EVALUATION OF A NEW PROCEDURE FOR ESTIMATING INCOME AND TAX AGGREGATES FROM ADVANCE DATA

John L. Czajka, Mathematica Policy Research, Inc.
Sharon M. Hirabayashi, Mathematica Policy Research, Inc.
Roderick J.A. Little, Datametrics Research, Inc. and University of California, Los Angeles
Donald B. Rubin, Datametrics Research, Inc. and Harvard University

A. INTRODUCTION

Individual and sole proprietorship tax returns filed with he Internal Revenue Service (IRS) in any tax year are processed and posted to a centralized data base called the Individual Master File (IMF). The Statistics of Income (SOI) Division of the IRS draws a probability sample of returns from the IMF and uses the data to produce extensive tabulations for government clients and for publication in SOI statistical series. The Office of Tax Analysis and the Joint Committee on Taxation require key tabulations by the end of each November, months before the statistics for the full year are available. To accommodate this need, the SOI Division prepares an Advance Data (AD) report from the returns sampled through late September. Sample observations are weighted to projections of total returns for the year, by sampling stratum.

For most income and tax variables, the advance estimates have tended to be very close to the final estimates prepared after the end of the processing year. For some variables, however, the advance estimates have differed from the final estimates by substantial margins. Furthermore, recent changes in tax regulations have contributed to an increase in the proportion of returns posted to the IMF after the AD sample close-out, thereby increasing the potential for error in the AD estimates. In view of these considerations there is a perceived need to improve the AD estimating methodology to project more accurately the number of late-posted returns and to adjust for income and tax differences between early and late returns in each stratum. This paper addresses the latter of these two needs.[1] We propose, apply and evaluate a new method of weighting the AD sample returns to improve their representation of complete year income and tax aggregates. The new procedures make use of propensity score methods developed by Rosenbaum and Rubin (1983).

B. ESTIMATION WITH ADVANCE DATA

The returns processed and posted to the IMF during a given calendar year, excluding a small subset (primarily residents of Puerto Rico and nonresidents), constitute the universe of the SOI Complete Report (CR) for the preceding tax year. The CR is based on a stratified probability sample of the returns present on the IMF at the end of the calendar year, with returns being sampled on a continuing basis throughout the year. The objective of the AD estimation is to anticipate the CR estimates of key income and tax items on the basis of returns sampled through late September. Given the projections of total returns, the remaining problem may be viewed as one of adjusting for unit nonresponse, where the missing units are tax returns posted to the IMF after the AD sample close-out. The distribution of returns by posting date is not random; hence the nonresponse adjustment must account for differences between early and late returns vis a vis the income and tax items for which advance estimates are required.

Current IRS Procedures

Following the customary method of adjustment for unit nonresponse, the AD estimation is based on a reweighting of the advance sample. Specifically, the returns in each sampling stratum (see below) are weighted up to a projection of the final population of returns in that stratum at the end of the calendar year--14 weeks after the AD close-out. The weight, representing the ratio of the projected full-year population to the advance sample count, may be expressed as the product of two components: the inverse of the sampling fraction (the base weight) and the projected growth in the population of returns between the AD close-out and the end of the year.

$$(1)\quad w = \frac{\text{Projected CR population}}{\text{AD sample count}}$$

$$= \frac{\text{AD population}}{\text{AD sample}} \times \frac{\text{Projected CR population}}{\text{AD population}}.$$

Variation in the projected growth by stratum (the second term in the bottom expression) is the mechanism by which the AD estimation methodology adjusts for differences between advance and late returns. For tax year 1981 the projected growth ratios ranged from 1.004 to 1.397.[2]

Because it is so central to the AD weighting procedure, the stratification of the SOI sample merits discussion. The principal stratifying dimension is a classification based on income and type of return. The sample design implemented with tax year 1982 provides for 29 categories or "sample codes." Twenty-seven of the categories represent a cross-classification of nine levels of income and three types of return: business; non-business, farm; and non-business, non-farm. The nine income levels represent combinations of total receipts (farm and business returns only) and the larger absolute value of a positive amounts total (PAT) and negative amounts total (NAT) computed over 19 income items. Two additional sample codes, set aside for high income nontaxable returns and business returns with high net profit or loss, complete the 29 strata. The stratification is concentrated on the upper tail of the income distribution. In tax year 1982, for example, 82 percent of the population of 95.6 million returns fell into the two lowest income, non-

business, non-farm sample codes, leaving 18 percent to be apportioned among the remaining 27 sample codes. The sampling rates varied from .02 percent in the largest sample code to 100 percent in the six highest income and two supplemental codes (Internal Revenue Service, 1984: 16).

In addition to sample code there is a geographic stratification with up to three levels in a given year. Returns from small states are sampled at a higher rate than those from large states to insure a minimum size for each state sample. However, this geographic stratification is unimportant in the reweighting of the advance sample, and we shall focus strictly on the sample code.

The proportion of returns posted after the AD close-out varies markedly across sample codes. The frequency of late posting rises monotonically with the income level of the code, except at the lowest levels, and business returns exhibit higher rates of late posting than non-business returns within the same income level. Table 1 illustrates this pattern with tabulations from the 1981 CR sample file. These tabulations use 1982 sample code definitions, consistent with the analyses reported below, rather than the 21 strata from which the 1981 sample was actually drawn. The percentage of returns posted after the AD close-out, in processing weeks or "cycles" 39-52, ranges from .8 and .9 percent in the lowest income non-business codes to over 32 percent in the highest income codes (all three types of returns). Business and non-business returns are most sharply differentiated at the lowest income level; their rates of late posting converge as income rises.

It is clear from Table 1 that the sample code is an excellent stratifier for reweighting the advance sample to estimate full year totals prior to the completion of processing. If the variation in late posting within sample codes were unrelated to the income and tax items for which AD estimates are prepared, then the use of this single stratifier would be sufficient. In fact, however, the performance of the AD estimates suggests that further stratification is required, as we shall see.

Performance of Advance Estimates

Table 2 presents a comparison of the AD and CR estimates of selected income and tax items for tax year 1981. Deviations of AD from CR estimates are expressed as absolute quantities (numbers of returns or millions of dollars) and as percentages of the CR estimates. The AD estimate of the total number of returns was .12 percent below the CR estimate. If the returns posted late were undifferentiated from same stratum returns posted prior to the AD close-out, and if the percentage error in the projections of total returns were constant across all the strata, then the percentage deviations for all of the items in the table would be -.12 percent. That the percentage deviations vary widely around this amount indicates that at least one of these conditions is not met.

In fact, IRS overprojected the total high income returns and underpredicted the low income returns in preparing the AD estimates for 1981 (Czajka, Little, and Rubin, 1984). In part as a result the AD overestimated adjusted gross income (AGI), one of the key components of the 1981 sample code definitions. Elsewhere we observe both small and large deviations, positive and negative, with no obvious pattern. The numbers of returns with interest and dividends are both overestimated by about .10 percent, but the interest Table 1 (1981 comparison) amounts are underestimated by .47 percent and dividend amounts overestimated by .64 percent. The number of returns with business net profit is underestimated by .14 percent and the amount of net profit by .55 percent. The number of returns with net loss is overestimated by .18 percent, but the magnitude of the losses is underestimated by nearly 6 percent. Net capital gains are underestimated by 4.52 percent while farm net income is underestimated by 1.0 percent. The number of returns with income tax is underestimated by only .06 percent, but the amount of tax is overestimated by .61 percent. The largest deviations for both numbers of returns and amounts are found on the additional tax items. The number of returns with tax for tax preferences is underestimated by 5.92 percent and the amount of tax by nearly double that. The error for minimum tax is somewhat smaller while that for alternative minimum tax is slightly greater.

Estimates of percentage error over a period of years would inform the user as to the expected precision of the advance estimates. By themselves, however, such error statistics do not provide a clear picture of the effectiveness of the AD methodology at projecting that which is unknown at the time the AD sample is closed out--namely, the income and tax on returns posted after the AD close-out. To examine this question, Table 3 describes the incidence of late postisng, and Table 4 compares the AD and CR estimates among late returns.

The incidence of late posting exhibits wide variation by type of income or tax item. The distribution of returns and amounts by posting period is reported in Table 3 for selected income and tax items in tax years 1981 and 1982 (these data are based on IRS tabulations of the total population of returns, so the year-end totals differ marginally from the CR estimates in Table 2). In 1981 1.43 percent of returns were posted late. In subpopulations this percentage varied from a low of .9 percent for returns with an overpayment or with unemployment compensation in their AGI, to a high of 15.71 percent for returns with alternative minimum tax. Variation was even greater for dollar amounts. In 1981 late-posted returns accounted for 1.81 percent of AGI, the percentage of other dollar amounts in late-posted returns varied from 1.07 percent for unemployment compensation in AGI to 28.65 percent for alternative minimum tax. Late posting increased by about 50 percent over all returns between 1981 and 1982, with some income and tax items showing larger increases and some smaller. This increase was due largely to the lengthening of the automatic extension from two months to four.

Table 4 reports the deviation of AD from CR estimates of selected income and tax items as a percentage of the corresponding late returns or

amounts. Whereas the AD projection of total returns was only .12 percent below the CR estimate for the year (Table 2), this 111,000 shortfall represented 8.15 percent of the returns actually posted late. Likewise, the AD estimate of total AGI exceeded the CR estimate by .38 percent, but this represented a 21 percent overestimate of the AGI on returns posted after AD close-out. For the 15 other items the error estimates range from less than one percent to nearly 50 percent for numbers of returns and from 8 percent to 51 percent for dollar amounts. The magnitudes suggest considerable room for improvement--much more so on some items than others. In addition, while Table 4 shows some items with comparable errors on number of returns and dollar amount (minimum and alternative minimum tax, itemized deductions, payments to an IRA), it includes other examples where the errors are in opposite directions (interest received, income tax before credits, balance due, overpayment). For items of the former kind, improving the AD estimate of late returns will improve the projected amount; for items of the latter kind, improving the projected number of returns, other things being equal, will actually increase the magnitude of the error on the amount.

Proposed New Method

Any proposal to alter the stratification of the AD file must take the current sample design as given. Therefore, changes to the weighting scheme must be implemented by poststratification. This together with the known substantial variation in late posting by sample code suggests searching out as potential new stratifiers variables that can improve the AD file's representation of the CR file within each of the present sample codes.

One known characteristic of late returns that is not obviously addressed by the current stratification is complexity. Late returns tend to include a larger number of schedules than do early returns (Sailer et al., 1982, Figure C). If this remains true within strata, then late, complex returns end up being represented by early, simple returns, thereby biasing the AD estimates of whatever items are related to the filing of additional schedules. One approach to improving the weighting of AD returns, therefore, is to add to the current stratification scheme a dimension for complexity, operationalized as the number of schedules attached to the basic tax form. One appeal of this approach is its simplicity, and we entertained it in our investigation of new weighting procedures because it would be relatively easy for IRS to implement. However, empirical investigation based on the 1981 CR microdata file found no clear evidence of a positive relationship between the number of schedules and the probability that a return will be posted late within individual sample codes. Apparently the current stratification already captures most of the relationship between the number of schedules and the lateness of posting.

Our second approach was more general in nature. Let x be a set of predictor variables available for early and late returns, and define the propensity to be posted late, p(x), as the proportion of cycle 39-52 returns in the population of returns with values x on the predictors. The theory of Rosenbaum and Rubin (1983), discussed in the context of survey nonresponse in David et al. (1983), shows that a) the distribution of x is the same for early and later filers within strata with constant values of p(x), and b) weighting cells defined with p(x) as one of the stratifiers yield unbiased estimates of the distributions of outcome variables. This theory suggests stratifying on an estimate of propensity to be posted late. Specifically, the procedure defines the binary variable ℓ with value 1 for current year late returns and 0 for current year early returns; calculates an estimate p(x) of p(x) by logistic regression of ℓ on x, using data from the previous year; forms 5 or 6 strata with grouped values of p(x); and then uses the grouped variable as an additional stratifier in the formation of weighting classes, just as complexity would have been used in the first approach. In view of the aforementioned importance of sample code as a stratifier, we propose calculating separate logistic regressions for each sample code, if this is practically feasible.

