The IRS Research Bulletin

Proceedings of the 2022 IRS / TPC Research Conference

Research, Applied Analytics & Statistics

Publication 1500 (Rev. 5-2023) Catalog Number 11546J Department of the Treasury **Internal Revenue Service** www.irs.gov

Papers given at the

12th Annual Joint Research Conference on Tax Administration

Cosponsored by the IRS and the Urban-Brookings Tax Policy Center

Held Virtually
June 16, 2022

Compiled and edited by Alan Plumley*
Research, Applied Analytics, and Statistics, Internal Revenue Service

*Prepared under the direction of Barry W. Johnson, Acting IRS Chief Research and Analytics Officer

Foreword

This edition of the IRS Research Bulletin (Publication 1500) features selected papers from the IRS-Tax Policy Center (TPC) Research Conference held virtually on June 16, 2022. Conference presenters and attendees included researchers from many areas of the IRS, officials from other government agencies, and academic and private sector experts on tax policy, tax administration, and tax compliance. Many people participated in this, our third fully virtual conference. Videos of the presentations are archived on the Tax Policy Center website to enable additional participation. Attendees participated in the discussions by submitting questions electronically as the sessions proceeded.

The conference began with welcoming remarks by Eric Toder, Co-Director of the Tax Policy Center, and by Melanie Krause, the IRS Chief Research and Analytics Officer. The remainder of the conference included sessions on using data from both operational and random audits both to measure the extent of noncompliance and to improve operational audit selection methods, balancing taxpayer burden vs. opportunities for noncompliance, ways to improve audit outcomes, and factors that influence compliance behavior. The keynote speaker was John M. Abowd, the Associate Director and Chief Scientist in the Research and Methodology Directorate of the U.S. Census Bureau, who offered his insights on data use policies.

We trust that this volume will enable IRS executives, managers, employees, stakeholders, and tax administrators elsewhere to stay abreast of the latest trends and research findings affecting tax administration. We anticipate that the research featured here will stimulate improved tax administration, additional helpful research, and even greater cooperation among tax administration researchers worldwide.

Acknowledgments

This IRS-TPC Research Conference was the result of preparation over a number of months by many people. The conference program was assembled by a committee representing research organizations throughout the IRS. Members of the program committee included: Alan Plumley, Brett Collins, and Aaron Katch (Research, Applied Analytics, and Statistics); Terry Ashley (Taxpayer Advocate); Ted Moorman (Criminal Investigation Division); and Rob McClelland (Tax Policy Center). In addition, Hailey Roemer and Ann Cleven from the Tax Policy Center oversaw numerous details to ensure that the conference ran smoothly.

This volume was prepared by Lisa Smith (layout and graphics) and Beth Kilss, Anne McDonough, and Sarah Swisher (editors), all of the IRS Statistics of Income Division. The authors of the papers are responsible for their content, and views expressed in these papers do not necessarily represent the views of the Department of the Treasury or the Internal Revenue Service.

We appreciate the contributions of everyone who helped make this conference a success.

Melanie Krause
IRS Chief Data and Analytics Officer

12th Annual IRS-TPC Joint Research Conference on Tax Administration

Contents

1

Balancing Audits: Enforcement vs. Measuring Noncompliance

Roy ◆ Collins ◆ Turk ◆ Payne

Rizzo ◆ Riddles ◆ Zhu ◆ Valliant ◆ Henry

Henderson ◆ Chugg ◆ Altenburger ◆ Ho ◆ Anderson

Guyton ◆ Turk ◆ Goldin

Improving Risk Models by Supplementing Random National Research Program Audits with Non-Random Operational Audits Using Statistical Controls for Bias

Ishani Roy, Brett Collins, Alex Turk, and Mark Payne (IRS, Research, Applied Analytics, and Statistics)[1]

I. Introduction

Measuring and predicting tax reporting compliance typically requires audit data to label the reporting behavior of the taxpayer. The Internal Revenue Service (IRS) National Research Program (NRP) provides reporting compliance data from random audits that thoroughly and objectively classify the issues on the tax return. The NRP audit data are representative of the underlying filing population, but detailed audits by revenue agents are costly and their random nature can result in a relatively large share of audits that end in small or no adjustments to total tax. This paper develops a framework for including labels from operational (OP) audits to supplement the results from random NRP audits to produce unbiased estimates of reporting compliance risk. The approach aligns with similar efforts to incorporate randomized and non-randomized data, including by Colnet *et al.* (2022) and Wiśniowski *et al.* (2020). Since OP audits have no additional opportunity cost, including the data from these audits promises to extend the capabilities of the risk model while keeping samples small and focused. We also investigate replacing part of the NRP data with appropriately chosen OP data to produce unbiased tax reporting predictions, which may be a useful way to extend the capabilities of the risk model in segments with low NRP coverage, or where NRP cases look like OP cases. This paper focuses on models for cases that involve claims of the Earned Income Tax Credit (EITC), which includes segments where NRP samples are already low.

II. Background

Beginning with the Audit Control Program in 1948, the IRS has used random audits to study the compliance characteristics of taxpayers and predict compliance risk. From 1962 to 1988, the IRS periodically conducted the Taxpayer Compliance Measurement Program, a program of in-depth audits involving large samples of randomly selected taxpayers. In 2001, the National Research Program began with 45,000 randomly selected audits and then in 2006 moved to a rolling sample of about 14,000 cases per year to reduce the exam workload and costs. Assuming comparability and accuracy across audits, these random audit programs provide a statistically valid representation of the compliance characteristics of the taxpayer population. They have been used to estimate voluntary compliance levels, develop audit selection strategies and to estimate the percentage and amounts of improper payments of nonrefundable credits such as the EITC, the Additional Child Tax Credit and the American Opportunity Tax Credit (pursuant to the Improper Payments Elimination and Recovery Act).

Recent reductions in the IRS's resources for enforcement and the consequent increase in the relative costs of the NRP have led to cutbacks in its scope, starting with Tax Year (TY) 2016 and an effort to redesign the program so that a smaller sample of random audits can be used to fulfill the IRS's various compliance measurement and modeling needs. In addition, the program redesign hopes to make it more agile in identifying emerging compliance risks and increase the frequency of updates to tax gap and improper payment estimates.

[1] The views in this paper are those of the authors only and do not necessarily reflect the positions of the IRS.

An NRP Redesign working group recommended several options for increasing the efficiency and usefulness of the random audits, including better aligning the sample size and design with IRS compliance priorities, applying machine learning and other artificial intelligence methods to identify compliance risk more efficiently and better align with the audit plan, and pursuing operation and analytic changes allowing integrated analysis of NRP and OP audit results. This paper represents one of a few different statistical approaches being considered for using some OP audit results to supplement random audit cases to make compliance estimates and develop audit selection procedures.

III. Methodology

A. Study Design and Data

The original data consist of TY2010–TY2015 NRP audits and 2010–2015 operational field (Revenue Agent (RA) and Tax Compliance Officer (TCO)) audits. The OP audits conducted by RAs and TCOs are in-person audits conducted with essentially the same procedures and auditors as the NRP audits. We exclude the correspondence audits from the model training data since they are not in-person and have a narrower focus. In general, the training sample is from TY2010–TY2014 and TY2015 is used as the validation year and is not used in model estimation.

The proposed framework for incorporating OP audit data along with NRP audit data for the development of the risk models has two key parts. The first part models the probability of a return to be audited by OP to construct a correction for selection bias that arises because of the differences between the selection processes for NRP and OP audits. For this model, we supplement the audit data with a random selection of non-audited taxpayers, sampling roughly equivalent numbers of taxpayers to the number of audited OP cases for each Activity Code (AC).

The second part of the framework includes a two-stage risk model that incorporates the bias correction term developed in the first part along with several other explanatory variables to predict outcomes such as the final tax adjustment and the probability of no change in tax. We test the models with different proportions of NRP and OP cases, to better understand how impactful adding OP cases to the data is to the results.

B. Weighting

The NRP data are stratified random samples and thus have sampling weights so that representative sample statistics can be calculated. Weights for model estimation for the NRP cases are calculated by dividing the base sampling weight by the average base weight for the AC. Thus, the weights for the NRP cases in an AC sum to the NRP sample size in that AC. The weights on the operational audit cases are set to 1 so that the OP weights also add up to the size of the OP sample in that AC. These weights are used in all model parameter estimation and calculations of averages for the validation year.

C. Audit Probability Model and Selection Bias Correction

Models for audit probability are estimated for each AC on TY2010–TY2014, with TY2015 reserved as the test sample. To generate controls, OP RA/TCO audit cases are grouped with a similar number of non-audited cases in the same AC randomly selected from the filing population. We use the popular framework of Heckman's selection bias correction to correct for the selection bias in OP RA/TCO cases compared to the NRP cases. To this end, we use a probit regression model for modeling probability p of field audit given the values of different explanatory variables (columns of X):

$$p=P(Audit \mid X)=\Phi(X\beta)$$

where Φ is the cumulative distribution function of a standard Gaussian variable. The variables in the model include a measure of distance between the IRS office that would conduct an audit and the taxpayer's zip code. This variable is expected to be uncorrelated with other return characteristics and acts as an instrument to provide information beyond the return characteristics that can help estimate the bias in selection of the group of

return characteristics available in the OP RA/TCO data. Other key variables include Discriminant Function (DIF) score and its square term, return characteristics such as wages, taxable income, passthrough income, and foreign income, as well as measures of consistency over time, including dummies for whether a taxpayer claimed the same children, or had the same 1099s as in the previous year. The squared DIF score is included to capture any degree of non-linearity with respect to the DIF score. We also include flowthrough income variables as dummies in the model.

The audit probability models are used to calculate the inverse Mill's ratio for the operational cases, a λ term that corrects for the selection bias when operational cases are combined with NRP cases. The λ term is set to 0 for the NRP cases. For the operational cases the λ term is related to the predicted probability of field audit "p" in the following way:

$$\lambda = \phi(z_p) \backslash p$$

where ϕ is the probability density function of a standard Gaussian distribution and z_p is the pth quantile of the standard Gaussian distribution.

D. Compliance Models

The compliance models predict the tax change (TC) following an audit, including cases where the return is accepted as filed (no change) and where assessments increased (positive change) or decreased (negative change). We build a model based on the hypothesis that the relationship between the expected amount of tax change following an audit and the return characteristics depend on the actual amount of tax change. To capture this relationship between the expected tax change, the return characteristics, and the true amount of tax change we propose using a piecewise model where the risk model of expected tax change on return characteristics is estimated at several level sets (classes) of the true tax change. This gives a two-stage modelling framework where the first stage models the probability of a return with specified characteristics to be associated with a particular class (low, medium, high) of the actual tax change. Then a second stage models the relationship between the nominal values of tax change and return characteristics for each of the three classes of actual tax change. Finally, the class probabilities are combined with the class-specific expected change to produce an expected tax change value given a set of return characteristics. Segmenting the distribution between three classes in this way allows us to better capture the differences between them, which may be lost if they were combined into a single model.

For each AC, the compliance model builds a two-stage risk model based on the training data and predicts tax change and probability of no change for the test data. The model uses three separate classes, returns with negative tax change, returns with no change, and returns with a positive tax change. ACs 270 and 271, which we are focusing on in this paper, no change is defined as a tax change between -$200 and $200, negative change is defined as tax change less than -$200 and positive change as tax change more than $200. The first stage of the compliance models is a multinomial logit that predicts the likelihood of each case falling into these three classes

$$log\left(\frac{P(C_j)}{P(C_1)}\right) = X\gamma_j + \gamma_{dj}D + \tau_j\lambda, \quad j = 2,3$$

where C_1, C_2 and C_3 are the three classes with C_1 treated as the reference class for the multinomial logit regression. Here D is a dummy indicating an NRP audit that is included to correct for the differences in the measurement process between the NRP and the OP audits and λ is the Heckman selection bias correction term defined above (Heckman (1979)). The term λ is defined as 0 for NRP audits. The fitted multinomial model predicts the probability of falling in each class, p_1, p_2, p_3 respectively, for each of the test cases in the 2015 data.

The second stage models the expected tax change (Tj), given the actual change for that return belongs to the jth class, using class specific linear regressions:

$$T_j = X\alpha_j + \alpha_{dj} D + \theta_j \lambda, j=1,2,3$$

The fitted regressions produce class specific predictions for the tax change variable for each test case, t_1, t_2, t_3, that are calculated setting D equal to one and λ equal to 0 for both NRP and operational cases in the test year.

For the test set, the final predicted tax change (t) for each case is an average of the class specific Predicted Tax Changes (PTC) weighted by the predicted class probabilities:

$$t = p_1 t_1 + p_2 t_2 + p_3 t_3$$

The second stage uses a stepwise regression approach to select regressors for each class, selecting from a pool of return characteristics including those that were used in the initial audit probability model. They include taxpayer characteristics such as filing status, income from schedules such as Schedule C and Schedule D, measures of consistency such as receiving W-2s from the same employer as in the previous year, DIF score, and selected DIF features that have proven to be effective at predicting audit results.

The methodology compares the performance of the risk model developed using only NRP data with those developed using a combined sample of NRP and OP cases. When the training data include only NRP cases, the variables D and λ drop out. Models are trained on TY2010–TY2014 data and tested on the holdout TY2015 data that include both operational and NRP audits. We compare models trained on only NRP cases with those trained on the pooled NRP cases and operational audits.

IV. Results

Our results focus on ACs 270 and 271, which both involve returns that include refundable credits, specifically claims for the EITC. These ACs, and particularly AC 271, are ones where NRP audits are already sparse and may be more likely to benefit from supplementing additional OP cases.

TABLE 1. Population Counts for Selected Activity Codes

Activity Code	TY 2019 Population (Cycle 202108)	NRP Training N	OP Training N
270	23,494,000	12,060	112,872
271	2,035,000	943	52,293

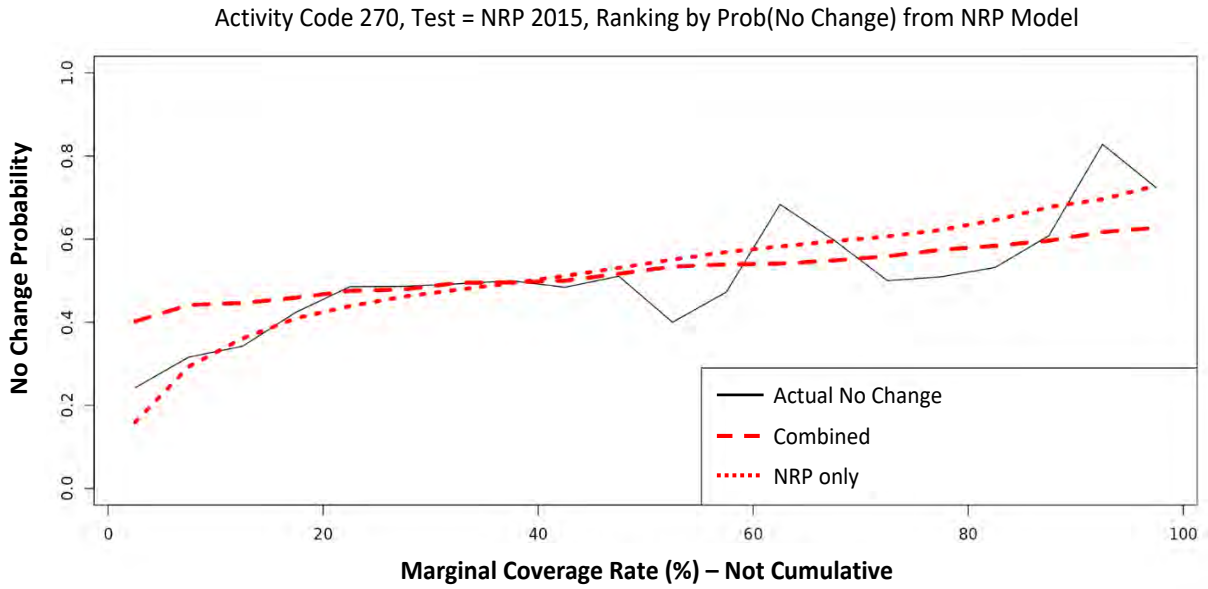
We present results using plots of predicted tax change and no change rates which compare our models built using only NRP data and using a combination of NRP and OP data against the actual outcomes. The plots use the NRP test data to group and rank into 5 percent bins, in some cases ranking by predicted outcomes and in other cases using the actual values. For each model we also test the sensitivity to the proportion of available NRP data that are included, such as 100 percent, 50 percent, or 25 percent.

E. Activity Code 270, Using Combined Data for Both Stages

AC 270 comprises cases with EITC claims, less than $200,000 in Total Positive Income (TPI), and either no attached Schedules C or F, or that include one or both schedules with total gross receipts under $25,000. These returns can be thought of as cases for individuals with low to moderate wage income who claim the EITC and have either no supplemental small business or farm income or a relatively small amount of nonwage income from these sources.

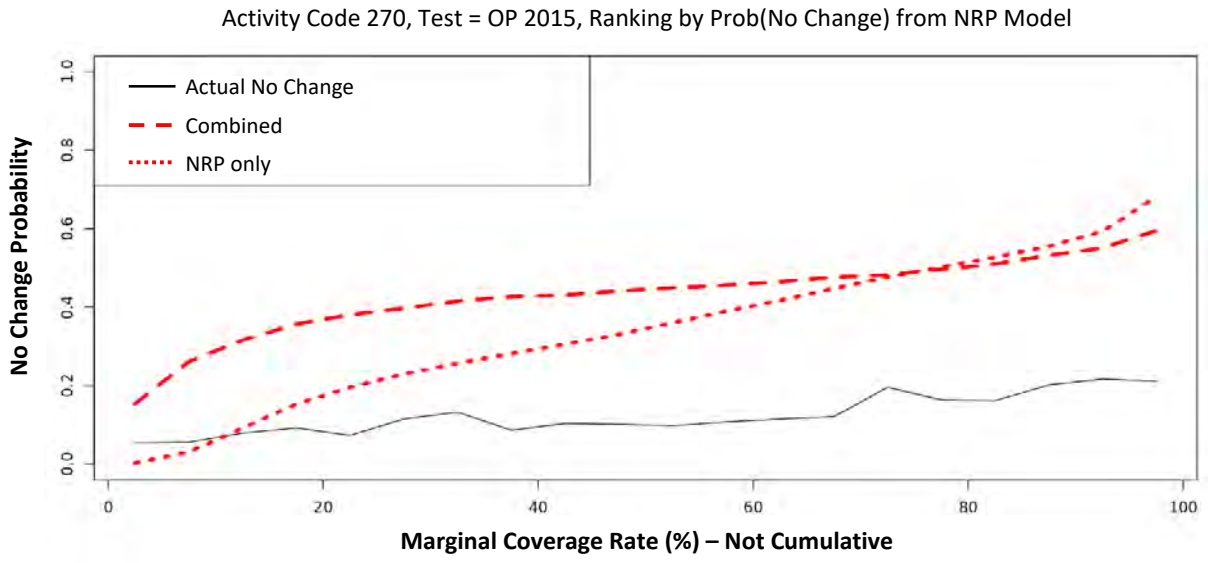
Figure 1 compares the predicted probability of no change for each model using the NRP data as a test. The data are ranked from the lowest to the highest probability of no change based on the predicted no change probability from the model with full set of NRP data as shown in the dotted red line. The dashed lines show the predictions for the combined NRP-OP model. The predicted no change rate from the NRP only model tracks the actual no change proportion very well throughout the range of the distribution whereas that from the combined model underestimates the actual no change proportion for high values of the proportion and overestimates it for low values of the proportion. However, the amount of under or over-estimation is reasonably small.

FIGURE 1. AC 270 Probability of No Change on NRP Data



When we compare the results of the no change predictions on OP data instead (Figure 2), we see that tested on the OP data the models appear to consistently over-predict the no change rates, which we would expect to be lower for OP cases since they are selected for their audit potential, rather than randomly as are the NRP cases. Part of this difference, however, may be because the data are still ranked by the NRP prediction (red dotted line), which will not be as well aligned with the OP test data.

FIGURE 2. AC 270 Probability of No Change on OP Data



To provide a comparison which is less dependent on how the data are ranked, we also plot the difference between the predicted and actual no change rates, with the combined models shown in the red solid line, and the NRP only model in the dashed line (Figures 3 and 4). Here the predicted no change rate is sorted according to each model's own predicted no change probabilities. While each bin may not include exactly the same cases for the different models, they represent the same part of the distribution, and no model is given an advantage

by being the one determining the ranking. Consistent with the results shown in Figure 1, Figure 3 suggests that both models perform relatively well when tested on the 2015 NRP data, with prediction errors staying close to zero across the distribution. When tested on the OP 2015 data however, Figure 4 confirms a tendency to over-predict no change rates, particularly for the combined models, which perform worse than the NRP-only models across the whole distribution.

FIGURE 3. AC 270 Error in Predictions for Probability of No Change on NRP Data

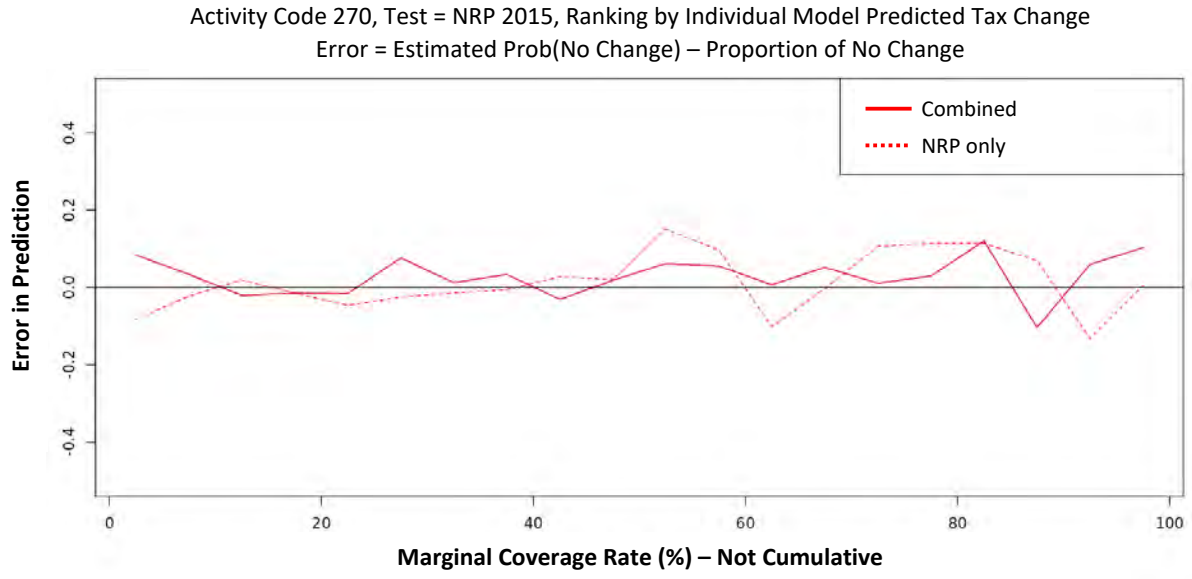
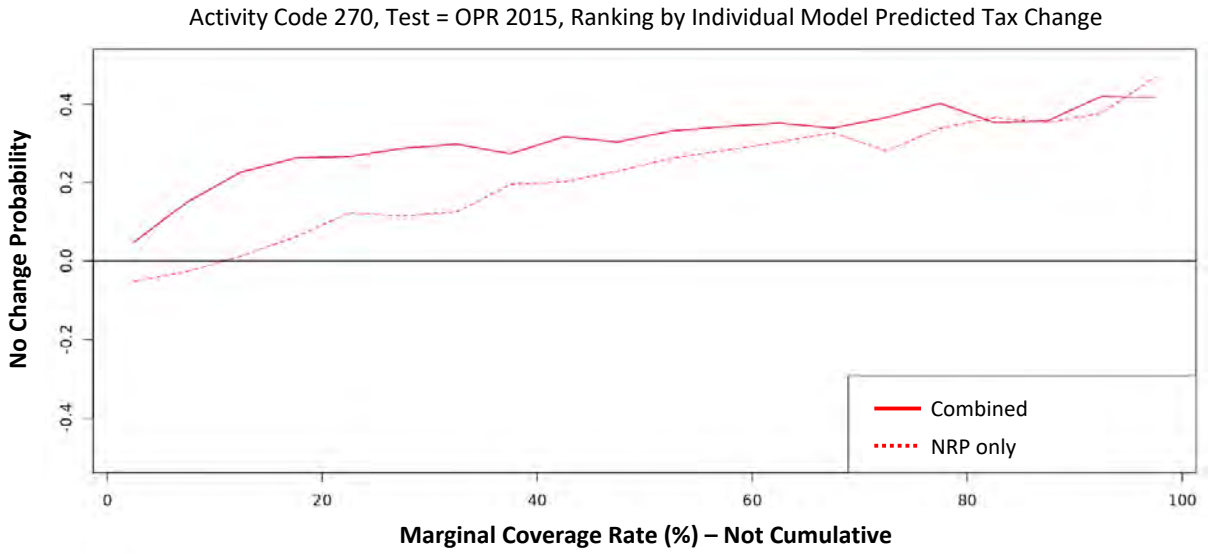


FIGURE 4. AC 270 Error in Predictions for Probability of No Change on OP Data



Turning to the tax change outcomes, Figure 5 shows the result of testing our model predictions against the actual tax change, using TY2015 NRP data as the test set. The ranking is based on the model trained with only NRP data. Both the NRP and NRP-OP combined models appear to result in very similar predictions, close enough that it is difficult to distinguish between the NRP and combined models. This implies that

supplementing NRP data with selected OP cases may not adversely affect tax change predictions for the 270s. There are slight differences at the top of the distribution where the NRP only model seems to track the actual tax change amount better than the predictions from the combined model. Figure 6 shows the same comparison using OP as the test data instead of NRP. While there continues to be little difference between the model built using NRP data alone (dotted line) and the model built using a mix of NRP and OP data (dashed line), both do seem to underpredict the tax change across the bottom part of the distribution. As with the no change predictions, however, this may partly be due to the way the NRP ranking compares with the OP test data, which is less aligned than the plots comparing the NRP ranking with the NRP test data. Another reason for the relative poorer performance in the OP test data could be because for prediction the NRP dummy D is set to one and the selection bias term λ is set to 0 for both NRP and operational cases in the test year. Thus, while the model is trained on both NRP and OP data, the prediction is calculated pretending the case is a random sample from the filing population even when the actual case belongs to the OP test data.

FIGURE 5. AC 270 Tax Change Predictions on NRP Data, Ranked by Predicted Tax Change Based on NRP-Only Model

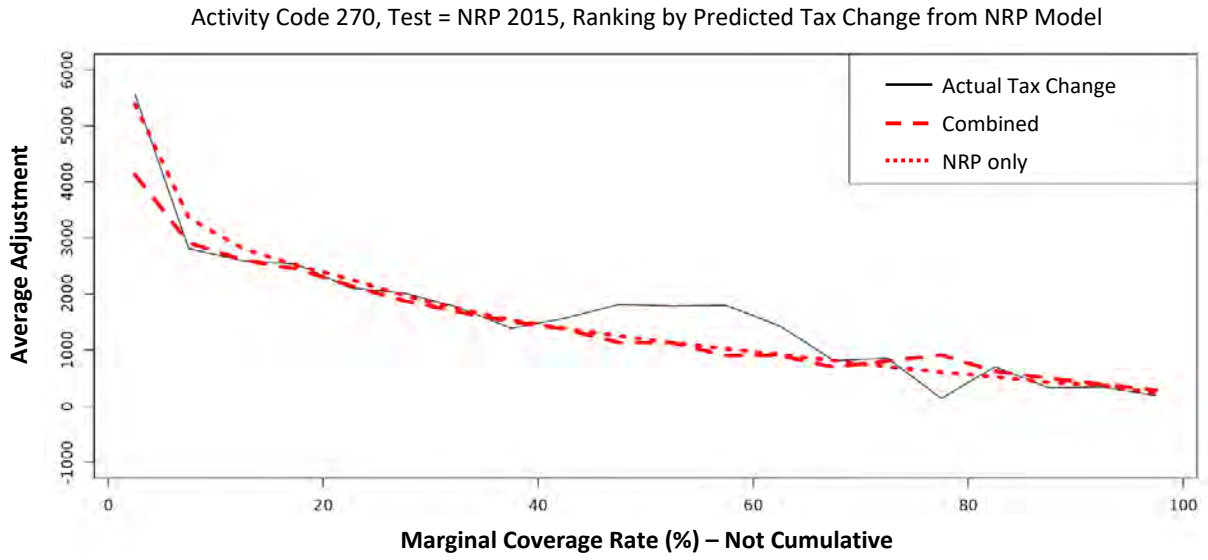
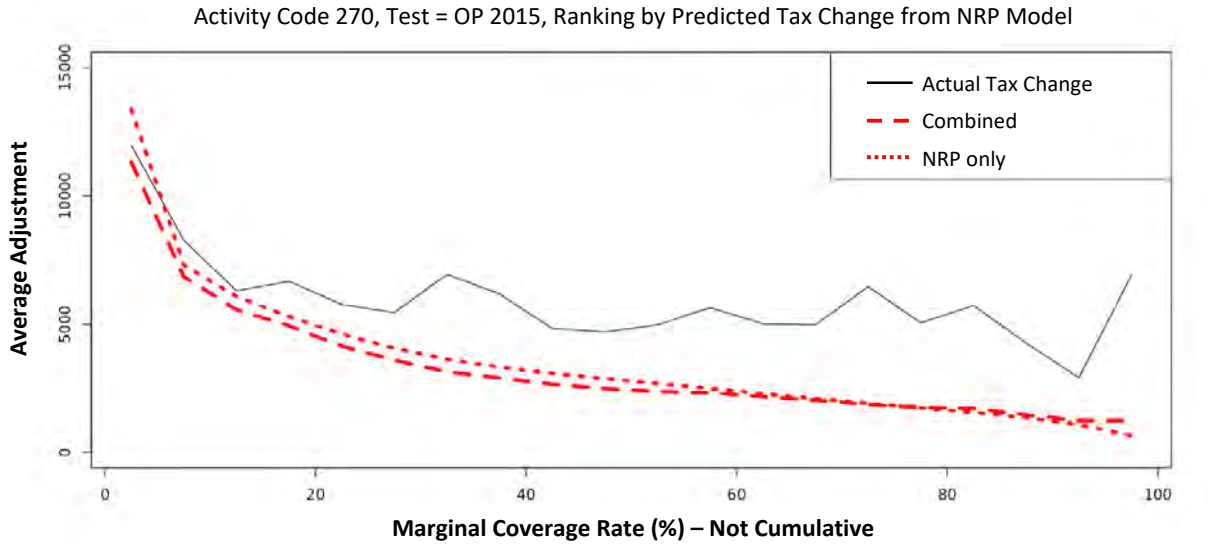


FIGURE 6. AC 270 Tax Change Predictions on OP Data, Ranked by Predicted Tax Change Based on NRP-Only Model



Like the plots for no change probability, to show the predicted tax changes from the different models in equal footing, we present the plots for the error in prediction, i.e., predicted tax change minus the actual tax change proportion. Figures 7 and 8 show the error distribution plots across the range of predictions for the NRP and the OP test data, respectively. The error plots confirm the similarity of the two models. It also corroborates the fact that expected tax change computed based on a random population model underpredicts the amount of tax change for the OP test data because the actual no change rate is lower than what is predicted.

