Compliance Measurement and Workload Selection with Operational Audit Data

by

Brian Erard

Prepared for the Internal Revenue Service Research Conference

George Washington University

June 11-12, 2002

B. Erard & Associates

A. Introduction

The degree of voluntary compliance is a fundamental, but elusive measure of the health of a tax administration. Indeed, difficulties in uncovering tax violations are the raison-d'être for tax noncompliance. In the face of measurement difficulties, most tax agencies have shied away from attempting to estimate the extent to which taxpayers voluntarily report their taxes, and instead have relied on proxy measures, such as the share of tax revenue that comes in without direct enforcement effort. A difficulty with such proxy measures is that they typically provide an incomplete and distorted perspective on the actual compliance environment. For example, the share of revenue coming in without direct enforcement effort could be brought to 100 percent by simply abandoning any attempts at enforcement; undoubtedly, however, such a practice would result in substantially less – not more – compliance.

An important exception to the reliance on proxy measures has been the approach undertaken by the Internal Revenue Service (IRS), which periodically through tax year 1988 conducted an ambitious Taxpayer Compliance Measurement Program (TCMP) for measuring the degree to which taxpayers voluntarily reported their tax obligations. This program, which involved intensive audits of a stratified random sample of federal income tax returns, allowed the IRS to measure at least with some plausible degree of precision the size of the income tax reporting gap.[1] However, the TCMP applied only to income taxes reported by individuals and small corporations; no comparable random samples were available for measuring the degree of income tax compliance for larger corporations or the degree of compliance associated with other federal taxes. Further, TCMP studies of individuals were conducted about every three years, which made it difficult to draw inferences about voluntary compliance during the interim periods between studies.[2]

Recently the IRS has announced a new program - the National Research Program (NRP), which like the TCMP will collect compliance information from a stratified random

[1] The precision of the TCMP estimates was influenced not only by the size of the data samples (about 50,000 in the case of individual income tax returns) but also by the need to make a substantial correction for undetected noncompliance.

[2] TCMP studies of small corporations were conducted less frequently.

sample of individual income tax returns on a periodic basis for use in generating measures of the income tax reporting gap. Also like the TCMP, studies under the new program will presumably be conducted only periodically, and they will not cover all federal taxes and taxpayer groups.

Although results from random audit studies for measuring voluntary compliance tend to be few and far between, the flow of data from operational audit programs tends to be continuous and plentiful. For example, the IRS regularly examines over 700,000 individual income tax returns each year under its operational audit program; in contrast, roughly 50,000 individual returns were examined every three years or so under the former TCMP. It therefore seems natural to explore potential ways to derive population measures of noncompliance from operational audit data. If a valid methodology can be employed, the large number of operational audit cases would permit the development of measures of noncompliance not only at the national level, but also potentially at the district or regional levels. A fundamental difficulty in using operational data is that returns selected for audit tend to be chosen not at random, but rather because they are considered likely to contain substantial amounts of unreported taxes. To the extent that workload selection methods are effective in identifying high risk returns for audit, levels of noncompliance will tend to be greater for audited returns than for unaudited returns, and this difference must be accounted for when developing projections of noncompliance for unaudited returns.

B. Erard & Associates, in collaboration with Professor Jonathan Feinstein of Yale University, has developed a statistical methodology to address this issue. Under prior contracts with the IRS, B. Erard & Associates has employed variants of this methodology to develop estimates of the estate tax reporting gap (Erard, 1999)[3] and to evaluate noncompliance with respect to the reporting of self-employment income by sole proprietors (Erard and Feinstein, 2001).

In Part B of this paper, we sketch out the methodology we have developed for the above studies of noncompliance and summarize the key findings. We then investigate possible extensions of our methodology in Part C. The first extension we consider is the modification

[3] For an extension of this methodology to analyze the determinants of noncompliance, see Eller, Erard, Ho (2000)

of our methodology to derive separate estimates of the magnitudes of noncompliance on multiple line items of a tax return. We then consider how the methodology might be extended to account for noncompliance that goes undetected during an audit. Finally, we consider ways that operational audit results could be used to update measures of noncompliance in the interim periods between random audit studies, such as those planned under the NRP.

Closely related to the issue of compliance measurement is the problem of workload selection. In particular, statistical models that can predict the likelihood and magnitude of noncompliance associated with alternative sets of tax return characteristics are potentially useful for identifying good candidates for audit as well as for developing measures of voluntary compliance. In the absence of random audit programs, federal and state tax agencies do frequently rely on operational audit data for the development of workload selection criteria, in some cases with the aid of statistical techniques such as regression and discriminant analysis. Typically, however, these agencies treat the underlying operational data as though they constitute a random and representative sample from the overall return population, failing to account for the fact that most returns in the sample were selected because they were considered to be at high risk of noncompliance. As a consequence, the estimated relationship within the operational audit sample between the magnitude of noncompliance and reported line item characteristics may not hold for returns within the general population, thereby weakening the effectiveness of the resulting workload selection criteria. In Part D of this paper, we consider how our methodology might be used to account for the unrepresentative nature of an operational audit sample, potentially allowing improved inferences to be drawn about the relationship between reported line item return characteristics and noncompliance within the general population, and hence more effective targeting of noncompliant taxpayers.