Decisions required in implementing this approach include a) the choice of predictors x and b) the choice of cut points for defining strata based on p(x). The specifics of both are detailed in our description of a test application below. Theoretical considerations, in the selection of predictors are laid out here.

Three characteristics determine good predictors x for inclusion in the logistic regression models:

(1) x should be a good predictor of the propensity to be posted late;

(2) x should be a good predictor of outcome variables $y_1 \cdots y_k$ tabulated in the AD report;

(3) the relationship between x and the propensity to be posted late should be relatively stable across adjacent years.

Variables that fail to satisfy (1) have minor influence on the weights and hence on the nonresponse adjustments. Variables that satisfy (1) but fail to satisfy (2) tend to increase the mean squared error of AD estimates by inflating their variance, without a compensating reduction in bias. Finally, variables that satisfy (1) and (2) but not (3) introduce bias because prior year data are used to estimate probabilities of late posting, which in turn determine the weights for current year data.

C. IMPLEMENTATION OF A NEW METHOD

The application and evaluation of the new weighting procedure entailed the use of 1981 CR sample microdata to estimate the propensity models, which were then applied to the early returns on the 1982 CR sample file to generate predicted propensities and construct new weight classes within the sample codes. To free the

results of any influence from projection error (which will be reduced in the future owing to new procedures introduced with the tax year 1983 AD estimates), we used full year CR estimates in place of the required projections. Estimates of a large set of income and tax aggregates prepared with the new weights and, alternatively, uniform weights within each sampling stratum (a simulation of the current method) were then compared with estimates based on the final CR weights to assess the relative accuracy of the current and proposed weighting schemes net of projection error.

Selection of Variables

Ideally, one would conduct a search over all plausible x variables, perhaps giving special attention to those which IRS intends to tabulate. With literally hundreds of possible x variables, however, it was not practical to do so. Instead, we employed a three-stage procedure involving, first, an a priori selection of a subset of all possible predictors; further screening of these variables on the basis of a stepwise OLS regression; application of stepwise logistic regression estimation within strata to derive the final set of stratum-specific models.

Drawing on its subject matter expertise, IRS provided a list of 28 items to be investigated as possible predictors of late posting. Conditions (2) and (3) for the choice of x variables, discussed in the preceding section, were addressed at this point. The 28 items were also believed to be related to late posting (condition 1). After further consideration, we excluded one of the 28 items, strongly related to late posting, because the nature of that relationship was known to have changed over time. For most of the remaining 27 items both an amount and a flag indicating whether the item was present on the return needed to be considered. In addition, some of the items could assume negative values, requiring at least one additional flag and amount to permit the identification of distinct effects of net income and net loss. In all, 64 variables were defined for empirical testing. This required a second stage of screening.

To reduce the number of variables sufficiently to allow us to estimate a sizeable number of logistic regressions without undue costs, we estimated by ordinary least squares (OLS) a forward stepwise regression of a dichotomous early/late indicator upon the 64 variables noted above. Because of the overall size of the microdata file (144,322 records) and the extremely skewed distribution of the dependent variable, we subsampled the early returns to reduce the number of observations and to more nearly equalize the numbers of early and late returns. Such subsampling on the dependent variable biases the OLS parameter estimates and the logistic regression intercept estimates. In the latter case the magnitude of the bias is a simple function of the sampling rate, so the intercepts can be corrected. In the former case there is no correction. However, in using the OLS procedure simply to screen out the weakest predictors of late posting, we judged the bias to be inconsequential.

Subsampling was done at the sample code level, to create analysis files for the eventual estimation of stratum-specific logistic regressions. Within each sample code we selected a fraction of early returns approximately equal to the number of late returns, except where the latter was much below 100. The combined subsample totaled just under 20,000 records, with late returns comprising 48.4 percent. The OLS regression was run on the full subsample, with dummy indicators for sample code being forced into the equation. On the basis of the regression results we dropped from further consideration 31 of the 64 variables, using fairly liberal criteria for retention. As a result, eight of the 28 items originally proposed by IRS were eliminated from any further representation whatsoever (i.e., neither as flags nor amounts).

Estimation of Propensity Models

The relationship between propensity tofile late and the predictors was found to vary greatly across sample code. Hence, final logistic regression models of the propensity toward late posting were calculated separately for 14 strata, obtained by collapsing the 29 sample codes on the basis of similar proportions of late returns in the complete file. The collapsed strata are shown in the column headings of Table 5. (Definitions of the sample codes are given in Table 1.)

The logistic regression models were developed through three rounds of alternative model estimation. Each round consisted of the estimation of an initial model using a prespecified set of predictors and the subsequent "stepping in" of additional predictors, subject to minimum statistical criteria for entry and retention. The initial model was defined with a common set of predictors across the 14 strata. The additional predictors included all of the remaining variables plus variables introduced into the analysis at this stage: stratum specific indicators of high and low income as well as several higher order interaction terms.

The objective behind including a common set of predictors in the equations for all 14 strata was to moderate the influence of sampling error upon the equation specifications across strata. Interstratum variation in the specifications was restricted to variables that exhibited net effects (at conventional significance levels) over and above the common predictors. In round one the common predictors comprised in eight of the first nine variables selected by the OLS procedure. In round two we expanded this set to include the sample code indicators (for strata combining two or more sample codes) plus 10 variables that had been stepped into the round one equations in at least four strata. At the same time we excluded from further consideration 7 variables that had been stepped into the round one equations in fewer than three strata. In the third and final round we dropped four variables from the common set and repeated the forward stepwise procedure.

The final 14 equations are reported in Table 5. The variable designations are spelled out in Table 6. The variables forced into the equations are listed in the top half of the table (CAPGAIN and above). Note that some of

the "forced in" variables were excluded from one or more equations. This occurred either because the variable (in each case a flag) was undefined in that stratum or because it was effectively constant. Except where specifically noted in the full variable name, a flag predictor distinguishes a nonzero amount (coded 1) from no amount (coded 0). The money amount variables are scaled in $100,000s. The intercept and sample code indicator coefficients have been adjusted to reflect the subsampling rates used for early returns (as explained above).

The coefficients exhibit substantial variation across strata (refer to Table 1 for sample code definitions). In part this can be attributed to the differing distributions of the variables across strata, together with the fact that the coefficients reported in the table are not standardized. This is especially true for income variables, where the amounts range from barely hundreds of dollars in the low income strata to perhaps millions of dollars in the highest income strata. The coefficients of these variables decrease substantially between the low and high income strata.

The most consistent predictors across strata are ITEMDUC, E-LOSS, TAXPREFFLG, NATGTPAT and PARTNRLOSFLG. With the exception of E-LOSS, the presence of an amount or the value of the amount was directly associated with the probability that a return would be posted late. Overseas returns had a high probability of being posted late in the four strata in which we included the OVERSEAS indicator. However, we found little evidence of variation in late posting by total income within strata. Only three of the final models include PAT indicators.

To ascertain whether our final model specification did indeed exhaust the ability of the full set of potential x variables proposed by IRS to predict late posting, we performed the following test. We combined the 14 subsamples into a single sample of close to 20,000 cases and regressed a dichotomous indicator of early/late posing on the following variables, using a forward stepwise OLS procedure:

- predicted propensity, constructed from the 14 equations;
- 65 proposed predictors drawn from the IRS list;
- PAT indicators; and
- 15 higher order interaction terms.

We forced in the predicted propensity score and allowed the remainder to be stepped in. From the results we sought to learn whether any variable added anything beyond the explanatory power captured in the propensity score.

The initial equation was estimated as:

$$\text{CYCLE} = .011 + .988P, \tag{2}$$

where CYCLE is coded 1 if late and 0 if early, and P is the propensity score. The equation produced an R^2 of 12.53 percent.

No other variable added appreciably to the explanatory power of the equation. The next four variables to enter and the associated R^2s were as follows:

Income averaging computation flag	12.58%
Foreign investment credit flag	12.62
PAT150K	12.65
Credit for tax on gasoline flag	12.67 .

The indicator for overseas returns, which we had excluded from many equations because of its very limited distribution, entered at the 13th step, raised the R^2 by only .02 percent, and fell short of statistical significance. Because of the enormous sample size, all of the preceding variables entered with statistically significant contributions, but the largest F was only 11.7, compared to 2795.7 for the propensity score. We concluded from these results that we had not overlooked any important predictors of late posting propensity among those potential predictors identified at the outset.

Calculation of Weights

The calculation of weights under the proposed new weighting scheme entails three steps:

(1) calculation of propensity scores for all observations in the AD sample, using the stratum-specific equations described above;

(2) assignment of each observation to one of K propensity classes defined for that observation's sample code; and

(3) calculation of a weight for each propensity class.

Each observation is assigned a weight corresponding to its sample code (stratum) and propensity class.

The propensity score for a given observation i in stratum j is

$$\hat{p}_{ij} = \frac{1}{1 + e^{-B_j X_{ij}}}, \tag{3}$$

where B_j is the vector of coefficients from Table 5 and X_{ij} the vector of values on the predictors.

To evaluate the propensity score method we computed two alternative sets of weights, based on two different approaches to step (3). The two methods utilize the same individual propensity scores and propensity classes. Method one defines the weight for propensity class k in stratum j to be:

$$w_{jk} = \frac{\hat{CR}^T_{jk}}{AD^S_{jk}}, \tag{4}$$

where $\hat{CR}^T_{jk}$ is the estimated CR population that would fall into propensity class k of stratum j, and AD^S_{jk} is the number of AD sample returns in the same class and stratum. Method two defines

a preliminary weight as:

$$(5) \qquad w'_{jk} = \sum_{i=1}^{AD^{s}_{jk}} \left(\frac{1}{1-p_{ijk}}\right) \Big/ AD^{s}_{jk} \ ,$$

where p_{ijk} is the predicted propensity for the ith observation in propensity class k of stratum j. The quantity $1/(1-p_{ijk})$ is the theoretical weight for an observation with a predicted propensity p_{ijk}; and the preliminary class weight is simply the mean of such individual weights for the sample observations in that class[3].

The preliminary weights are rescaled so that the weighted sum of the sample observations by stratum (that is, summed over propensity classes) equals the projected CR population of that stratum:

$$(6) \qquad w^{*}_{jk} = (w'_{jk}) \frac{\hat{CR}T_{j.}}{\sum_{k} (w'_{jk})(AD^{s}_{jk})} \ .$$

Method two assigns relatively greater weight to the higher propensity classes than does method one, the more so the higher the propensity scores. Method two yields monotonically increasing weights, whereas the weights computed under method one will not necessarily increase (and could decrease) between a given propensity class and the next higher class.

Applying method one in practice would entail projecting the total number of late returns for each propensity class within each sample code--i.e., disaggregating the projected sample code total. Unlike the projections by sample code, projections for individual propensity classes would have to be made without the benefit of periodic tabulations of returns processed during the current year. One way to do this would be to compute for each sample code the proportion of late returns by propensity class for the previous year and apply this distribution to the projected current year total. The assumption of stability in this distribution (that is, the distribution of late returns by percentile of AGI) between successive years does not seem unreasonable. In the evaluation, however, we used the actual 1982 sample estimates of late returns in each propensity class to construct the weights. We did so in order that the performance of method one not be weakened by errors in the projections of the distribution of late returns by propensity class. This is consistent with our desire to evaluate the three methods in their purest form, divorced from projection error, as the refinement of projection methods is independent of the method of calculating weights. This choice may afford a comparative advantage to method one relative to method two and the current IRS procedure in this evaluation, as method one weights will incorporate more information about actual late returns in 1982 than either of the other sets of weights.