FIGURE 7. AC 270 Error in Predictions for Tax Change on NRP Data

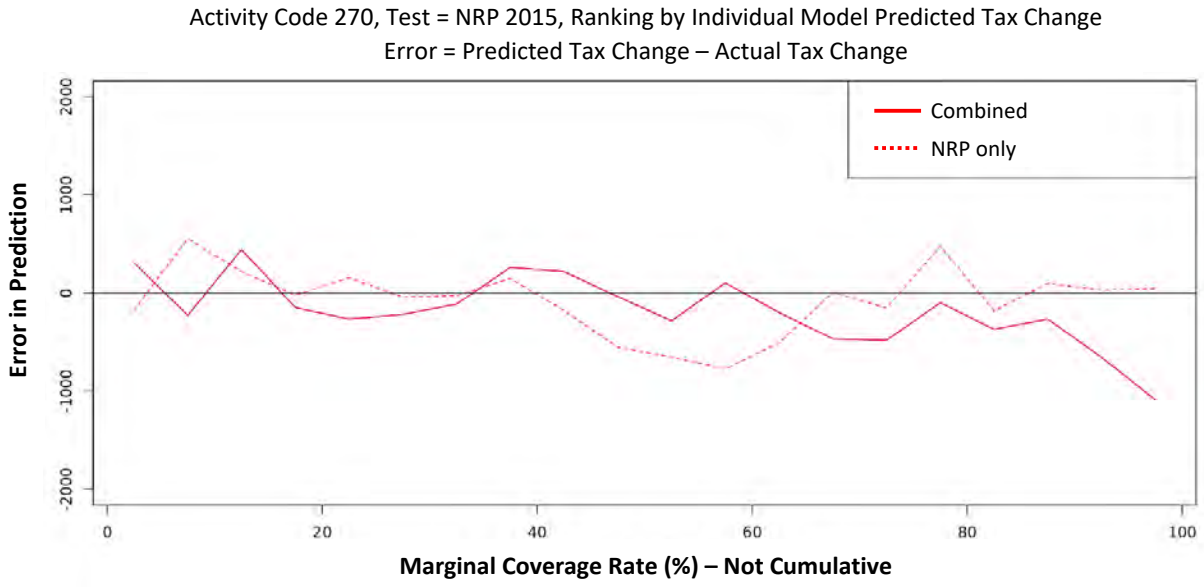
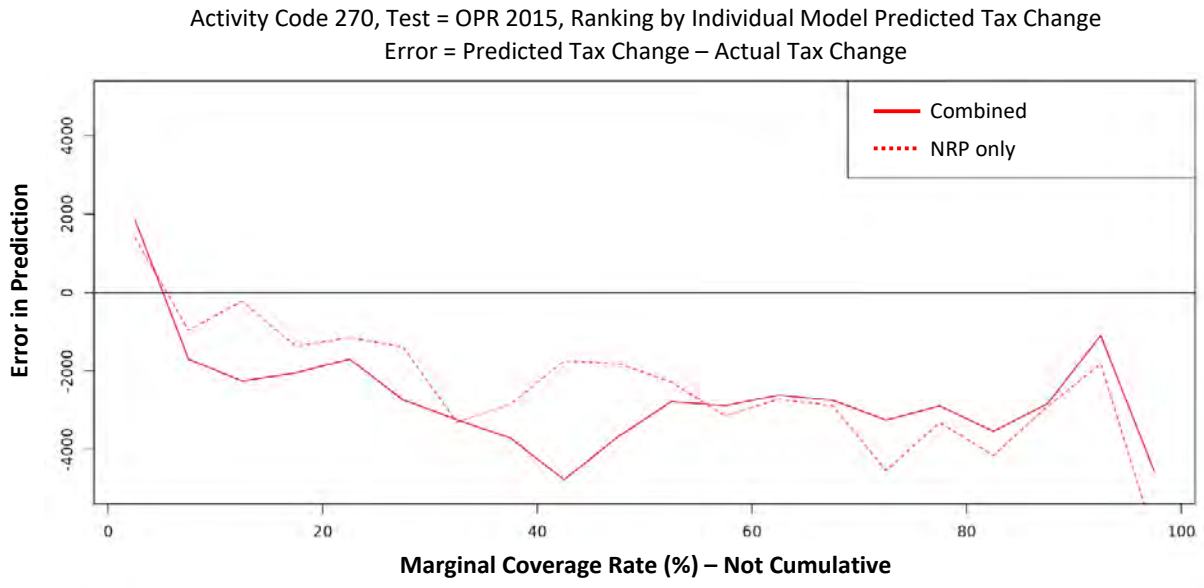


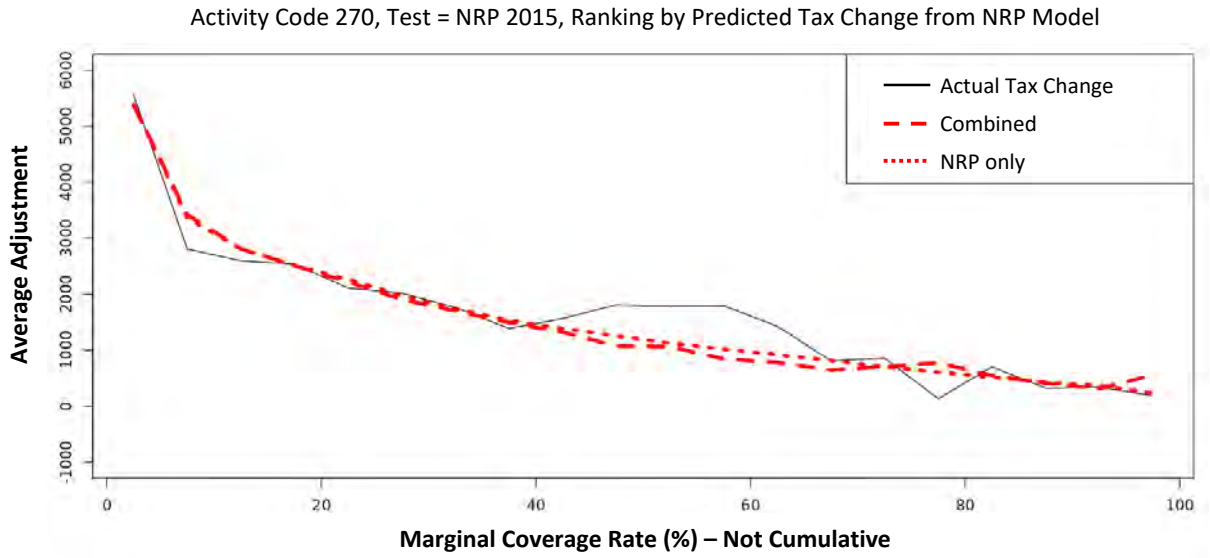
FIGURE 8. AC 270 Error in Predictions for Tax Change on OP Data



F. Activity Code 270, Using Combined Data Only in the Class-Specific Regression

The NRP only model performs better than the combined model in estimating the no change class. This is because the selection of the OP cases is not random, and the no change cases are generally selected out in the OP selection process. Hence, there is a bias in the no change probability estimation when OP data are present in the training set. We decided to use the combined data model only for the regression part of the two-stage risk model, with the multinomial logit classification of the change/no change classes done based on only the NRP part of the data. This is one way to test whether differences in selection on the operational side, such as screen outs that remove cases from the pool of potential audits are impacting the classification models enough to shift the results. Figure 9 shows the result for the tax change value, predicted and actual, for the models with only NRP data as well as NRP and OP data combined. The ranking is based on PTC from the NRP only model.

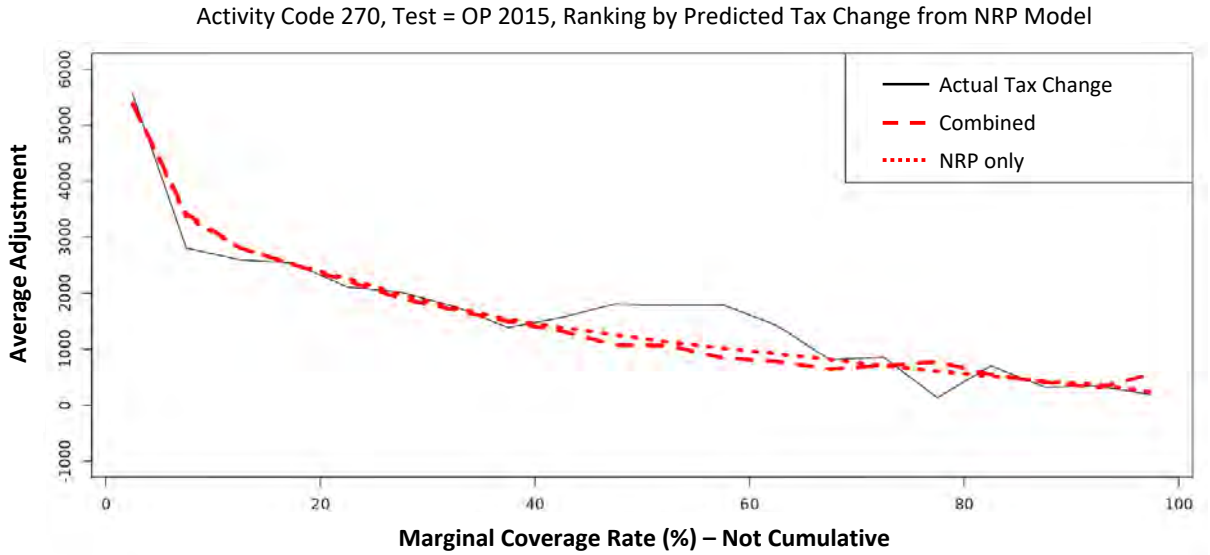
FIGURE 9. AC 270 Tax Change Predictions on NRP Data, Ranked by Predicted Tax Change Based on NRP Only Model. (The combined data are used only in the second stage of the two-stage risk model.)



The PTC from the NRP only model and that from the combined model are nearly perfectly aligned now, particularly at the high tax change range. The difference observed between the PTC from the two models at the high tax change bin in Figure 5 can be attributed to the misclassification of the no change class from the combined model due to the influence of the OP data. The OP data can supplement and provide additional information to the NRP informed model once it has been determined that there is a non-zero change. Thus, we decided to use a two-stage model with the combined data used only for the conditional class specific for subsequent analysis.

The plot for the OP data where the two-stage model is using the combined data only in the class specific regression is given in Figure 10. Again, both models perform worse for the OP test data since the predictions assume that the predicted population is aligned with the NRP data. The predictions from the two models seem to be closer, particularly at the top of the tax change bracket, compared to those in Figure 6.

FIGURE 10. AC 270 Tax Change Predictions on OP Data, Ranked by Predicted Tax Change Based on NRP Only Model. (The combined data are used only in the second stage of the two-stage risk model.)



Figures 11 and 12 present the error in prediction of the tax change amount where the error curves are based on the ranking obtained using the PTC from the individual models. For the NRP test data in Figure 11, both models seem fairly consistent with each other as well as accurate with the error typically within $1,000 of the actual average tax adjustment. Figure 12 shows the same information using the OP data as the test set, and suggests the model accuracy drops for the OP cases, which we are mostly under-predicting by $2,000-$4,000 across the distribution. It is worth noting that the models still produce similar results, however, showing that we do not lose any performance by including the OP cases in the risk models.

FIGURE 11. AC 270 Error in Predictions for Tax Change on NRP Data When Combined Data Used Only in the Second Stage of the Risk Model

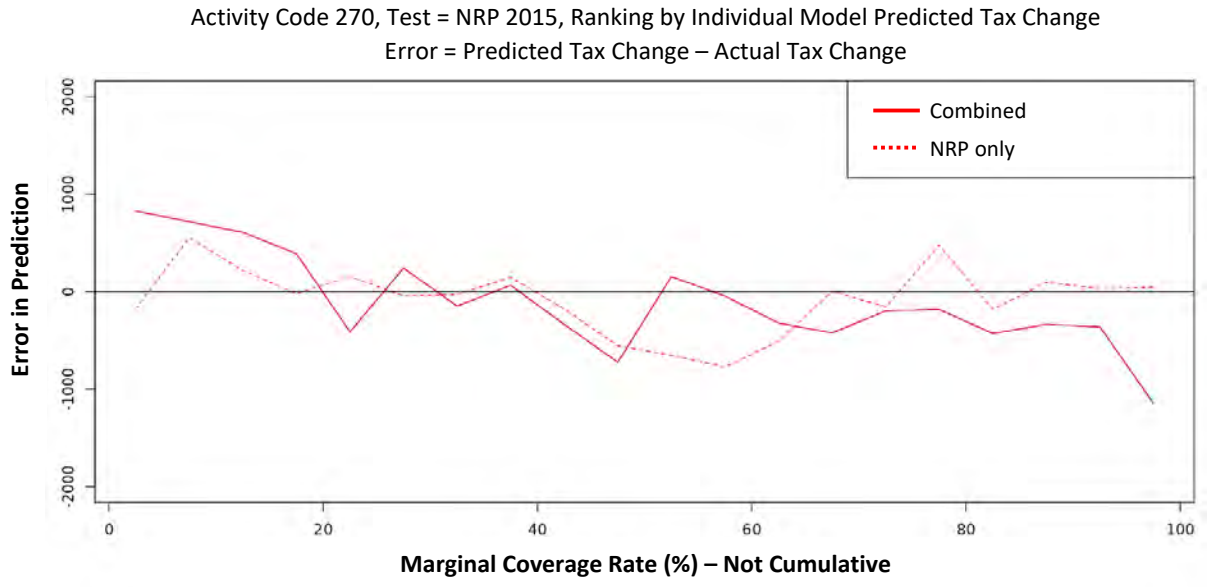
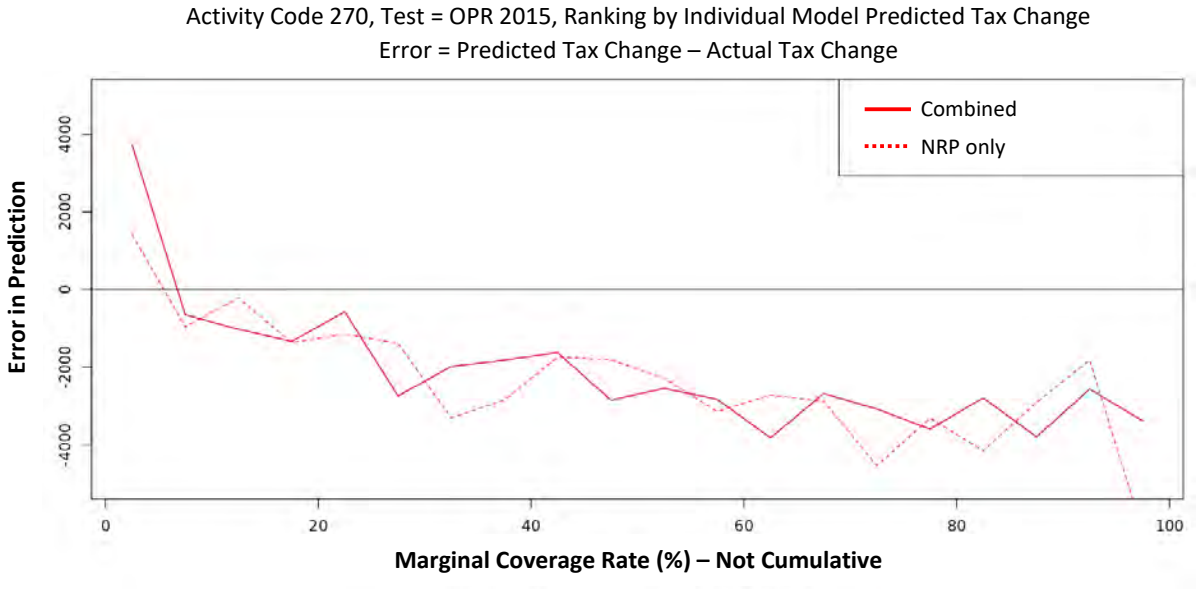


FIGURE 12. AC 270 Error in Predictions for Tax Change on OP Data When Combined Data Used Only in the Second Stage of the Risk Model



We also investigate whether the predictive risk model can give similar performance when only a fraction of the NRP data are used, particularly when the OP data are used to supplement in the second stage of the model. Such a model would provide the necessary cost effectiveness in the NRP design while still being able to produce unbiased tax reporting predictions using the additional information provided by the appropriately chosen OP training cases used in the second stage of the model. This may be a useful way to extend the capabilities of the risk model in segments with low NRP coverage, or where NRP cases look like OP cases.

To this end, we trained the risk models including only a fraction of the NRP sample in the training data. The randomly sampled fraction is used by itself to build the NRP-only risk model or combined with the OP data in the second stage of the risk model to build the model for the combined data. The proportions used were 50 percent and 25 percent, and the results were compared with the results from the training data with 100 percent NRP data. The reduced NRP sample size included is N=6,030 NRP cases for the 50 percent NRP data model, and only N=3,015 cases for the 25 percent NRP data model. Figures 13 and 14 provide the no change probability plots for all the models with different proportions for the NRP data when the test data are NRP and OP, respectively. Note that since only NRP data are used in the first stage of the risk model, the predicted probabilities of no change are exactly on top of each other for the NRP-only and the combined model. Figures 15 and 16 provide the predicted tax change amount for the NRP and OP test data, respectively. The models with different sample fractions seem to provide very similar predictions for the NRP test data, indicating the possibility of using only a fraction of the NRP sample size without compromising the prediction capabilities.

To gauge the uncertainty associated with the subsampling process, we used repeated sampling to construct a 90 percent confidence interval for the predictions based on the 50 percent NRP sample (shown in blue in Figure 15). The plots with the confidence band (separate plots for the predicted tax change based on the NRP only model and those based on the combined model) are shown in Figures A7 and A8 in the appendix. The ventiles are ranked based on the predicted tax change from the NRP only model. For each ventile, the red line shows the predicted tax change based on the model with 100 percent NRP data while the blue line shows the mean of 100 replications of the model with 50 percent NRP data. The grey band is the pointwise 90 percent band based on the 100 replications. The plots show stable predictions for most of the distribution, but a bit more uncertainty at the highest and lowest ventiles.

Though the models tested on NRP data seem quite stable, for the OP test data there seems to be greater variation in terms of the sample fraction, particularly for the high tax change bracket. But since the OP data are predicted based on the random population assumption and the data are ranked by the NRP-only model, it is not clear how the reduction of NRP sample is affecting the quality of prediction for the OP test cases.

FIGURE 13. AC 270 Probability of No Change on NRP Data, Ranked by Predicted Probability of No Change Based on the 100% NRP-Only Model

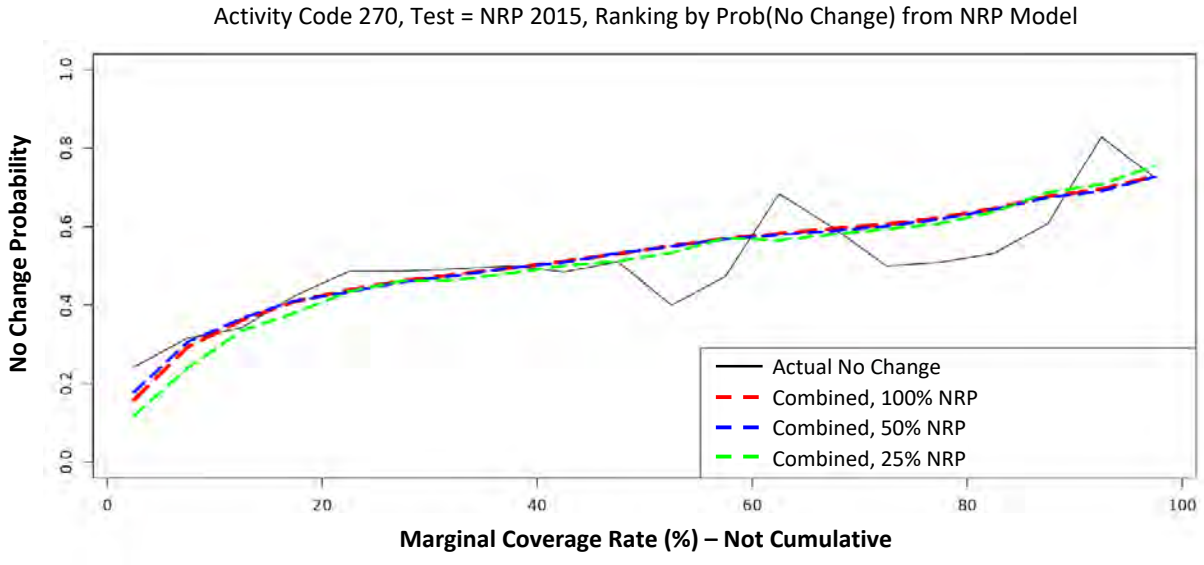


FIGURE 14. AC 270 Probability of No Change on OP Data, Ranked by Predicted Probability of No Change Based on the 100% NRP-Only Model

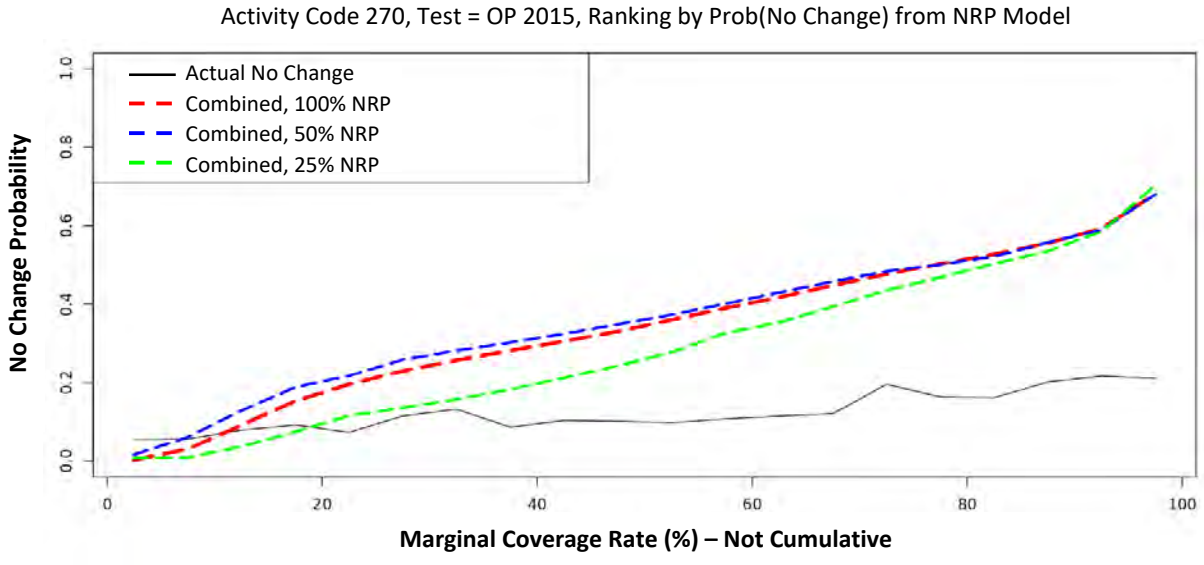


FIGURE 15. AC 270 Tax Change Predictions on NRP Data, Ranked by Predicted Tax Change Based on the 100% NRP-Only Model

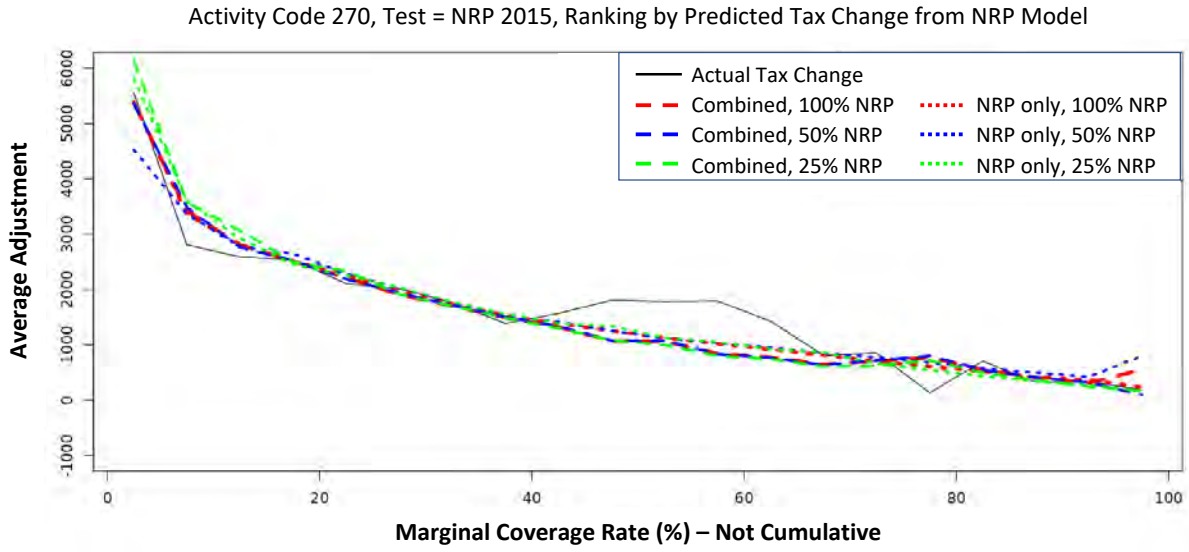
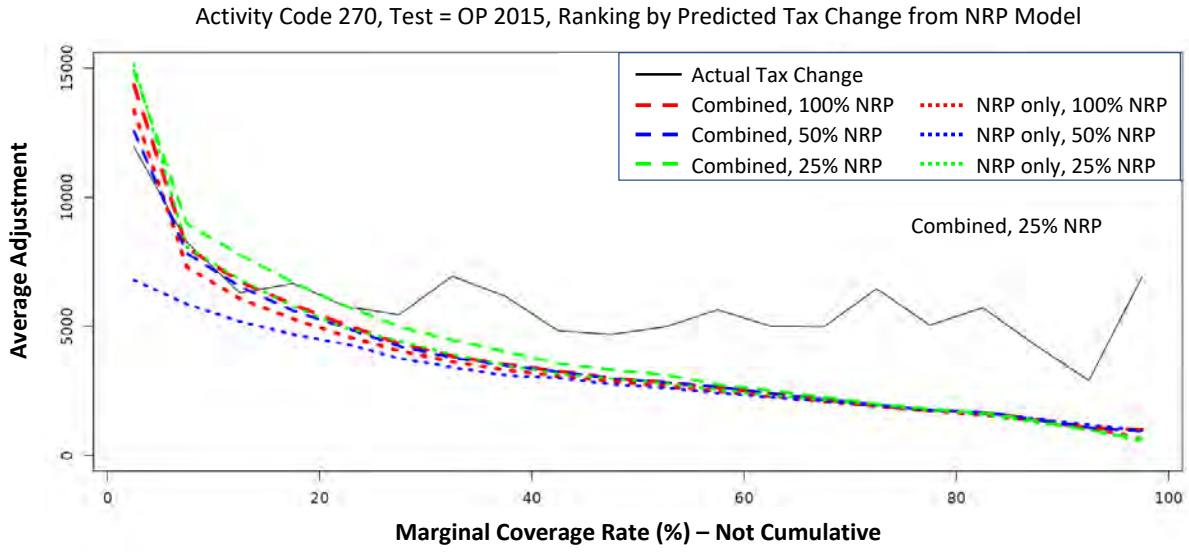


FIGURE 16. AC 270 Tax Change Predictions on OP Data, Ranked by Predicted Tax Change Based on the 100% NRP-Only Model



As another check on the sensitivity of the models to the proportion of the NRP data used to train them, we compare the mean predicted tax change and the Root Mean Squared Error (RMSE) for models that used all available NRP data, half the available NRP data, and a quarter of the available data. Because of extreme observations, particularly at the upper end of the distributions, we trimmed 2 percent of data at each of the distributions for computation of the mean and the RMSE. This includes both the combined models and the models that used only NRP data, for both the OP test year and the NRP test year. For AC 270, there is little change in the mean prediction and error when we vary the amount of NRP data used, with few exceptions. This is true when the combined data are used in both stages of the risk model (Table 2) as well as when it is used only in the first stage (Table 3).