In Part E, we offer some concluding remarks regarding the usefulness of our methodology for both compliance measurement and workload selection.

B. Methodology of B. Erard & Associates

To estimate the tax reporting gap using operational audit data, it is necessary to develop a methodology to accurately extrapolate from measures of compliance for a population of audited taxpayers to the much larger population of filers not subjected to audit in a given year. The fundamental challenge associated with using operational audit data to estimate the overall extent of noncompliance in a population is the fact that, because of the way returns are selected, those subjected to operational audits are relatively more likely to understate taxes by substantial amounts. In general, returns selected for operational audits will tend to differ from returns not selected both in terms of their recorded and unrecorded characteristics. and both types of differences may be relevant to compliance.

B1. Accounting for differences in recorded characteristics

As a result of examination selection criteria, such as the Discriminant Function (DIF) score for individual income tax returns, the composition of audited and unaudited returns tends to differ in systematic ways. For example, returns reporting certain sources of income (e.g., rents) or expenses (e.g., depreciation) may receive greater audit coverage than returns not reporting these items. One can control for the extent to which noncompliance tends to be larger on returns reporting such characteristics through regression techniques. For example, regression equations describing the likelihood and magnitude of noncompliance can be specified that include dummy explanatory variables (regressors) for the presence of various types of recorded line item information. The estimated coefficients of these regressors would then allow one to account for the extent to which having or not having, for example, rental income, influences the likely level of noncompliance.

B2. Accounting for differences in unrecorded characteristics

Audited and unaudited returns also tend to differ in terms of certain unrecorded characteristics, because returns that meet selection criteria based on recorded information are not automatically subjected to audit. Typically, an experienced examiner known as a "classifier" first reviews the return and makes a determination whether the return should be audited, and if so, what issues should be investigated. The classifier will normally have access to information at this stage that is not recorded for later analysis. For example, a classifier will typically be able to review letters, receipts, and other supporting information that is attached to the return. Such information is not normally transcribed onto IRS databases during return processing. To the extent that the decision whether to audit a return is influenced by unrecorded information, there is the potential for noncompliance prediction formulae based on audited returns to produce misleading estimates when applied to unaudited returns with similar recorded characteristics. This phenomenon, known as "selection bias" in the econometric literature, results from an inability to control directly for unrecorded characteristics related to noncompliance that result in some returns being selected for audit and others not being selected.

To illustrate this phenomenon, consider the following hypothetical scenario. Based on past experience showing a problem with improper claims for charitable donations, all returns claiming more than $1,000 in donations are flagged for possible audit and sent to a classifier. Half of the flagged returns contain legitimate receipts that substantiate the reported claims, so the classifier determines that only the half not containing receipts should be examined. The audit results from these returns indicate that charitable donations are overstated by an average of 25 percent. An analyst, unaware of the fact that the unaudited returns claiming substantial donations all contain legitimate receipts, incorrectly infers that the claims on these returns are also overstated by 25 percent. The source of this mistaken inference is selection bias, or the failure to account for unrecorded aspects of the audit selection decision (in this case, the presence of receipts) that materially impact on noncompliance.

In our methodology, we control for possible selection bias through an econometric model of sample selection. In particular, we begin by constructing a data sample containing both returns that have been subjected to operational audits and returns that have not been audited. We then estimate a specification describing the likelihood that a given return will be audited jointly with a set of equations describing the likelihood and magnitude of noncompliance on the return, thereby allowing us to control for the role of audit selection in observed compliance outcomes. The estimated parameters from these equations are then used to predict for each unaudited return in our sample the likelihood and magnitude of noncompliance on the return. By aggregating our predictions over all returns, we are able to arrive at an estimate of the overall magnitude of noncompliance in the population.

As will be discussed more fully below, the equations describing noncompliance can be specified in different ways, depending on the issues of primary concern. For example, in modeling estate tax noncompliance, we have employed a specification that allows for both overstatements and understatements of tax liability, because tax overstatements are relatively common occurrence on estate tax returns. In contrast, in modeling compliance on self-employment returns, we have accounted only for understatements of liability, because overstatements were not a major issue in this application. In the case of sole proprietors, however, we employ separate specifications for noncompliance on revenue and expense items, because reporting violations on such items appear to be driven by different factors.