With regard to the definition of propensity classes, we decided first of all to create six classes in each stratum. The potential impact of small weight classes on the variance of the final AD estimates suggested something like equal sized classes. Because of the much greater error in current AD estimates of money amounts than in numbers of returns (recall Table 2), there seemed more to gain by defining the classes on the basis of aggregate income rather than numbers of sample returns. In other words, the boundary between the first (lowest) and second propensity classes would be that propensity score which corresponded to the first sextile of cumulative AGI for that stratum; the boundary between the second and third classes would be that propensity score which corresponded to the second sextile of cumulative AGI; and so on. Upon reviewing distributions of propensity scores for early and late returns we determined that the distribution of weights could be improved by reducing the size of the highest class and, to compensate, enlarging the lowest class. Accordingly, we fixed the boundaries between propensity classes at three-, five-, seven-, nine-, and eleven-twelfths of the cumulative AGI distribution.

The boundaries among the propensity classes for all 29 strata are shown in Table 7. The weights computed for methods one and two and the simulated current method are reproduced in Table 8. These weights are expressed as multipliers--the growth ratios of equation (1) above--showing the relative increase in weight over and above the simple inverse of the sampling fraction. Thus a weight of 1.2 implies that to make the AD estimate of the full year population, returns in that class must represent 20 percent more returns than they did when sampled.

It may be noted that in a few of the strata we collapsed two or more propensity classes into a single class. We did so whenever the range of propensity scores for a given class turned out to be extremely small. In implementing the calculation of cut points between classes we had divided the range of possible propensity scores (that is, from zero to one, excluding the endpoints) into 81 discrete intervals.[4] We used discrete intervals primarily so that we could easily review the distributions of scores, but we chose to define the propensity classes in terms of these same discrete intervals. It happened that in some cases the lower bounds of two or more propensity classes fell into the same discrete interval. When this occurred we simply combined the classes. The most extreme example is provided by sample code 40, where four of the six propensity classes were assigned to the same discrete interval, .0075 to .0100.

D. RESULTS

To evaluate the new methods we computed three sets of advance estimates of selected income and tax items by applying the alternative weight multipliers in Table 8 to the 1982 CR sample returns posted prior to cycle 39. We then compared these estimates with tabulations based on all returns on the CR file. The results are reported below.

The selected income and tax items on which the evaluation is based were chosen because of their prominence in the AD tabulations circu-

lated by IRS. They include variables present in the propensity equations (and therefore contributors to the propensity scores that factor into the new weights) as well as items with no obvious relation to any of the variables in the models. This distinction is potentially important. We expect the propensity score approach to produce improved estimates of outcome variables that happen to be included among the x variables in the models because we know that their relationships to late posting have been incorporated into the propensity scores. We anticipate improvements in the other variables as well, because the propensity scores are intended to represent the tendency toward late posting generally. However, for variables not tested as possible predictors of late posting we have no prior empirical evidence to suggest that their relationships with late posting are indeed captured in the propensity scores. Accordingly our expectations for improved performance among these variables are less strong. Where the new procedures do not produce significantly improved estimates, we would recommend that such variables be tested as possible predictors in a future application of the new method.

Variables Represented in the Propensity Models

Absolute and percentage deviations from CR estimates are reported in Table 9 for the three alternative advance estimates of selected income and tax items that were included in some form in the propensity models. The table reports as well the CR estimates from which the deviations are measured.

For most income and tax amounts the advance estimates based on propensity score methods one and two lie substantially closer to the CR estimates than do those based on the simulated current method. The new methods also yield closer approximations to the CR estimates of numbers of returns on items where the "current" method produces deviations of one percent or more. There is generally little difference among the alternative advance estimates of numbers of returns where the current method and CR estimates are themselves very close.

The most dramatic improvements to the advance estimates of money amounts occur on dividends, itemized deductions, total tax preferences and alternative minimum tax. On total dividends, where the current method yields an excess of 737 million dollars or 1.36 percent over the CR estimate, the method one estimate falls within 76 million dollars (0.14 percent) and the method two estimate within 28 million dollars (0.05 percent). For total tax preferences a deviation of 205 million dollars or 13.50 percent obtained with current method is reduced to 16 and 52 million dollars, respectively, under methods one and two. Here even the estimates of the number of returns are much closer to the CR estimate with methods one and two than with the current method. Under the current method the number of returns with additional tax for tax preferences is underestimated by 16.7 thousand or 7.43 percent; the new methods reduce this to 5.2 and 2.9 thousand. The alternative minimum tax, a component of total tax preferences, presents a comparable picture with respect to both the number of returns and the monetary amount. For itemized deductions the current method yields a result that departs from the CR estimate by close to 1.4 billion dollars or .48 percent. The new methods approximate the CR estimate to within 366 million and 221 million dollars, respectively.

For business net profit the new methods and the current method yield very comparable estimates. However, the current method underestimates the magnitude of the net loss amount by 1.0 billion dollars or 5.75 percent while the new methods deviate from the CR estimate by only 228 million and 179 million dollars. Curiously, this combination of better loss estimates and comparable profit estimates produces less accurate estimates of net profit less loss under the new methods relative to the current. This happens because the current method estimates of net profit and net loss diverge from the CR estimates in opposite directions (i.e., one understates a positive amount while the other understates a negative amount), and the deviations largely nullify each other when the two estimates are summed.[5] This kind of result is not repeated for any other of the net profit less loss amounts reported in Table 9, however.

For partnership income we again find no appreciable differences among the advance estimates of net income, whereas the new methods yield markedly smaller deviations than the current method (by one-half to two-thirds) on net loss. Here, unlike business income, the relative magnitudes of the deviations on income and loss are such that the two new methods lead to substantially better advance estimates of CR net income less loss than does the current method. The current method differs from the CR by 2.6 billion dollars on a net income less loss amount of 899 million dollars. The new methods deviate from the CR estimate by 569 million and 220 million dollars, respectively.

On the whole, the new methods yield no improvement over the current method on amounts of capital assets sales. Similarly, the new methods show little improvement relative to the current method on employee business expenses--an item on which the current method yields results very close to the CR estimates of both number of returns and money amount.

Variables Not Represented in the Propensity Models

Absolute and percentage deviations from CR estimates are reported in Table 10 for the three alternative advance estimates of selected income and tax items that were not represented directly in the propensity models. The variables not included in the propensity models but selected for this evaluation constitute major income and tax items frequently included in summary reports of advance estimates or highlighted in publications based on the CR tabulations. The results reported in Table 10 show that the gains in accuracy seen in Table 9 for variables included in the propensity models do indeed generalize to variables that were not used to construct propensity scores, although they do not generalize to all the variables we examined. On advance estimates of money amounts the propensity score methods often show sizable improvements over the

current method; rarely do the new methods result in less accurate estimates. For numbers of returns, on the other hand, we find generally little or at best modest improvement.

Where the propensity score methods yield significantly better estimates of money amounts, the margin of improvement over the current method (expressed as the proportionate reduction in the deviation from the CR estimate) generally range between 50 and 75 percent. Improvements of this magnitude may be seen on AGI; net losses on farm, estate or trust, Small Business Corporation, and other income; taxable income; and alternative measures of income tax. On AGI the estimate based on the current method exceeds the CR estimate by 9.7 billion dollars, or .53 percent. The propensity score methods reduce this deviation to 4.6 billion and 2.9 billion dollars. On farm net loss the current method understates in absolute magnitude the CR estimate by 334 million dollars or 1.87 percent. The propensity score methods reduce this deviation to 81 million and 42 million dollars. The improvement for net profit less loss is comparable, with the percentage deviation being reduced from 3.51 percent to .80 and .59 percent.

On Small Business Corporation income, the new methods produce estimates with somewhat greater deviations than the current method on net profit, but they compensate with substantially smaller deviations on net loss. The net effect is such that whereas the current method yields an estimate of the small and volatile net profit less loss that is over one billion dollars wide of the mark, the new methods produce estimates that are within 441 and 279 million dollars. On other income the new methods generate slightly improved estimates of net profit and much more substantially improved estimates of net loss. On net income less loss, the current method understates the magnitude of the 10.3 billion dollar aggregate loss by 2.0 billion dollars or 19.2 percent. The new methods reduce the margins on other income profit and loss to 921 and 568 million, respectively, or 8.9 and 5.5 percent.

Perhaps not surprisingly, the improvements for taxable income are very comparable to those registered for AGI. Here there are parallel improvements for both number of returns and money amount. Comparable improvements are realized for income tax before and after credits, total income tax, and total tax liabilities. On tax after credits, the current method produces an estimate that is high by 2.8 billion dollars or 1.02 percent. Propensity score method one reduces the error to less than 1.2 billion, and method two lowers it still further to 1.0 billion, or .36 percent of the CR estimate. Even the estimated number of returns shows proportionate improvements on this order, although the current method estimate is itself very close to the CR estimate, differing by less than one-tenth of one percent.

Small improvements are recorded for interest paid deduction and for payments to an IRA, and smaller still for salaries and wages, medical and dental expenses, and taxes paid deduction. On other items--unemployment compensation, pensions and annuities, alimony, and deduction for a two-earner couple--the propensity score methods show little or inconsistent improvement relative to the current method. On exemptions and contributions deductions the new methods fare somewhat worse than the current method.

Within Stratum Comparisons

The weighted AD file is used routinely to estimate not only aggregates over all taxpayers but also distributions by detailed AGI class. It is of interest, therefore, to consider to what extent the comparative advantage of the propensity score methods extends to the subaggregate level. Because they use more weight classes, estimates based on the propensity score methods might exhibit less bias but greater variance than comparable estimates based on the current method. Another issue is whether the relative superiority of the propensity score methods increases with the frequency of late posting. We have seen evidence of this in the comparative performance of the alternative advance estimates of items with substantial versus little late posting, but there the results may reflect the ability of our propensity models to capture the tendency toward late posting in those particular variables. Comparisons by sample code provide more general evidence in that sample codes with relatively high rates of late posting overall will exhibit late posting in all variables. Finally, the application of the propensity score method entailed the estimation of separate equations by groups of sample codes, with differential results. It remains to be seen whether there is any evidence to support changes in the specifications of the models or the grouping of sample codes.

Table 11 presents the CR estimate and percentage deviations from that estimate by sample code for three of the items from Table 9 and three from Table 10. The items selected are dollar amounts, and they include items on which the propensity score methods performed substantially better than the current method as well as items on which the propensity score methods produced no aggregate improvement. By including the latter items we seek to determine whether the lack of improvement on those items characterizes all sample codes or whether it reflects systematically better performance in some sample codes and worse performance in others.

As reported above, the aggregate AD estimate of total dividends received is 1.36 percent above the CR, whereas the estimate based on propensity score method one (PSM-1) falls within .14 percent and method two (PSM-2) within .05 percent. Reviewing the results by sample code we find, first of all, that the simulated AD estimate tends to overstate the CR estimate by an increasing amount as the sample code income level, and with it the rate of late posting, rises (that is, from codes ending in 0 to codes ending in 8). There is no such pattern among the propensity score estimates, although there is a suggestion of increasing variance (the estimates being somewhat farther from the CR estimates in both the positive and negative direction as sample code rises). Within the non-business, farm returns (50s codes) the current AD method performs as well as method two

overall and, except for the two highest codes, also as well as method one. Within the 40s and 60s codes the propensity score methods perform at least as well relative to the AD method in the high income sample codes as they do in the low income codes.

The AD estimates of AGI exhibit to an even more pronounced degree the pattern observed for gross dividends. The propensity score methods also show evidence of increasing distance from the CR estimates with rising sample code in the non-business non-farm and the business codes, but the growth is less rapid than for the AD estimates. As a result the propensity score estimates increase their advantage over the current AD estimates as the propensity for late posting rises. Within the non-business, farm returns, method two falls increasingly below the CR estimate as sample code increases while method one exhibits no clear pattern.