TABLE 2. AC 270 Mean Tax Change and Root Mean Squared Error Using Combined Model in First Stage

Mean and RMSE of Tax Change	Combined Models			NRP Models			Actual
Proportion NRP Data Used	100%	50%	25%	100%	50%	25%	NA
Mean NRP	1,658	1,628	1,517	1,765	1,722	1,655	1,743
Mean OP	3,097	3,045	2,859	3,318	3,432	3,265	4,734
RMSE NRP	2,552	2,553	2,549	2,502	2,498	2,498	0
RMSE OP	4,618	4,619	4,643	4,763	4,879	4,677	0

TABLE 3. AC 270 Mean Tax Change and Root Mean Squared Error Using NRP-Only Model in First Stage

Mean and RMSE of Tax Change	Combined Models			NRP Models			Actual
Proportion NRP Data Used	100%	50%	25%	100%	50%	25%	NA
Mean NRP	1,731	1,763	1,714	1,765	1,722	1,655	1,743
Mean OP	3,583	3,715	3,496	3,318	3,432	3,265	4,734
RMSE NRP	2,601	2,614	2,641	2,502	2,498	2,498	0
RMSE OP	4,813	4,859	4,949	4,763	4,879	4,677	0

We also checked the performance of the different models when they are compared with actual tax change with respect to an oracle ranking based on the actual tax change from the test data. The results are given in the appendix (Figures A1 and A2) and show while the models underpredict the tax change at the top bracket due to the extreme outliers, the PTC generally show an upward trend indicating that a selection based on the PTC will generally select the actual high tax change cases at the top part of the distribution.

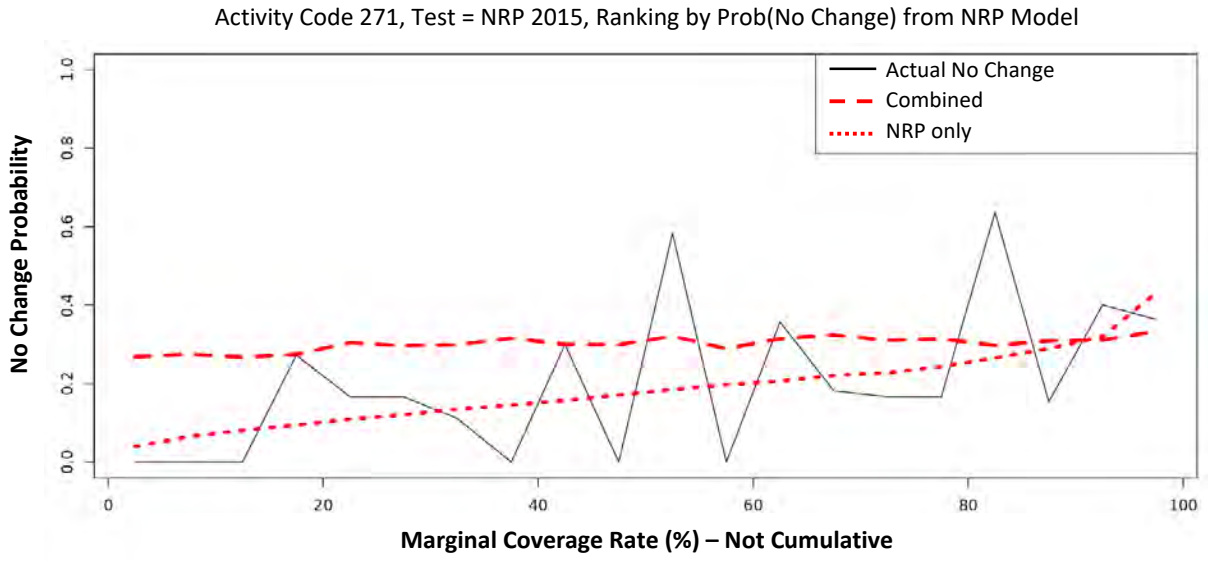
The appendix also contains additional plots showing the error in estimation for the no change probability and tax change for the models trained with different proportions of NRP data (Figures A3–A6).

H. Activity Code 271

AC 271 comprises cases with EITC claims, less than $200,000 in Total Positive Income (TPI), and at least one Schedules C or F with total gross receipts above $24,999. These returns can be thought of as cases for individuals with low to moderate wage income who claim the EITC and have a relatively larger amount of non-wage income from small business or farm income reported on Schedules C or F. There are very few NRP cases selected for the 271s, so it's an area where models may benefit the most from supplementing the limited number of NRP cases with OP ones.

As in AC 270, for 271 the estimates of no change probabilities are better (closer to the actual proportion of no change) when the classification model (multinomial logit) is trained on the NRP data rather than the combined data. This is clear in the no change probability estimates from the two models when tested on the NRP data (Figure 17). The data are ranked based on the no change probability estimates from the model trained only on NRP data. The no change probability estimates are biased when OP data are included in the training sample.

FIGURE 17. AC 271 Probability of No Change on NRP Data



Therefore, we use the two-stage risk model with the combined data used only in the second stage of the model. We performed the simulation experiments with the different sample fractions of NRP data used in training, as well. Given that the number of NRP cases in AC 271 is already low, taking further fractions of the data drastically reduced the NRP sample size and the results show a higher degree of variability.

Figures 18 and 19 show the no change probability plots for the NRP test data and the OP test data, respectively. The models are based on training data where different proportions of NRP data are used as well as those where OP data are combined with NRP in the second stage regression. Because we are using NRP data for classification in all models, the estimated no change probabilities are the same for both the models trained with only NRP and those where OP data are used in addition to NRP for the regression.

For the NRP test data (Figure 18), the model based on 100% NRP does quite well in predicting the no change class probabilities across the different bins. There is a slight deterioration when only half the samples are used. Using only 25 percent of the NRP data results in larger deviations between the estimated and the actual no change probabilities, which is expected given the small size of the NRP sample for AC 271.

For the OP test data (Figure 19), all models seem to be bias and underpredict the no change probability for most of the distribution.

FIGURE 18. AC 271 Probability of No Change on NRP Data (The combined data are used only in the second stage of the two-stage risk model.)

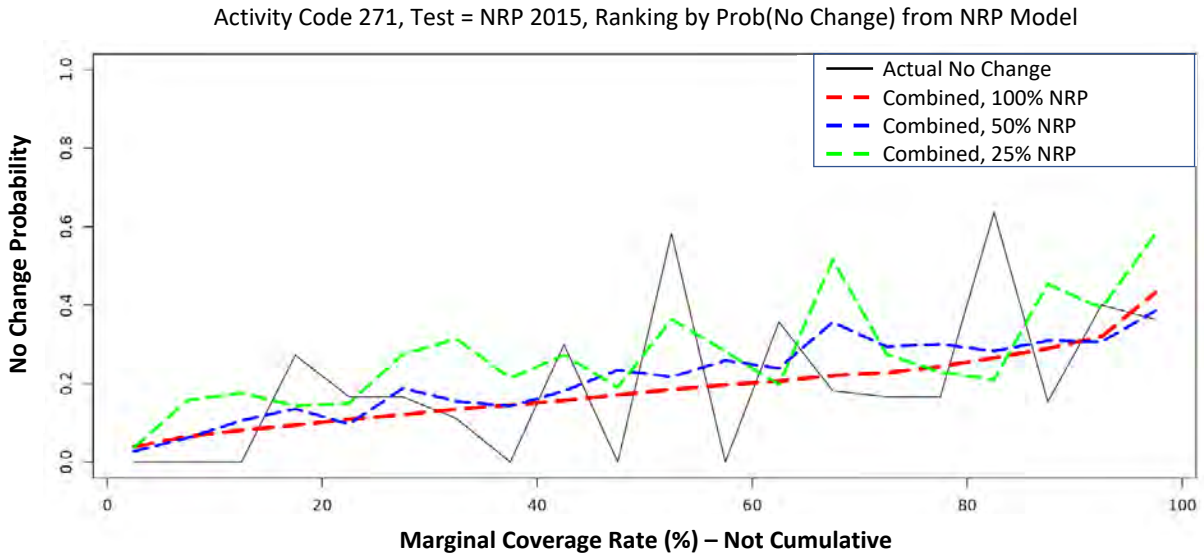
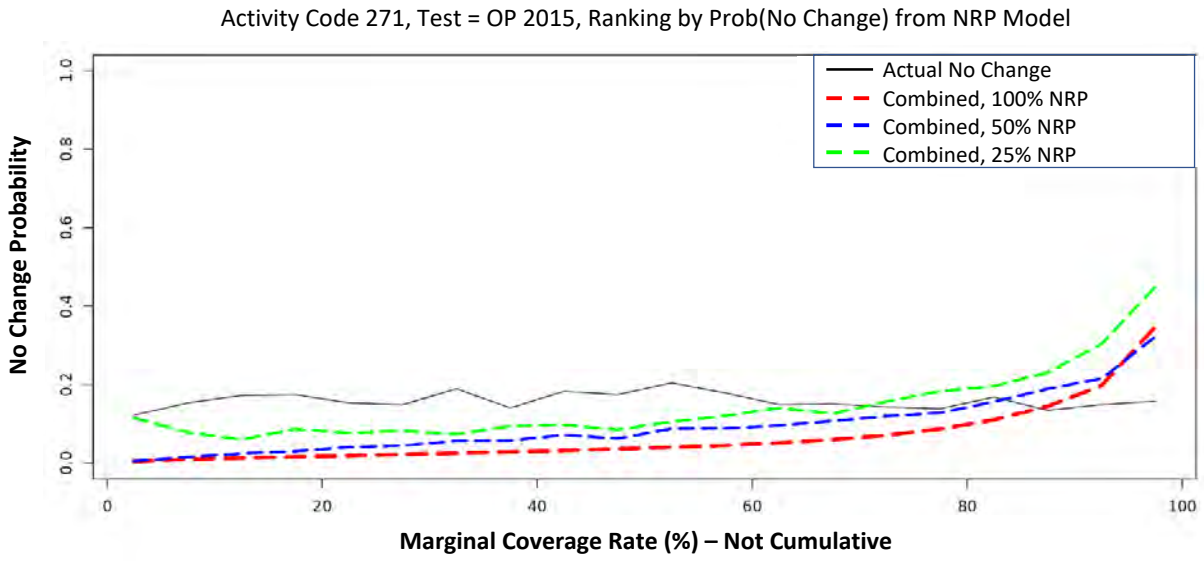


FIGURE 19. AC 271 Probability of No Change on OP Data (The combined data are used only in the second stage of the two-stage risk model.)



Next, we present the predicted tax change plots when the ranking is done based on the model that uses 100 percent of the NRP data and no OP data. Figures 20 and 21 show the predicted tax change plots for the NRP test data and the OP test data, respectively. The plots for PTC based on training data with fractional NRP samples are also shown in the same plot. The PTC curves from the different models are close for the NRP test data, but not as close as they were for AC 270. There is a greater amount of variability in the plots based on fractional NRP samples, with models based the 100 percent NRP data (red lines) performing better than those based on subsamples. The greater variation is most likely due to the reduction in sample size for AC 271. The prediction for the OP test data shows greater variability for the models trained on different proportions of NRP data, with the model based on 100 percent NRP data performing the best. The models using the combined data

are all biased and the change in the fraction of NRP data used does not seem to affect the prediction as much. This is because the NRP data size is very small compared to that of the OP data, giving the OP data great influence in the model. However, when we are predicting for the OP test data, we are treating the cases as random samples from the filing population though they are actually cases selected by the OP audit process. Moreover, the ranking is based on the PTC from the NRP only model.

FIGURE 20. AC 271 Tax Change Predictions on NRP Data, Ranked by PTC From the Model With 100% NRP Data

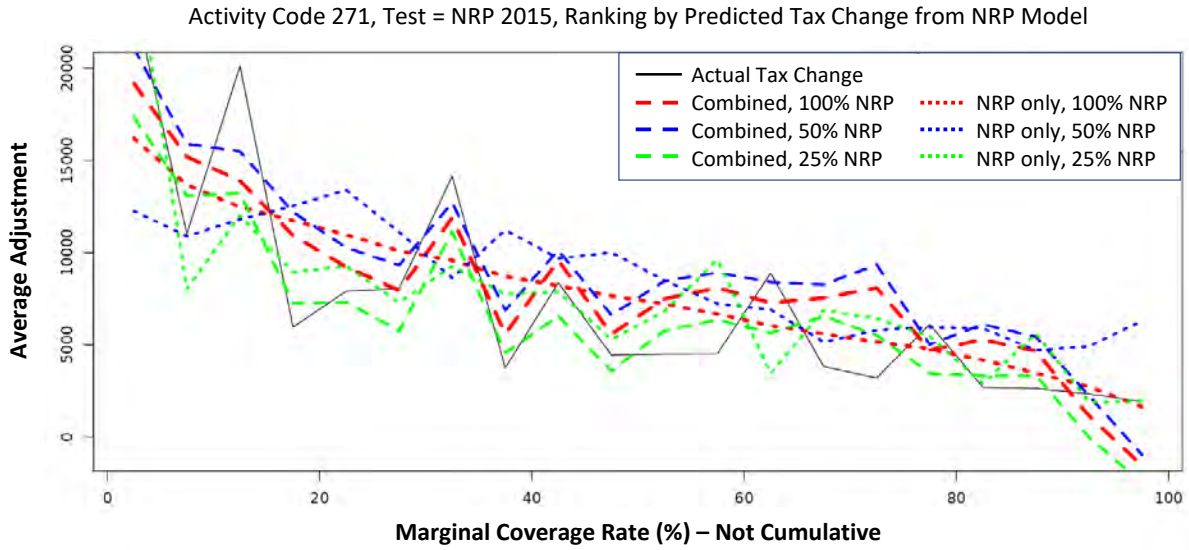
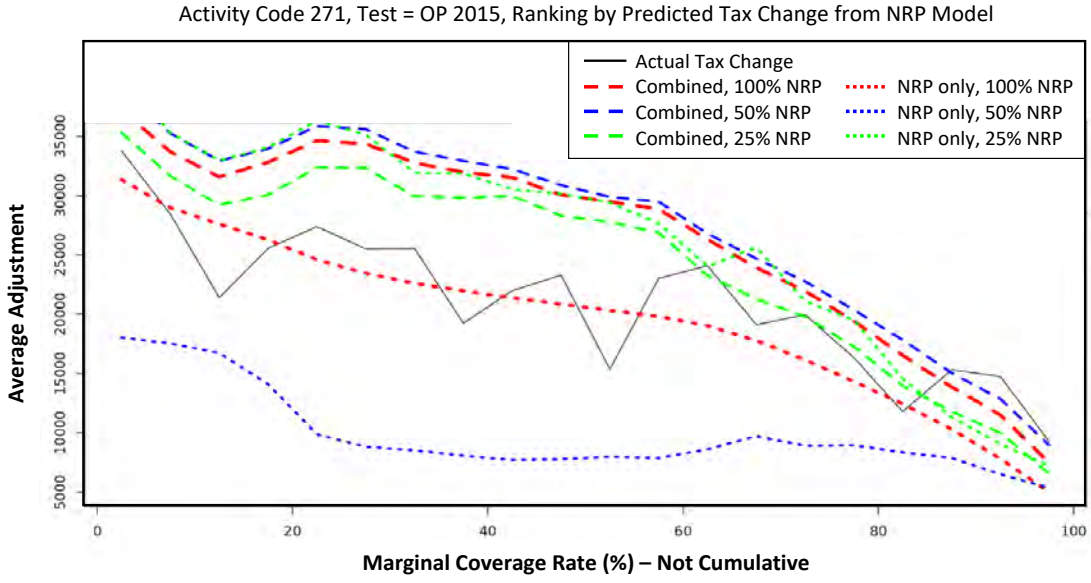


FIGURE 21. AC 271 Tax Change Predictions on OP Data, Ranked by PTC From the Model With 100% NRP Data



Next, we show the plots for the error in prediction. For clarity we only present the plots for the models based on 100% NRP data, using the individual models to sort the bins, rather than just the NRP data. Figure 22 shows the error plot for the NRP test data. The NRP-only model performs better than the combined model,

which over-predicts at the top of the distribution and underpredicts at the bottom. This loss of accuracy and the large ups and downs in the prediction error for each bin are likely driven by the low NRP sample size for the 271s. For the OP test data (Figure 23), the performance is similar to that observed for AC 270, with the combined model having a large upward bias.

FIGURE 22. AC 271 Error in Predictions for Tax Change on NRP Data When Combined Data Used Only in the Second Stage of the Risk Model

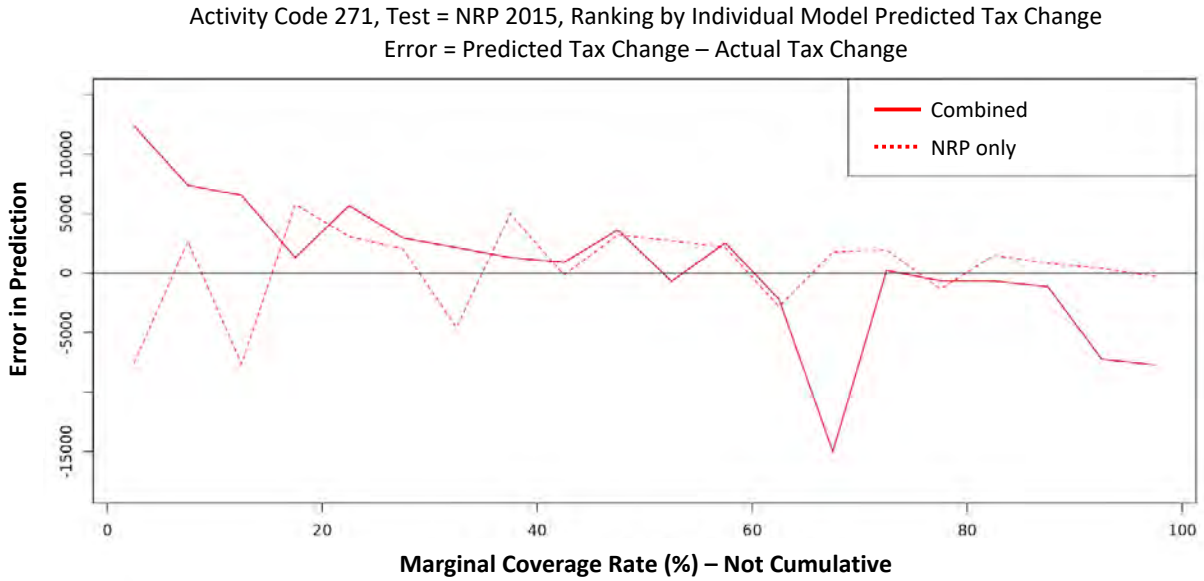
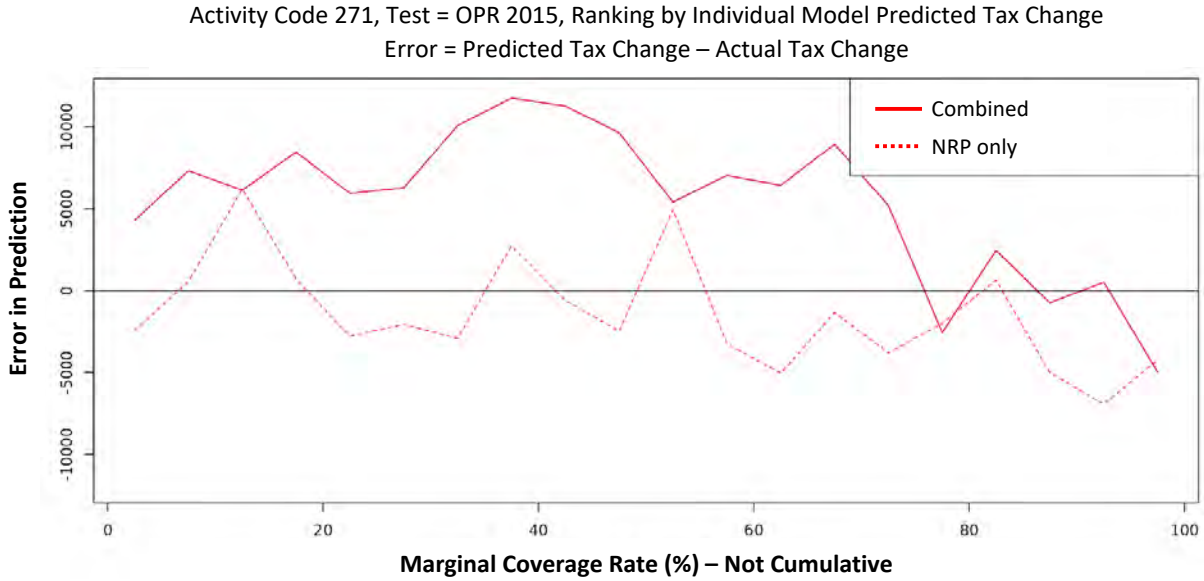


FIGURE 23. AC 271 Error in Predictions for Tax Change on OP Data When Combined Data Used Only in the Second Stage of the Risk Model



We also checked the performance of the different models when they are compared with actual tax change with respect to an oracle ranking based on the actual tax change from the test data. The results are given in the appendix (Figures B1 and B2). Additional plots showing the error in estimation for the no change probability and tax change for the models trained with different proportions of NRP data are given in the appendix in Figures B3-B6.

As with AC 270, for AC 271 we compare the means and root mean squared error for each mode, when combined data are used for both stages of the risk model (Table 4) and when it is used only in the first stage (Table 5). Despite the means being more sensitive to the different NRP sample sizes (which result from the small number of 271 NRP cases), the errors remain quite stable across the combined models, with more variation in the NRP-only models.

TABLE 4. AC 271 Mean Tax Change and Root Mean Squared Error Using Combined Model in First Stage

Mean and RMSE of Tax Change	Combined Models			NRP Models			Actual
Proportion NRP Data Used	100%	50%	25%	100%	50%	25%	NA
Mean NRP	6,854	5,626	6,006	7,617	6,393	5,531	6,195
Mean OP	21,890	20,958	21,697	19,658	16,886	18,917	16,040
RMSE NRP	9,072	9,036	9,059	7,966	9,073	8,992	0
RMSE OP	25,381	25,182	25,378	24,940	25,036	26,454	0

TABLE 5. AC 271 Mean Tax Change and Root Mean Squared Error Using NRP-Only Model in First Stage

Mean and RMSE of Tax Change	Combined Models			NRP Models			Actual
Proportion NRP Data Used	100%	50%	25%	100%	50%	25%	NA
Mean NRP	8,153	9,133	6,476	7,617	6,393	5,531	6,195
Mean OP	26,537	27,581	24,401	19,658	16,886	18,917	16,040
RMSE NRP	9,733	10,169	9,553	7,966	9,073	8,992	0
RMSE OP	27,093	27,690	27,227	24,940	25,036	26,454	0

V. Conclusions

In this paper we demonstrate methods of combining random and non-random audit data to estimate reporting compliance risk models for individual returns with the EITC. There appear to be many areas where adding the non-random/OP cases to NRP is a viable approach and simple measurement error and selection bias control results in similar estimates overall and across the risk distributions. However, the simple controls may not be sufficient for estimating the (conditional) proportion that have no tax adjustment. The OP audit data may induce bias into estimated likelihoods of "no tax change" and thus, randomly selected and consistently labeled data (like NRP) are critical to estimate these parts of the risk distributions (extensive margin). However, the results suggest that the simple selection bias and measurement error controls may be sufficient to estimate the expected amount of adjustment, conditional on there being an adjustment (intensive margin). This of course assumes that we have some randomly selected and consistently labeled data for the model training. More rigorous approaches that 1) weight operational cases based on the quality of the label and 2) better account for compliant cases 'screened out' of the operational audit pipeline prior to audit may be able to account for the biases.

For AC 270 cases, models generally reach similar predictions even if only 25 percent of available NRP data are used for training. For AC 271, where the NRP sample is already quite low, the models suffer if the sample is made smaller, but supplementing the sample with OP cases seems to improve the models' stability and accuracy for some segments of the distribution. Since the OP cases are lower cost compared with NRP audits, incorporating the OP cases into risk models may be beneficial, particularly in areas where the NRP sample is already low or nonexistent.

References

Colnet, Bénédicte, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang (2022). "Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review." arXiv:2011.08047.

Heckman, James (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47, pp. 153–161.

Ho, Chih-Chin, and Alex Turk (1999). "Measuring Tax Reporting Compliance: A Trichotomous Choice Model." *Turning Administrative Systems Into Information Systems*, Statistics of Income Division, Internal Revenue Service, 1998–1999, pp. 91–95.

Lee, Lung-Fei (1983). "Generalized Econometric Models with Selectivity." *Econometrica*, Vol. 51, No. 2, pp. 507–512.

Wiśniowski, Arkadiusz, Joseph W Sakshaug, Diego Andres Perez Ruiz, and Annelies G Blom (2020). "Integrating Probability and Nonprobability Samples for Survey Inference." *Journal of Survey Statistics and Methodology*, Vol. 8, No. 1, pp. 120–147.

Appendix A

A. Activity Code 270

The following two plots show how the predicted tax changes from the two-stage models compare with the actual tax change for the NRP and the OP test data when cases are ranked by their actual tax change. The ranking by actual tax change is of course an oracle ranking, unavailable in practice. It is interesting to see, however, that while the predicted tax change amount severely underpredicts the actual amount for bins with large outliers, the overall trend in the ranks for the predicted changes align with those for the actual tax change.

FIGURE A1. AC 270 Tax Change Predictions on NRP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by the Actual Tax Change.)

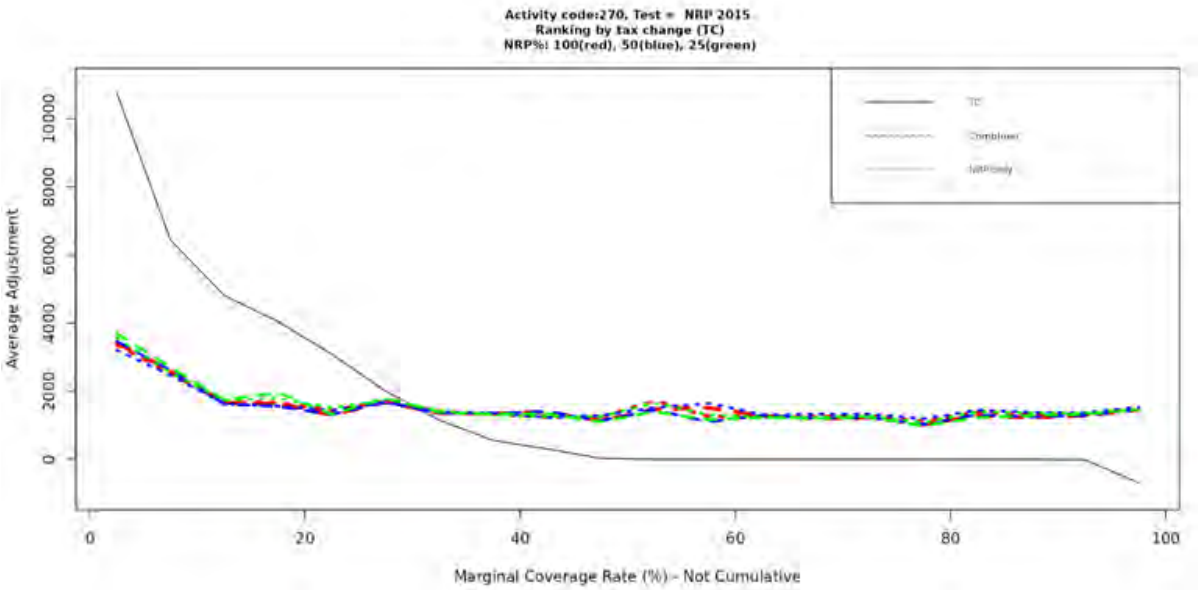
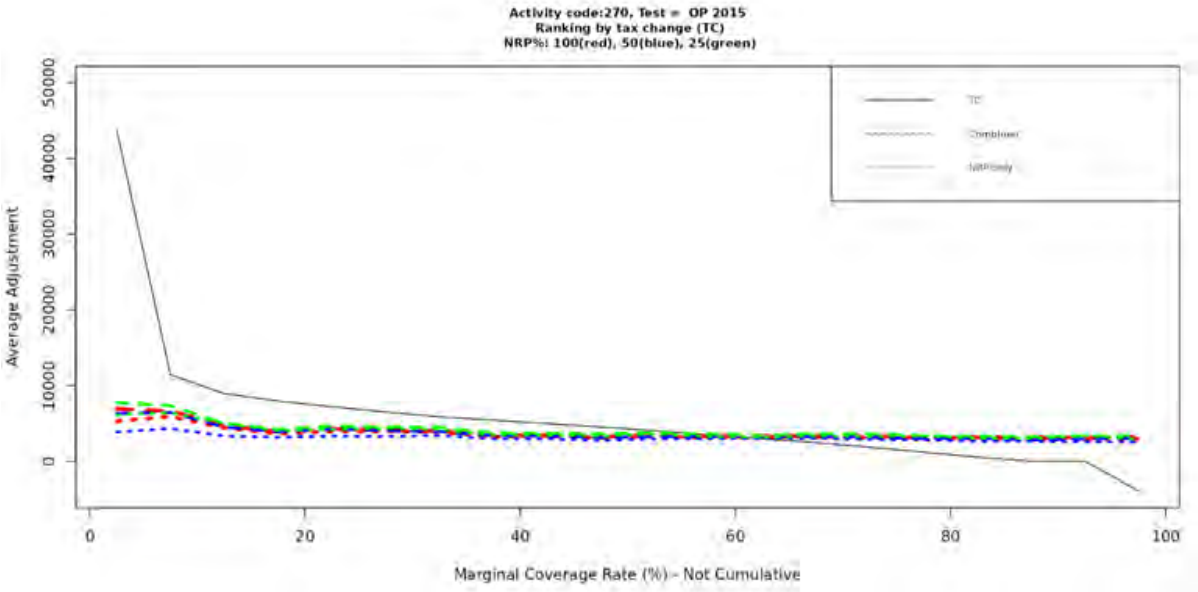


FIGURE A2. AC 270 Tax Change Predictions on OP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by the Actual Tax Change.)