In each of our illustrations below, our model contains an audit expression of the form:

$$A^* = \beta_A' X_A + \epsilon_A, \tag{1}$$

where the term A^* represents an index of the likelihood that a return with recorded characteristics X_A will be audited. The term ϵ_A represents a normally distributed random disturbance, and β_A is a vector of coefficients to be estimated. From our data sample containing both audited and unaudited returns, we can deduce whether A^* is greater than zero (indicated by whether an audit has been performed). In the statistical literature, Equation (1) is referred to as a probit specification.

A key feature of our methodology is to allow the error term ϵ_A in Equation (1) to be correlated with the error terms in our expressions referring to different forms of noncompliance on the return. If selection bias due to unobservable characteristics is a problem, these correlations should be positive, indicating that returns selected for an operational audit are more likely to possess significant noncompliance than returns with similar recorded characteristics that are not selected for audit. By estimating the correlations, we are able to test explicitly the hypothesis of selection bias. We can correct for such bias if it is found to be present by incorporating the correlation terms into our expressions for predicting the magnitude of noncompliance on returns not subjected to audit.

B3 Illustration #1: Estate Tax Reporting Gap

To estimate the estate tax reporting gap, we combined line item tax return information from the IRS Estate Tax Return Sample (ETRS) with audit details from IRS Estate Post-Audit Study (EPAS) file for calendar year 1992. The resulting estimation sample included $4,193$ returns, $1,374$ of which had been audited. Sample weights were developed to make these $4,193$ returns representative of the overall population of $59,176$ returns that were filed during the calendar year. Table 1 provides weighted statistics on the audit assessments recorded for the returns in our sample that were subjected to audit. Nearly one quarter the estimated $10,209$ returns that were audited received a negative assessment (indicating an overstatement of tax liability). Only about 10 percent had no adjustment as a result of the audit, and the remaining 66 percent received a positive assessment (indicating an understatement of tax liability). Understatements were about twice as large in magnitude as overstatements. Over all audited returns, the average assessment was $54,739.

B3.1 Modeling noncompliance

In developing our model of estate tax noncompliance, we felt it was important for our specification to be able to capture the most salient features of the audit assessment distribution within our sample. In particular, we wanted to develop a specification that allowed for significant numbers of negative, zero, and positive assessments. In addition, we wanted to be able to account for the skewed nature of the assessment distribution. Specifically, the audit sample contains a large number of returns with relatively small assessments in absolute value and a small number of returns with extremely large assessments. Ultimately, we developed a model containing three separate equations to describe the distribution of noncompliance. The first equation was a probit specification for the likelihood that the assessment would be positive:

$$P^* = \beta_P' X_P + \epsilon_P, \tag{2}$$

where P^* is an index of the likelihood that the assessment is positive, X_P is a vector of line item return characteristics used as explanatory variables, β_P is a vector of coefficients to be estimated, and ϵ_P is a normally distributed random disturbance term. From the audit data, we can deduce whether the assessment is positive (implying $P^* > 0$) or non-positive (implying $P^* <= 0$).

Should the audit assessment turn out to be positive (indicating that taxes have been understated), it is necessary to describe the magnitude of the additional taxes assessed. To account for the skewed nature of the assessment distribution, we employ the following log-normal specification for the magnitude of the assessment:

$$ln(R) = \beta_R' X_R + \epsilon_R, \tag{3}$$

where ln(R) represents the natural log of the tax assessment, X_R is a vector of line item characteristics used as explanatory variables, β_R is a vector of coefficients to be estimated, and ϵ_R is a random normal disturbance term.

Should the assessment instead turn out to be non-positive, it is necessary to be able to account whether the assessment is negative (indicating a tax overstatement) or zero. Further, if the assessment is negative, it is desirable for the specification to be able to predict the magnitude of the assessment. The following displaced log-normal distribution was employed for this purpose:

$$ln(M * +D) = \beta_M' X_M + \epsilon_M, \tag{4}$$

where M^* is an index of the likelihood that the assessment is negative, X_M is a vector of line item characteristics used as explanatory variables, β_M is a vector of coefficients to be estimated, and ϵ_M is a random normal disturbance term. The term D (also estimated) is the "displacement parameter," which indicates how far the lower bound of the log-normal distribution is shifted below the ordinary bound of zero. Under this specification, we observe no tax change as a result of the audit whenever M^* falls between $-D$ and zero. Otherwise, we observe a tax reduction in the amount of M^*.