On itemized deductions the strong aggregate performance of the propensity score methods relative to the AD method is seen to be the result of superior performance within a small subset of sample codes together with small overestimates in very large strata counterbalancing the underestimates in most other strata. The propensity score methods provide closer approximations to the CR estimates in code ranges 42-45 and 60-63. Within the farm strata the AD method is, if anything, somewhat closer to the CR estimates than are the propensity score methods.

On total tax preferences the propensity score methods produce much better estimates than the AD method through sample codes 43-46 and 62-67, which account for most of the aggregate amount of total tax preferences. Within the farm codes there is again basically no advantage to any of the methods.

On salaries and wages and estate or trust net profits--items on which the aggregate estimates of the AD and two propensity score applications are fairly comparable--the sample code-specific estimates are themselves very similar (especially on estate/trust). The top two or three codes for each type of return generate the largest errors, but below that there is little differentiation by sample code in the average magnitudes of the errors.

This brief survey of results by sample code suggests several conclusions. For those items where the aggregate estimates based on propensity score methods are substantially closer to the CR estimates than are the simulated AD estimates, we do not find this level of performance repeated in every sample code. On the other hand, neither do we find that propensity score methods achieve their lower overall bias with higher variance at the sample code level. In approximately three-quarters of the sample codes the estimates based on the propensity score methods deviate less from the CR estimates than do the estimates based on the current AD procedures. Moreover, the relative superiority of the new methods tends to be most pronounced in those sample codes where the current method produces the largest deviations from the CR estimates--namely, those sample codes with high proportions of returns posted late. For items where the three advance estimates of aggregate amounts differ little from each other, the sample code estimates tend to be similar as well.

Estimates among non-business, farm returns tend not to show the level of improvement as we find among non-farm returns. This may be attributable to the propensity equations. Because the samples within the farm codes are quite small, we combined non-business, farm returns with non-business, non-farm returns by income level (based on comparable rates of late posting) in estimating the equations, and we did not attempt to estimate farm-specific interactions other than through the intercepts. Perhaps another strategy for aggregating farm returns would have yielded better predictions of late posting among farm returns. This would be an issue for future research.

Discussion

In assessing the implications of the evaluation, it is very significant to note that the gains in accuracy realized with the propensity score methods are by no means limited to variables that were used to construct the propensity scores. One appeal of the propensity score approach to weighting is that an adequate model of the underlying propensity phenomenon should permit improved estimates of whatever observed variables exhibit such tendencies. In the present study the empirical development of propensity models was restricted by an a priori selection of potential covariates. Yet the method yielded improved estimates of key variables outside this basic set. Refinements of the method in this particular context would broaden the set of variables searched to develop the models, adding variables for which (further) improvements are desired.

Another aspect of the results that surprised us was the marginal superiority of method two over method one. Method one used the actual complete report estimates of the numbers of returns in each propensity class to construct the weights by propensity class. Method two, like the simulated current method, used only the sample code population estimates, relying on the propensity scores themselves to generate the differential weights. Yet the results presented in Tables 9 and 10 show the following performance of the two methods:

Method of Comparison	# of Returns	Amount
Variables Included in Propensity Models		
Method 2 better than 1	5	9
Method 1 better than 2	11	7
Variables Not Included in Models		
Method 2 better than 1	18	22
Method 1 better than 2	11	11

On the whole, the estimates based on method two are superior to those based on method one. Since method two is more directly applicable in practice than is method one, these results suggest that the gains observed in this evaluation should indeed be realizable in practice.

E. SUMMARY

This paper has proposed and tested a new method of preparing annual estimates of income and tax statistics from an advance or truncated sample. The problem examined here is one of adjusting for unit nonresponse, where the missing units are tax returns yet to be added to the sample, and the missing returns are known to differ from the advance returns on the variables for which advance estimates are desired. No information is available from the missing returns for the current year, but all returns are available for previous years. Currently IRS projects the total number of late returns by sampling stratum and computes stratum-specific weights to inflate the sampled returns to the projected totals. The proposed new method involves differentially weighting the returns within the sample strata on the basis of their estimated propensity to be late. Individual propensity scores are computed on the basis of logistic regression models of late posting estimated on the full-year sample from the previous year. Within each sample stratum the advance returns are then divided into six propensity classes, and each class is assigned a weight translating the propensity into a projected increase in the number of returns of that type.

Weights for the new method were estimated using tax year 1981 data and then applied to data from tax year 1982. Estimates were prepared from the advance sample using two different methods of reweighting the propensity classes. Deviations of these estimates from the complete year estimates were compared with deviations based on assigning a single weight per stratum (the current method). The new methods produced estimates that were generally much closer to the final estimates for variables included in the propensity equations. Estimates of dollar amounts showed much greater improvement proportionally than estimates of the number of returns on which those income or tax items were present. Currently, there is greater error on amounts than on numbers of returns. Because the propensity equations are intended to represent the general tendency to be posted late, improvements are expected for variables not included in the equations as well as those that are included. Such improvements are limited by the extent to which the equations do indeed represent a general propensity. We registered improvements on several variables that were not included in the equations. Variables on which no improvements were recorded would be prime candidates for inclusion in future model development, as they reflect aspects of late posting not captured in the models presented here.

The results demonstrate the value of reweighting on the basis of propensity scores. The methods tested here could be applied on a regular basis to the preparation of advance estimates and provide sizeable error reductions for income and tax items that are poorly estimated from advance data using the current methodology. With further research and development the method could be tailored to provide improvements where they are most needed while maintaining or even improving upon the accuracy of key variables for which the current magnitudes of error are already comparatively low.

FOOTNOTES

[1] Improvements to the projection methodology were the subject of another research effort under this same project. The results of that effort are detailed in Czajka (1984a, 1984b, 1985).

[2] Personal communication from Ray Shadid, IRS, Statistics of Income Division.

[3] The implied individual weight grows very rapidly as p approaches unity. To avoid excessively high weights for the top propensity class we computed the class weight as the inverse of one less the mean propensity.

[4] The intervals were as follows: zero to .0300 in increments of .0025 (12 intervals); .030 to .040 in increments of .005 (2 intervals); .04 to .40 in increments of .01 (37 intervals); and .4 to 1.0 in increments of .02 (30 intervals). We divided the intervals this way because of the extreme bunching of propensity scores at low values in strata with low mean propensities.

[5] Note that the net loss amount was included among the predictors in the propensity equations whereas the net profit amount was not. Net profit was tested but rejected as a predictor at an early stage of the model development. A possible explanation is that the relationship between net business income and late posting changed between 1981 and 1982. In further research using 1982 data we found net profit to be a significnat predictor of late posting in a number of strata.

ACKNOWLEDGMENTS

This research was performed under a contract with the Internal Revenue Service. The authors wish to thank Fritz Scheuren, Director of the Statistics of Income Division, for his valuable input to the research, the assistance provided by members of his staff, and his helpful comments on an earlier draft. The assistance of Lucia Wesley, who typed the manuscript, is also gratefully acknowledged.

REFERENCES

Czajka, John L. "Evaluation of the Projections of 1983 Individual Tax Returns." Memorandum. Washington: Mathematica Policy Research, Inc., 1985.

_____. "Projections of Total Tax Returns from an Advance Date: Evaluation of Alternatives." Project Report. Washington: Mathematica Policy Research, Inc., 1984a.

_____. "Revised Projections of Total Tax Returns by Sample Code, Tax Year 1983." Memorandum. Washington: Mathematica Policy Research, Inc., 1984b.

Czajka, John L., Donald B. Rubin and Roderick J.A. Little. "Modeling the Propensity Toward Late Posting of Tax Returns: An Application to Tax Year 1981." Project Report. Washington: Mathematica Policy Research, Inc., 1984.

David, Martin, Roderick J.A. Little, Michael Samuhel and Robert Triest. "Imputation Models Based on the Propensity to Respond." 1983 Proceedings of the Business and Economics Section. Washington: American Statistical Association.

Internal Revenue Service. Statistics of Income--1982: Individual Income Tax Returns. Washington: U.S. Government Printing Office, 1984.

_____. Statistics of Income--1981: Individual Income Tax Returns. U.S. Government Printing Office, 1983.

_____. "Advance Data: 1981 Individual Income Tax Returns." Mimeograph. Washington: Internal Revenue Service, 1982.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika, vol. 70, 1983, pp. 41-55.

Sailer, Peter, Charles Hicks, Dave Watson and Dan Trevors. "Results of Coverage and Processing Changes to the 1980 Individual Statistics of Income Program." 1982 Proceedings of the Section on Survey Research Methods. Washington: American Statistical Association.

TABLE 1

UNWEIGHTED SAMPLE COUNTS OF TOTAL RETURNS AND LATE RETURNS (CYCLES 39+) BY SAMPLE CODE: TAX YEAR 1981

Sample Stratum		Non-business, Non-farm Returns				Non-business, Farm Returns				Business Returns			
		Sample Code	Total Returns	Cycle 39+ Returns	Percent Cycle 39+	Sample Code	Total Returns	Cycle 39+ Returns	Percent Cycle 39+	Sample Code	Total Returns	Cycle 39+ Returns	Percent Cycle 39+
High Income Nontaxable Returns										28	241	58	24.1
Schedule C Net Profit or Loss of $200,000 or More										38	4,390	1,397	31.8
Individual Returns Not Coded 28 or 38:													
Larger of Absolute Value of PAT or NAT	Total Receipts												
<$20,000	<$200,000	40	29,840	261	0.9	50	714	6	0.8	60	10,230	350	3.4
$20,000 <$50,000	<$500,000 $200,000 <$500,000	41	19,426	170	0.9	51	774	13	1.7	61	12,511	337	2.7
$50,000 <$100,000	<$1,000,000 $500,000 <$1,000,000	42	7,462	141	1.9	52	555	15	2.7	62	7,240	369	5.1
$100,000 <$200,000	<$5,000,000 $1,000,000 <$5,000,000	43	5,752	227	3.9	53	942	57	6.1	63	7,044	549	7.8
$200,000 <$500,000	<$10,000,000 $5,000,000 <$10,000,000	44	5,912	420	7.1	54	1,112	88	7.9	64	4,054	553	13.6
$500,000 <$1,000,000	<$20,000,000 $10,000,000 <$20,000,000	45	9,992	1,209	12.1	55	1,631	218	13.4	65	4,545	874	19.2
$1,000,000 <$2,000,000	<$30,000,000 $20,000,000 <$30,000,000	46	4,141	667	16.1	56	749	129	17.2	66	2,330	606	26.0
$2,000,000 <$5,000,000	<$50,000,000 $30,000,000 <$50,000,000	47	1,287	333	25.9	57	265	64	24.2	67	698	222	31.8
$5,000,000 and Over	Any Amount $50,000,000 and Over	48	280	100	35.7	58	61	20	32.8	68	144	55	38.2

SOURCE: 1981 SOI Individual/Sole Proprietorship sample file; tabulations prepared by Mathematica Policy Research.

NOTE: The sample codes here reflect the 1982 rather than 1981 design. The 1982 stratum definitions were applied to the corresponding variables on the 1981 Complete Report sample file. The replication is complete except for the substitution of the larger of interest income or taxable interest and dividends for the sum of interest income and taxable dividends, the latter of which is not recorded separately on the 1981 file.