The following four plots provide the plots for error in estimation for the no change probability and for the tax change amount for the NRP and the OP test data for AC 270. Plots include models trained with different proportion of the NRP data.

FIGURE A3. AC 270 Error in No Change Probability Predictions on NRP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

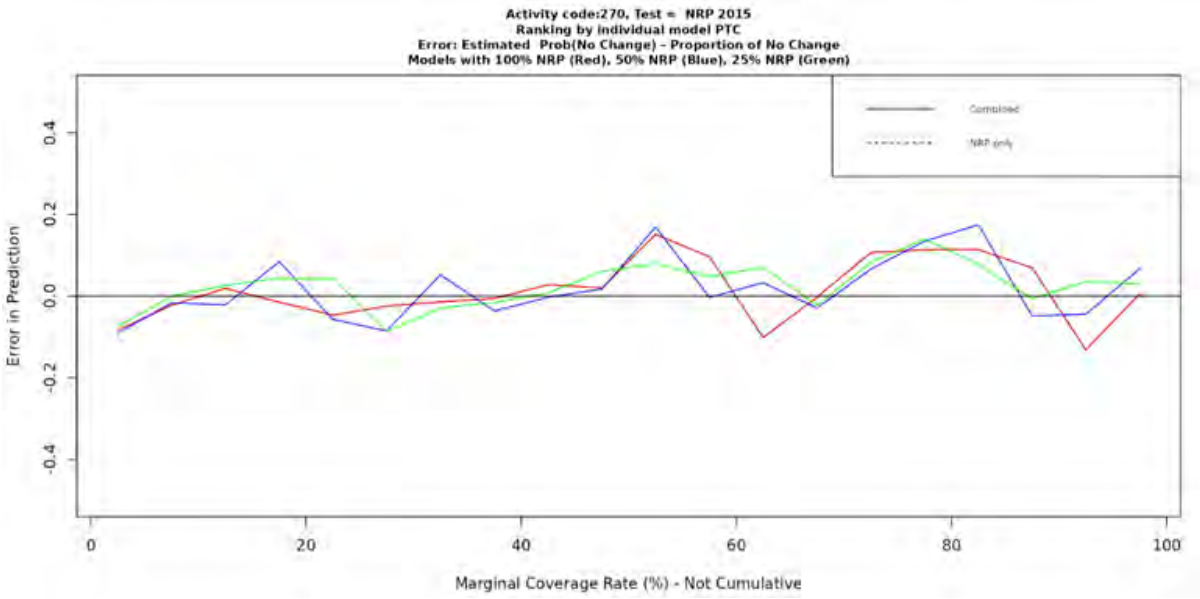


FIGURE A4. AC 270 Error in No Change Probability Predictions on OP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

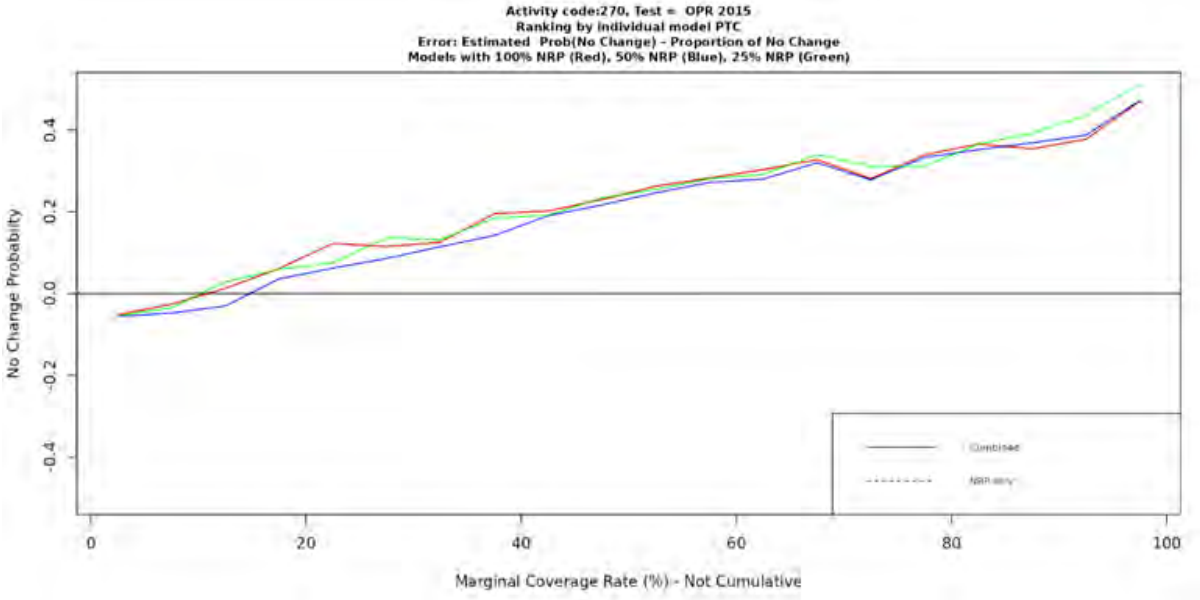


FIGURE A5. AC 270 Error in Tax Change Predictions on NRP Data for Models Trained With Different Proportions of NRP Data. Cases Are Ranked by Probability Estimates From Respective Models.

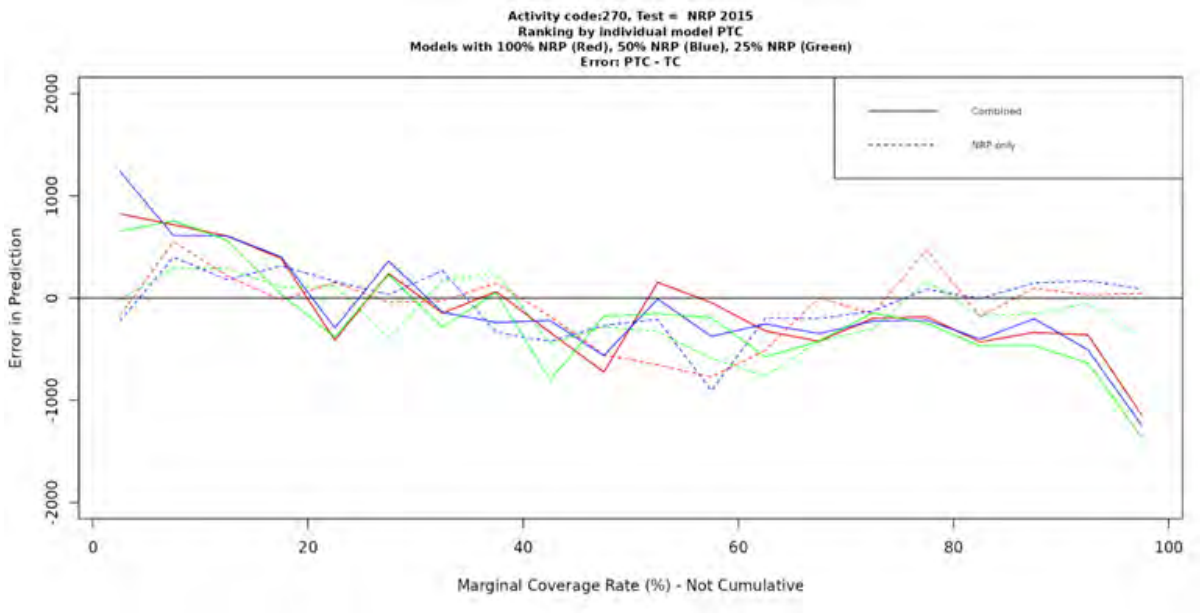


FIGURE A6. AC 270 Error in Tax Change Predictions on OP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

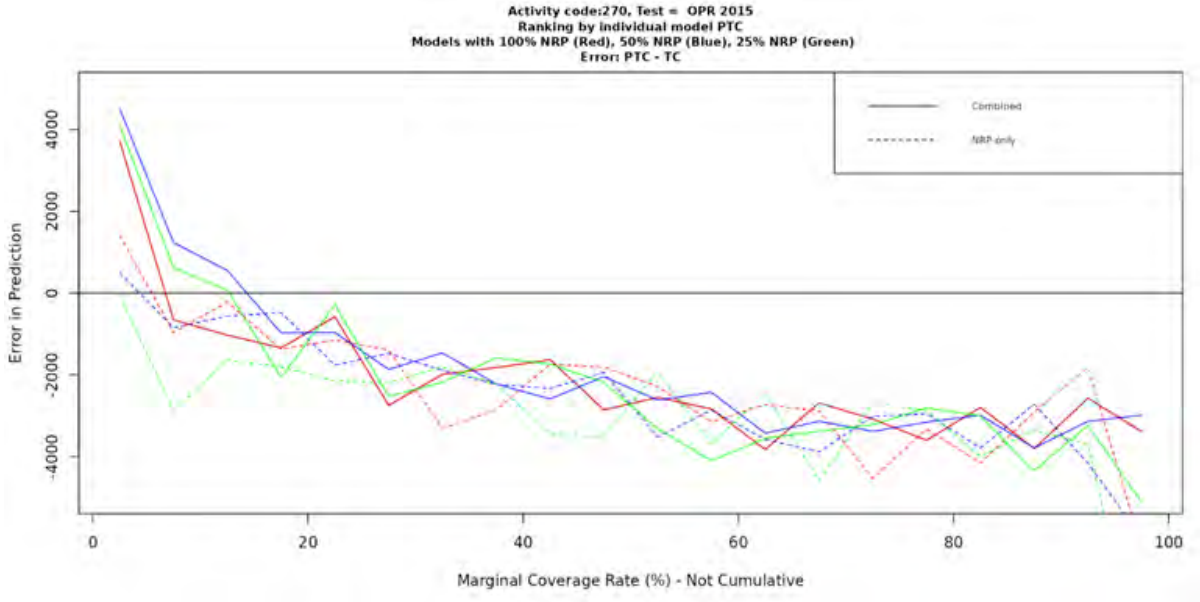


FIGURE A7. AC 270 90% Confidence Interval for Predicted Tax Change in Combined NRP/OP Model, Ranked by Predicted Tax Change of NRP Only Model

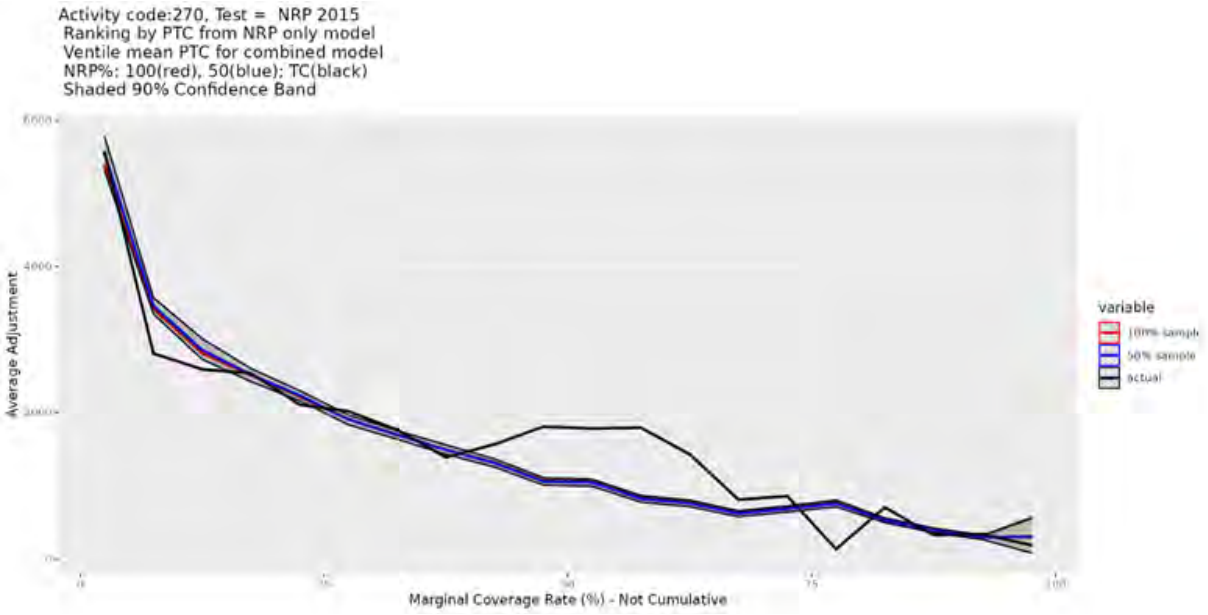
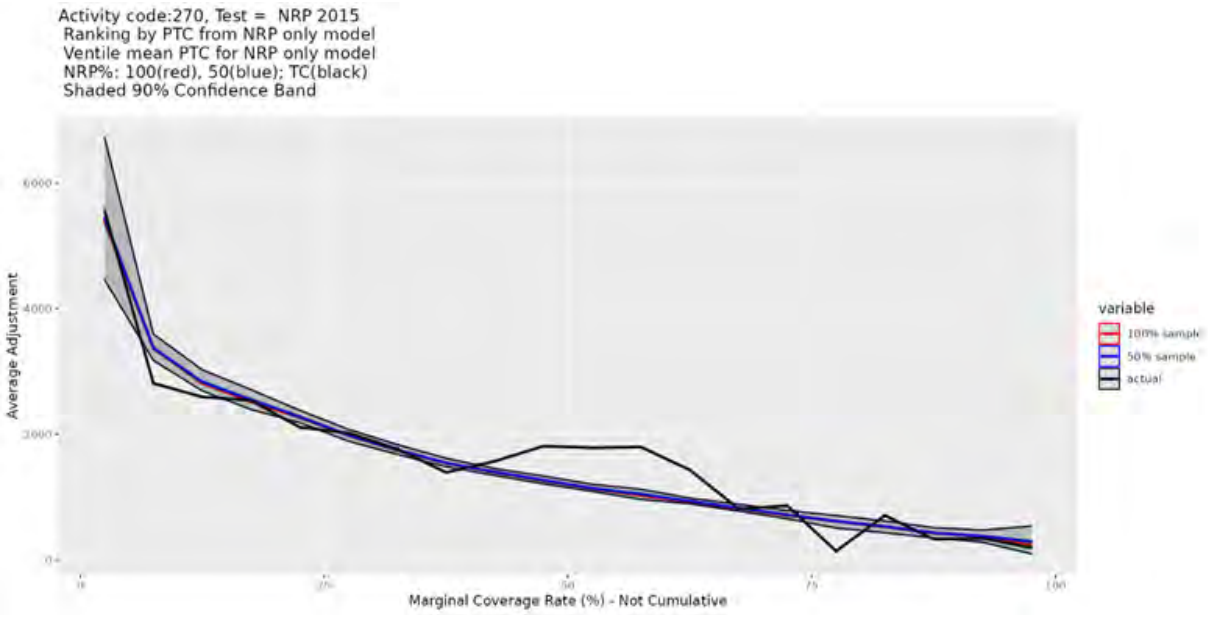


FIGURE A8. AC 270 90% Confidence Interval for Predicted Tax Change in NRP Only Model, Ranked by Predicted Tax Change of NRP Only Model



Figures A9 and A10 show the 90 percent confidence bands for the predicted tax change based on repeated subsampling of the NRP sample at the 50 percent rate when the ventiles are ranked by the PTC from the combined model.

FIGURE A9. AC 270 90% Confidence Interval for Predicted Tax Change in Combined NRP/OP Model, Ranked by Predicted Tax Change of Combined NRP/OP Model

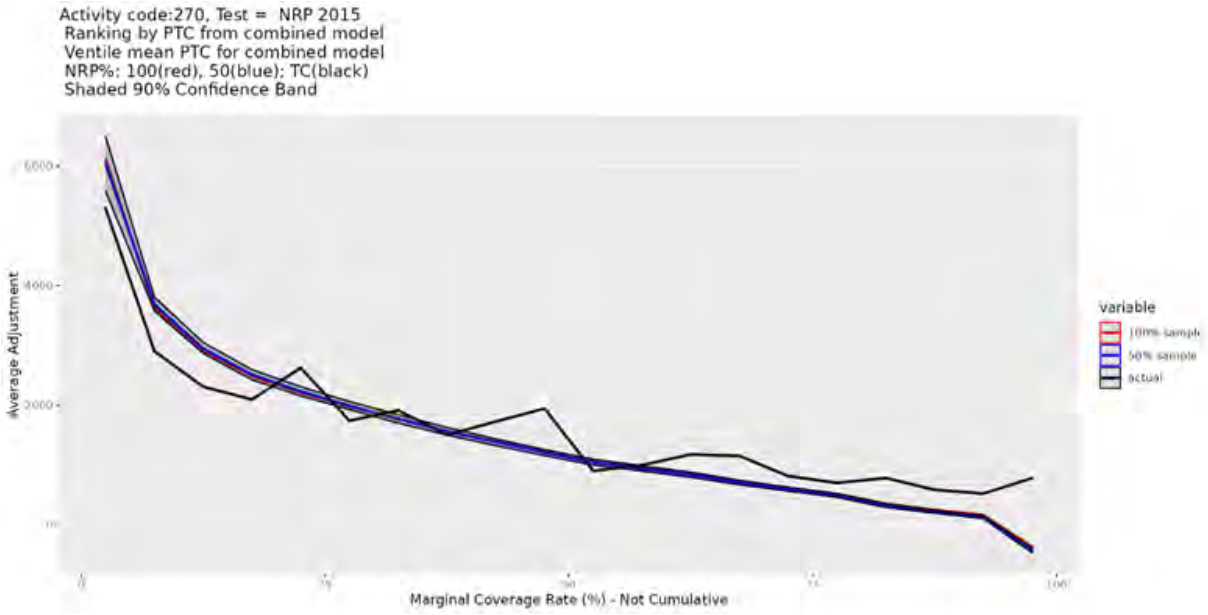
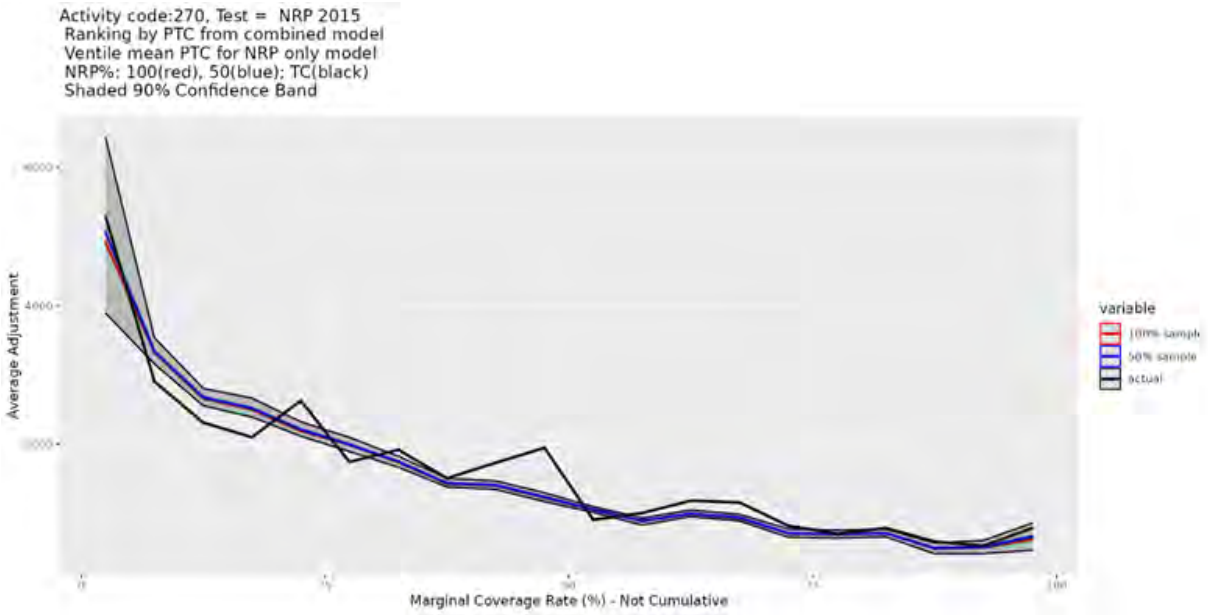


FIGURE A10. AC 270 90% Confidence Interval for Predicted Tax Change in NRP Only Model, Ranked by Predicted Tax Change of Combined NRP/OP Model



Appendix B

B. Activity Code 271

The following two plots show how the predicted tax changes from the two-stage models compare with the actual tax change for the NRP and the OP test data when cases are ranked by their actual tax change. The ranking by actual tax change is of course an oracle ranking, unavailable in practice. It is interesting to see, however, that while the predicted tax change amount severely underpredicts the actual amount for bins with large outliers, the overall trend in the ranks for the predicted changes generally align with those for the actual tax change, with some exceptions.

FIGURE B1. AC 271 Tax Change Predictions on NRP Data, Ranked by the Actual Tax Change

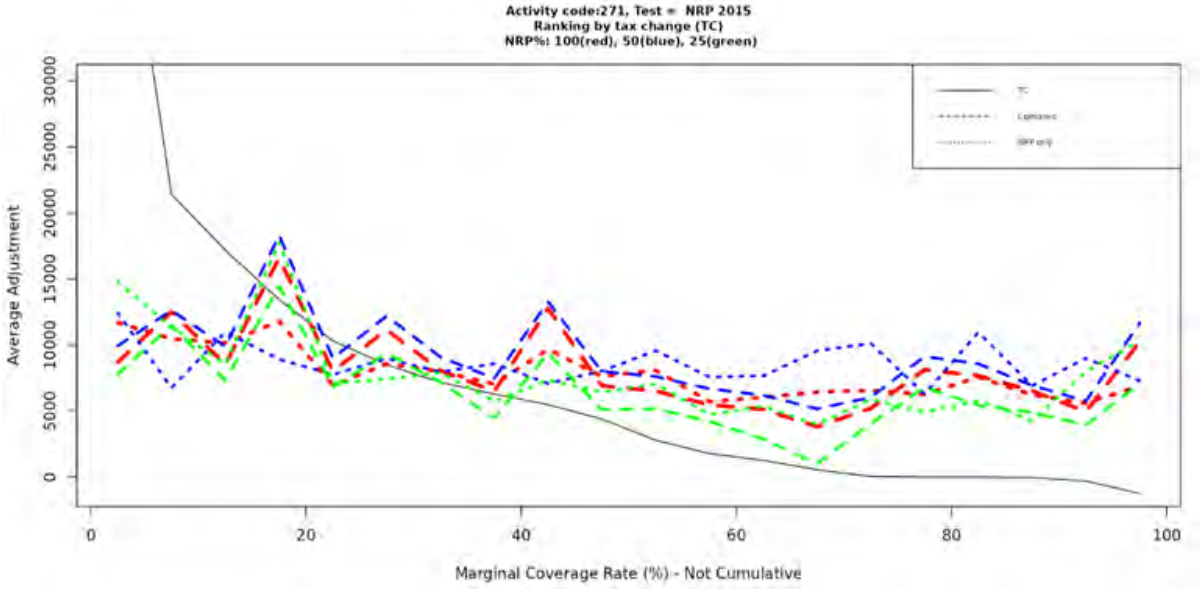
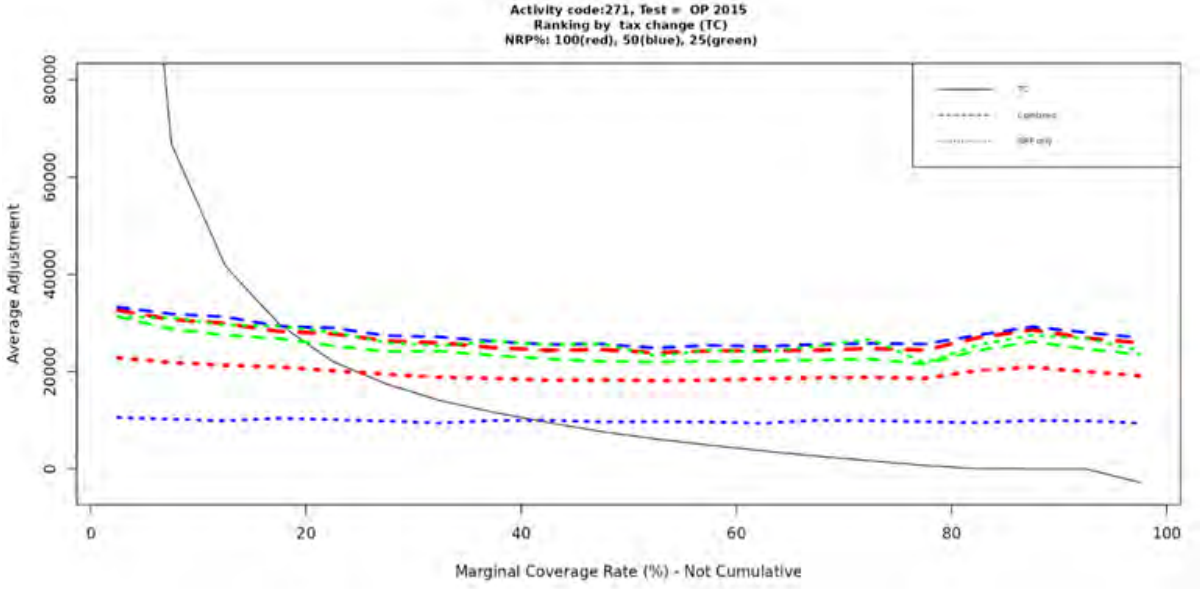


FIGURE B2. AC 271 Tax Change Predictions on OP Data, Ranked by the Actual Tax Change



The following four plots provide the plots for error in estimation for the no change probability and for the tax change amount for the NRP and the OP test data for AC 271. Plots include models trained with different proportion of the NRP data.