B3.2 Variable selection

The influence of recorded taxpayer characteristics on noncompliance is captured through the estimated coefficients of the explanatory variables of the model (i.e., the coefficients of X_P, X_R, and X_M). To select these explanatory variables, we began by identifying a set of 61 recorded line item return characteristics that potentially might impact on compliance. We then employed a backward variable selection methodology to select a separate set of explanatory variables for each equation of the model. Under this procedure, we first estimated a given equation using all 61 characteristics as explanatory variables. We then examined the t-statistic associated with each variable. The t-statistic provides a measure of the contribution of a given explanatory variable to the fit of an equation. The lower the t-statistic, the smaller the contribution. We eliminated the variable associated with the smallest t-statistic and repeated estimation of the equation with the remaining 60 variables. We again eliminated the variable associated with the smallest t-statistic, and re-estimated the model with the remaining 59 explanatory variables. This process was repeated until all remaining explanatory variables had a t-statistic that exceeded a minimum threshold value.[4]

B3.3 Accounting for the role of audit selection

To account for differences in noncompliance among audited and unaudited returns that cannot be explained by recorded line item characteristics, we incorporated the audit selection equation [Equation (1)] into our specification. The explanatory variables for this equation (X_A) were chosen using a variable selection procedure similar to the one described above for the noncompliance equations. With the explanatory variables selected for each of the equations of the model, we proceeded to estimate the full set of equations

[4] The threshold was set at the 10 percent critical value for the t-distribution.

jointly, allowing correlations between the disturbance of the audit equation (ϵ_A) and the disturbances of the noncompliance equations (ϵ_P and ϵ_R).[5] As discussed previously, these correlations account for unrecorded taxpayer characteristics (such as the presence of supporting documentation submitted along with a return) that impact on both the likelihood of an audit and the assessed level of noncompliance should an audit take place.

B3.4 Results of Analysis

We used the estimated parameters of the model to predict, for each unaudited return in the sample, the value of the assessment that would have been made if it had been subjected to examination. In the case of examined returns, the actual audit assessment was used as the measure of noncompliance. In a small number of cases, the assessment results were unavailable for returns that had been examined. For these cases, the predicted assessment was computed in a manner similar to that employed for unaudited returns. Table 2 summarizes the results of the analysis. Our model estimates that estate taxes were understated on net by $1.5 billion on calendar year 1992 returns, implying an average net understatement of $25,600. As expected, returns selected for audit account for a disproportionate share of all noncompliance. Specifically, audited returns (with observed assessments) represent about 17 percent of the population, but account for 34 percent of the total estimated tax gap.

B4. Illustration #2: Underreporting of Business Income by Sole Proprietors

As an illustration of how operational audit data can be used to estimate the individual income tax gap, we developed a model of underreporting for one of the major elements of the tax gap – underreported self-employment income – and we applied our model to data covering two separate business audit classes from one IRS district.

[5] For tractability, we restricted the correlation between ϵ_A and ϵ_M (the disturbance of the equation describing the magnitude of overstated taxes) to zero; this seems reasonable given that estate tax audit selection is presumably focussed on identifying cases of tax understatements, not overstatements.

In our empirical illustration, we focus on two separate forms of noncompliance: under-reporting of self-employment revenue and over-reporting of self-employment expenses. However, our approach is quite general, allowing other forms of noncompliance to be paired in different applications; for example the amounts of noncompliance detected on each of two different tax forms or return line items. As discussed below in Part C, the model is also flexible in that it can be generalized to allow for more than two separate forms of noncompliance.

B4.1 Modeling noncompliance

We restrict ourselves to considering only two possible cases for these dependent variables: (i) perfect compliance – zero noncompliance detected, or (ii) positive noncompliance detected. Thus, unlike our previous estate tax example, we exclude the third possibility: detection of a taxpayer mistake indicating that the taxpayer has overstated his true liability. We exclude this possibility mainly because overpayments are relatively uncommon in our data samples and those that do occur tend to be relatively small in magnitude.

Our two expressions for noncompliance represent a modified version of the well-known tobit specification. Let R refer to the magnitude of under-reporting of a revenue or income source on a return, and let E refer to the magnitude of over-reporting of an expense or deduction source on the return. Each of these variables is either zero or positive; thus, as noted above, we exclude the possibility of an error by the taxpayer resulting in an overpayment of taxes. In our estimation we focus only on reporting on Schedule C, and thus in our application of the model R refers to under-reporting of Schedule C gross business revenues and E refers to over-reporting of Schedule C gross business expenses. Let X_R represent all recorded return characteristics expected to be significantly associated with the under-reporting of revenue; similarly, let X_E represent all recorded return characteristics expected to be significantly associated with the over-reporting of expenses. Lastly, let ϵ_R represent a random disturbance term reflecting unrecorded factors that influence under-reporting of revenue that are not included in X_R, and let ϵ_E represent a random disturbance reflecting unrecorded factors that influence over-reporting of expenses by the taxpayer that are not included in X_E.