TABLE 2

COMPARISON OF ADVANCE DATA AND COMPLETE REPORT ESTIMATES
OF SELECTED INCOME AND TAX ITEMS, 1981

Item	Number of Returns				Money Amount (Millions of Dollars)			
	Advance Data	Complete Report	Absolute Deviation AD-CR	Percentage Deviation (AD-CR)/CR	Advance Data	Complete Report	Absolute Deviation AD-CR	Percentage Deviation (AD-CR)/CR
Total Number of Returns	95,284,813	95,396,123	-111,310	-0.12%	--	--	--	--
Adjusted Gross Income Less Deficit	--	--	--	--	1,779,359.5	1,772,604.3	6,755.2	0.38%
Salaries and Wages	84,199,098	84,208,807	-9709	-0.01	1,489,233.7	1,486,100.5	3,133.2	0.21
Interest Received	49,706,957	49,656,550	50,407	0.10	139,901.5	140,559.4	-657.9	-0.47
Gross Dividends Received	16,500,857	16,482,018	18,839	0.11	48,470.2	48,161.5	308.7	0.64
Business Income (Schedule C)								
Net Profit	6,525,440	6,534,688	-9,248	-0.14	68,154.5	68,531.4	-376.9	-0.55
Net Loss	3,042,256	3,036,721	5,535	0.18	-14,543.4	-15,459.8	916.4	-5.93
Net Profit Less Loss	9,567,696	9,571,409	-3713	-0.04	53,611.1	53,071.6	539.5	1.02
Sales of Capital Assets (Schedule D)								
Net Gain	6,982,760	7,025,861	-43,101	-0.61	33,143.1	34,713.1	-1,570.0	-4.52
Net Loss	2,450,097	2,459,126	-9,029	-0.37	-3,861.9	-3,894.6	32.7	-0.84
Net Gain Less Loss	9,432,857	9,484,987	-52,130	-0.55	29,281.2	30,818.5	-1,537.3	-4.99
Farm Income (Schedule F)								
Net Profit	980,843	983,543	-2,700	-0.28	8,445.8	8,532.1	-86.3	-1.01
Net Loss	1,648,580	1,657,711	-9,131	-0.55	-15,877.3	-16,344.1	466.8	-2.86
Net Profit Less Loss	2,629,423	2,641,254	-11,831	-0.45	-7,431.5	-7,812.0	380.5	-4.87
Estate or Trust Income Less Loss	760,731	770,524	-9,793	-1.27	3,875.0	3,965.8	-90.8	-2.29
Partnership Net Profit Less Loss	3,680,368	3,752,209	-71,841	-1.91	2,074.6	-137.6	2,212.2	*
Small Business Corporation								
Net Profit Less Loss	740,227	778,408	-38,181	-4.91	-444.9	-817.2	372.3	-45.56
Pensions and Annuities in AGI	8,173,818	8,157,475	16,343	0.20	52,079.9	51,886.4	193.5	0.37
Unemployment Compensation in AGI	2,251,651	2,245,250	6,401	0.29	2,316.7	2,315.3	1.4	0.06
Payments to an IRA	3,434,455	3,415,053	19,402	0.57	4,773.0	4,750.2	22.8	0.48
Itemized Deductions	31,515,624	31,571,246	-55,622	-0.18	255,370.6	256,448.0	-1,077.4	-0.42
Employee Business Expenses	6,894,190	6,919,044	-24,854	-0.36	14,890.5	15,065.4	-174.9	-1.16
Taxable Income	89,805,882	89,851,304	-45,422	-0.05	1,415,615.5	1,410,880.7	4,734.8	0.34
Income Tax Before Credits	78,974,594	79,011,548	-36,854	-0.05	295,662.8	293,590.0	2,072.8	0.71
Income Tax After Credits	76,601,906	76,635,184	-33,278	-0.04	284,232.9	282,302.0	1,930.9	0.68
Additional Tax for Tax Preferences								
Total	236,058	250,908	-14,850	-5.92	1,617.1	1,827.0	-209.9	-11.49
Minimum Tax	119,546	125,721	-6,175	-4.91	513.1	565.6	-52.5	-9.28
Alternative Minimum Tax	127,427	137,113	-9,686	-7.06	1,103.9	1,261.3	-157.4	-12.48
Total Income Tax	76,682,212	76,724,724	-42,512	-0.06	285,849.9	284,129.0	1,720.9	0.61
Tax Due at Time of Filing	22,799,649	23,007,970	-208,321	-0.91	35,110.0	34,907.6	202.4	0.58
Total Overpayment	70,145,777	70,028,741	117,036	0.17	56,525.9	56,635.0	-109.1	-0.19

SOURCE: Internal Revenue Service (1982, 1983).

*Because of the magnitude of the absolute deviation compared to the small base estimate a percentage deviation would not be meaningful.

TABLE 3
DISTRIBUTION OF SELECTED INCOME AND TAX TOTALS BY POSTING PERIOD,
TAX YEARS 1981 AND 1982
(Money amounts in millions of dollars)

Item	Tax Year 1981				Tax Year 1982			
	Complete Year Total	Distribution by Period*		Percent Posted Late	Complete Year Total	Distribution by Period*		Percent Posted Late
		Posted Early	Posted Late			Posted Early	Posted Late	
Number of Returns	95,520,265	94,155,237	1,365,028	1.43%	95,608,581	93,518,967	2,089,614	2.19%
Adjusted Gross Income Less Deficit	1,770,644.2	1,738,585.5	32,058.7	1.81	1,845,330.1	1,791,744.3	53,585.8	2.90
Salaries and Wages								
Number of Returns	83,663,168	82,720,851	942,317	1.13	82,608,556	81,197,737	1,410,819	1.71
Amount	1,473,658.2	1,450,614.2	23,044.0	1.56	1,546,189.4	1,507,306.7	38,882.7	2.51
Interest Received								
Number of Returns	49,377,509	48,637,686	739,823	1.50	52,469,253	51,236,970	1,232,283	2.35
Amount	141,294.6	136,562.1	4,732.5	3.35	156,287.5	148,902.1	7,385.4	4.73
Gross Dividends								
Number of Returns	16,386,013	16,065,465	320,548	1.96	16,548,470	16,015,124	533,346	3.22
Amount	48,395.5	45,827.1	2,568.4	5.31	52,336.9	48,270.3	4,066.6	7.77
Business Net Profit Less Loss								
Number of Returns	9,922,886	9,511,927	419,959	4.14	10,491,304	9,815,980	675,324	6.44
Amount	52,996.5	50,276.1	2,720.4	5.13	50,779.1	45,766.6	5,032.5	9.91
Net Capital Gain Less Loss								
Number of Returns	8,237,487	7,972,277	265,210	3.22	8,441,657	8,017,589	424,068	5.02
Amount	30,557.6	25,746.2	4,811.4	15.75	35,614.2	28,599.6	7,014.6	19.70
Farm Net Income Less Loss								
Number of Returns	2,711,844	2,653,244	58,600	2.16	2,705,546	2,615,667	89,879	3.32
Amount	-7,963.4	-6,897.9	-1,065.5	13.38	-9,811.7	-8,281.3	-1,530.4	15.60
Pensions and Annuities in AGI								
Number of Returns	7,100,197	6,995,055	105,142	1.48	7,734,992	7,556,868	178,124	3.32
Amount	46,961.4	46,323.5	637.9	1.36	53,787.6	52,586.1	1,201.5	2.23
Unemployment Compensation in AGI								
Number of Returns	2,200,505	2,180,620	19,885	0.90	5,475,638	5,408,078	67,560	1.23
Amount	2,298.0	2,273.3	24.7	1.07	7,467.9	7,359.9	108.0	1.45
Payments to an IRA								
Number of Returns	3,458,180	3,392,768	65,412	1.89	11,948,697	11,642,622	306,075	2.56
Amount	4,795.3	4,692.1	103.2	2.15	28,145.4	27,336.0	809.4	2.88
Itemized Deductions								
Number of Returns	32,111,348	31,516,340	595,008	1.85	33,819,331	32,827,438	991,893	2.93
Amount	258,712.4	249,196.6	9,515.8	3.68	288,397.8	272,508.5	15,889.3	5.51
Taxable Income								
Number of Returns	89,804,116	88,617,454	1,186,662	1.32	89,748,370	87,927,175	1,821,195	2.03
Amount	1,410,334.6	1,384,379.4	25,955.2	1.84	1,469,516.8	1,425,030.9	44,485.9	3.03

Table 3 (continued)

Item	Tax Year 1981				Tax Year 1982			
	Complete Year Total	Distribution by Period*		Percent Posted Late	Complete Year Total	Distribution by Period*		Percent Posted Late
		Posted Early	Posted Late			Posted Early	Posted Late	
Income Tax Before Credits								
Number of Returns	78,788,823	77,732,789	1,056,034	1.34	79,279,999	77,641,042	1,638,957	2.07
Amount	290,269.7	282,589.2	7,680.5	2.65	282,898.2	270,766.1	12,132.1	4.29
Total Tax Credits								
Number of Returns	16,752,399	16,401,778	350,621	2.09	16,258,726	15,717,421	541,305	3.33
Amount	7,553.0	6,912.3	640.7	8.48	7,253.8	6,451.8	802.0	11.06
Investment Credit								
Number of Returns	4,616,328	4,420,689	195,639	4.24	4,405,837	4,109,037	296,800	6.74
Amount	3,999.4	3,621.4	378.0	9.45	4,000.6	3,434.2	566.4	14.16
Minimum Tax								
Number of Returns	116,749	100,768	15,981	13.69	109,181	88,902	20,279	18.57
Amount	545.7	415.6	130.1	23.84	464.5	321.5	143.0	30.79
Alternative Minimum Tax								
Number of Returns	125,393	105,694	19,699	15.71	130,819	103,148	27,671	21.15
Amount	1,087.0	775.6	311.4	28.65	1,012.5	666.0	346.5	34.22
Balance Due								
Number of Returns	23,047,734	22,407,638	640,096	2.78	20,244,531	19,237,670	1,006,861	4.97
Amount	35,443.7	33,688.7	1,755.0	4.95	29,556.3	27,226.4	2,329.9	7.88
Overpayment								
Number of Returns	69,886,019	69,274,305	611,714	0.88	72,493,767	71,590,187	903,580	1.25
Amount	56,439.2	55,061.5	1,377.7	2.44	63,619.8	61,488.7	2,131.1	3.35

SOURCE: Internal Revenue Service, Statistics of Income Division, unpublished tabulations from the Individual Master File; compiled by Mathematica Policy Research, Inc.

*For Tax Year 1981 "early" returns were posted as of September 18, 1982. For Tax Year 1982 early returns were posted as of September 15, 1983.

TABLE 4

DEVIATION OF ADVANCE DATA FROM COMPLETE REPORT ESTIMATES EXPRESSED AS A PERCENTAGE OF LATE POSTED RETURNS AND AMOUNTS: TAX YEAR 1981

Item	Percentage Error on Late Posted Number of Returns	Percentage Error on Late Posted Money Amounts
Total Number of Returns	-8.15%	--
Adjusted Gross Income Less Deficit	--	21.07%
Salaries and Wages	-1.03	13.51
Interest Received	6.81	-13.90
Gross Dividends Received	5.88	12.02
Business Net Profit Less Loss	-0.88	19.83
Net Capital Gain Less Loss	-19.66	-31.95
Farm Net Income Less Loss	-20.19	-35.71
Pensions and Annuities in AGI	15.54	30.33
Payments to an IRA	29.66	22.09
Itemized Deductions	-9.35	-11.32
Taxable Income	-3.83	18.24
Income Tax Before Credits	-3.50	26.99
Minimum Tax	-38.64	-40.35
Alternative Minimum Tax	-49.17	-50.55
Balance Due	-32.55	11.53
Overpayment	19.13	-7.92

SOURCE: Tables 2 and 3. Entries are absolute deviations from Table 2 divided by quantities posted late from Table 3, then multiplied by 100 percent.