FIGURE B3. AC 271 Error in No Change Probability Predictions on NRP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

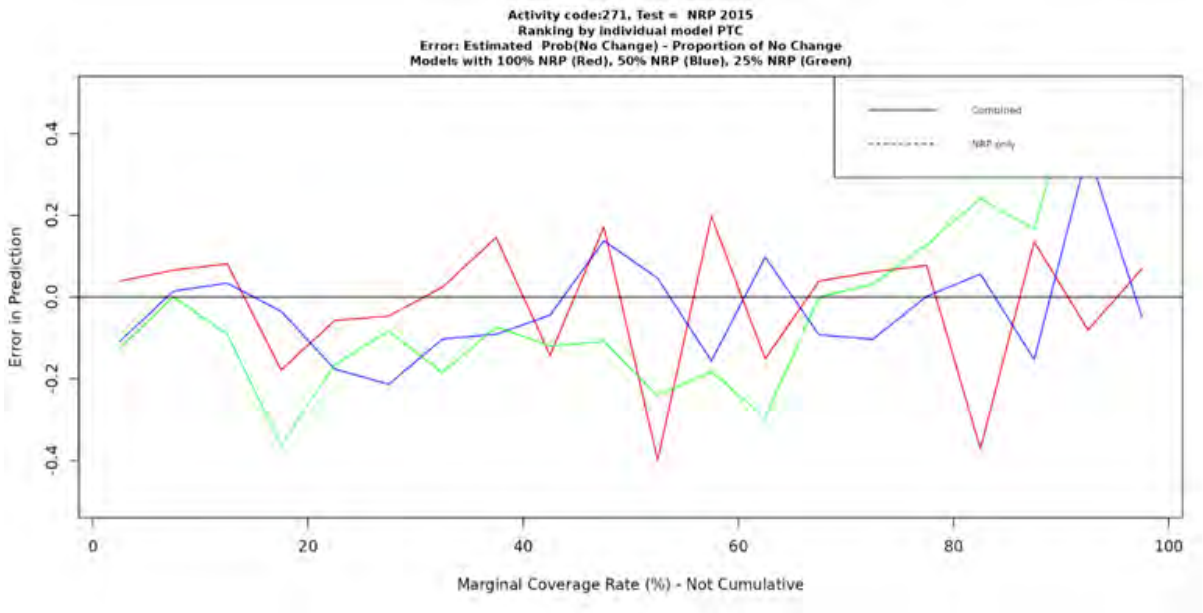


FIGURE B4. AC 271 Error in No Change Probability Predictions on OP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

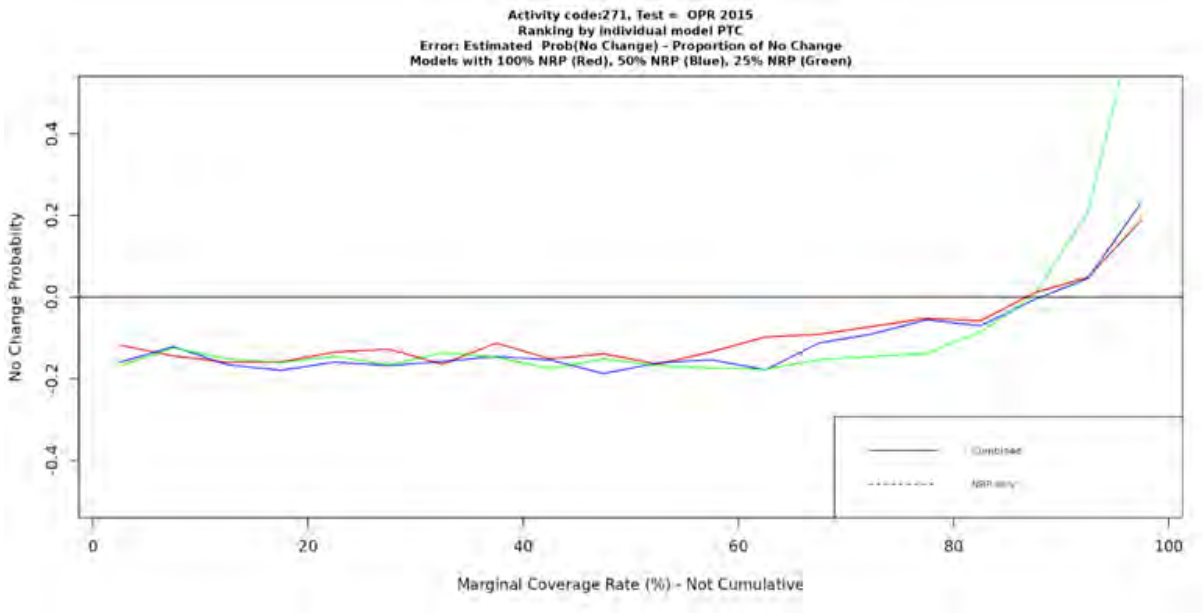


FIGURE B5. AC 271 Error in Tax Change Predictions on NRP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

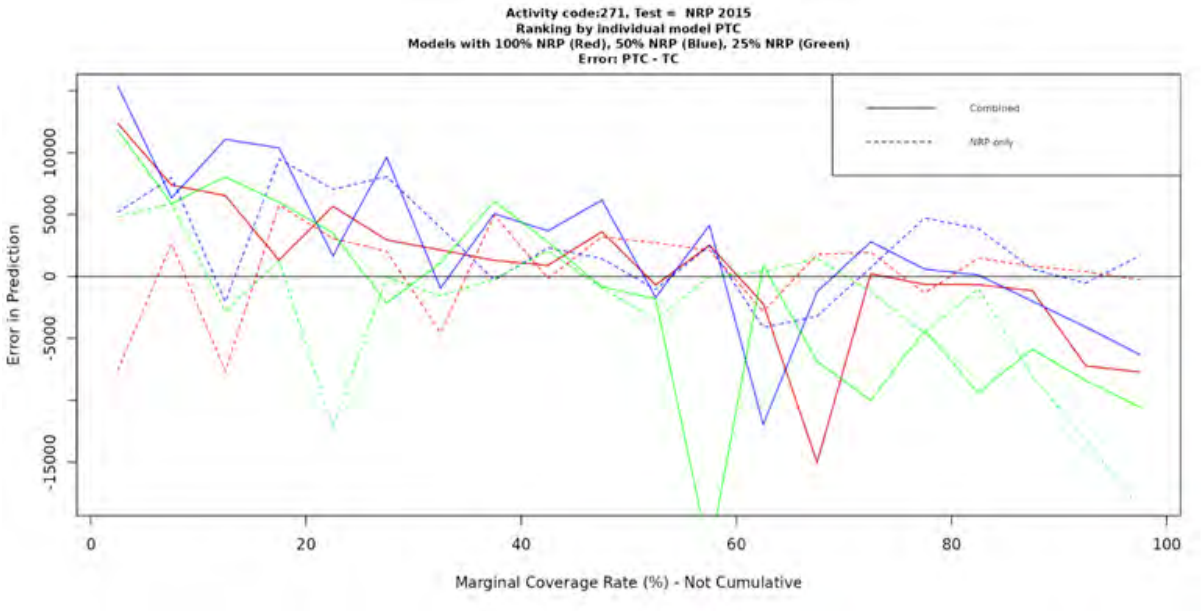
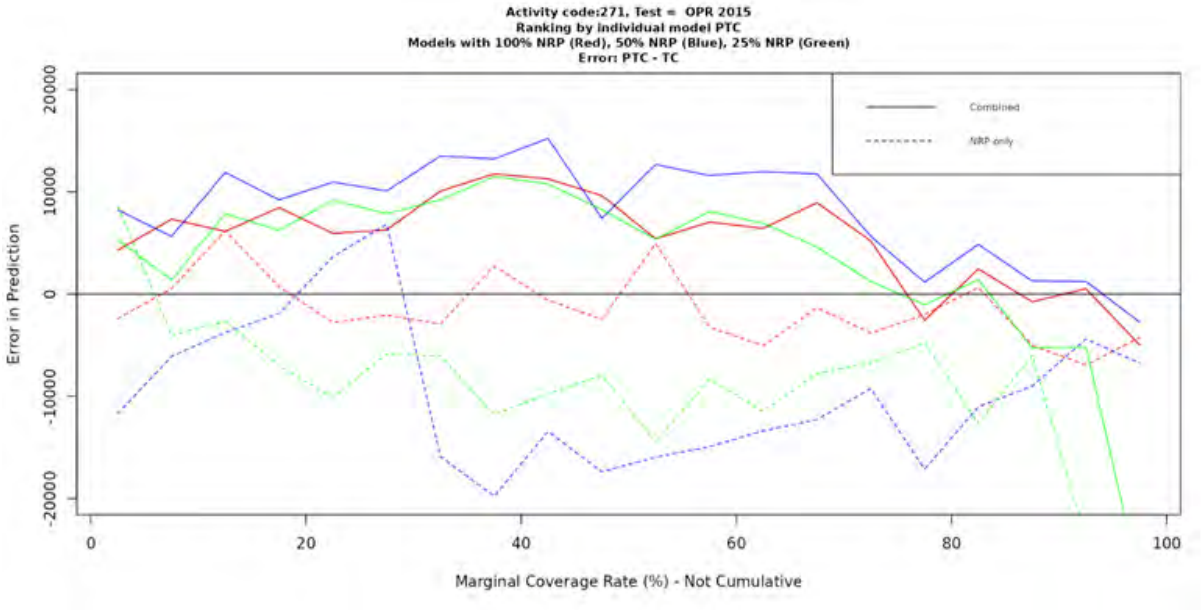


FIGURE B6. AC 271 Error in Tax Change Predictions on OP Data for Models Trained With Different Proportions of NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)



Next, we present the plots for the combined model when the OP data are used in both stages of the risk model. The first four plots are for the estimated no change probabilities and the predicted tax change (for NRP and OP test data) and the next four plots are for the error in estimation of the no change probability and the

error in prediction of tax change (for NRP and OP test data). Each plot shows the results for all the models trained with different proportions of the NRP data.

FIGURE B7. AC 271 Probability of No Change on NRP Data, Ranked by Predicted Probability of No Change Based on the Combined Model

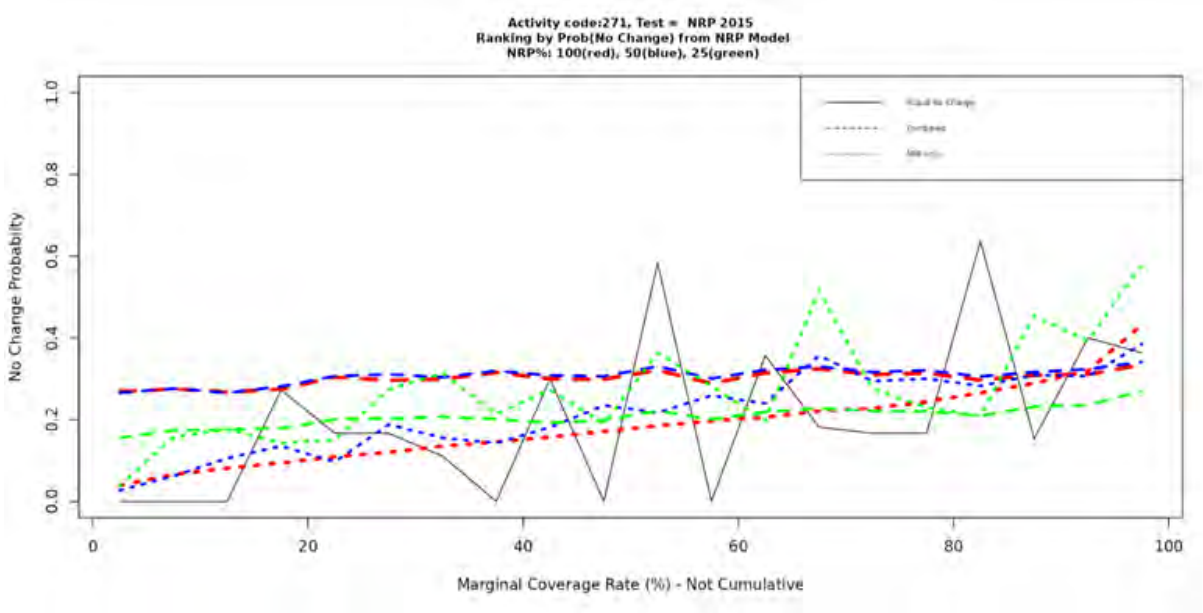


FIGURE B8. AC 271 Probability of No Change on OP Data, Ranked by Predicted Probability of No Change Based on the Combined Model

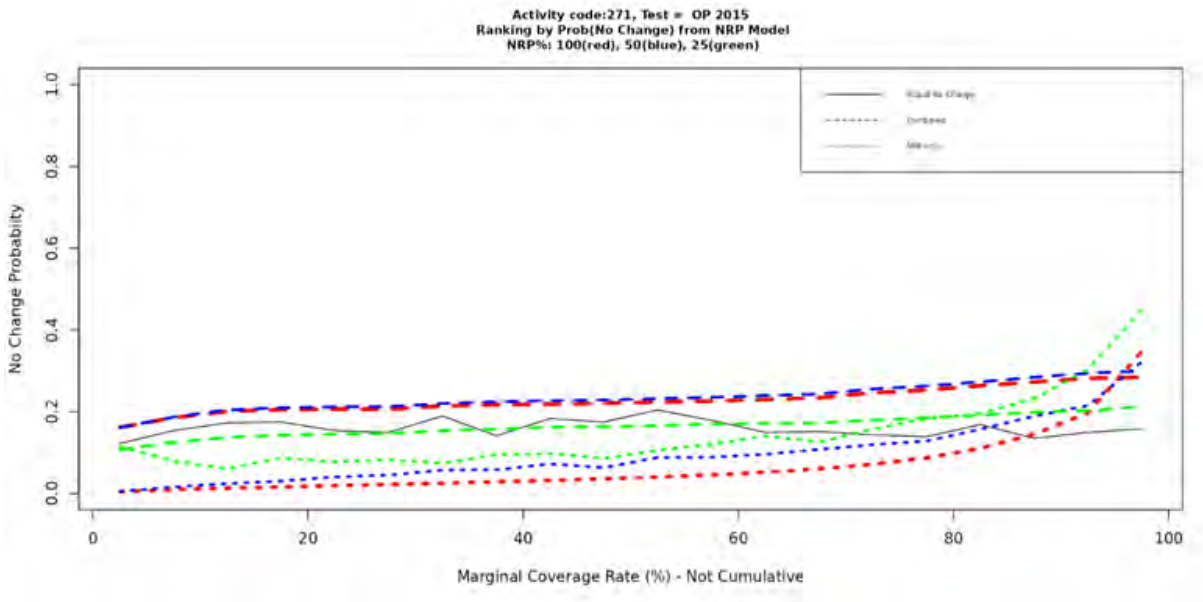


FIGURE B9. AC 271 Tax Change Predictions on NRP Data, Ranked by Predicted Tax Change Based on the Combined Model

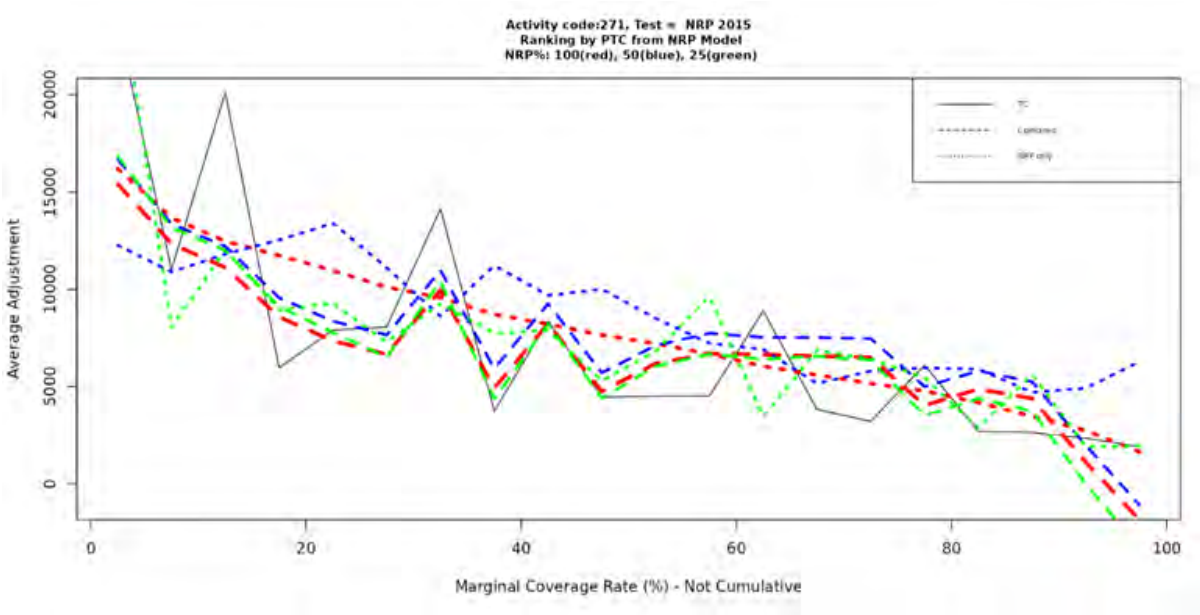


FIGURE B10. AC 271 Tax Change Predictions on OP Data, Ranked by Predicted Tax Change Based on the Combined Model

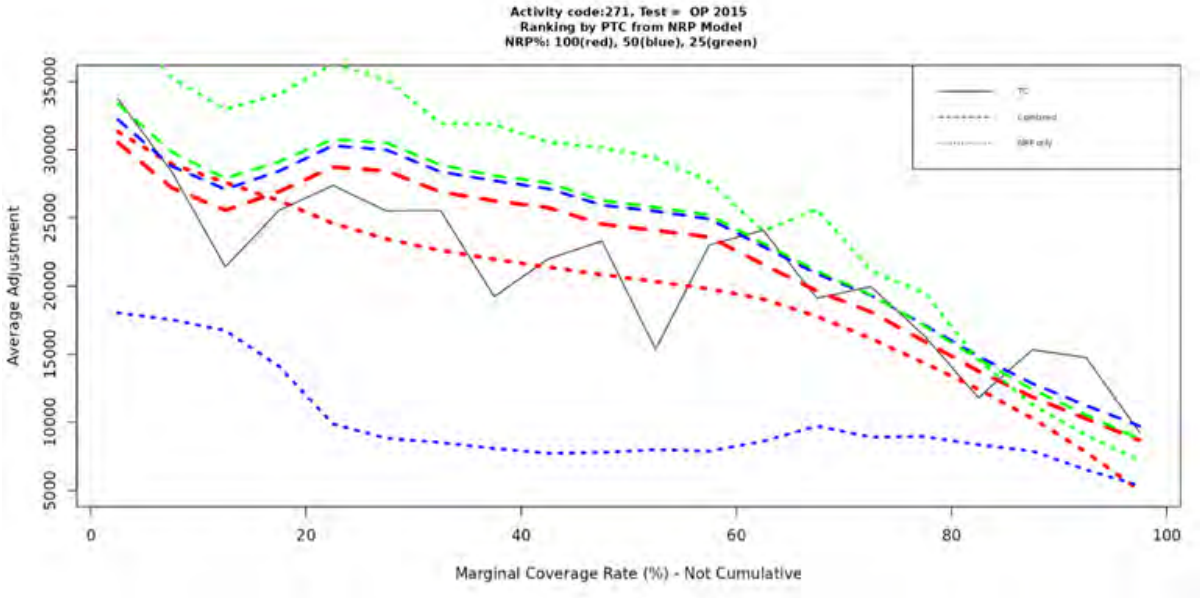


FIGURE B11. AC 271 Error in No Change Probability on NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

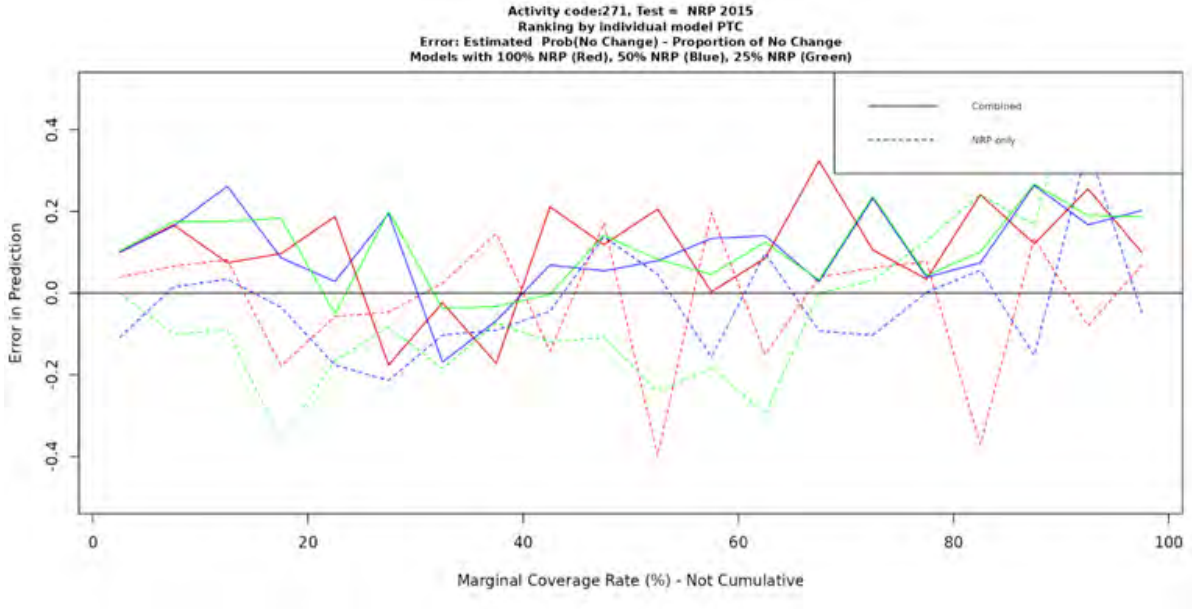


FIGURE B12. AC 271 No Change Probability on OP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

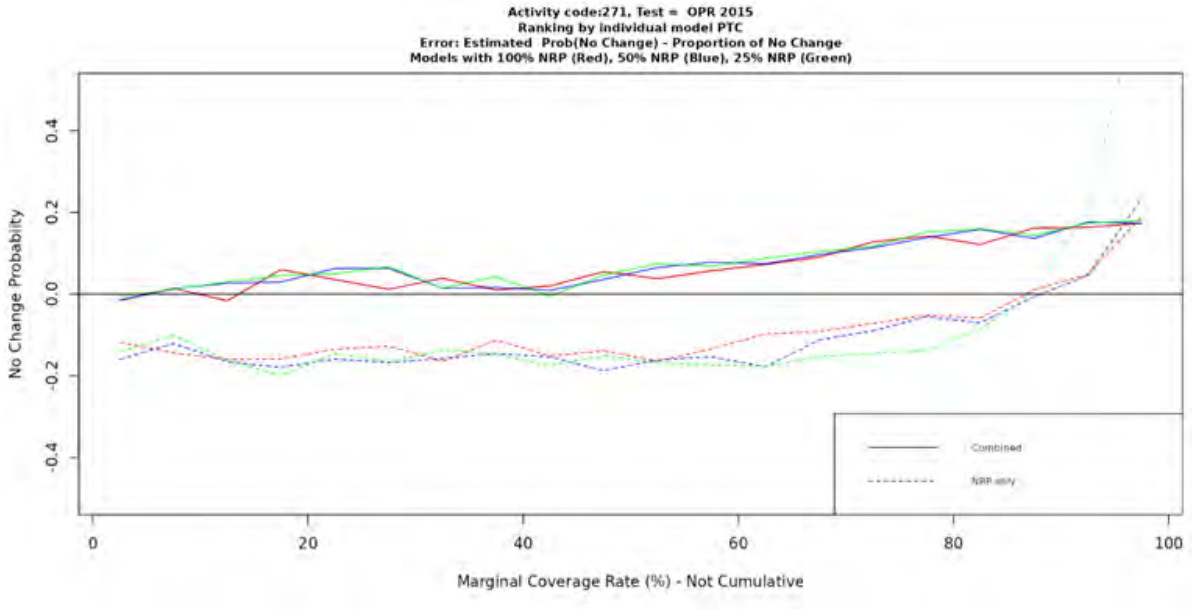


FIGURE B13. AC 271 Error in Tax Change Predictions on NRP Data (Cases Are Ranked by Probability Estimates From Respective Models.)

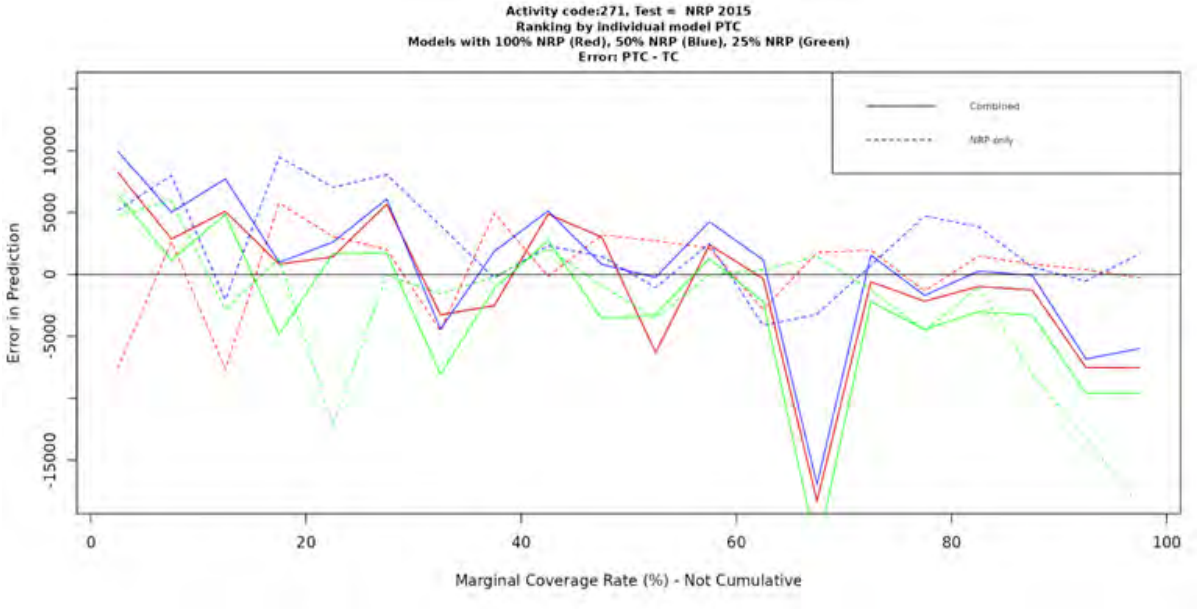
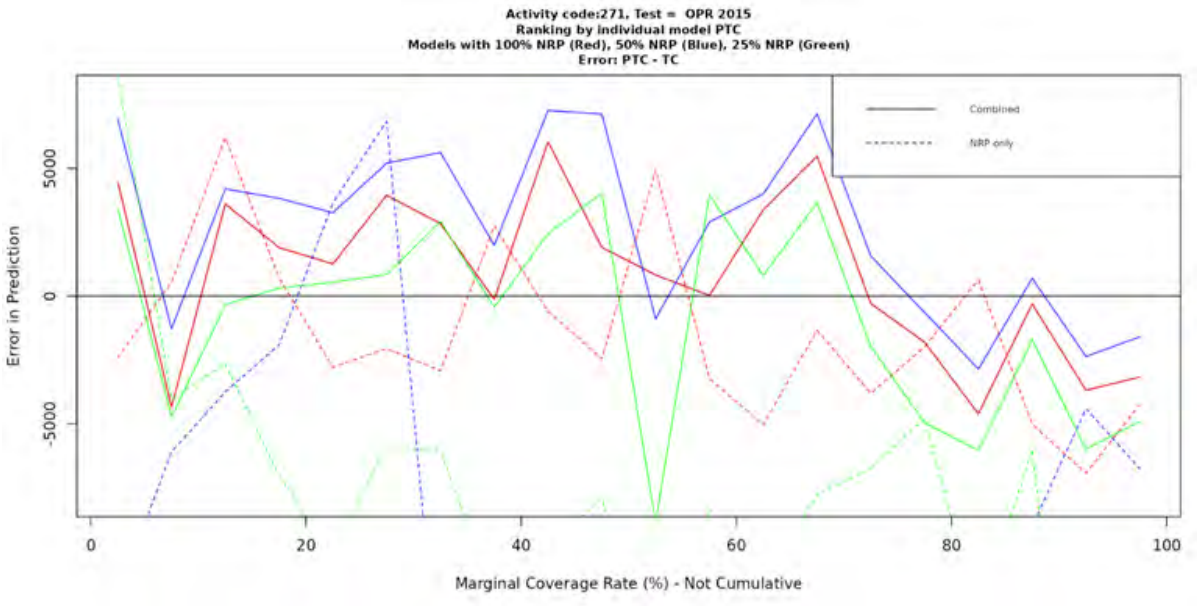


FIGURE B14. AC 271 Error in Tax Change Predictions on OP Data (Cases Are Ranked by Probability Estimates From Respective Models.)



Augmenting National Research Program Tax Change Estimates by Incorporating Operational Audit Information: A New RAAS Research Initiative

Lou Rizzo, John Riddles, Xiaoshu Zhu, Richard Valliant (Westat), and Kimberly Henry (IRS, RAAS)

Introduction

The National Research Program (NRP) is a long-standing initiative of the Internal Revenue Service (IRS) that conducts audits on a probability sample of tax returns. NRP's individual probability sample and comprehensive audit process results in an average tax change estimate that has a designated level of sampling error and limited sources of non-sampling error (measurement error, selection effects).

In recent years, reductions in the NRP sample sizes have resulted in a reduction in the precision of tax change estimators. This research initiative studies the possibility of leveraging a large source of auxiliary information in operational (OP) audits conducted each year within normal IRS tax compliance assurance operations. This source of data is several orders of magnitude larger in size than the NRP samples each year, but its value as a supplemental estimator of tax change is questionable, without careful adjustment. The greatest adjustment required when combining these returns with NRP cases needs to address the fact that the OP tax returns are not selected via a probability mechanism, but according to multi-objective, IRS operational procedures.

The current NRP is a stratified random sample drawn from the relevant tax return universe. These sampled returns are then audited. Continuing reductions in resources have led to reductions in NRP sample sizes over time. The challenge is to maintain the goals of NRP, including but not limited to tax gap estimation, in this difficult cost environment.

One partial solution to this is to use some of the IRS's OP audits. Major challenges of combining OP and NRP audits are that the operational audits are not audited like NRP cases and the operational audits are not performed on a random sample from the tax return population. Simple design-based methods cannot be used to provide national-level estimates from OP audits, in contrast with NRP audits.

For purposes of this paper, the Tax Years (TY) 2013–2015 NRP dataset was augmented with operational audits. The OP records were incorporated in estimates by creating a propensity score. This pseudo-probability represents the probability the record has of being audited based on a model rather than a known, explicit sample design. This is not a true probability but treated as such in estimation. For example, see Elliott and Valliant (2017), Valliant (2020), and Chen *et al.* (2019).

We carried out a microanalysis of the differences between the NRP audits and the operational-only audits. We matched the two sets of audits by placing them in a common set of cells determined by subpartitioning each NRP substratum, using the tax return information used for NRP strata definitions. Within each of these cells, the NRP and operational audits have the same values of tax return classification items for the subpartitioning fields. Thus, conditional on the subpartitioning fields and assuming no measurement or selection difference, the two sets of audits should be comparable and any difference then can be viewed as a sign of measurement error[1] or selection effects.

[1] As the focus of this research is on selection bias adjustments, measurement error is not explicitly adjusted for. However, the proposed composite estimators may account for measurement error between the data sources.

For estimation weights, we assign the NRP sampling weights to the NRP component and new weights to the OP component by analyzing the propensity of being OP, conditional on being in the NRP-OP dataset. The NRP-OP data together is assumed to represent the full population of all tax returns, after weighting the NRP observations with their sampling weight. This propensity of being an operational audit was estimated using a machine learning program. Recent references for comparing differing machine learning algorithms to estimate probabilities of inclusion in the particular context of online surveys include Ferri-Garcia and Rueda (2020) and Castro-Martin *et al.* (2020).

We develop one estimator, which treats the operational audits as self-representing in the tax return universe for a given year. Being absent of selection bias, the NRP sample covers the entire Form 1040 tax filing population with its probability-based sample weights. The OP returns are self-representing and given a weight of 1. These operational audits overlap with the NRP universe, so to avoid double-counting the portion of the universe that the OP audits represent, we use random forest modeling to compute the propensity a given return in the NRP sample would have been operationally audited. We then adjust the NRP sample weights by one minus this estimated propensity. Tax return items serve as model predictors. These estimators are conservative, making very weak assumptions, but are not very different in properties with the simple, unadjusted NRP-based weighted estimates.

