our results indicate that there
984 is substantial persistence in filing behavior. Thus, once a ghost is brought into the
985 system, he is likely to remain in the system.
986 Identifying ghosts and encouraging them to file is a challenging task. The results
987 of this study indicate that only 57 percent of the potential nonfiler population could
988 be located through an intensive search. However, locatable nonfilers apparently
989 account for a disproportionate share of all unpaid taxes. Thus, a substantial portion
990 of the nonfiler tax gap is at least potentially collectable. The extent to which it is
953

951 [41]This is a return-based estimate, meaning that it represents the number of returns that should have
952 been filed but were not.

cost-effective and/or socially desirable to search out nonfilers and recover taxes is an important question for future research.

9. Uncited reference

Graeber et al., 1992

Acknowledgements

An earlier draft of this paper was completed while the first author was an Associate Professor at Carleton University. The current draft was largely completed while he was on sabbatical as a Visiting Associate Professor and Office of Tax Policy Research Fellow at the University of Michigan. We are grateful to the Internal Revenue Service for providing us with access to the data used in this analysis. We would particularly like to acknowledge Jeff Colson, Dennis Cox, Carol Sattler, Arthur Sparrow, and Joel Stubbs for many helpful discussions about the data. We also thank Jim Alm, Matt Murray, Joel Slemrod, and two anonymous referees for valuable comments on an earlier draft. The first author thanks the Social Sciences and Humanities Research Council of Canada for financial assistance. Any opinions expressed in this paper are those of the authors; they do not necessarily represent the views of the Internal Revenue Service.

References

Andreoni, J., Erard, B., Feinstein, J.S., 1998. Tax compliance. Journal of Economic Literature 36, 818–860.

Allingham, M.G., Sandmo, A., 1972. Income tax evasion: a theoretical analysis. Journal of Public Economics 1 (3/4), 323–338.

Alm, J., Bahl, R., Murray, M.N., 1991. Tax base erosion in developing countries. Economic Development and Cultural Change 39 (4), 849–872.

Blumenthal, M., Slemrod, J., 1992. The compliance cost of the U.S. individual income tax system: a second look after tax reform. National Tax Journal 45 (2), 185–202.

Cowell, F.A., 1990. Cheating the Government: The Economics of Evasion. M.I.T. Press, Cambridge.

Cowell, F.A., Gordon, J.P.F., 1995. Auditing with ghosts. In: Fiorentini, G., Peltzman, S. (Eds.), The Economics of Organised Crime. Cambridge University Press, Cambridge, pp. 185–186.

Crane, S.E., Nourzad, F., 1993. An empirical analysis of factors that distinguish those who evade on their tax return from those who do not file a return. Mimeo, Marquette University.

Cross, R.B., Shaw, G.K., 1982. On the economics of tax aversion. Public Finance 37, 36–47.

Engel, E.M.R.A., Hines Jr., J.R., 1999. Understanding tax evasion dynamics. N.B.E.R. Working Paper No. W6903.

Erard, B., 1992. The influence of tax audits on reporting behavior. In: Slemrod, J. (Ed.), Why People Pay Taxes: Tax Compliance and Enforcement. The University of Michigan Press, Ann Arbor, pp. 95–114.

1030 Erard, B., 1997. Self-selection with measurement errors: a microeconometric analysis of the decision to
1031 seek tax assistance and its implications for tax compliance. Journal of Econometrics 52 (2),
1032 163–197.
1033 Graeber, M.J., Nichols, B.L., Sparrow, D., 1992. Characteristics of delinquent returns. U.S. Department
1034 of the Treasury, Internal Revenue Service, The IRS Research Bulletin, Publication 1500, pp. 38–46.
1035 Manski, C., Lerman, S., 1977. The estimation of choice probabilities from choice-based samples.
1036 Econometrica 45.
1037 Mantel, N., Brown, C., 1973. A logistic reanalysis of Ashford and Bowden's data on respiratory
1038 symptoms in British coal miners. Biometrics 22, 649–665.
1039 Morimune, K., 1979. Comparisons of normal and logistic models in the bivariate dichotomous analysis.
1040 Econometrica, 47.
1041 Murphy, K.M., Topel, R.H., 1985. Estimation and inference in two-step econometric models. Journal of
1042 Business and Economic Statistics 3, 370–377.
1043 Nerlove, M., Press, S.J., 1983. Univariate and multivariate log-linear and logistic models. Rand
1044 Corporation Working Paper Number R-1306-eda/nih.
1045 Simon, C.P., Witte, A.D., 1982. Beating the System: The Underground Economy. Auburn House,
1046 Boston.
1047 Slemrod, J., 1995. A general model of the behavioral response to taxation. Mimeo, University of
1048 Michigan.
1049 Yitzhaki, S., 1974. A note on income tax evasion: a theoretical analysis. Journal of Public Economics 3
1050 (2), 201–202.
1051 U.S. Internal Revenue Service, 1996. Federal Tax Compliance Research: Individual Income Tax Gap
1052 Estimates for 1985, 1988, and 1992. U.S. Department of the Treasury, Internal Revenue Service,
1053 Publication 1415 (Rev. 4-96), Washington, D.C.