TABLE 5

LOGISTIC REGRESSION COEFFICIENTS OF PROPENSITY MODELS, BY SAMPLE CODE

Predictor Variable	28	38	40 41 50 51	42 43 52 53	44 54	45 55	46 56	47 48 57 58	60 61	62 63	64	65	66	67 68
Intercept	-1.999	-1.160	-4.818	-3.844	-2.965	-1.967	-1.689	.108	-3.336	-2.622	-2.167	-2.000	-2.233	-1.278
2nd Code	--	--	-.239	.349	-.127	-.054	.009	-.032	-.394	.059	--	--	--	.030
3rd Code	--	--	-.458	-.078	--	--	--	-.180	--	--	--	--	--	--
4th Code	--	--	.085	.591	--	--	--	-1.145	--	--	--	--	--	--
ITEMDUCFLG	.864	-.637*	.163	-.386	-.411	-1.041*	-.534	-1.721*	-.384*	-.338	-.363	-.145	-.434	-.533
PARTNRLOSFLG	.343	.333*	1.194*	.111	-.070	.211*	.063	.290	-.086	.289*	.135	.340*	.280	.618*
ALTMINTAXFLG	.626	.615*		1.334*	.415	.277	-.422	.484*		.174	.129	.215	.124	.242
TAXPREFFLG	-.254	.325*		.151	1.581*	.174	.315*	.382*		.348	.239	.215*	1.005*	.462*
DIVGT400	-.422	-.055	-.214	-.121	-.303	-.346*	-.380*	.360	-.219	-.474*	-.394*	-.117	-.482*	.165
E-INCOMEFLG	1.227	.446*	.450*	.349*	.330	.376*	.045	-.209	.251	.364*	.277	.256	.418	-.553
INTGT400	-.047	-.468*	-.124	-.183	-.360	-.140	-.471	-1.157*	-.486*	-.609*	-.162	-.218	.434	
ITEMDUC	.010	.018	.754	1.26*	.733*	.213*	.152*	.011	8.29*	1.72*	.349*	.096	.110*	.010
DIVINCOME	-.250	-.028*	3.17	-.448	-.255	-.190*	-.113*	-.003	4.33	-.517	.208	-.189*	-.095*	-.010
E-LOSS	-.757	-.037*	1.10	-.275	-.302*	-.121*	-.075*	-.003	-5.03	-1.28*	-.026	-.110*	-.056*	-.012
BUSEXPFLG	.079	.256	.535*	-.203	.394*	.314*	.540*	-.092	-.438*	-.294	-.175	.231	.252	.470
NATGTPAT	.083	.222	1.804*	.812*	1.006*	.314	.826*	.074	1.165*	.767*	.457	.594*	.515	.456
C-LOSS	-.351	-.024*	--	--	--	--	--	--	-2.24	-.465	-.984*	-.347*	-.191	-.325
CAPGAIN	-.085	.002	4.93	.715	.190	.095*	.018	.003	10.81*	.474	.219	.086*	.010	-.001

Continued

Table 5 (continued)

Predictor Variable	28	38	40 41 50 51	42 43 52 53	44 54	45 55	46 56	47 48 57 58	60 61	62 63	64	65	66	67 68
OVERSEAS				2.678*	1.532*	2.837*	1.456*							
PARTNRLOS	-.742						.051*				.271*		.054*	
ALTMINTAX		-.077*		-13.85*				.149*						
E-LOSSFLG	-1.110*							.384*		-.344*				
INVSTCRDFLG					.397*	.276*	.317*							
ENERGYFLG						-.398*				-.421*		-.691*		
TAXPREF								.066*						.064*
C-LOSSFLG									-.395*			-.249*		
JOBCREDFLG												.805*		.881*
GASTAXFLG		-.573*												
PARTNRGAINFLG						.189*		.383*						
THEFTLOS									92.18*					
PAT75K										.278*				
PAT750K		.277*												
PAT3500K*CODE58								2.077*						
PAT7500K*CODE58								-2.401*						
PARTNR*TAXPRF					-1.266*								-.778*	
DIV*ALTMIN							.810*							
N	119	2893	1133	1013	1032	2916	1614	1033	1377	1820	1136	1791	1180	604

SOURCE: 1981 SOI Individual/Sole Proprietorship sample file; estimates prepared by Mathematica Policy Research.

* Statistically significant at the .05 level. Statistical significance was not assessed for adjusted intercept and sample code coefficients.

TABLE 6

MNEMONIC DESIGNATIONS FOR PREDICTIONS IN FINAL MODELS

MNEMONIC	Full Variable Name
ITEMDUCFLG	Total itemized deductions flag
PARTNRLOSFLG	Partnership net loss flag
ALTMINTAXFLG	Alternative minimun tax flag
TAXPREFFLG	Total tax preference flag
DIVGT400	Flag for dividend income greater than $400
E-INCOMEFLG	Schedule E net income flag
INTGT400	Flag for interest income greater than $400
ITEMDUC	Total itemized deductions amount
DIVINCOME	Dividend income amount
E-LOSS	Schedule E net loss amount
BUSEXPFLG	Employee business expenses flag
NATGTPAT	Flag for negative amounts total exceeding positive
C-LOSS	Schedule C net loss amount
CAPGAIN	Net capital gain amount reported on schedule D
OVERSEAS	Return from overseas
PARTNRLOS	Partnership net loss
ALTMINTAX	Alternative minimum tax amount
E-LOSSFLG	Schedule E net loss flag
INVSTCRDFLG	Investment credit flag
ENERGYFLG	Residential energy credit flag
TAXPREF	Total tax preferences amount
C-LOSSFLG	Schedule C net loss flag
JOBCREDFLG	Jobs credit flag
GASTAXFLG	Credit for tax on gasoline flag
PARTNRGAINFLG	Partnership net gain flag
THEFTLOS	Casualty and theft loss amount
PAT75K	Flag indicating PAT > $75,000
PAT750K	Flag indicating PAT > $750,000
PAT3500K	Flag indicating PAT > $3,500,000
PAT7500K	Flag indicating PAT > $7,500,000
PAT7500K*CODE58	Product of indicator PAT7500K and sample code 58 indicator
PAT3500K*CODE58	Product of indicator PAT3500K and sample code 58 indicator
PARTNR*TAXPRF	Product of partnership net loss flag and tax preferences flag
DIV*ALTMIN	Product of flag for dividend income greater than $400 and alternative minimum tax flag

TABLE 7

LOWER BOUNDS OF PROPENSITY CLASSES TWO THROUGH SIX, BY SAMPLE CODE

Sample Code	Propensity Class				
	2	3	4	5	6
28	.1600	.2300	.2900	.4000	.5600
38	.1800	.2300	.2900	.4000	.6600
40	.0075	*	*	*	.0100
41	.0050	.0075	*	*	.0125
42	.0125	.0150	*	.0175	.0225
43	.0225	.0275	*	.0350	.0600
44	.0275	.0400	.0500	.0700	.1500
45	.0500	.0700	.0900	.1200	.2200
46	.0600	.0700	.1000	.1600	.3200
47	.1000	.1200	.1700	.2800	.4600
48	.1100	.1400	.2300	.4000	.6000
50	.0050	*	*	.0075	.0275
51	.0075	.0100	*	.0125	.0400
52	.0125	.0150	.0175	.0225	.0400
53	.0275	.0400	.0500	.0800	.1200
54	.0400	.0500	.0800	.1100	.1600
55	.0600	.0900	.1300	.1800	.2400
56	.0700	.0900	.1600	.2100	.3000
57	.1000	.1300	.2300	.3300	.5200
58	.0350	.0600	.1100	.2400	.6000
60	.0225	.0275	*	*	.0500
61	.0150	.0175	.0225	.0275	.0500
62	.0275	.0350	.0400	.0600	.0900
63	.0400	.0500	.0700	.1000	.1900
64	.0800	.0900	.1200	.1600	.2500
65	.1100	.1400	.1600	.2200	.3300
66	.1300	.1600	.2000	.2700	.4600
67	.1700	.1900	.2400	.3300	.4200
68	.1600	.1900	.2600	.3500	.5600

SOURCE: Mathematica Policy Research, Inc.

NOTE: The lower bound of the first propensity class is zero.

*This propensity class is combined with the preceding class(es).

TABLE 8

ALTERNATIVE WEIGHT MULTIPLIERS FOR INFLATING 1982 ADVANCE 1040 RETURNS
TO PROVIDE ADVANCE ESTIMATES OF COMPLETE YEAR DATA: BY SAMPLE CODE AND PROPENSITY CLASS

Sample Code	Uniform Weight Multiplier	Weight Multipliers by Propensity Class											
		Method One						Method Two					
		1	2	3	4	5	6	1	2	3	4	5	6
28	1.303	1.214	1.321	1.293	1.441	1.257	1.571	0.998	1.097	1.184	1.337	1.683	3.207
38	1.380	1.159	1.291	1.403	1.607	1.861	2.724	1.165	1.255	1.352	1.530	1.999	4.384
40	1.010	1.021	1.008*	1.008*	1.008*	1.008*	1.018	1.009	1.010*	1.010*	1.010*	1.010*	1.018
41	1.013	1.011	1.012	1.010*	1.010*	1.010*	1.027	1.008	1.010	1.012*	1.012*	1.012*	1.024
42	1.029	1.017	1.014	1.024*	1.024*	1.029	1.085	1.020	1.022	1.025*	1.025*	1.028	1.066
43	1.080	1.047	1.052	1.071*	1.071*	1.105	1.199	1.050	1.057	1.063*	1.063*	1.076	1.283
44	1.105	1.030	1.041	1.083	1.128	1.168	1.288	1.050	1.062	1.074	1.091	1.137	1.357
45	1.174	1.045	1.072	1.115	1.169	1.279	1.531	1.068	1.098	1.121	1.151	1.226	1.584
46	1.270	1.059	1.090	1.184	1.254	1.473	1.920	1.116	1.146	1.168	1.223	1.380	1.998
47	1.376	1.098	1.565	1.257	1.314	1.508	2.202	1.127	1.158	1.194	1.307	1.614	2.457
48	1.495	1.081	1.426	1.509	1.256	2.156	2.143	1.089	1.139	1.206	1.463	1.967	3.561
50	1.015	1.012	1.016*	1.016*	1.016*	1.017	1.018	1.009	1.010*	1.010*	1.010*	1.017	1.046
51	1.015	1.000	1.013	1.012*	1.012*	1.015	1.038	1.005	1.007	1.009*	1.009*	1.015	1.070
52	1.038	1.020	1.017	1.039	1.041	1.045	1.072	1.026	1.029	1.031	1.035	1.043	1.078
53	1.101	1.073	1.063	1.065	1.245	1.080	1.149	1.052	1.064	1.076	1.097	1.138	1.253
54	1.150	1.110	1.158	1.084	1.108	1.178	1.415	1.074	1.088	1.110	1.147	1.189	1.466
55	1.218	1.066	1.131	1.255	1.240	1.306	1.427	1.092	1.128	1.168	1.231	1.316	1.544
56	1.269	1.176	1.164	1.250	1.326	1.292	1.594	1.103	1.145	1.202	1.285	1.392	1.733
57	1.311	1.148	1.209	1.378	1.436	1.333	1.500	1.007	1.048	1.140	1.281	1.601	2.339
58	1.452	1.154	1.200	1.154	1.421	1.765	2.167	0.970	0.991	1.046	1.159	1.457	4.677

Continued

Table 8 (continued)

Sample Code	Uniform Weight Multiplier	Weight Multipliers by Propensity Class											
		Method One						Method Two					
		1	2	3	4	5	6	1	2	3	4	5	6
60	1.050	1.021	1.035	1.058*	1.058*	1.058*	1.091	1.032	1.038	1.049*	1.049*	1.049*	1.110
61	1.045	1.026	1.016	1.036	1.062	1.056	1.093	1.023	1.027	1.031	1.036	1.047	1.164
62	1.078	1.039	1.048	1.049	1.082	1.122	1.174	1.046	1.057	1.063	1.077	1.101	1.180
63	1.146	1.070	1.094	1.101	1.177	1.215	1.309	1.071	1.085	1.102	1.132	1.197	1.463
64	1.229	1.117	1.134	1.174	1.275	1.377	1.483	1.131	1.152	1.176	1.223	1.309	1.658
65	1.356	1.148	1.251	1.280	1.392	1.538	1.826	1.196	1.255	1.290	1.347	1.502	1.853
66	1.447	1.208	1.251	1.337	1.474	1.833	1.841	1.197	1.268	1.320	1.408	1.638	2.495
67	1.630	1.289	1.496	1.542	1.470	2.010	2.529	1.354	1.435	1.495	1.650	1.873	2.410
68	1.779	1.258	1.211	1.594	1.789	2.500	3.333	1.337	1.430	1.509	1.719	2.076	4.270

SOURCE: Prepared by Mathematica Policy Research, Inc., from the 1982 SOI 1040 Complete Report microdata file.