We develop composite estimators that linearly combine NRP estimators and operational return estimators (unweighted), where the linear factor is based on relative precision. This is only allowable in fully-equated strata, in which we confirm the NRP and the OP distributions align based on a statistical test. We use a Kolmogorov-Smirnov (K-S) test of the equality of the distributions, as well as t-tests for means, standard deviations, and selected percentiles to confirm full-equating status for a number of NRP strata. When confirmed, we assume the OP estimators are unbiased, as well as the NRP estimators, and generate the unbiased composite estimates based on relative variances alone.

Another estimator we develop is a sliding scale composite estimator, which combines the NRP and OP estimators for a stratum, whether or not full equating holds. In contrast to the basic composite estimator, the sliding scale sets composite weights by MSE rather than variance and assumes the unequated OP estimator may be biased. The squared bias is estimated by taking the squared difference of NRP and OP estimates minus the variance.

To gauge whether or not the NRP and OP audit data are comparable enough to combine their information in composition, K-S tests were applied on the distributions of tax change from each audit data source. A small set of fully equated substrata, identified within this research, proceeded directly to composite estimation with results provided. Partial equating can lead to a form of composite estimation (composite estimation within the partially equated domain and use of the NRP records for the complement domain). Partial equating is possible for about a third of the substrata, also summarized and presented here.

Where full equating cannot be justified, in some cases, we identify partial equating, in which NRP and OP distributions pass the K-S tests of distributional equality within a stratum's subdomain. Then composite estimation can be applied within these subdomains, with the NRP sample providing the estimator for the subdomain complement within the stratum. These subdomains are based on the use of Discriminant Function (DIF) scores, which are model-based values assigned to each tax return representing (based on the model) how strongly an operational audit is warranted. The subdomain may be all NRP and OP returns with a DIF value greater than an operationally assigned DIF cutoff (which varies in a complex way across particular return domains).

One preliminary conclusion that we reached was that low-tax-change tax returns did not tend to appear in the operational audit dataset. The tax change at low percentiles, such as the 5th and 10th, tended to be higher for the OP as compared to NRP. This indicates that the OP audits are missing part of the universe at the low-end of tax change, as to be expected under the OP audits' risk-based selection. However, for analysis, preliminary results indicated that sole use of the propensity-weighted OP audits produced positively biased estimates of tax change. Within particular NRP strata, the OP audits represented a segment of the audited population in which propensity weighting adjustments alone were not sufficient to remove the biases.

To overcome this, additional operational audit case selections called surveyed discretionary (SD) cases were incorporated into the analysis via an auxiliary training dataset. These cases were tentatively designated for operational audits but were identified as most likely having little tax change potential during a pre-audit assessment, which led to these cases being set further back in the OP audit queue. This set of SD cases sheds light on a subset of the tax return universe not covered by operational audits.

As no audit was conducted for these cases, we do not know the tax change value for these returns. We assume that had they been audited, a portion might have non-zero tax changes. Our approach was to impute tax-change values from the NRP cases in the same NRP substratum from the discrete distribution bounded above by the 5th (weighted) percentile of the NRP tax-change values in the substratum (including tax-change zeroes). Drawing randomly with-replacement from this discrete NRP tax-change distribution provided a set of imputed values, including zeroes and small positive values (up to the 5th percentile for NRP within the substratum). Imputed SD cases of zero tax change are set aside, and SD cases imputed with positive tax change values are included with the positive operational audits, augmenting this operational audit set. The operational audit tax-change values augmented by these positive imputed-SD cases is called the SD-augmented OP distribution.

In addition to identifying fully-equated and partially-equated NRP substrata, we study these proposed composite estimators over the 2013–2015 NRP strata. A small set of fully-equated substrata proceeded directly to composite estimation, and the results from this composite estimation are provided, along with further results for partially-equating substrata. Partial equating can lead to a form of composite estimation within the partially equated domain and using NRP records for the complement domain. Partial equating is possible for about one-third of the substrata. Alternative estimators for no-equating strata were also developed and compared. Our analysis demonstrates there are NRP strata and substrata where the tax change is comparable enough to combine operational audits with NRP audits, but it has to be done carefully and conservatively. Our initial results show this combination cannot be done blindly; without adjustment for potential return selection biases, our estimates of tax change will be positive biased. This is not surprising, since OP audits are selected with filters to select noncompliant returns but must be accounted for making estimations to the general filing population.

Data and Methods

Data

In this study, we created a joint NRP-OP dataset: the current NRP data set augmented with operational audits. The OP records will be incorporated in estimates by creating a propensity score. This pseudo-probability, representing the probability the record has of being audited, is based on a model rather than a known explicit sample design.

NRP Data

Initial work applying the propensity model and evaluating results for TY 2015 demonstrated that the NRP sample size for that year are small in some substrata. This led to larger standard errors and more variation, so, we used a three-year dataset. We mixed TYs 2013–2015, so that we do not have a clean estimator of a single tax year, but mixed NRP and OP data to make equivalent comparisons between the NRP and OP data. The result of this is we have a three times larger (in general) NRP sample size.

OP Audit Data

The operational audit (OP) dataset is very large,[2] so we drew a subsample of the three-year OP set. We do not need the full set, given OP one-year sample sizes are generally quite large, much larger than the NRP sample sizes even after subsampling. If we subsample randomly, then we have unbiased estimates of the full universe. The goal in subsampling the OPs was to reduce their sample size closer to the number of returns in

[2] It is about 2.4 million records before one-third sampling. This is not in itself an enormously large dataset, but the time to run complex, computing resource-intensive programs in the context of a research project makes it uncomfortable. Scaling up to the full data set should not be a difficult process during production, where the number of different programs is much less.

each substratum obtained in one year's sample. We randomly sampled one third of the three-year dataset TYs 2013–2015 for major strata 270 through 281, and randomly sampled one-half of the two TY OP data sets 2014 and 2015 for the Other major stratum. Major stratum Other did not exist until TY 2014. NRP returns were not subsampled in any year. The overall dataset can be seen as a weighted average of the three tax years together (each tax year receiving weight close to one-third[3]). The sample sizes by major strata before and after subsampling are presented in Table 1.

TABLE 1. NRP and OP Sample Sizes by Major Strata

Strata	NRP Sample Size	OP Sample Size After Subsampling	OP Sample Size Before Subsampling	OP Subsampling Rate
270	7,002	399,236	1,197,618	33.3%
271	541	17,962	53,844	33.4%
272	6,367	129,320	387,920	33.3%
273	3,180	110,161	330,460	33.3%
274	3,068	80,416	241,223	33.3%
275	2,735	21,460	64,350	33.3%
276	2,479	10,813	32,399	33.4%
277	2,220	11,031	33,063	33.4%
278	2,663	4,546	13,620	33.4%
279	4,231	30,072	90,179	33.3%
280	2,851	28,443	85,316	33.3%
281	3,903	17,747	53,220	33.3%
Other	1,361	62,274	124,510	50.0%
Total	42,601	923,481	2,707,722	

Zero tax-change and negative tax-change returns also exist. The NRP sample, being a completely representative sample from the tax universe, has significant percentages of both types of returns.

Table 2 below presents the weighted percentages of positive, zero, and negative tax changes at the major stratum level (using the NRP totals for each substratum as weights across the substrata), for the NRP sample and for the SD-augmented operational audits. Note that this includes the imputed values for the SD operational audits.

Overall, the NRP percentage of negative tax-change returns is 7.5 percent, whereas the OP percentage is 4.4 percent. This difference is highly significant (t-statistic -13.9). The NRP percentage of zero tax-change returns is 43.8 percent.

We estimated the OP percentage of zero tax-change returns in two ways. Our first estimate uses our best imputations for the SD returns. This estimate is 33.3 percent and significantly different from the NRP percentage of zero-tax change returns (a t-statistic of -26.4). Our imputations are only imputations, so we also generated an estimate of OP percentage of zero tax-change returns based on assuming all SD returns are zero (resulting in an upper-bound for the percentage of zero tax-change returns). This estimate is 41.7 percent. This is still smaller than the NRP percentage at 43.8 percent, but here the t-statistic is only -5.4. Including the SD imputations brings the augmented dataset's coverage closer to that of the NRP sample.

[3] The share of each tax year in the full dataset is proportional to the total number of returns that tax year (which does not vary much across the three tax years).

TABLE 2. NRP and SD-Augmented OP Percentages of Positive, Zero, and Negative Tax Changes

Major Stratum	NRP Neg	NRP Zero	NRP Pos	OP Neg	OP Zero	OP Pos
270	6.2%	38.9%	54.9%	2.7%	32.9%	64.4%
271	13.8%	6.9%	79.2%	7.2%	43.2%	49.7%
272	8.2%	57.1%	34.7%	5.5%	42.6%	51.9%
273	9.0%	25.1%	66.0%	3.2%	27.5%	69.4%
274	9.1%	21.6%	69.5%	2.8%	36.0%	61.2%
275	7.5%	13.0%	79.5%	9.0%	39.8%	51.2%
276	9.0%	12.7%	78.3%	12.8%	36.8%	50.4%
277	7.8%	10.8%	81.4%	15.0%	40.8%	44.2%
278	10.9%	24.5%	64.6%	11.4%	55.9%	32.7%
279	13.9%	28.9%	57.2%	14.0%	57.1%	28.9%
280	13.6%	16.3%	70.2%	13.9%	52.6%	33.6%
281	18.1%	28.3%	53.7%	19.9%	59.2%	20.9%
Other	7.8%	23.9%	68.3%	6.9%	43.3%	49.8%

Dependent Variable

Our dependent variable is the average change in income tax, defined as the difference in computed income taxes reported on Form 1040 and the associated tax amount determined within an audit. Our analysis compares the difference between mean values of the tax change from the NRP sample of audits and the tax change from the operational audits. Results from the major strata, defined by Exam Activity Code, and graphical results at the substratum level, are presented.

Statistical Analysis

Propensity Score

To estimate propensity scores, we used a machine learning algorithm CFOREST in R. Random forests were first developed in Breiman (2001), and are an elaboration of trees, which are cell structures generated to be homogeneous in propensity within the cells, and heterogeneous in propensity across the cells. See also Hastie *et al.* (2009). Trees tend to be very sensitive to slight perturbations in the data, our motivation for using forests. A random forest is a large number of trees based on resampling from the original dataset (and sampling from the predictors as well). The forest estimate is based on a mean value of propensities across the trees, generating a more stable estimate.

Random forests, in their original form, sometimes are subject to overfitting. So, we used a modification of random forests implemented in the R package CFOREST which reduces the size of the constituent trees (e.g., Hothorn *et al.* (2006) and Hothorn and Zeileis (2015)).

The underlying dataset for running CFOREST was developed based on the procedure defining clusters $c = 1, \ldots, C$ of NRP sampling substrata using ordered NRP sampling rates. We draw operational records within these clusters, using the mean NRP sampling rate, generated a cluster dataset with roughly the correct population proportions, and ran the CFOREST program to generate propensity estimates for the cluster. This drawing procedure was bootstrapped to provide greater stability.

Table 3 below presents the distribution of operational propensity estimates by major NRP strata. The OP propensities show a substantial amount of variation within some of the major strata. Stratum 272 has the largest coefficient of variation (CV = 330 percent) of the propensities, while stratum 277 at CV = 55 percent has

the smallest. This contrasts to the selection probabilities in NRP, which by design have limited or no variation within the NRP strata prior to nonresponse adjustment.

TABLE 3. Distribution of Operational Propensity Estimates, by Major Stratum

Major Stratum	Return Total Count	Mean of Predicted Prop.	Std Dev of Predicted Prop.	CV of Predicted Prop.	Min. of Predicted Prop.	1st Qrt of Predicted Prop.	Median of Predicted Prop	3rd Qrt of Predicted Prop.	95th Perc of Pred. Prop.	Max. of Predicted Prop.
270	23,898,238	1.4%	1.8%	131%	0.0%	0.2%	0.7%	1.9%	4.3%	22.2%
271	1,334,259	1.5%	1.4%	91%	0.2%	0.4%	0.8%	2. 5%	4.1%	16.7%
272	76,187,820	0.1%	0.3%	330%	0.0%	0.1%	0.0%	0.1%	0.4%	6.8%
273	14,461,352	0.6%	1.3%	213%	0.02%	0.1%	0.2%	0.6%	2.7%	14.7%
274	9,827,635	0.7%	1.6%	238%	0.01%	0.1%	0.1%	0.4%	3.1%	18.6%
275	2,680,147	0.5%	0.6%	118%	0.0%	0.1%	0.3%	0.6%	1.7%	7.8%
276	739,791	0.9%	0.6%	71%	0.1%	0.4%	0.7%	1.3%	2.1%	6.8%
277	574,019	1.5%	0.8%	55%	0.3%	0.9%	1.3%	1.9%	3.0%	8.3%
278	1,181,166	0.3%	0.3%	76%	0.1%	0.2%	0.3%	0.4%	0.8%	3.7%
279	4,724,458	0.4%	0.5%	124%	0.0%	0.1%	0.2%	0.4%	1.9%	4.9%
280	1,957,898	1.1%	1.1%	101%	0.1%	0.5%	0.8%	1.3%	3.3%	16.5%
281	505,216	2.0%	2.0%	101%	0.2%	0.8%	1.3%	2.3%	5.8%	23.1%
Other	7,511,262	0.7%	0.9%	127%	0.1%	0.2%	0.4%	0.8%	2.7%	21.8%

NOTES: Pred. Prop. = predicted propensity; Qrt = quartile.

Composite Estimation for Tax Change

This section presents the alternative estimators compared in our analysis. Simple composite estimators were developed first and hold when the OP and NRP tax change is comparable within an NRP stratum or substratum. More complex composite estimators are introduced when full equality does not hold.

The following theory is presented for combining the NRP and SD-augmented OP cases in the estimation of tax change. Define h as the substratum indicator (e.g., one of the 87 substrata within the 13 major strata for the NRP). We define:

T_h = total of some outcome variable y in substratum h;
$\hat{t}_{NRP}(G)$= estimate of total for a general domain G;
U_h = universe of returns in substratum h;
A_h = subset of U_h that SD-augmented OP represent; if SD-augmented OP cases only represent themselves, then A_h is the set of sample OP cases itself within the subdomain.

Equating NRP and OP Values Within an NRP Stratum

When A_h is all of U_h (i.e., NRP and SD-augmented OP audits equate completely) is called the case of full equating, and we can proceed directly to composite estimation combining the NRP and SD-augmented OP estimators based on variance considerations alone. The case where A_h is a proper subset of U_h is the case of partial equating, in which we do only composite estimation (between NRP and SD-augmented OP) over A_h, and use only NRP values for $U_h - A_h$.

Substrata Extensions to the Composite Estimator

Substratum Composite Estimator for Fully-Equated Strata
The forms of the fully-equated substratum composite estimator for the total and the mean (tax change in our application, but any variable applies) respectively for substratum h are

$$\hat{T}_h = \alpha_h \, \hat{t}_{NRP}(U_h) \;+\; (1-\alpha_h)\hat{t}_{OP}(U_h)$$
$$\hat{\bar{T}}_h = \alpha_h \, \hat{\bar{t}}_{NRP}(U_h) \;+\; (1-\alpha_h)\hat{\bar{t}}_{OP}(U_h)$$

$\hat{t}_{NRP}(U_h)$ $\left(\hat{\bar{t}}_{NRP}(U_h)\right)$ is the NRP estimator for the total (mean) in substratum universe U_h, $\hat{t}_{OP}(U_h)$ $\left(\hat{\bar{t}}_{OP}(U_h)\right)$ is the OP estimator for the total (mean) for substratum universe U_h, and α_h is a weight between 0 and 1.

Ideally, α_h would be proportional to the inverse of the variance of $\hat{t}_{NRP}(U_h)$. In this case, α_h is proportional to the size of the NRP sample substratum relative to the size of the OP sample. We computed α_h as based on the NRP sample size n_h^{NRP} divided by the sum of the NRP and OP sample sizes $n_h^{NRP} + n_h^{OP}$ within stratum h. Thus, our factor is proportional to the inverse of the relative variance, with the additional assumption that the variance from a simple random sample is appropriate (i.e., no design effects). As we are operating within a substratum where the NRP weights are generally equal, and treating the OP units as from an equal probability sample, this is a reasonable approach. Because the estimated population number of returns in a substratum is a constant in all equal probability samples, the factor α_h is also the same for totals and means.

The composite estimator is unbiased with respect to random sampling in NRP and pseudo-random sampling in OP. It would also be model-unbiased under this superpopulation model:

$$E_M\left(y_{hi}\right) = \begin{cases} \mu_{1h} & i \in U_h - A_h \\ \mu_{2h} & i \in A_h \end{cases}$$

where EM denotes expectation with respect to the superpopulation model and μ_{1h}, μ_{2h} are the NRP and OP superpopulation means within each substratum. The model posits that the mean tax change differs in $U_h - A_h$ and A_h and that returns within each of those subsets of substratum h have a common mean. Unbiasedness in either sense depends on being able to correctly separate the returns in a substratum into $U_h - A_h$ and A_h. Our results, using a distributional statistical test, identify NRP strata and substrata where this holds.

Substratum Composite Estimator for Partially-Equated Strata
Within the designated domain of partial equating A_h, we do composite estimation to obtain final tax-change estimates for the substratum. The composite estimator for substratum h is

$$\hat{T}_h = \hat{t}_{NRP}(U_h - A_h) \;+\; \alpha_h \, \hat{t}_{NRP}(A_h) \;+\; (1-\alpha_h)\hat{t}_{OP}(A_h)$$

We compute α_h based on the effective NRP sample size for domain A_h, $\tilde{n}_h^{NRP}(A_h)$, divided by the sum of the effective NRP and the OP sample sizes $\tilde{n}_h^{NRP}(A_h) + n_h^{OP}(A_h)$ within domain A_h. The effective sample size $\tilde{n}_h^{NRP}(A_h)$is the nominal sample size $n_h^{NRP}(A_h)$ divided by the Kish design effect (see Kish (1992)) for unequal weights (one plus the coefficient of variation of the weights squared) over the weights in A_h.

Estimators for Augmented Self-Representing OP Audits
In this section, we discuss NRP estimators augmented by self-representing OP audits (i.e., each OP audit receives a weight of 1). Here, the SD-OP returns represent only themselves and are not weighted-up. These are the most conservative estimators, making minimal assumptions about the nature of the OP audits, but provide only a small improvement beyond the simple weighted NRP estimators.

These estimators and the sliding scale composite estimators previously described are probably the only options to improve on NRP estimators in substrata where equating between NRP and OP proves impossible. We computed these estimators for all the substrata in this section, including those which are dealt with in preceding sections, to provide a full picture.

In this estimation approach, we make the OP returns a self-representing set of audits (all weights being 1), calling this set $OPSR_h$. The estimate from this self-representing domain is $\hat{t}_{OP}(OPSR_h)$, where all operational audits have an equal weight. We need to find a space for this domain within the NRP estimate, which covers the entire universe including $OPSR_h$. One technique is to find an operational complement for each NRP return one-by-one, setting OP-complement weights for the NRP returns equal to their NRP sample weights multiplied by $1-\hat{\pi}_{hi}^{(OP)}$, where $\hat{\pi}_{hi}^{(OP)}$ is the estimated propensity of NRP return hi being an operational audit (estimated by the propensity model).

For example, if a particular NRP audit has almost no chance of being an operational audit, we bring it in almost with its full weight. If, on the other hand, it has a 10 percent chance of being an operational audit, then we downweight it in the NRP estimate by 10 percent. In an extreme case, suppose all tax returns with tax changes greater than a given threshold have a 100 percent chance of being operationally audited. These would be certainly in the operational set and, under this approach, would be dropped from the NRP sample (one minus their propensity of being audited would in this case be zero). The idea is to prevent double-counting in this probabilistic sense of any returns, i.e., any NRP returns with a high propensity of being operational are systematically and accordingly downweighted.

If we write $s_h(NRP)$ as the NRP sample within substratum h, with NRP sample weights $w_{hi}^{(NRP)}$ and tax change value y_{hi}, then our operational-complement-adjusted NRP estimator of the total and mean respectively are as follows:

$$\hat{t}_{NRP}(U_h - OPSR_h) = \sum_{s_h(NRP)} \left\{ w_{hi}^{(NRP)} \left(1 - \hat{\pi}_{hi}^{(OP)}\right) y_{hi} \right\}.$$

$$\bar{t}_{NRP}(U_h - OPSR_h) = \frac{\sum_{s_h(NRP)} \left\{ w_{hi}^{(NRP)} \left(1-\hat{\pi}_{hi}^{(OP)}\right) y_{hi} \right\}}{\sum_{s_h(NRP)} \left\{ w_{hi}^{(NRP)} \left(1-\hat{\pi}_{hi}^{(OP)}\right) \right\}}.$$

The form of the overall estimator for substratum h is

$$\hat{T}_h = \hat{t}_{NRP}(U_h - OPSR_h) + \hat{t}_{OP}(OPSR_h).$$

To better understand this estimator, suppose we consider its form if the NRP was an equal probability sample within each stratum, rather than a stratified estimator with varying stratum sampling rates. Then, in this special case $w_{hi}^{(NRP)}$ would be $N_h/n_h(NRP)$ where N_h is the population size and $n_h(NRP)$ is the NRP sample size.

Viewing the OP sample as an equal probability sample, we have $\hat{\pi}_{hi}^{(OP)} = n_h(OP)/N_h$, and in this special case then $w_{hi}^{(NRP)}\left(1-\hat{\pi}_{hi}^{(OP)}\right) = (N_h - n_h(OP))/n_h(NRP)$, so that each NRP case equally represents the tax return population minus the OP cases.

Sliding Scale Estimator

Scale composite estimators are an alternative to those presented in the preceding sections. We dynamically combine the NRP and OP samples within substrata, giving more or less weight to the OP cases, depending on whether the OP sample accurately represents an entire substratum or not.

Define the following: $\bar{T}_h$= population mean in substratum h, $\hat{\bar{t}}_{NRP}(U_h) =$ estimator of the mean in h from NRP, and $\hat{\bar{t}}_{OP}(U_h)$ estimator of the mean in h from OP. We define the composite estimator of the mean in substratum h as we did earlier in the fully equated case: a linear combination of the NRP estimator and the OP estimator, with α_h as the linear share of the NRP estimator and $(1-\alpha_h)$ as the share of the OP estimator.

What is different here is that we develop this universe composite estimator whether or not equating has been confirmed. The OP estimator is treated as an equal whether or not it is equated to the NRP estimator. The sliding scale estimator works by generating the α_h inversely proportional to the estimated MSEs of the two estimators, rather than to the variances. The NRP estimator is treated as unbiased. Since equating is not presupposed here, the OP estimator is not assumed to be unbiased, so the measured MSE includes squared-bias and variance terms.

In terms of formulas then the composite estimator of $\bar{T}_h$ is

$$\hat{\bar{T}}_h = \alpha_h \hat{\bar{t}}_{NRP}(U_h) + (1-\alpha_h)\hat{\bar{t}}_{OP}(U_h), \text{ where}$$

$$\alpha_h = \frac{1/mse\left(\hat{\bar{t}}_{NRP}(U_h)\right)}{1/mse\left(\hat{\bar{t}}_{NRP}(U_h)\right) + 1/mse\left(\hat{\bar{t}}_{OP}(U_h)\right)}$$

$$1-\alpha_h = \frac{1/mse\left(\hat{\bar{t}}_{OP}(U_h)\right)}{1/mse\left(\hat{\bar{t}}_{NRP}(U_h)\right) + 1/mse\left(\hat{\bar{t}}_{OP}(U_h)\right)}$$

Since the NRP estimator is unbiased, its bias is zero and its mean squared error (MSE) is equal to its variance. The OP estimator may be biased and the MSE is a summation of squared bias and variance. We do not provide details here as to how the MSE is estimated for the OP estimator, but in brief, we estimate it by taking a one-degree-of-freedom squared bias (the square of the difference between biased mean and the unbiased mean) and subtract out an appropriate variance component estimate (as the one-degree-of-freedom squared bias estimate contains the real squared bias intermixed with variance components).

Hybrid Estimator

We also consider a new hybrid estimator, defined to be the fully-equated composite estimator for fully-equated substrata, the partially-equated composite estimator for partially-equated substrata, and the NRP estimator for no-equated substrata.

Results

Tax Change Estimates

Table 4 presents the estimated mean for the tax change using the NRP records alone (using NRP sample weights with no further adjustments) for each of the thirteen 2015 NRP major strata. We adjusted the NRP sample weights through poststratification where the post-strata were related to the NRP strata, but with some modifications. This poststratification should help in improving the efficiency of the estimator, as we established that our post-strata cells correlate well with NRP tax change.

The overall TY 2015 NRP weighted mean is $1,217 per tax return, which has a high degree of precision (standard error being only $67; CV = 5.5 percent). Note that this includes zero tax-change returns and negative tax-change returns. The estimated, average amounts of tax change range from a low of $271 for stratum 272 to a high of $15,511 for stratum 277. The averages for most major strata are estimated fairly precisely with nine of thirteen strata means having CVs of 11 percent or less. The estimated mean for stratum 275 is notably imprecise having CV = 46.5 percent (arising from extreme outliers: see Table 4).

TABLE 4. Estimates of Tax Change from NRP Records Alone Using NRP Weights for TY 2015 Major Strata

Major Stratum	Sample Size	NRP Weighted Mean ($)	Std Error of Mean ($)	CV of Mean (%)	Lower Bound 95% CL for Mean	Upper Bound 95% CL for Mean	Sum of Weights (est of total returns)	Total $ Value (billions)
270	2,172	1,771	57	3.2	1,659	1,884	23,553,072	41.7
271	169	7,958	849	10.7	6,294	9,623	1,320,484	10.5
272	2,086	291	23	7.9	246	336	76,085,502	22.1
273	1,041	1,467	117	8.0	1,238	1,697	14,356,577	21.1
274	934	1,288	80	6.2	1,132	1,444	9,743,186	12.5
275	879	7,057	3,278	46.5	632	13,482	2,664,310	18.8
276	727	6,160	479	7.8	5,220	7,099	731,809	4.5
277	636	15,511	2,817	18.2	9,989	21,034	565,964	8.8
278	855	1,883	162	8.6	1,566	2,201	1,178,086	2.2
279	1,339	1,661	207	12.5	1,256	2,067	4,708,198	7.8
280	876	4,139	353	8.5	3,447	4,831	1,937,794	8.0
281	1,193	6,568	1,581	24.1	3,469	9,667	495,136	3.3
Other	835	1,992	216	10.8	1,569	2,415	7,436,308	14.8
All	13,742	1,217	67	5.5	1,087	1,348	144,776,427	176.2

Table 5 presents the weighted distribution of tax change for the thirteen major strata for the TY 2015 NRP sample (using the base weights). For the mean, 25th percentile, the median, and the 75th percentile, these are approximately unbiased estimates of what the distribution would look like if the entire tax return universe were audited using NRP audit methodology.

TABLE 5. Estimates in Dollars of the Distribution of Tax Change from NRP Records Alone, Using NRP Weights, for TY 2015 Major Strata

Major Stratum	NRP Weighted Mean	Weighted 25th Percentile	Weighted Median	Weighted 75th Percentile
270	1,771	0	124	3,073
271	7,958	32	3,904	9,336
272	291	0	0	83
273	1,467	0	247	1,508
274	1,288	0	240	1,365
275	7,057	69	1,322	4,213
276	6,160	140	1,911	6,879
277	15,512	302	3,053	10,804
278	1,883	0	213	1,545
279	1,661	0	39	696
280	4,139	0	514	3,377
281	6,567	0	34	2,587
Other	1,992	0	303	1,578
All	1,217	0	0	639

In Table 6, we present the differences between the NRP-Only Weighted NRP Audit Mean and the OP-Only Weighted Mean using CFOREST Propensity Weights With Poststratification[4] (the last column in Table 4), including the difference of means, an estimated standard error of the difference assuming independence between the two means, a t-statistic (difference divided by standard error), and a two-sided p-value for this t-statistic assuming the normal distribution (of the null hypothesis of no difference).

In general, the OP poststratified (PS) Weighted Mean tracks considerably higher than the NRP Weighted Mean. Overall, the OP PS Weighted Mean is more than two times higher than the NRP Weighted Mean, and this difference is highly significant. Major Stratum 272 shows the greatest difference: the NRP Weighted Mean is only $291 whereas the OP Weighted Mean is $2,288. On the other hand, there are Major Strata 275 and 277, where the NRP Weighted Mean and the OP PS Weighted Mean are very close. In Major Strata 275, the OP PS Weighted Mean is actually smaller than the NRP Weighted Mean, though this difference is not statistically significant (the null hypothesis of no difference is not rejected). Clearly, a selection bias exists in the operational auditing structure, implying that the OP data by themselves cannot be considered another random sample from the tax return universe.