Our specification for noncompliance is as follows:

$$ln(R^* + D_R) = X_R\beta_R + \epsilon_R \tag{5}$$

$$ln(E^* + D_E) = X_E\beta_E + \epsilon_E, \tag{6}$$

In these expressions R^* and E^* are latent variables measuring the propensity of the taxpayer to understate revenue and overstate expenses, respectively. We observe no assessment of additional self-employment revenue from an audit if R^* is less than or equal to zero. Otherwise, we observe an assessment of R^* additional dollars. Similarly, we observe no change in self-employment expenses as a result of an audit if E^* is less than equal to zero. Otherwise, expenses are reduced by E^* dollars. The terms β_R and β_E represent coefficients to be estimated. In a standard tobit specification, one would assume that the conditional distributions of R^* and E^* (given the explanatory variables X_R and X_E) are normal. As in the estate tax illustration, however, the distributions of understated self-employment revenue and overstated self-employment expenses tend to be highly skewed rather than having the "bell shape" that is characteristic of the normal distribution. We therefore employ a displaced log-normal specification for each form of noncompliance. The expression $ln(\bullet)$ represents the natural log function, and the parameters D_R and D_E are displacement parameters (also to be estimated) that influence the amount of mass under the displaced log-normal distribution that pertains to compliant reports (i.e., cases where $R = 0$ and $E = 0$, respectively).

B4.2 Variable Selection

The explanatory variables X_R and X_E in Equations (5) and (6) control for recorded line item characteristics of the income tax return that are associated with the two forms of noncompliance. To select the explanatory variables for our analysis, we again began by identifying a large number of line item characteristics that were potentially associated with noncompliance. We then employed a variable selection methodology separately for each of the two equations to select the final variables for analysis. However, for this project, we implemented a proprietary methodology that we have developed to perform the variable selection. The advantage of this methodology is that it selects explanatory variables on the basis of their performance outside the estimation sample, in the general population of returns. In contrast, traditional variable selection procedures, such as the backward selection procedure used for our estate tax gap study, choose variables on the basis of their performance within the same sample used for estimation. This can lead to the selection of variables that perform well within the estimation sample, but perform relatively poorly in the general population of returns.

Our approach involves drawing bootstrap samples from the original data sample. Each bootstrap sample consists of a set of returns drawn at random with replacement from the original sample of data. The number of cases selected for each bootstrap sample is the same and equal to the number of returns in the original data sample. Under this approach, the original data sample serves as a surrogate for the underlying return population, and the bootstrap samples are treated as random samples drawn from this population. Our procedure begins by drawing a series of independent bootstrap samples from our original sample of audited returns. For each bootstrap sample, we fit our displaced log-normal tobit model to all possible univariate specifications from the set of candidate regressors. The parameter estimates for each specification are then employed to predict the magnitude of noncompliance on each return in the original data sample. The predicted magnitude is compared to the actual magnitude, and an out-of-sample measure of the mean-squared prediction error (the squared difference between the two magnitudes) is computed. The results are averaged over all bootstrap trials, providing us with a measure of the average out-of-sample mean-squared error associated with each univariate specification. The candidate regressor with the smallest measure by this criterion is selected for inclusion in our model.

We then draw a second series of independent bootstrap samples. For each bootstrap sample, we fit our displaced log-normal tobit model to all possible specifications involving two explanatory variables that include the previously selected regressor and one of the remaining candidate explanatory variables. The average out-of-sample measures of the mean-squared prediction error for each of these specifications over all trials is compared, and the candidate variables associated with the smallest measure by this criterion are selected as the best two variable specification.

Subsequent series of bootstrap trials are used to sequentially select the best specifications associated with 3 to 15 regressors. To choose how many variables to include in our final specification, we compare the average mean-squared errors associated with the best model of each size (from the best univariate model to the best 15 variable model), and we select the model (and associated model size) with the smallest value by this criterion.

B4.3 Accounting for the role of audit selection

To account for differences in noncompliance among audited and unaudited returns that cannot be explained by recorded line item characteristics, we incorporated the audit selection equation [Equation (1)] into our specification. The explanatory variables for this equation (X_A) were chosen using an out-of-sample variable selection procedure similar to the one described above for the noncompliance equations. We allowed a correlation between the disturbance of the audit selection equation (ϵ_A) and the disturbances of the two noncompliance equations (ϵ_R and ϵ_E) to account for the role of unrecorded characteristics (such as the presence of supporting documentation with a return) that impact on both audit selection and noncompliance.

B4.4 Two-Stage Estimation Strategy

With the explanatory variables chosen for each of the equations of our model, the next step was to estimate the model's parameters. To reduce the complexity associated with this process, we developed a two-stage estimation procedure. In the first stage we performed a standard probit estimation of the audit selection equation using the method of maximum likelihood.