*Two or more adjacent classes were combined to form a single weight class.

TABLE 9

ABSOLUTE AND PERCENTAGE DEVIATIONS FROM COMPLETE REPORT FOR THREE ALTERNATIVE ADVANCE ESTIMATES OF SELECTED INCOME AND TAX ITEMS USED AS PREDICTORS OF LATE POSTING: TAX YEAR 1982
(Money amounts are in millions of dollars.)

Income or Tax Item	Complete Report Estimate	Absolute Deviation from Complete Report			Percentage Deviation		
		Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2	Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2
Domestic and Foreign Dividends Received							
Total dividends							
Number of returns	17,185,359	-27,838	-19,159	-40,145	-0.16	-0.11	-0.23
Amount	54,031.3	737.2	-75.5	27.7	1.36	-0.14	0.05
Dividends in AGI							
Number of returns	13,171,051	-24,282	-21,070	-39,393	-0.18	-0.16	-.030
Amount	52,129.0	739.0	-72.8	32.3	1.42	-0.14	0.06
Interest Received[a]							
Number of returns	52,841,028	80,161	121,227	70,195	0.15	0.23	0.13
Amount	157,008.0	-192.1	87.1	-66.8	-0.12	0.06	-0.04
Total Itemized Deductions							
Number of returns	33,431,485	14,956	13,230	30,931	0.04	0.04	0.09
Amount	284,462.6	-1,378.0	-366.5	-220.8	-0.48	-0.13	-0.08
Total Tax Preferences[b]							
Number of returns	225,034	-16,717	-5,166	-2,943	-7.43	-2.30	-1.31
Amount	1,516.3	-204.7	-16.3	51.5	-13.50	-1.08	3.40
Alternative Minimum Tax[b]							
Number of returns	132,022	-11,331	-3,528	-969	-8.59	-2.67	-0.73
Amount	1,065.9	-145.9	11.1	74.8	-13.69	1.05	7.02
Business Income (Schedule C)[c]							
Net Profit							
Number of returns	6,760,145	-33,322	-20,594	-32,720	-0.49	-0.30	-0.48
Amount	68,644.1	-1,806.2	-1,784.5	-1,780.3	-2.63	-2.60	-2.59
Net Loss							
Number of returns	3,330,902	21,661	11,644	23,461	0.65	0.35	0.70
Amount	-18,071.9	1,039.6	227.8	-178.6	-5.75	-1.26	0.99
Net Profit Less Loss							
Number of returns	10,091,047	-11,660	-8,950	-9,260	-0.12	-0.09	-0.09
Amount	50,572.3	-766.5	-1,556.8	-1,958.9	-1.52	-3.08	-3.87

Table 9 (continued)

Income or Tax Item	Complete Report Estimate	Absolute Deviation from Complete Report			Percentage Deviation		
		Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2	Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2
Sales of Capital Assets (Schedule D)[d]							
Net Gain							
Number of returns	5,854,430	-46,855	-25,407	-21,757	-0.80	-0.43	-0.37
Amount	38,087.4	-788.1	560.2	1,083.7	-2.07	1.47	2.85
Net Loss							
Number of returns	2,513,482	-32,482	-35,766	-37,365	-1.29	-1.42	-1.49
Amount	-4,113.3	79.6	86.0	89.6	-1.93	-2.09	-2.18
Net Gain Less Loss							
Number of returns	8,367,912	-79,337	-61,173	-59,122	-0.95	-0.73	-0.71
Amount	33,974.1	-708.5	646.2	1,173.3	-2.09	1.40	3.45
Partnership Income[e]							
Net Profit							
Number of returns	1,817,944	-47,033	-37,457	-38,271	-2.59	-2.06	-2.11
Amount	27,364.3	-1,208.7	-1,181.7	-1,347.7	-4.42	-4.32	-4.93
Net Loss							
Number of returns	2,126,409	-55,866	-22,456	-27,371	-2.63	-1.06	-1.10
Amount	-28,263.7	3,811.0	1,750.7	1,127.4	-13.48	-6.19	-3.99
Net Profit Less Loss							
Number of returns	3,944,353	-102,899	-59,913	-61,642	-2.61	-1.52	-1.56
Amount	-899.4	2,602.3	569.0	-220.3	-289.33	-63.26	24.49
Employee Business Expenses[a]							
Number of returns	7,057,975	-6,342	5,842	7,615	-0.09	0.08	0.11
Amount	16,276.2	-74.6	-62.1	-57.1	-0.46	-0.38	-0.35

SOURCE: Prepared by Mathematica Policy Research from the 1982 SOI Complete Report microdata file.

[a]Only the flag was used as a predictor. The amount was tested but rejected.
[b]The recipiency flag appears in propensity equations for all strata. The amount appears in selected strata.
[c]Only the net loss amount was used as a predictor. The net income amount was tested but rejected.
[d]The total gain or loss amount was included in the propensity equations as a single predictor. Separate gain and loss variables were tested but rejected.
[e]A loss flag appears in the propensity equations for all strata. A net profit flag and net loss amount appear in selected strata. The net profit amount was tested but rejected.

TABLE 10

ABSOLUTE AND PERCENTAGE DEVIATIONS FROM COMPLETE REPORT FOR THREE ALTERNATIVE ADVANCE ESTIMATES OF SELECTED INCOME AND TAX ITEMS NOT USED AS PREDICTORS OF LATE POSTING: TAX YEAR 1982
(Money amounts are in millions of dollars.)

Income or Tax Item	Complete Report Estimate	Absolute Deviation from Complete Report			Percentage Deviation		
		Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2	Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2
Adjusted Gross Income (AGI) Less Deficit	1,852,072.4	9,748.7	4,594.0	2,903.0	0.53	0.25	0.16
Salaries and Wages							
Number of returns	83,105,532	170,836	117,162	148,571	0.21	0.14	0.18
Amount	1,564,948.3	4,295.7	3,559.5	3,561.4	0.27	0.23	0.23
Unemployment Compensation							
Total							
Number of returns	10,105,079	57,860	51,625	56,460	0.57	0.51	0.56
Amount	19,818.4	108.9	98.3	110.6	0.55	0.50	0.56
Included in AGI							
Number of returns	5,347,634	32,900	30,784	30,699	0.62	0.58	0.57
Amount	7,089.1	12.7	12.0	11.5	0.18	0.17	0.16
Farm Income (Schedule F)[a]							
Net Profit							
Number of returns	932,986	4,191	3,287	1,200	0.45	0.35	0.13
Amount	7,994.2	10.9	-2.3	-16.1	0.14	-0.03	-0.20
Net Loss							
Number of returns	1,755,632	-5,113	-2,580	472	-0.29	-0.15	0.03
Amount	-17,822.8	334.1	81.3	-42.1	-1.87	-0.46	0.24
Net Profit Less Loss							
Number of returns	2,688,618	-922	708	1,672	-0.03	0.03	0.06
Amount	-9,828.6	344.9	79.1	-58.1	-3.51	-0.80	0.59
Pensions and Annuities in AGI							
Number of returns	8,824,875	38,287	54,554	29,021	0.43	0.62	0.33
Amount	60,122.9	182.4	207.8	85.2	0.30	0.35	0.14
Alimony Received							
Number of returns	316,617	1,251	1,724	1,385	0.40	0.54	0.44
Amount	1,946.1	-122.2	-118.1	-120.2	-6.28	-6.07	-6.18
Estate or Trust Income[b]							
Net Income							
Number of returns	797,391	-25,870	-23,286	-23,160	-3.24	-2.92	-2.90
Amount	6,088.8	-319.0	-340.9	-351.8	-5.24	-5.60	-5.78
Net Loss							
Number of returns	61,810	-5,397	-4,764	-4,548	-8.73	-7.71	-7.36
Amount	-342.7	37.5	22.7	18.0	-10.95	-6.61	-5.26
Net Income Less Loss							
Number of returns	859,201	-31,267	-28,050	-27,708	-3.64	-3.26	-3.22
Amount	5,746.2	-281.5	-318.3	-333.8	-4.90	-5.54	-5.81
Small Business Corporation[b]							
Net Profit							
Number of returns	416,549	-13,482	-12,115	-12,191	-3.24	-2.91	-2.93
Amount	5,580.3	-186.4	-324.8	-296.3	-3.34	-5.82	-5.31
Net Loss							
Number of returns	421,219	-31,291	-25,037	-23,375	-7.43	-5.94	-5.55
Amount	-6,426.4	1,227.2	765.7	574.8	-19.10	-11.91	-8.94
Net Profit Less Loss							
Number of returns	837,768	-44,773	-37,152	-35,567	-5.34	-4.43	-4.25
Amount	-846.1	1,040.8	440.9	278.5	-123.01	-52.11	-32.92
Other Income							
Net Income							
Number of returns	3,703,370	-28,463	-24,862	-25,754	-0.77	-0.67	-0.70
Amount	7,641.0	-424.5	-380.2	-373.0	-5.56	-4.98	-4.88
Net Loss							
Number of returns	553,778	-38,502	-27,939	-24,953	-6.95	-5.05	-4.51
Amount	-17,942.3	2,403.8	1,303.9	941.2	-13.40	-7.27	-5.25
Net Income Less Loss							
Number of returns	4,257,148	-66,965	-52,801	-50,707	-1.57	-1.24	-1.19
Amount	-10,301.3	1,979.3	920.7	568.2	-19.21	-8.94	-5.52

Table 10 (continued)

Income or Tax Item	Complete Report Estimate	Absolute Deviation from Complete Report			Percentage Deviation		
		Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2	Simulated Advance Data Estimate	Propensity Score Method 1	Propensity Score Method 2
Payments to an IRA							
Number of returns	12,101,016	101,150	84,888	77,179	0.84	0.71	0.64
Amount	28,273.8	238.9	182.7	165.7	0.84	0.65	0.59
Deduction for Two-Earner Couple[c]							
Number of returns	21,689,907	136,655	126,726	124,770	0.63	0.58	0.58
Amount	9,047.7	61.3	54.9	55.2	0.68	0.61	0.61
Taxable Income							
Number of returns	89,716,570	67,444	39,341	17,128	0.08	0.04	0.02
Amount	1,473,318.0	6,846.5	2,976.0	2,313.8	0.46	0.20	0.16
Income Tax							
Tax Before Credits							
Number of returns	79,348,582	50,762	27,285	1,417	0.06	0.03	0.00
Amount	283,926.4	2,549.3	946.7	789.1	0.90	0.33	0.28
Tax Credits							
Number of returns	18,882,528	-18,745	-20,400	-26,094	-0.10	-0.11	-0.14
Amount	7,854.2	-267.5	-208.3	-215.6	-3.41	-2.65	-2.75
Tax After Credits							
Number of returns	76,959,571	72,677	46,981	20,573	0.09	0.06	0.03
Amount	276,072.2	2,816.9	1,155.0	1,004.7	1.02	0.42	0.36
Total Income Tax							
Number of returns	77,033,977	63,653	42,539	17,645	0.08	0.06	0.02
Amount	277,588.5	2,612.2	1,138.7	1,056.2	0.94	0.41	0.38
Total Tax Liabilities							
Number of returns	78,680,509	3,076	-5,507	-28,404	0.00	-0.01	-0.04
Amount	284,699.0	2,353.7	897.7	811.7	0.83	0.32	0.29
Selected Itemized Deductions							
Medical and Dental Expenses							
Number of returns	21,980,291	48,403	46,155	58,787	0.22	0.21	0.27
Amount	21,704.6	-275.5	-217.5	-205.6	-1.27	-1.00	-0.95
Taxes Paid							
Number of returns	33,079,548	28,700	26,653	44,074	0.09	0.08	0.13
Amount	88,032.1	498.1	452.3	452.5	0.57	0.51	0.51
Interest Paid							
Number of returns	30,242,954	29,373	35,097	50,923	0.10	0.12	0.17
Amount	121,822.6	-1,795.3	-954.2	-845.7	-1.47	-0.78	0.69
Contributions							
Number of returns	30,509,886	66,578	63,254	78,795	0.22	0.21	0.26
Amount	33,467.9	282.7	338.2	325.6	0.84	1.01	0.97
Exemptions							
Total Number	232,188,753	72,683	133,721	114,854	0.03	0.06	0.05
Age or Blindness	14,211,172	-12,845	19,949	-28,421	-0.09	0.14	-0.20
Other	217,977,581	85,528	113,772	143,275	0.04	0.05	0.07

SOURCE: Prepared by Mathematica Policy Research from 1982 SOI Complete Report microdata file.