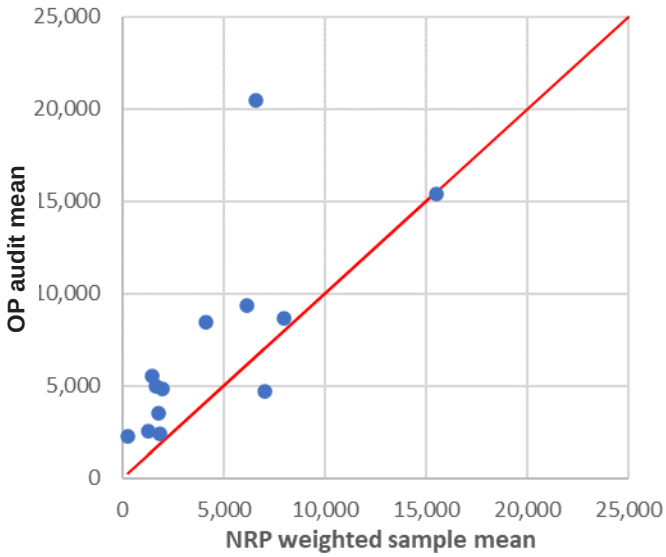
TABLE 6. Differences Between NRP-Only and OP-Only Mean Tax Change with Poststratified Weights, TY 2015 Major Strata

Major Stratum	NRP Weighted Mean ($)	OP Weighted Mean ($)	Difference ($)	t-statistic	Two-sided p-value
270	1,771	3,503	1,732	11.51	0.000
271	7,958	8,689	731	0.73	0.466
272	291	2,288	1,997	5.09	0.000
273	1,467	5,556	4,089	10.20	0.000
274	1,288	2,546	1,258	5.60	0.000
275	7,057	4,700	-2,357	-0.72	0.474
276	6,160	9,344	3,185	1.03	0.302
277	15,511	15,420	-91	-0.03	0.975
278	1,883	2,438	555	2.58	0.010
279	1,661	4,952	3,290	7.93	0.000
280	4,139	8,413	4,274	6.76	0.000
281	6,568	20,474	13,906	3.25	0.001
Other	1,992	4,817	2,825	10.09	0.000
All	1,217	3,315	2,098	9.21	0.000

Figure 1 shows the weighted means with the OP audit-based means on the horizontal axis, the NRP weighted means on the vertical. A 45-degree reference line shows that most of the means are unequal.

[4] This is the inverse of the CFOREST propensity estimate, poststratified to post-strata control totals.

FIGURE 1. Mean Tax Change, Weighted NRP vs. OP Audits, Tax Year TY 2015 Major Strata



Incorporating Zero and Negative Tax-Change Returns.
The NRP sample, being a representative sample from the tax universe, has significant percentages of both zero and negative tax-change of returns. Tables 7 through 9 present the weighted percentages of zero and negative tax changes at the major stratum level (using the NRP totals for each substratum as weights across the substrata), for the NRP sample and for the SD-augmented operational audits.

Table 7 presents the weighted percentages of negative tax changes at the major stratum level. Along with this is a standard error based on taking the square-weighted average of substratum variances for NRP and OP.[5] There is a standard error of the difference between the two percentages (assuming independence), and a t-statistic then for testing the null hypothesis of no difference. In most cases, the OP percentage of negative tax returns is smaller than the corresponding NRP percentage. For the grand total, 7.52 percent of NRP returns had a negative tax change, with 4.43 percent of OP returns having a negative tax return, a significant difference (t-value being -13.94).

The t-statistics for differences in percentages in all 13 major strata are larger in absolute value than 1.96, the usual 0.05 significance level. That is the NRP and SD-augmented OP estimates of percentages of returns with negative tax change are different. This is one piece of evidence that complete equating at the major stratum level will not be possible.

Table 8 presents estimates for the percentages of zero tax changes in each major stratum. Included are the unbiased NRP percentages and using SD-augmented OP percentages using our best imputations of SD values (some imputed as zeroes and some small positive values). This is our best estimate of the OP percentage of zero tax change returns if all SD cases were audited. Table 9 presents a different, related set of estimates for the percentages of zero tax changes. Included again are the unbiased NRP percentages, but here using SD-augmented OP percentages assuming that all SDs are zeroes. This creates an upper bound for the percentage of zero tax change OP returns.

[5] Note that this is the correct answer for NRP, which is a stratified sample design, but for OP, which is not based on a probability sample, it is a model-based variance conditioning on substratum final sample sizes and substratum population sizes. As such, it is a lower-bound for the true variance (for OP), making the t-statistics upper bounds.

TABLE 7. NRP and SD-Augmented OP Percentages of Negative Tax Changes

Major Stratum	NRP Weight (one year)	NRP Percent Negative Tax Change		OP Percent Negative Tax Change		OP-NRP Percent Negative Tax Change		
		Percent	Standard Error	Percent	Standard Error	Percent	Standard Error	t-statistic
270	23,898,238	5.2	0.3	2.2	0.1	-2.9	0.3	-9.6
271	1,334,259	7.4	1.3	3.1	0.1	-4.3	1.3	-3.4
272	76,187,820	7.0	0.3	5.3	0.1	-1.7	0.4	-4.7
273	14,461,352	8.6	0.6	3.5	0.1	-5.2	0.6	-8.2
274	9,827,635	7.8	0.6	2.7	0.1	-5.1	0.6	-8.7
275	2,680,147	8.7	0.6	5.8	0.3	-3.0	0.6	-4.8
276	739,791	8.4	0.6	6.3	0.3	-2.1	0.7	-3.3
277	574,019	8.3	0.6	6.8	0.3	-1.5	0.7	-2.2
278	1,181,166	11.7	0.7	6.5	0.5	-5.2	0.9	-6.0
279	4,724,458	14.0	0.7	7.2	0.2	-6.9	0.7	-10.0
280	1,957,898	13.0	0.8	5.7	0.1	-7.2	0.8	-9.3
281	505,216	18.2	0.7	12.0	0.3	-6.3	0.7	-8.7
Other	7,511,262	10.1	1.4	2.9	0.1	-7.2	1.4	-5.3
Total	145,583,261	7.5	0.2	4.4	0.1	-3.1	0.2	-13.9

TABLE 8. NRP and SD-Augmented OP Percentages of Zero Tax Changes, Best Estimate Using the SD Imputations

Major Stratum	NRP Weight	NRP Percent Zero Tax Change		OP Percent Zero (w/0 SD) Tax Change		OP (w/0 SD) - NRP Percent Negative Tax Change		
		Percent	Standard Error	Percent	Standard Error	Percent	Standard Error	t-statistic
270	23,898,238	40.5	0.6	20.9	0.2	-19.6	0.6	-30.9
271	1,334,259	10.0	1.5	22.2	0.3	12.2	1.6	7.7
272	76,187,820	57.5	0.6	41.0	0.2	-16.5	0.7	-24.3
273	14,461,352	26.8	0.9	26.6	0.2	-0.2	0.9	-0.2
274	9,827,635	21.1	0.9	23.7	0.3	2.6	0.9	2.8
275	2,680,147	12.5	0.7	25.9	0.4	13.4	0.8	16.8
276	739,791	10.2	0.6	22.7	0.5	12.5	0.8	15.7
277	574,019	10.4	0.7	26.0	0.5	15.6	0.8	19.3
278	1,181,166	23.7	1.0	34.9	0.9	11.2	1.3	8.5
279	4,724,458	29.1	0.9	41.7	0.3	12.5	0.9	13.5
280	1,957,898	15.6	0.8	34.1	0.3	18.4	0.9	20.9
281	505,216	26.9	0.8	44.3	0.4	17.4	0.8	19.9
900	7,511,262	21.9	1.6	19.6	0.2	-2.4	1.6	-1.5
Total	145,583,261	43.8	0.4	33.3	0.1	-10.5	0.4	-26.4

NOTE: All t-statistics are significantly different from 0 except major strata 273 and Other.

TABLE 9. NRP and SD-Augmented OP Percentages of Zero Tax Changes, Upper Bound

Major Stratum	NRP Weight	NRP Percent Zero Tax Change		UB OP Percent Zero Tax Change		UB OP to NRP Percent Negative Tax Change		
		Percent	Standard Error	Percent	Standard Error	Percent	Standard Error	t-statistic
270	23,898,238	40.5	0.6	22.0	0.2	-18.5	0.6	-29.1
271	1,334,259	10.0	1.5	31.5	0.4	21.5	1.6	13.6
272	76,187,820	57.5	0.6	51.1	0.2	-6.4	0.7	-9.5
273	14,461,352	26.8	0.9	35.9	0.2	9.1	0.9	9.6
274	9,827,635	21.1	0.9	32.3	0.3	11.2	0.9	11.8
275	2,680,147	12.5	0.7	35.4	0.5	22.9	0.8	28.1
276	739,791	10.2	0.6	31.5	0.5	21.3	0.8	25.9
277	574,019	10.4	0.7	35.2	0.5	24.8	0.8	30.0
278	1,181,166	23.7	1.0	41.0	0.9	17.3	1.3	13.0
279	4,724,458	29.1	0.9	57.6	0.3	28.5	0.9	30.6
280	1,957,898	15.6	0.8	47.4	0.3	31.8	0.9	35.8
281	505,216	26.9	0.8	57.4	0.4	30.4	0.9	34.9
900	7,511,262	21.9	1.6	25.7	0.3	3.8	1.6	2.3
Total	145,583,261	43.8	0.4	41.7	0.1	-2.1	0.4	-5.4

NOTE: All t-statistics are significant. This is additional evidence that major strata probably cannot be full equated.

Kolmogorov–Smirnov Tests to Identify Fully/Partial Equated Strata

This material summarizes our comparisons of the NRP and SD-augmented OP distributions. The distributions were compared using the K-S Test, and by comparing the means, standard deviations, and critical percentiles. Unweighted analyses were appropriate because within a substratum the NRP weights are virtually always equal, and we anticipate that equal weights within substrata will also be efficient for OP returns. The analysis was done for the most recent available TY 2015, and also for TYs 2013–2015,[6] over the 87 NRP substrata in the most recent stratification.

The fully equated substrata allowed for simple composite estimation between the NRP and OP estimators. Of the remaining substrata, we looked for subdomains in which partial equating could take place. We utilized the DIF score assigned to each return to generate these subdomains, due to its correlation with OP audit selection.

Within each of the substrata, we attempted to find a subset such that the tax change field for NRP and the equivalent field for OP audits are comparable, determined by a K-S test using this DIF score. These groups are determined primarily by finding the smallest cutoff for the DIF score such that the K-S test has a p-value >= 20 percent comparing the NRP and OP returns above the cutoff, when possible. We include all OP records in the equated domain group but only NRP records with the DIF score above the determined cutoff to identify candidate equated domains.

The next step in this algorithm is to do a K-S test of NRP vs. OP for each substratum, where each substratum is subset to records in any candidate domain. The K-S test equality does not necessarily hold up at the substrata level, so we apply this second filter. Any substratum that does not pass the K-S test (K-S p-value greater than 20 percent) using the defined DIF cutoffs (which differ by substratum crossed with area-level) is not considered to be equated. We retain substrata within the major stratum that pass the K-S equating test with the defined subgroups as equated substrata (NRP and OP are equated within the substrata within the defined DIF subgroups).

[6] Using three years of data increases the sample sizes, but risks mixing in slightly differing distributions across the tax years.

After working through this algorithm, we found equated domains for 11 out of 87 substrata. Summarized by major stratum in Table 10, no consistent patterns appear for the occurrence of matching tax change distributions but the fact any exist with such a conservative test is encouraging.

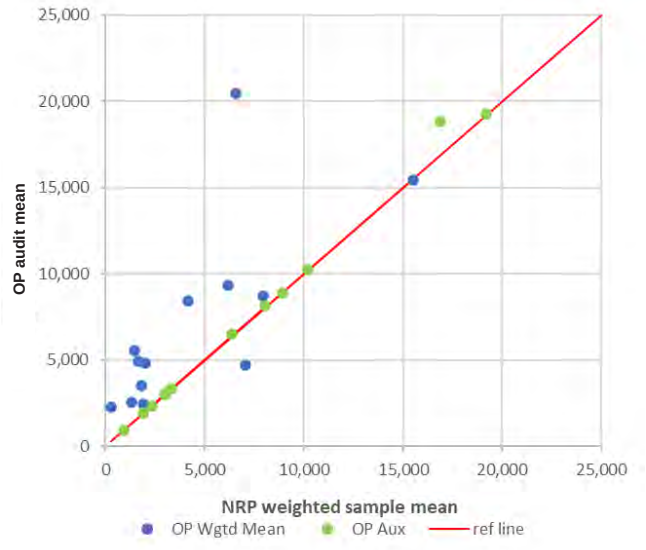
TABLE 10. Categorizations of Matching Status (OP vs. NRP) for Major Strata

Major Stratum	2013–2015 Weighted Sum	Matching NRP to OP Rates
270	40,323,602	0 out of 15
271	3,668,463	1 out of 6
272	80,818,157	0 out of 8
273	28,364,297	0 out of 4
274	20,942,993	0 out of 4
275	6,499,596	1 out of 4
276	1,869,590	2 out of 5
277	1,464,803	2 out of 5
278	2,347,967	0 out of 3
279	7,387,984	1 out of 6
280	3,877,587	0 out of 3
281	744,698	0 out of 3
Other	8,261,944	4 out of 21

Estimators With Augmented Self-Representing OP Audits

Figure 2 presents the NRP and OP weighted estimators (positive tax-change values only) and the self-weighting OP-augmented estimator. Note that the OP values are also all positive. One can see that the new estimators are very close in general to the old estimators, though they are slightly larger in most cases (probably as the OP audits tend to be missing some of the smaller positive tax-change values). Omitted standard errors were generally slightly smaller, indicating a gain in precision from adding in the OP audits.

FIGURE 2. Comparison of NRP and OP Weighted Estimators and OP-Augmented ("OP Aux") NRP Estimators, by Major Strata



Sliding Scale Composite Estimator Results

Table 11 shows results for the sliding scale composite estimator by major stratum.

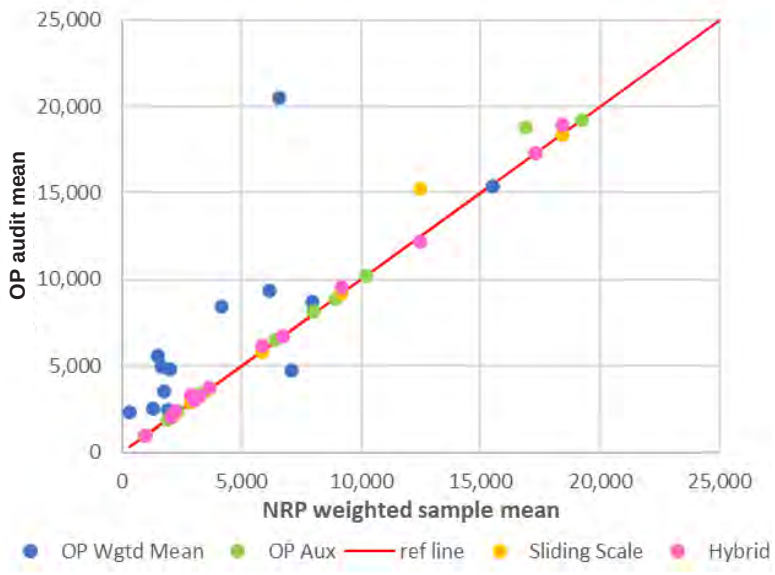
TABLE 11. Sliding Scale Composite Estimators for Means: All-Universe

Major stratum	NRP		OP		Sliding Scale		
	Weighted Mean ($)	Mean Standard Error ($)	Mean ($)	Standard Error ($)	Alpha (%)	Mean ($)	RMSE* ($)
270	3,235	49	5,019	320	99.9	3,237	49
271	12,456	716	15,486	1,062	7.5	15,257	1,021
272	966	47	4,470	442	100.0	967	47
273	2,218	108	8,028	772	100.0	2,220	108
274	2,035	243	5,554	2,946	99.5	2,051	242
275	5,845	1,351	7,157	427	9.1	7,038	407
276	9,185	957	14,235	5,998	97.5	9,310	945
277	18,409	1,870	27,950	7,224	96.2	18,776	1,834
278	3,030	136	5,853	419	99.8	3,036	136
279	3,628	352	13,713	1,219	99.9	3,640	352
280	6,717	503	18,078	1,367	99.8	6,739	502
281	17,278	2,058	90,236	13,535	99.9	17,336	2,057
Other	2,887	204	6,278	501	99.6	2,899	204
Total	2,607	90	6,694	1,594	100.0	2,609	90

*Root mean squared error

Figure 3 provides a summary of the major stratum estimates, including the NRP estimator, the OP propensity-weighted estimator, the OP self-representing auxiliary estimator, the hybrid estimator, and the sliding scale estimator. The OP self-representing auxiliary estimator and the hybrid estimator are fairly close to the NRP estimator (conservative), with the sliding scale sometimes stronger in the direction of the OP audits.

FIGURE 3. Summary of Estimators at Major Stratum Level



Comparing Among Alternative Estimators

In this section, we compare and contrast the various estimation approaches outlined in the preceding sections. There are three basic sets of substrata that will be handled separately: fully equated substrata, partially equated substrata, and others (neither fully nor partially equated).

Figures 4 and 5 present the proportion of each type of estimator from the NRP estimator (totals and means, respectively). For the simple linear composite estimators, this is just the alpha factor. For the more complicated estimators, such as the OP self-representing estimator, it is an implicit alpha, which is a linear approximation for the share of the NRP estimator.

The proportions of alternative composite-based totals to the NRP weighted equivalent are shown by NRP substratum, ordered by their equating status. Results are presented for the fully-equated substratum, the partially-equated substrata, and the other substrata. For the fully-equated substrata, there is a fully-equated composite estimator, but no partially-equated composite estimator. For the partially-equated substrata, there is a partially-equated composite estimator, but no fully-equated composite estimator. For the other substrata ("3-no") there are no fully-equated or partially-equated composite estimators. The sliding scale and OP self-representing estimators are provided for all substrata.

As expected, fully equated substrata have very low proportions on the NRP estimate, allowing more utilization of the OP data, but much less to none in the substrata where the distribution of tax change significantly varies between the NRP and OP data. In addition to identifying for which substrata this holds, our application indicates the estimators are suitable candidates to consider against the current NRP estimator or a naïve combination of the two data sources. In substrata where the NRP and OP data align, more of the OP audit data contribute more to the estimation, and the OP data are less utilized when they differ too much from the NRP.

FIGURE 4. Proportion of Estimated Total from NRP Estimator for All Types of Estimators

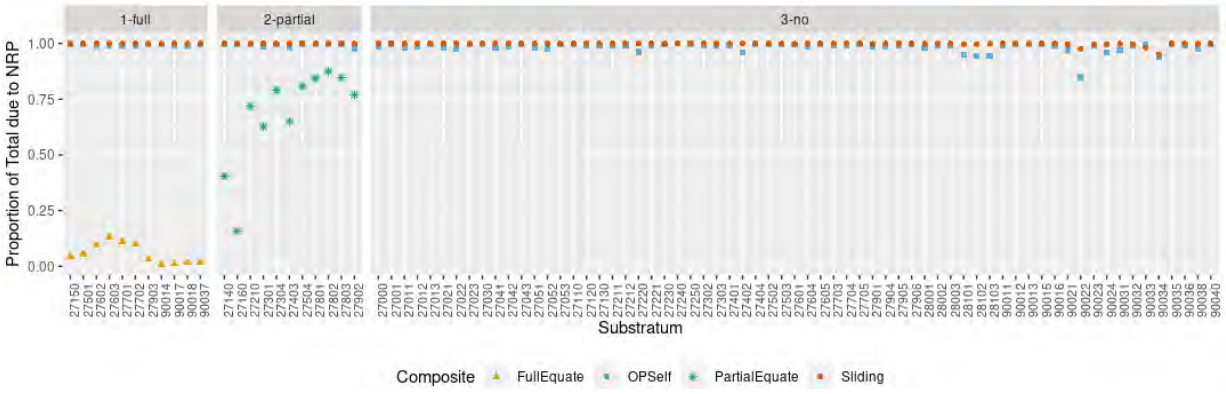
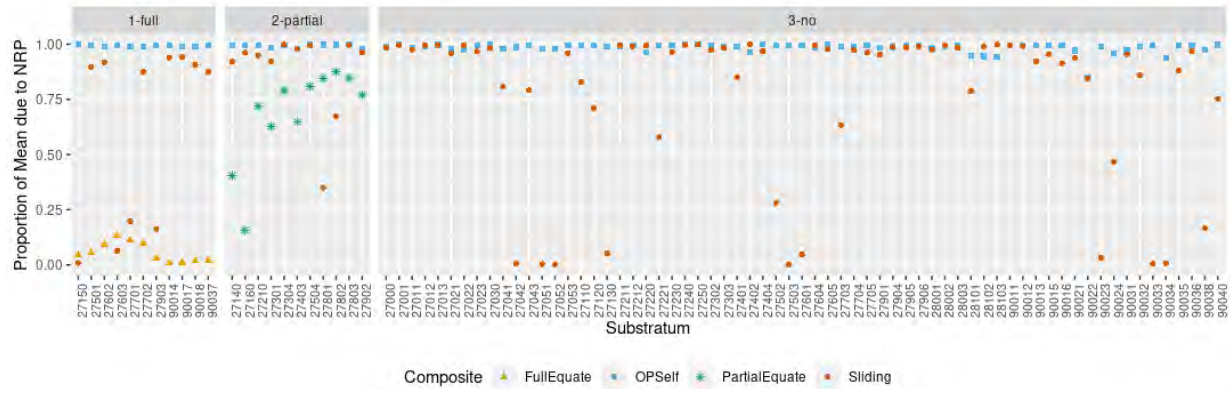


FIGURE 5. Proportion of Estimated Mean from NRP Estimator for All Types of Estimators



Figures 6 and 7 provide the actual composite estimators for totals and means respectively (utilizing the same separation of substrata as Figures 4 and 5).

FIGURE 6. Final Estimates of Totals for All Types of Estimators

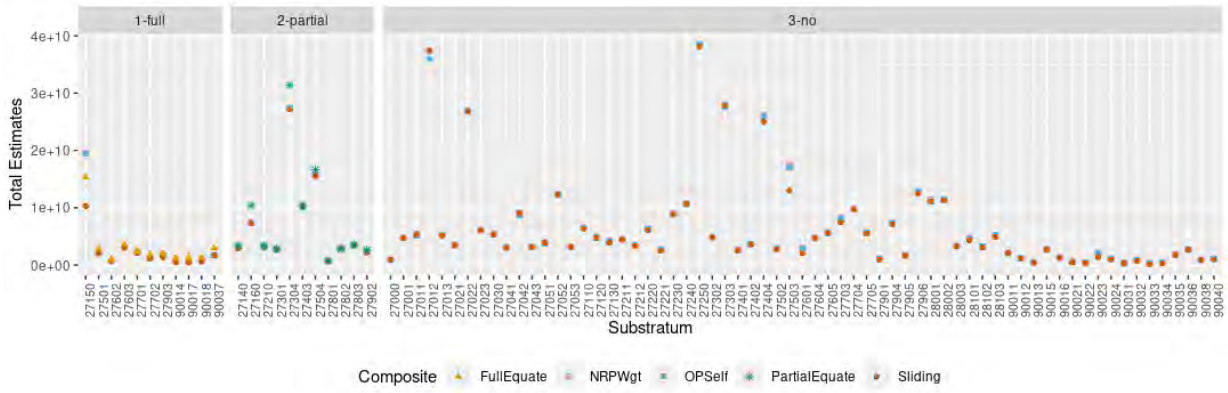
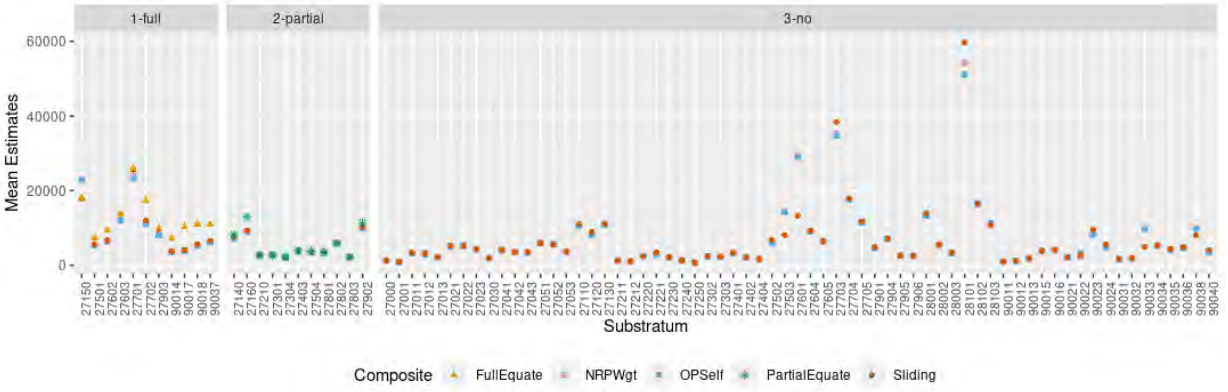


FIGURE 7. Final Estimates of Means for All Types of Estimators



The full-equating and partial-equating show significant differences from the NRP estimates. The sliding scale also sometimes does, but generally it closely tracks the NRP estimate. The OP Self-Representing generally tracks the NRP estimate.

Conclusions and Future Considerations

Given the detected systematic difference between NRP and operational audits, it doesn't seem useful to put together composite estimators, or collect the NRP and OP estimates and weight them up using the NRP sample weight and the OP inverse propensity to respond. Any operational audits that are based on a pseudo-random selection process conditional on tax return items can be added to the NRP universe using their inverse OP propensities, but operational audits based on a nonignorable selection process might have to be treated as self-representing (a weight of 1). Assigning them any other weight implies that the OP record can represent other non-OP records as well as itself, and this seems problematic unless the selection issue can be addressed.

We have developed four different procedures for augmenting NRP estimators with OP auxiliary data. Three of them include designating a portion of the NRP universe within the substratum as being equated with some or all of the OP audits. The first case is fully equating where the OP distribution is found to match well

the NRP distribution. In such cases, the NRP and OP distributions are apples to apples, and we have developed composite estimators to put these NRP and OP estimates together. We implemented full equating for substrata where the full equating matching could be justified. In these cases, the composite estimator is generally a substantial improvement over the simple NRP estimator, measured by RMSE.

Another case is partial equating, which is a portion of the NRP universe that is matched with the OP audits. We do this by using the DIF score and designating a cutoff for it. Above this cutoff, the NRP and OP audits can be equated. We equate the NRP universe above the DIF cutoffs with all OP audits.

There are two further estimators that combine NRP and unequated OP estimators directly. The first augmented OP self-representing estimator brings in the OP estimators as self-representing: the OP audits represent themselves with a weight of 1. The NRP estimators are then modified to leave a space for the OP estimates by multiplying their sample weights by a factor based on their propensity to be operational (the factor is one minus this propensity). The second sliding scale composite estimator is a full-scale composite estimator that does not require equating, but compensates for this by reducing the influence of the OP estimate by making its share inversely proportional to its estimated MSE, not its estimated variance.

Table 12 summarizes the procedures that have been compared in this research.

TABLE 12. Research Summary

NRP Substratum Type	Estimation Type	Precision of Estimate (Compared to NRP Alone, Which is Low)
Fully Equated	Full Composite Estimation	High
Partially Equated	Composite Estimation for Equating Domain	Medium
No Equating	NRP Augmented by Self-Representing OP	Low
	Sliding Scale Composite Estimation	Varies

Bringing in the OP surveyed cases (following imputation) allowed us to represent the lower part of the OP tax change distribution and to successfully equate a much larger set of substrata, and this represents progress. The partially-equated substrata will also have composite estimators which improve on the NRP estimator. The sliding scale composite estimator is another alternative to these with somewhat similar results. The OP self-representing domain estimator is a very conservative alternative.

One issue that needs to be addressed in development of these estimators is the existence of negative tax change returns, zero tax change returns, along with the positive tax change returns which are the subject of most of our research. The negative tax change returns are a fairly small percentage of the return universe, and they could be handled within each NRP substratum by either simply using the weighted NRP returns, or by possibly supplementing these with OP returns if equating could be confirmed (within particular NRP substrata, as for the positive tax change returns).

The zero tax change returns require no equating certainly (the right answer is zero for each and every return of this type). The important issue is the percentage of these among all returns. It may be best to simply use the NRP weighted percentage, as an unbiased estimator of the true percentage. The operational audit process is not designed to provide zero tax change returns, as each operational zero tax change audit is in a sense a failure, i.e., a wasted effort. Thus, it is probably necessary to rely on the NRP weighted percentage alone as the only estimate of this with any hope of being accurate.

For the positive tax change returns, there are two issues to confront. The first is selection, which is the fact that operational auditing is designed to, as much as is possible, find the high-positive tax change returns. Our mechanism to counter selection bias was equating when the operational distribution matches the NRP

distribution. If so, we have full equating and have confidence supplementing NRP data with OP cases. If not, then partial equating, based on a known selection mechanism such as the DIF score, can be utilized. Using the DIF score is an important part of overcoming selection, but there may exist further predictors to tell us whether a particular return has a higher probability of being audited than a measured model probability based on its tax return items.

Another issue is outliers. There are many remote outliers in tax change, and these outliers are a considerable part of the final estimate. As a probability sample, the NRP sample will only include these by chance, so the NRP sample will always be limited in getting the far ends of the tax change distribution represented with any precision. The OP audits can fill in a good portion of these outliers, if well-handled.

References

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Castro-Martin, L., M. Rueda, and R. Ferri-Garcia (2020). "Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques. Mathematics 8, 879, doi: 10.3390/math8060879.

Chen, Y., P. Li, and C. Wu (2019). "Doubly Robust Inference with Non-probability Survey Samples." *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2019.1677241.

Elliott, M. R., and R. Valliant (2017). "Inference for Nonprobability Samples." Statistical Science 32(2), 249–264.

Ferri-Garcia, R., and M. Rueda (2020). "Propensity Score Adjustment Using Machine Learning Classification Algorithms to Control Selection Bias in Online Surveys." PLOS One https://doi.org/10.1371/journal.pone.0231500.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. www.springer.com.

Hothorn, T., K. Hornik, and A. Zeileis (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

Hothorn, T., and A. Zeileis. (2015). partykit: "A Modular Toolkit for Recursive." *Journal of Machine Learning Research*, 16, 3905–3909. Retrieved from http://jmlr.org/papers/v16/hothorn15a.html

Kish, L. (1992). "Weighting for Unequal Pi." *Journal of Official Statistics* 8 (2): 183–200.