In the second stage, we separately estimated the parameters of Equations (5) and (6) (including the error correlation terms used to account for sample selection), conditional on the values of the audit selection parameters estimated in the first stage. We generated our estimates by maximizing a separate conditional likelihood function for each of these two equations with respect to the parameters of each equation.

We used the full set of estimated parameters from our model to predict, for each unaudited return, the amount of understated revenue and overstated expenses. In the case of examined returns, the actual audit assessment was used as the measure of noncompliance.

B4.5 Results of analysis

We estimated our model separately on data samples from two business exam classes – total gross receipts $< \$25{,}000$ (exam class 35) and $\$25{,}000 \leq$ total gross receipts $< \$100{,}000$ (exam class 36) – from the Chicago district. These samples each included line item return information for a group of audited and unaudited returns from the IRS Individual Returns Transaction File (IRTF). Details on examination results for the audited returns was matched to these samples from the Examination Operational Automation Database (EOAD). The sample from exam class 35 contained $1,132$ returns, including 221 that had been audited. The sample from exam class 36 contained $1,678$ returns, including 342 that had been audited.

To evaluate the performance of our estimated specifications for noncompliance with respect to business revenue and expenses, we compare in Table 3 the predicted frequencies and average magnitudes of understated revenue and overstated expenses on the audited returns in each class with the actual audit results. For both classes, the actual and predicted frequencies are extremely similar, both with respect to understated revenue and overstated expenses. The average actual and predicted magnitudes of overstated expenses are also extremely similar. Our expressions for predicting the average amount of understated revenue are somewhat less precise, over-predicting this form of noncompliance by about 7.7 percent in exam class 35 and under-predicting it by about 4.5 percent in exam class 36.

Tables 4 and 5 present further evidence on the fit of our estimated noncompliance specifications for exam classes 35 and 36, respectively. The tables summarize the performance of these specifications in predicting whether a given return will receive an adjustment to business revenue or expenses as the result of an audit. The results indicate that our specifications successfully predict whether a given return will receive an adjustment to business expenses over 60 percent of the time and to business revenue over 80 percent of the time.

The average predicted likelihood and magnitude of noncompliance over all unaudited returns in each sample are presented in Table 6. Comparing the results in this table with those presented in Table 3, the average predicted likelihood and magnitude of both forms of noncompliance tend to be lower on unaudited returns than they are on audited returns. This is intuitive, suggesting that those business class returns selected for audit tend to be less compliant than those not selected.

The one exception to this finding is the results for understated revenues in exam class 36. In this case, the average predicted probability of noncompliance is 22.3 percent for unaudited returns, compared to 21.3 percent for audited returns. Furthermore, the average predicted magnitude of noncompliance on unaudited returns is $4,056, or approximately 8 percent larger than the average predicted magnitude on audited returns ($3,750). This result is most likely a reflection of the audit selection strategy employed by the IRS, which is focussed on the overall tax change rather than the adjustments associated with particular forms of noncompliance. Observe that the overall predicted understatement of self-employment income (understated revenue plus overstated expenses) is substantially higher on audited returns than unaudited returns in exam class 36 ($13,551 compared to $10,567). Thus, while the returns selected for audit may have less potential for an adjustment to business revenue than unaudited returns, they more than make up for this in terms of their potential for an adjustment to business expenses.

C. Extensions of Methodology

We discuss three ways in which our model might be extended in future work: (i) allowing for more than two forms of noncompliance; (ii) controlling for undetected noncompliance; and (iii) application as an interim estimation strategy during periods between random audit studies.

C1. Allowing for multiple forms of noncompliance

It is straightforward to extend our model to allow for additional forms of noncompliance using the two stage estimation strategy presented in our illustration concerning noncompliance by sole proprietors in Section B4. A new equation would be incorporated into our model for each additional form of noncompliance. The parameters for each of these equations (including the correlation between the error term of the equation and the error term of the audit selection equation [Equation (1)]) would then be estimated separately from all of the other parameters of the model during the second stage of the estimation process. A separate conditional likelihood function would be maximized to estimate the parameters associated each new equation.

C2. Accounting for undetected noncompliance

Incomplete detection is an important issue for estimating the tax gap from audit data of any kind. However, it is likely to be especially important when the data are based on ordinary operational audits, because the results of such audits typically involve a limited number of issues on the return rather than a comprehensive examination of the entire return.

Our model can be extended to address the issues of incomplete detection through application of the method of detection controlled estimation (DCE).[6] However, in order to carry out a detection controlled analysis, the following supplementary information would be required. First, as in earlier applications of DCE, it is essential to be able to divide returns into groups such that all the returns in a given group are known to have been audited by the same examiner (or group of examiners). This division does not require the actual names of examiners to be included on the data base; all that would be required is a code that uniquely identified the returns investigated by different examiners. Second, in order to understand how comprehensive each audit has been, it is necessary to identify which line items have been examined, regardless whether they have been adjusted.[7] It would be even more useful if the order in which the line-items are examined could be coded in the audit data base. In that case, a very rich model of the audit process could be developed that takes into account the factors driving the decision by examiners whether to explore additional issues beyond those assigned by the classifier.