[a]Flags and amounts were tested but rejected as predictors of propensity to be posted late.
[b]This income source is included on Supplemental Income Schedule E, which encompasses rent, royalties, partnerships, small business corporations, estate and trust income, and selected other sources. A Schedule E income flag and loss amount appear in the propensity equations for all strata. The net income amount was tested but rejected.
[c]This variable was not available prior to 1982.

TABLE 11

PERCENTAGE DEVIATIONS FROM COMPLETE REPORT FOR ALTERNATIVE ADVANCE ESTIMATES OF SIX INCOME AND TAX AMOUNTS, BY SAMPLE CODE
(Amounts are in millions of dollars.)

Sample Code	Sample Size	Total Dividends Received				Total Itemized Deductions				Total Tax Preferences			
		Complete Report Estimate	Percentage Deviation			Complete Report Estimate	Percentage Deviation			Complete Report Estimate	Percentage Deviation		
			AD	PSM-1	PSM-2		AD	PSM-1	PSM-2		AD	PSM-1	PSM-2
Total	77,574	54,031.3	1.36	-0.14	0.05	284,462.6	-0.48	-0.13	-0.08	1,516.3	-13.50	-1.08	3.40
28	201	11.9	-4.1	0.1	29.4	49.3	-21.2	-20.1	-13.0	0.0	0.0	0.0	0.0
38	9,331	787.0	-12.7	-9.0	-4.9	1,170.9	-11.3	-4.7	-0.5	234.1	-20.4	1.7	17.2
Non-business, Non-farm													
40	10,289	6,766.8	-0.5	0.1	-0.4	27,200.0	-0.7	-0.6	-0.5	0.0	0.0	0.0	0.0
41	7,708	11.972.5	-1.6	-1.4	-1.4	121 420.7	0.0	0.1	0.1	9.7	1.3	2.7	2.4
42	3,660	9,325.3	0.0	-0.4	-0.3	43,373.0	-0.4	-0.1	-0.2	46.4	-1.5	3.7	1.9
43	1,961	4,774.5	2.5	1.0	1.1	12,887.3	-0.5	0.2	0.1	133.6	-7.2	-4.6	-2.5
44	2,288	4,119.7	4.0	1.0	1.9	6,379.4	-1.2	0.6	0.4	117.0	-8.6	2.5	5.1
45	2,318	1,750.7	5.8	-1.4	-0.3	1,891.1	-0.9	0.0	-0.4	76.9	-15.2	0.1	0.3
46	1,942	1,028.7	11.1	0.3	3.1	988.7	-1.6	-0.6	-1.4	55.7	-15.9	3.9	1.8
47	1.286	832.1	12.8	5.0	3.3	802.8	5.2	4.8	3.5	51.9	0.1	29.2	40.2
48	287	522.1	10.9	-1.4	-4.0	540.5	-2.1	-3.5	-5.8	30.6	-27.4	-1.7	35.4
Non-business, Farm													
50	3,453	112.7	-1.5	-1.5	-1.7	805.1	0.4	0.5	0.9	0.1	1.5	1.7	2.4
51	3,443	281.0	-3.7	-3.8	-3.6	2,873.7	0.1	0.0	-0.1	6.3	-6.8	-6.6	-6.4
52	859	332.7	0.1	-0.7	-0.5	1,401.4	-1.1	-1.5	-1.6	23.1	-9.9	-8.5	-8.5
53	370	251.9	5.3	4.2	4.0	839.7	-8.3	-7.0	-8.8	23.8	-9.3	-8.8	-8.6
54	356	347.7	-1.4	-2.6	-4.5	652.2	-8.1	-7.7	-7.7	37.4	-22.6	-14.0	-11.6
55	448	207.9	4.1	-3.8	-3.2	242.4	-2.4	-4.0	-4.8	21.7	-5.0	1.3	2.3
56	346	188.8	1.0	-4.7	-8.9	140.4	-10.0	-11.3	-12.6	11.4	-2.3	1.5	3.8
57	251	114.8	8.4	4.2	-4.2	103.8	0.3	0.2	-7.4	11.1	-3.1	2.2	14.5
58	73	74.5	8.4	3.8	12.9	76.5	-20.1	-26.7	-36.2	6.3	-65.6	-57.7	-27.9
Business													
60	6,062	696.4	0.9	0.2	0.6	7,959.5	-1.4	-0.9	-0.8	0.3	5.0	8.7	10.4
61	6,715	2,123.1	-1.7	-1.2	-0.3	26,808.9	-0.6	0.1	0.7	11.0	-1.7	1.4	5.9
62	3,145	2,243.3	2.0	0.1	0.5	14,254.5	-0.8	-0.2	-0.4	85.8	-7.8	-1.2	-0.9
63	2,281	1,653.9	2.9	0.0	-0.2	5,890.0	-1.9	-0.3	0.1	115.0	-14.2	-6.8	-2.4
64	3,180	1,572.8	7.1	8.6	8.6	3,187.7	-0.8	1.2	0.7	168.7	-13.7	-3.9	-2.8
65	3,367	796.7	8.8	-0.4	1.4	1,155.3	-1.1	-1.5	-1.7	101.2	-12.6	-0.2	-0.9
66	1,132	510.3	8.6	-0.7	-1.9	640.7	-2.5	-0.9	-3.2	67.9	-10.3	4.2	5.7
67	686	379.0	7.6	-1.7	-0.3	445.7	-0.7	-3.6	-3.4	41.5	-22.2	-5.9	-8.8
68	136	252.2	23.6	4.6	6.8	281.4	6.0	8.6	5.6	27 5	-48.6	-30.0	-32.4

TABLE 11 (continued)

Sample Code	Sample Size	AGI Less Deficits				Salaries and Wages				Estate or Trust Net Profit			
		Complete Report Estimate	Percentage Deviation			Complete Report Estimate	Percentage Deviation			Complete Report Estimate	Percentage Deviation		
			AD	PSM-1	PSM-2		AD	PSM-1	PSM-2		AD	PSM-1	PSM-2
Total	77,574	1,852,072.4	0.53	0.25	0.16	1,564,948.3	0.27	0.23	0.23	6,088.8	-5.24	-5.60	-5.78
28	201	95.2	-11.4	-11.6	-11.3	47.6	-4.6	-4.2	-6.6	2.2	6.5	6.5	28.7
38	9,331	3,167.3	25.6	12.9	9.9	1,015.0	-14.7	-8.3	-5.9	158.1	-21.3	-18.2	-20.4
Non-business, Non-farm													
40	10,289	454,540.7	0.0	0.1	0.1	381,244.8	0.2	0.1	0.2	644.9	-12.8	-12.1	-12.1
41	7,708	779,232.4	0.1	0.1	0.1	712,926.7	0.1	0.1	0.1	1,318.4	-2.2	-1.9	-1.8
42	3,660	210,402.5	0.2	0.1	0.1	184,773.5	0.0	0.0	0.0	1,225.2	-3.8	-2.6	-3.1
43	1,961	58,434.8	0.9	0.4	0.1	45.971.8	0.2	0.3	0.2	594.4	0.6	0.7	0.5
44	2,288	29,797.3	2.9	1.6	1.4	19,993.4	0.3	0.5	0.2	378.0	-1.6	-2.4	-2.5
45	2,318	8,580.5	7.0	2.9	2.8	4,367.9	1.6	7.0	0.7	187.8	-1.4	-5.2	-5.0
46	1,942	4,200.2	11.8	2.2	2.3	1,547.8	3.0	1.3	0.9	95.0	-3.0	-10.0	-9.8
47	1,286	3,065.9	20.1	8.8	4.5	710.2	-0.5	-5.6	-7.7	54.1	2.3	-8.4	-10.0
48	287	1,595.3	37.2	20.1	4.7	273.3	-3.6	-10.0	-16.6	31.1	-71.2	-71.9	-74.3
Non-business, Farm													
50	3,453	7,199.9	0.0	0.0	-0.6	4,550.9	-0.1	-0.1	-0.3	9.3	1.5	1.6	1.5
51	3,443	20,907.5	0.5	0.2	-0.3	17,890.9	0.3	0.2	0.0	32.8	0.4	0.3	0.3
52	859	7,468.8	0.6	-0.4	-0.6	6,276.0	0.7	0.1	0.2	35.7	1.7	2.0	1.6
53	370	2,534.7	-4.3	-5.8	-9.4	2,428.3	-1.3	-1.5	-3.0	29.7	-0.4	-3.8	-3.6
54	356	1,518.0	4.8	2.4	-1.6	1,635.6	2.4	0.6	0.0	66.8	-15.3	-13.2	-12.7
55	448	658.8	12.3	0.6	-2.2	511.9	-0.6	-3.0	-3.7	26.1	-1.0	-7.1	-6.6
56	346	345.0	12.0	-0.3	-10.1	195.3	-0.3	-2.6	-3.9	21.9	-31.0	-35.2	-37.5
57	251	247.6	21.8	8.1	-27.6	93.1	-1.3	-2.6	-9.8	14.5	-10.5	-17.3	-27.1
58	73	158.8	16.8	-3.6	-27.1	87.0	9.5	2.0	-2.8	7.5	-82.1	-85.5	-87.6
Business													
60	6.062	38,732.6	0.7	0.2	0.1	22,317.8	2.0	1.5	1.5	69.9	5.0	6.9	7.7
61	6,715	119,184.1	0.5	0.4	0.2	91,840.2	1.1	1.0	0.9	198.6	-7.1	-6.6	-6.6
62	3,145	57,931.3	0.7	0.2	0.2	40,194.9	2.0	1.9	1.8	207.3	-6.1	-6.8	-6.6
63	2,281	22,222.2	2.4	0.7	-0.1	12,963.1	1.3	1.5	1.3	164.6	-4.6	-6.0	-6.7
64	3,180	11,539.1	7.1	3.6	3.0	7,513.1	2.4	1.2	0.9	206.7	-6.5	-7.5	-7.8
65	3,367	3,945.7	12.7	5.2	5.8	2,146.3	2.7	1.3	1.2	133.3	-5.4	-11.6	-10.9
66	1,132	2,164.1	16.8	6.6	1.5	819.4	0.1	-1.2	-2.3	91.7	-20.9	-27.3	-27.4
67	686	1,417.6	21.6	3.8	6.6	433.0	-0.4	-5.4	-4.8	45.7	-22.8	-29.7	-29.3
68	136	784.5	30.2	4.6	3.3	179.5	8.3	9.5	16.3	17.4	52.1	34.2	34.9

SOURCE: Prepared by Mathematica Policy Research from the 1982 SOI Complete Report microdata file.

NOTE: Sample codes are defined in Table 1. Reported sample size is the number of returns in the advance sample.