[6] See Feinstein (1990, 1991) and Erard (1997) for discussion and applications of this methodology.

[7] Unfortunately, such information is not currently collected as part of the EOAD used in our illustration concerning noncompliance by sole proprietors in Section B4.

If such data were to be made available, it would be relatively straightforward to extend our methodology to include expressions describing the detection process. The best such model is the fractional detection model described in Feinstein (1991). As discussed in that article, it is possible to fit specifications in which the detection rate is allowed to differ for each form of noncompliance. Such a model allows for many different patterns of detection. For example, it can address the case where some examiners are unusually good at detecting certain forms of noncompliance but only average at detecting other forms.

C3. Updating Random Audit Study Results

A third way our methodology could be extended would be to use it as a means to update compliance estimates based on random audit studies, such as those planned under the NRP. Such studies are typically performed only periodically, and it is desirable to update compliance estimates during interim periods to reflect changes in the compliance environment. We sketch below how our methodology might be used to update the results from an NRP baseline study.

The first step would be to use our methodology to estimate the level of noncompliance using operational audit data from the NRP base year. The results would be compared against independent estimates based on the NRP.[8] Then, assuming the results were not the same, one or more calibration factors would be developed to bring the results of our methodology in line with the NRP results.[9]

Our methodology would then be applied to operational audit data from the subsequent year to estimate the level of noncompliance in that year. Finally, the calibration factor(s) from the base year would be applied to the estimate for the subsequent year to produce our updated estimate of noncompliance.

[8] Both sets of estimates could be adjusted to account for undetected noncompliance.

[9] Different adjustment factors might be used for different sources of noncompliance or different classes of taxpayers.

D. Workload Selection

In addition to its usefulness as a tool for measuring compliance, our methodology is also suitable for workload selection. Although federal and state tax agencies sometimes apply statistical techniques, such as regression or discriminant analysis, to develop audit selection criteria based on operational audit results, typically these agencies treat the underlying data as though they constitute a random and representative sample from the overall return population, failing to account for the fact that most returns in the sample were selected because they were considered to be at high risk of noncompliance. As a consequence, the relationship estimated between the magnitude of noncompliance and reported line item characteristics within the operational audit sample may not hold for returns within the general population, which will tend to have lower levels of noncompliance on average. This failure to account for the unrepresentative nature of the data sample being used for formula development can weaken the effectiveness of the resulting workload selection formulae in identifying the best audit candidates in the overall return population.

Our methodology explicitly takes into account the selected nature of the operational audit sample and corrects for the effects of audit selection in the formulae used to predict noncompliance. In addition, our methodology employs a unique approach for selecting predictor variables based on their out-of-sample performance. By correcting for the non-representative nature of the audit sample and through the selection of more appropriate predictor variables, our methodology has the potential to outperform models typically used by federal and state tax agencies for workload selection.

To employ our methodology for workload selection, one would proceed in a manner analogous to tax gap estimation, selecting sets of explanatory variables for predicting which returns in the data sample would be audited and the likelihood and magnitude of noncompliance. One would then estimate our compliance model using the data sample and compute the appropriate formula based on the results for predicting the amount of noncompliance associated with any given return. This formula would be then be applied to returns from the general population, and the returns would then be sorted according to the predicted magnitude of noncompliance. The returns predicted to contain the largest amounts of noncompliance would then be forwarded to a classifier for further audit consideration.

E. Concluding Remarks

Our methodology provides a means of extrapolating from examination results achieved on returns subjected to operational audits the likely amount of noncompliance present on returns that have not been audited. In performing this extrapolation, we are able to control for differences in both recorded and unrecorded characteristics of returns that lead some returns to be selected for examination while others are not.

We find that our out-of-sample variable selection methodology performs well and that we are able to identify, in a parsimonious model, the observable characteristics of returns associated with audit selection and noncompliance. While it is not possible to directly confirm that our aggregate estimates of noncompliance on unaudited returns are accurate, we are able to assess the performance of our specification in predicting noncompliance on examined returns. The results indicate that our model performs well in this regard.

Extensions of our methodology are feasible, and may be used to address multiple forms of underreporting and undetected noncompliance. In addition, our methodology could be employed as a means to update compliance estimates based on random audit studies from a prior year.

Our methodology is also well-suited for use in developing audit selection criteria. Unlike models currently employed by federal and state tax authorities for operational audit data, our methodology corrects for the fact that returns within the estimation sample are more likely to require an assessment than returns within the general population. This correction should permit greater accuracy in the prediction of which returns in the general population are most in need of examination. In addition, our out-of-sample variable selection technique provides an effective means for choosing among competing predictor variables for inclusion in workload selection formulae.

Table 1: Distribution and Magnitude of Assessments

Audit Assessment	*Number of Cases*	*Percentage of Returns*	*Average Assessment*
Negative	2,414	23.65	-$48,270
Zero	1,018	9.97	$ 0
Positive	6,777	66.38	$99,662
Total	10,209	100.00	$54,739

Table 2: Breakdown of Projected Estate Tax Reporting Gap

Cases	*Projected Aggregate Understatements ($ Thousands)*	*Projected Aggregate Overstatements ($ Thousands)*	*Projected Aggregate Net Understatements ($ Thousands)*
Audit Cases with Known Results*	675,396	116,516	558,880
Audit Cases with Unknown Results	92,642	14,668	77,974
Non-Audit Cases	1,032,972	153,827	879,145
All Cases	1,801,011	285,012	1,515,999

*Actual assessment results used.

Table 3: Actual and Predicted Sch. C Noncompliance on Examined Returns

	TGR < $25K		*$25K<=TGR<$100K*	
	Understated Revenue	*Overstated Expenses*	*Understated Revenue*	*Overstated Expenses*
Actual % of returns with:	17.2	54.3	21.1	62.3
Predicted % of returns with	17.2	55.5	21.3	62.4
Actual mean amount of:	$1,365	$4,502	$3,920	$9,800
Predicted mean amount of:	$1,470	$4,492	$3,750	$9,801

Table 4: Classification Tables, TGR < $25K

Understated Revenue on Audited Returns

	Actual Compliant	Actual Noncompliant	
Predicted Compliant	183	38	221
Predicted Noncompliant	0	0	0
	183	38	221

Percent correctly classified: 82.8

Overstated Expenses on Audited Returns

	Actual Compliant	Actual Noncompliant	
Predicted Compliant	50	34	136
Predicted Noncompliant	51	86	85
	101	120	221

Percent correctly classified: 61.5

Table 5: Classification Tables, $25K <= TGR < $100K

Understated Revenue on Audited Returns

	Actual Compliant	*Actual Noncompliant*	
Predicted Compliant	260	58	318
Predicted Noncompliant	10	14	24
	270	72	342

Percent correctly classified: 80.1

Overstated Expenses on Audited Returns

	Actual Compliant	*Actual Non-Compliant*	
Predicted Compliant	46	31	77
Predicted Noncompliant	83	182	265
	129	213	342

Percent correctly classified: 66.7

Table 6: Predicted Sch. C Noncompliance on Unexamined Returns

	TGR < $25K		*$25K<=TGR<$100K*	
	Understated Revenue	*Overstated Expenses*	*Understated Revenue*	*Overstated Expenses*
Predicted % of returns with	15.8	41.1	22.3	51.0
Predicted mean amount of:	$1,298	$2,712	$4,056	$6,511

F. References

Alm, James, Brian Erard, and Jonathan S. Feinstein. (1996) "The Relationship Between State and Federal Tax Audits," in *Empirical Foundations of Household Taxation*, ed. Martin Feldstein and James Poterba. Chicago: University of Chicago Press, 235–273.

Erard, Brian. (1997) "Self-Selection with Measurement Errors: A Microeconometric Analysis of the Decision to Seek Tax Assistance and its Implications for Tax Compliance," *Journal of Econometrics*, **81**, 319-356.

Erard, Brian. (1999) "Estate Tax Underreporting Study," Report prepared by B. Erard & Associates for the IRS Economic Analysis and Modeling Group, Order number TIRNO-98-P-00406, March 4.

Erard, Brian and Jonathan S. Feinstein. (2001) "Estimating the Federal Income Tax Gap Using Operational Audit Data," Report prepared by B. Erard & Associates for the IRS Economic Analysis and Modeling Group, Order number TIRNO-00-P-01128, November 6.

Eller, Martha Britton, Brian Erard, and Chih-Chin Ho. (2000) "The Magnitude and Determinants of Federal Estate Tax Noncompliance," in *Rethinking Estate and Gift Taxation*, ed. William G. Gale, James R. Hines, Jr., and Joel Slemrod. Washington, DC: The Brookings Institution, 375 - 410.

Feinstein, Jonathan S. (1990) "Detection Controlled Estimation," *Journal of Law and Economics*, **33:1**, 233–76.

Feinstein, Jonathan S. (1991) "An Econometric Analysis of Income Tax Evasion and its Detection," *Rand Journal of Economics*, **22:1**, 14–35.

Heckman, J. (1979) "Sample selection bias as a specification error," *Econometrica*, **47**, 153–61.