Valliant, R. (2020). "Comparing Alternatives for Estimation from Nonprobability Samples." *Journal of Survey Statistics and Methodology, 8*, 231–263.

Integrating Reward Maximization and Population Estimation: Sequential Decision-Making for Internal Revenue Service Audit Selection

Peter Henderson, Ben Chugg, Brandon Anderson,[1] Kristen Altenburger, Jacob Goldin,[2] and Daniel E. Ho (Stanford University), and Alex Turk and John Guyton (IRS, RAAS)

We introduce a new setting that we call optimize-and-estimate structured bandits. (Bandit algorithms learn to pick the best option from a set of choices by learning from previous decisions.) Here, a policy must select a batch of arms, each characterized by its own context, that would allow it to maximize reward and maintain an accurate (ideally unbiased) population estimate of the reward. This setting is inherent to many public and private sector applications and often requires handling delayed feedback, small data, and distribution shifts. We demonstrate its importance on real data from the Internal Revenue Service (IRS). The IRS performs yearly audits of the tax base. Two of its most important objectives are to identify suspected misreporting and to estimate the "tax gap"—that is, the global difference between the amount paid and the true amount owed. Based on a unique collaboration with the IRS, we cast these two processes as a unified optimize-and-estimate structured bandit. We provide a novel mechanism for unbiased population estimation that achieves rewards comparable to baseline approaches. This approach has the potential to improve audit efficacy, while maintaining policy-relevant estimates of the tax gap. This has important social consequences given that the current tax gap is estimated at nearly half a trillion dollars. We suggest that this problem setting is fertile ground for further research, and we highlight its interesting challenges. The results of this and related research are currently being incorporated into the continual improvement of the IRS audit selection methods.

Link to the current version of the full paper: https://arxiv.org/abs/2204.11910.

[1] Also affiliated with IRS, Research, Applied Analytics, and Statistics
[2] Also affiliated with U.S. Department of the Treasury

2

Burden vs. Opportuniuty

Black ◆ Goldin ◆ Hess ◆ Ho ◆ Lester ◆ Paul ◆ Portz Johns

The Spiderweb of Partnership Tax Planning[1]

Emily Black (Carnegie Mellon University), Jacob Goldin, Ryan Hess, Daniel E. Ho, Rebecca Lester,[2] and Mansheej Paul (Stanford University), and Annette Portz (IRS, RAAS)

A central economic question relates to the amount, type, and form of taxation to which businesses are subject. Much of the prior literature examines these issues for a subset of U.S. businesses that operate as "C" corporations. However, C corporations represent only 16 percent of all business returns as of 2020. Recent work demonstrates that the rise of other, noncorporate entities is a critical factor for estimating business tax revenue collections (Cooper *et al.* (2016)), explaining the declining corporate sector labor share (Smith *et al.* (2021a)), and studying rising inequality (Smith *et al.* (2019 and 2021b)). This paper examines the tax planning of one large group of noncorporate firms—partnerships. Specifically, we study how organizational flexibility afforded by these entities facilitates business tax planning. We use confidential, anonymized administrative tax data to evaluate the size, ownership, and connections among partnership entities. We show that approximately 75 percent of partnership groups are "simple" structures composed of one single partnership owned directly by individual taxpayers. In contrast, the most complex organizations resemble webs of ownership, with clusters of overlapping partners. Preliminary work leverages machine learning (ML) approaches to predict partnership noncompliance. When using firm characteristics based on the prior literature, we find that Ridge models have higher accuracy rates. Interestingly, inclusion of the additional partnership features does not appear to improve the Ridge model accuracy. In contrast, we find substantial improvement in the Random Forest model, confirming the predictive values of these features and suggesting a nonlinear relation between these features and tax planning. Future analyses will include additional features and alternative ML approaches to improve upon this baseline.

References

Cooper, Michael, John McClelland, James Pearce, Richard Prisinzano, Joseph Sullivan, Danny Yagan, Owen Zidar, and Eric Zwick (2016). "Business in the United States: Who Owns It, and How Much Tax Do They Pay?" In: *Tax Policy and the Economy* 30.1, pp. 91–128.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick (2019). "Capitalists in the Twenty-First Century." In: *The Quarterly Journal of Economics* 134.4, pp. 1675–1745.

Smith, Matthew, Danny Yagan, Owen M. Zidar, and Eric Zwick (2021a). *The Rise of Pass-Throughs and the Decline of the Labor Share*. Tech. rep. National Bureau of Economic Research.

Smith, Matthew, Owen M. Zidar, and Eric Zwick (2021b). *Top Wealth in America: New Estimates and Implications for Taxing the Rich*. Tech. rep. National Bureau of Economic Research.

[1] We thank John Guyton, Thomas Hertz, Ron Hodge, Larry May, Keisha Miller, Eric Tressler, and Alex Turk from the Internal Revenue Service (IRS) for providing us with the relevant data, sharing their expertise, and supporting our research. The IRS provided confidential tax information to the authors under the Joint Statistical Research Program. All opinions are those of the authors and do not reflect the views of the IRS.

[2] Corresponding author; rlester@stanford.edu. Mailing address: 655 Knight Way, Stanford, CA 94305.

Distribution of the Tax Years 2011–2013 Individual Income Tax and Self-Employment Tax Underreporting Tax Gap

Drew Johns (IRS, Research, Applied Analytics & Statistics)[1]

1. Introduction

Tax Gap estimates historically have been reported according to the type of tax, while the major components within a type of tax have reflected sources of misreporting, such as line items or groups of line items. Since Tax Year (TY) 2006, the Internal Revenue Service (IRS) through the National Research Program (NRP) has conducted annual studies of individual income tax reporting compliance. Data from these studies have provided the richest source of reporting compliance data and, therefore, allowed a more detailed breakout of the tax gap for individual income tax underreporting than can be made for other components. Because of this additional detail, this tax gap component also prompts more questions. Recently, there has been interest in understanding the distribution of the individual income tax underreporting tax gap by tax return characteristics. This paper presents estimates of the distribution of the individual income tax and self-employment (SE) tax underreporting tax gap for the TYs 2011–2013 time period.

This increased interest in the distribution of the individual income tax gap primarily is the result of two areas of inquiry. The first concerns the allocation of IRS resources and, in particular, the allocation of examination resources. Associated with this is speculation about the income distribution of the tax gap estimates and whether the IRS allocates too many resources toward issues and tax returns on the lower end of the income distribution. The second area of inquiry is an economic policy question about the extent to which income inequality has increased over time. There is interest in whether accounting for tax noncompliance affects the measures of inequality. This paper focuses on the first area of inquiry–how the distribution of the tax gap compares to the distribution of examination resources.[2]

Table 1 shows the tax gap estimates for the TYs 2011–2013 time period. The individual income tax underreporting tax gap for the TYs 2011–2013 time period is $245 billion. The self-employment tax underreporting tax gap is $45 billion. The uncollected Social Security and Medicare tax underreporting tax gap is $1 billion. Self-employment tax and uncollected Social Security and Medicare tax are also reported on the individual income tax return and included in the calculation of total tax on that return. Therefore, this report includes underreported self-employment tax in estimating the distribution of the underreporting tax gap for the individual income tax return. The sum of the individual income tax underreporting tax gap, the self-employment tax underreporting tax gap and the uncollected Social Security and Medicare tax underreporting tax gap is $290 billion.[3]

IRS stratifies tax returns based on tax return characteristics into mutually exclusive groups for examination planning and tracking. These groups are called activity codes. For individual income tax returns, the primary stratifiers are the level of reported Total Positive Income (TPI), the level of reported Total Gross Receipts (TGR) and presence of the Earned Income Tax Credit (EITC). The characteristics and groups were chosen such that returns within a given group are relatively similar in the compliance characteristics, while having relatively different compliance characteristics than other groups. This paper includes estimates of noncompliance

[1] The author is an economist in the Compliance Modeling Lab within the Knowledge Development & Application Division. The views expressed in this paper do not necessarily represent the views of the Department of the Treasury or the IRS.

[2] There is often a perception that the allocation of examination resources should be allocated proportional to the tax gap. However, there is little reason to expect that such an allocation is necessarily optimal for maximizing direct revenue, voluntary compliance, or social welfare.

[3] The sum of the three components is $290 billion and not $291 billion due to rounding.

by activity code and compares those estimates to the allocation of examination resources, controlling for the type of examination.

The paper also looks at the distribution of noncompliance by three different levels of income: reported income, examiner-determined income, and Detection Controlled Estimation (DCE) adjusted income. Taxpayers who underreport their income often fall into a lower income percentile when ranked by reported income relative to their percentile when ranked by examiner-determined or DCE-adjusted income. This fact has implications for tax administration and resource allocation since only reported income for taxpayers is observed, absent an examination. In addition, a significant portion of the estimated tax gap reflects noncompliance that would unlikely be detected on an examination. Therefore, the tax gap does not necessarily reflect the amount of direct revenue that could be recovered through increased enforcement activity.

2. Data and Methodology

2.1 National Research Program

The foundation of the individual income tax underreporting tax gap estimates is data from the IRS NRP individual income reporting compliance studies, supplemented with estimates of undetected income using an econometric technique called Detection Controlled Estimation. NRP designs and administers reporting compliance studies for the IRS.[4] The NRP reporting compliance studies are examination programs where returns are selected for audit (examination) in a statistical manner that allows one to draw inferences about the population from the results of those audits. The purpose of a given NRP audit is to ascertain the correctness of the return examined and determine the correct liability.

The first NRP study of individual income tax reporting compliance consisted of a stratified random sample of about 45,000 TY 2001 individual income tax returns filed during Calendar Year (CY) 2002.[5] That study served as the basis for the TY 2001 individual income tax underreporting tax gap estimates. Beginning with TY 2006, the IRS began smaller annual studies of approximately 14,000 individual income tax returns a year. The annual studies can be combined over several years to provide compliance estimates at a similar level of reliability as a single-year larger study. Data for a given tax year generally are available for analysis purposes about three years after the returns are filed.

NRP uses a process called classification to determine the type of audit for each return selected and the mandatory issues to be examined.[6] The classification process compares information return documents (Forms W-2, Forms 1099, etc.) with the actual tax return to identify discrepancies and items that appear large, unusual, or questionable. Some line items on the return, typically those that cannot be verified through informationreturns, are always classified as mandatory to audit. In the case of simpler returns where information can bereconciled with third-party information and there appears to be a low likelihood that items are missing from the return, taxpayers are not audited and not even contacted. Returns that have only a small number of simplerissues identified in classification are routed to campus correspondence examination where the examinations can be handled through telephone calls, faxes, and traditional mail. More complicated returns are assigned to one of two types of audits that involve face-to-face interaction with an examiner: either an office audit handled by a Tax Compliance Officer (TCO) or a field audit handled by a Revenue Agent (RA) who may visit the taxpayer's place of business.

The classification process, by selecting an appropriate audit technique and set of issues, serves to reach an appropriate balance among the objectives of ensuring the taxpayer reported the correct liability, obtaining comprehensive and reliable information about reporting compliance, and taxpayer and examiner effort involved in an examination. The number of mandatory issues on an NRP-selected audit typically exceeds the

[4] NRP conducts more than just individual reporting compliance studies. It should be assumed for the remainder of this chapter that references to an NRP study refer to an individual reporting compliance study, unless explicitly stated otherwise.

[5] The TY 2001 individual reporting compliance study consisted of returns with tax periods ending between July 2001 and June 2002, the overwhelming majority of which ended on December 31, 2001, and were filed in early 2002.

[6] Examples of issues include line items on the return, filing status, number of dependents, and whether an activity is engaged in for profit or as a hobby.

TABLE 1. Tax Years 2011–2013[1] Tax Gap Estimates

[Money amounts are in billions of dollars]

Tax Gap Component	TYs 2011–2013	Share of Gross Tax Gap
Estimated Total True Tax	**$2,683**	
Gross Tax Gap	$441	100%
Voluntary Compliance Rate	83.6%	
Enforced and Other Late Payments	$60	
Net Tax Gap	$381	
Net Compliance Rate	85.8%	
Nonfiling Gap	**$39**	**9%**
Individual Income Tax	$31	7%
Self-Employment Tax	$6	1%
Estate Tax	$2	[2]
Underreporting Gap	**$352**	**80%**
Individual Income Tax	**$245**	**56%**
Non-Business Income	$57	13%
Business Income	$110	25%
Adjustments, Deductions, Exemptions	$20	4%
Filing Status	$5	1%
Other Taxes [4]	$1	[2]
Unallocated Marginal Effects [5]	$10	2%
Credits	$42	10%
Corporation Income Tax	**$37**	**8%**
Small Corporations (assets under $10M)	$11	2%
Large Corporations (assets of $10M or more)	$26	6%
Employment Tax	**$69**	**16%**
Self-Employment Tax	$45	10%
Uncollected Social Security and Medicare Tax	$1	[2]
FICA and Unemployment Tax	$24	5%
Estate Tax	**$1**	**[2]**
Underpayment Gap	**$50**	**11%**
Individual Income Tax	$38	9%
Corporation Income Tax	$5	1%
Employment Tax	$6	1%
Estate Tax	[3]	[2]
Excise Tax	[3]	[2]

[1] The estimates are the annual averages for the TYs 2011–2013 timeframe.

[2] Less than 0.5 percent.

[3] Less than $0.5 billion.

[4] Other taxes include the Alternative Minimum Tax and taxes reported in the "Other Taxes" section of the Form 1040, except for self-employment tax and unreported Social Security and Medicare tax (which are included in the employment tax gap estimates).

[5] Unallocated marginal effects is the difference between (1) the estimate of the individual income tax underreporting tax gap where underreported tax is calculated based on all misreporting combined and (2) the estimate of the individual income tax underreporting tax gap based on the sum of the tax gaps associated with each line item where the line item tax gap is calculated based on the misreporting of that item only. There may be a difference whenever more than one line item has been misreported on the same return and the combined misreporting results in a higher marginal tax rate than when the tax on the misreported amounts is calculated separately.

Detail may not add to total due to rounding.

number of issues that would have been examined had the return been selected through another IRS compliance risk-based return selection processes. The audits selected through these latter programs generally are more limited in the scope of issues covered compared with those covered in the audits selected under NRP. NRP audits, therefore, are more complete audits, which is beneficial for ascertaining the accuracy of the return and determining the correct tax liability. Examiners also have the discretion to expand the audit to include non-classified issues, typically whenever information is uncovered during the audit that causes the examiner to question those issues.

2.2 Detection Controlled Estimation

Not all underreported income is detected by every audit, even ones of the scope and quality of NRP audits. This was confirmed by the 1976 IRS Taxpayer Compliance Measurement Program (TCMP) individual income tax reporting compliance study, which was the last IRS reporting compliance study to audit taxpayers without the auditors having the use of third-party information return documents. The IRS later compared the information return documents to the audit findings and found that for every $1.00 of detected unreported income that was reported on information documents, an additional $2.28 went undetected. As a result of that study, the IRS began multiplying the portion of income detected without the use of information documents by a multiplier, typically 3.28, to estimate the individual income tax underreporting tax gap

In the late 1980s, Jonathan Feinstein developed an econometric technique for estimating undetected income that he termed "Detection Controlled Estimation" or DCE (Feinstein (1990, 1991)). The assumption underlying the methodology was that examiners have varying abilities for detecting income that can be observed through patterns in the data collected from taxpayer audits. Feinstein explained that the observed audit adjustment actually reflects the product of the true (unobserved) unreported income and the propensity of the examiner to detect unreported income. Feinstein's application of the methodology to TCMP data resulted in comparable estimates for the amount of undetected income as the IRS was assuming based on the 1976 TCMP study.

The TY 2001 tax gap estimates incorporated DCE estimation for the first time. The original DCE methodology focused on estimating overall noncompliance for a given return. The first iteration of DCE using the TY 2001 NRP data involved DCE estimation for two categories of income separately for two categories of taxpayers. The results were then synthesized down to four "multipliers" that were then applied to positive adjustments made on face-to-face audits.

Further research determined that there was an opportunity to expand DCE estimation to allow for greater variability in the average detection rates across line items. The IRS Office of Research contracted with Dr. Brian Erard (B. Erard and Associates) and Professor Feinstein (Yale School of Management) to extend and refine Professor Feinstein's original DCE methodology. The current DCE methodology used for the TY 2011–2013 individual income tax underreporting tax gap estimates provides microlevel estimates of undetected income at the return and issue level. Appendix 1 provides greater detail on the current methodology.

2.3 Compliance Measures

The definition of the compliance measures in this paper are listed below.

Examiner-Determined

Examiner-determined estimates are based on the amounts of misreporting determined by the examiner and are not adjusted for undetected income. This amount could be interpreted as the expected direct enforcement revenue, excluding any indirect effects, if all noncompliant returns in the population were audited.

DCE-Adjusted
DCE-adjusted estimates are based on the examiner-determined net misreporting plus DCE estimates of undetected income. The published tax gap estimates are based on the DCE-adjusted estimates. Since the estimated amount of the tax gap associated with income that is unlikely to be detected on an audit is more than the estimated amount that is likely to be detected, the tax gap should not be interpreted as the amount of money that could be collected if all noncompliant returns in the population were audited.

Underreporting Tax Gap
The underreporting tax gap is defined as the net amount of misreported tax that should have been reported on timely filed returns.

Net Misreported Amount (NMA)
The Net Misreported Amount, or NMA, is a concept associated with the underreporting tax gap. The NMA is the dollar amount of misreporting associated with a particular tax return or schedule line item. Although most often the NMA reflects an amount of income, expense, or similar line item that has been misreported; the NMA is also defined for the amount of tax or credit misreported. Since amounts reported on tax return and schedule lines can be either positive or negative and can be overstated or understated, the actual computation depends on whether the line item is an income (or tax) item or an offset item (such as a deduction, expense, or credit).

For an income or tax item, the NMA is calculated as the sum of all amounts understated minus the sum of all amounts overstated. In general, income items are underreported in the aggregate, so the NMA for income items generally is positive.

For an offset item, the NMA is calculated as the sum of all amounts overstated minus the sum of all amounts understated. In general, offset items are overstated in the aggregate, so the NMA for offsets typically is positive. For this concept, the word *net* refers to the offsetting of overstated and understated amounts and not the subtraction of enforced and other late payments.

Net Misreporting Percentage (NMP)
The Net Misreporting Percentage (NMP) for a given line item is the NMA divided by the sum of the *absolute values of the amounts that should have been reported*. For most return or schedule line items, amounts that should have been reported can be positive only. However, amounts can be either positive or negative for business-related net income and certain other lines. So, for those line items where amounts can be negative, the denominator of the NMP is not the net of positive and negative amounts, but instead it is the total of all the amounts disregarding the sign in the calculation—that is, it is the sum of the absolute values. The NMP is a complement to the NMA.

Net Overclaim Percentage (NOP)
The net overclaim percentage for a given line item is the NMA divided by the amount reported.

2.4 Income and Tax Definitions

Definitions of income related terms are listed below.

Total Positive Income (TPITG)
In general, TPI is the sum of all positive amounts shown for the various sources of income reported on an individual income tax return and, thus excludes losses.[7] The definition and calculation of TPI in this report, unless otherwise specified, differs slightly from the definition used generally by the IRS and is being denoted with the subscript $_{TG}$.

[7] https://www.irs.gov/statistics/soi-tax-stats-irs-data-book-glossary-of-terms.

Total Gross Receipts (TGR)
TGR are the sum of gross receipts from farm and nonfarm businesses calculated by adding the positive values of gross receipts and other income from Schedule C and gross income (which can be positive or negative) from Schedule F. Schedule C is used to report profit or loss from nonfarm sole proprietorships. Schedule F is used to report profit or loss from farming. If a taxpayer reports farm and nonfarm income, the return is classified by the larger source of income.[8]

Reported TPITG
Reported TPI_{TG} is the level of TPI as reported by the taxpayer.

Examiner-Detrmined TPITG
Examiner-determined income TPI_{TG} is the level of TPI based on income which would have been determined by IRS examiners.

DCE-Adjusted TPITG
DCE-adjusted income TPI_{TG} is the level of total positive after adjusting examiner-determined income for DCE estimates of income undetected by IRS examiners.

Tax plus SE Tax After Refundable Credits
Tax plus SE Tax after Refundable Credits is the individual income tax plus SE tax minus refundable credits. This amount can be either positive or negative. Underreporting the Tax plus SE Tax after Refundable Credits increases the tax gap.

Tax plus SE Tax Before Nonrefundable and Refundable Credits
Tax plus SE Tax Before Nonrefundable and Refundable Credits is the individual income tax plus SE tax before subtracting nonrefundable or refundable credits. This amount can only be positive. Underreporting Tax plus SE Tax Before Nonrefundable and Refundable Credits increase the tax gap.

Total Nonrefundable and Refundable Credits
Total Nonrefundable and Refundable Credits is the sum of nonrefundable and refundable credits. This amount can only be positive. Overreporting total nonrefundable and refundable credits increases the tax gap.

2.5 Activity Codes

IRS stratifies tax returns based on tax return characteristics into mutually exclusive groups called activity codes for examination planning and tracking. For individual income tax returns, the primary stratifiers are the level of *reported* TPI, the level of *reported* Total TGR and presence of the EITC. The characteristics and groups were chosen such that returns within a given group are relatively similar in the compliance characteristics, while having relatively different compliance characteristics than other groups. There are 12 domestic individual income tax return activity codes, although an additional two activity codes were implemented beginning with Filing Year 2022.

[8] Ibid.

TABLE 2. Individual Income Tax Return Activity Code Definitions (Filing Years 2007–2021)

Activity Code	Definition
270	EITC > 0; TGR < $25,000
271	EITC > 0; TGR ≥ $25,000
272	TPI < $200,000 and No Schedule C, E, F, or Form 2106
273	TPI < $200,000; No Schedule C/F; Schedule E or Form 2106 present
274	Schedule C; TGR < $25,000; TPI < $200,000
275	Schedule C; $25,000 ≤ TGR < $100,000; TPI < $200,000
276	Schedule C; $100,000 ≤ TGR < $200,000; TPI < $200,000
277	Schedule C; $200,000 ≤ TGR; TPI < $200,000
278	Schedule F Not Classified Elsewhere; TPI < $200,000
279	No Schedule C/F; $200,000 ≤ TPI < $1,000,000
280	Schedule C/F; $200,000 ≤ TPI < $1,000,000
281*	TPI ≥ $1,000,000

TPI = Total Positive Income

TGR = Total Gross Receipts

EITC = Earned Income Tax Credit

*In Filing Year 2022, the IRS is in the process of changing Form 1040 Activity Codes from 12 to 14 Activity Codes. This will be done by splitting Activity Code 281 into three new Activity Codes:

282: TPI ≥ $1,000,000 and < $ 5,00,000

283: TPI ≥ $ 5,000,000 and < $10,000,000

284: TPI ≥ $10,000,000

3. Distribution of Individual Income Tax Underreporting Tax Gap by Activity Code

Tables 3 and 4 show the distribution of the individual income tax plus SE tax underreporting tax gap by activity codes. Table 3 shows the tax gap in billions of dollars. Table 4 then shows the shares as a percentage of the total individual income tax plus SE tax underreporting tax gap. The tax gap is then further broken down into two components:

1. Tax before nonrefundable and refundable credits, including SE Tax, and
2. Total nonrefundable and refundable credits.

Breaking the tax gap down into these two components facilitates the interpretation of compliance rates as shown in Table 5. Unlike tax after refundable credits, the reported, examiner-determined and DCE-adjusted amounts for these components cannot be negative. Therefore, the NMP and NOP ratio measures are more easily interpreted for these two components.

Impact of DCE on the Estimates

The DCE-adjusted estimate of the individual income tax plus SE tax underreporting tax gap for is $290.4 billion, consistent with the published tax gap estimates. The estimate based only on examiner-determined misreporting is $139.1 billion. Therefore, the DCE-adjusted estimate for the overall individual income tax plus SE tax underreporting tax gap is more than double the estimate based on examiner-determined misreporting. Nearly all of the increase in the estimate tax gap from the DCE adjustment is attributable to tax before any nonrefundable or refundable credits. The DCE adjustment for undetected income has only a small impact on the estimated tax gap associated with misreported credits. The small impact on credits that does exist is due to the change in eligibility and the calculation of those credits based on the increase in income.

Returns Reporting TPI of $1 Million or More

Returns that reported TPI of $1 million or more fall into activity code 281. These returns accounted for $7.5 billion, or 2.6 percent of the DCE-adjusted individual income tax plus SE tax underreporting tax gap. This DCE-adjusted tax gap estimate is more than three times the estimate based on examiner-determined misreporting.

EITC versus Returns that Claimed EITC

IRS Publication 1415 (Rev. 9-2019) shows that EITC accounts for 6 percent of the $441 billion TY 2011–2013 gross tax gap and 11 percent of the individual income tax underreporting tax gap. EITC accounts for 9 percent of the individual income tax plus SE tax underreporting tax gap. These numbers reflect the share of the tax gap associated with EITC as a line item. The share of the tax gap associated with returns that claimed EITC is much higher. Although EITC as a line item accounts for 9 percent of the underreporting tax gap for individual income tax plus SE tax, returns that claimed EITC account for 28 percent. If looking at just the individual income tax plus SE tax underreporting tax gap associated with misreporting likely to be detected by an examiner, returns that claimed EITC account for nearly 36 percent.

Returns that claimed EITC fall into activity codes 270 and 271. Although the tax gap associated with the EITC line item is $27 billion, the tax gap, including SE tax, associated with returns that claimed EITC is $82 billion. Returns that claim EITC claim other credits and the tax gap for all credits associated with returns that claimed EITC is nearly $35 billion. The tax gap associated with underreported tax, before credits, is $47 billion for returns that claimed EITC. These numbers suggest that it might be misleading to look at only the tax gap associated with EITC as a line item when considering the allocation of resources to various activity codes.

TABLE 3. Distribution of TYs 2011–2013 Individual Income Tax plus Self-Employment Tax Underreporting Tax Gap by Activity Code ($ Billions)

Activity Code	Tax After Refundable Credits, Including SE Tax		Tax Before Nonrefundable and Refundable Credits, Including SE Tax		Total Nonrefundable and Refundable Credits	
	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted
270	$39.6	$58.9	$10.4	$27.8	$29.2	$31.1
271	$10.1	$23.2	$8.0	$19.5	$2.1	$3.7
272	$21.6	$36.3	$16.1	$30.8	$5.5	$5.5
273	$17.4	$38.8	$16.4	$37.8	$1.0	$1.0
274	$12.2	$34.7	$11.7	$34.2	$0.5	$0.5
275	$8.3	$23.1	$8.2	$23.1	$0.1	$0.1
276	$4.2	$10.5	$4.1	$10.5	[1]	[1]
277	$6.3	$14.6	$6.3	$14.6	$0.1	$0.1
278	$2.6	$8.3	$2.6	$8.3	[1]	[1]
279	$6.0	$15.5	$6.1	$15.5	-$0.1	-$0.1
280	$8.3	$19.0	$8.2	$18.9	$0.1	$0.1
281	$2.4	$7.5	$2.4	$7.5	[1]	[1]
Total	$139.1	$290.4	$100.5	$248.5	$38.5	$41.9

[1] Less than 0.05 percent or $0.05 billion

DCE–Detection Controlled Estimation.,

TABLE 4. Distribution of TYs 2011–2013 Individual Income Tax plus Self-Employment Tax Underreporting Tax Gap by Activity Code (Shares)

Activity Code	Tax After Refundable Credits, Including SE Tax		Tax Before Nonrefundable and Refundable Credits, Including SE Tax		Total Nonrefundable and Refundable Credits	
	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted
270	28.5%	20.3%	10.4%	11.2%	75.7%	74.1%
271	7.3%	8.0%	7.9%	7.9%	5.5%	8.9%
272	15.5%	12.5%	16.0%	12.4%	14.2%	13.0%
273	12.5%	13.3%	16.3%	15.2%	2.6%	2.4%
274	8.8%	12.0%	11.7%	13.8%	1.3%	1.2%
275	5.9%	8.0%	8.1%	9.3%	0.2%	0.2%
276	3.0%	3.6%	4.1%	4.2%	0.1%	0.1%
277	4.5%	5.0%	6.2%	5.9%	0.2%	0.1%
278	1.9%	2.9%	2.6%	3.3%	0.1%	0.1%
279	4.3%	5.3%	6.1%	6.2%	-0.1%	-0.1%
280	6.0%	6.5%	8.1%	7.6%	0.2%	0.2%
281	1.7%	2.6%	2.4%	3.0%	-0.1%	[1]
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[1] Less than 0.05 percent or $0.05 billion.

DCE–Detection Controlled Estimation.

Compliance Rates by Activity Code

Table 5 shows the NMP for tax after refundable credits and tax before nonrefundable and refundable credits, approximately reflecting the share of true amount of tax that was misreported. Table 5 also presents the NOP for total nonrefundable and refundable credits, approximately reflecting the share of the dollar amount of credit claims that was misreported.

The estimated NMP for tax plus SE tax after refundable credits is lowest for activity code 281 (TPI of $1 million or more) followed by 279 (TPI between $200,000 and $999,999 and without Schedule C or F) and 272 (TPI under $200,000; no EITC and not Schedule C, E, F or Form 2106). The NMP for tax plus SE tax after refundable credits is highest for activity codes 270 and 271, both reflecting returns that claimed EITC followed by 277 (TGR >= $200,000 and TPI <= $200,000).

The three activity codes with the highest NMPs (271, 270, 277) and the three with the lowest NMPs (281, 279, 272) are the same regardless of whether we consider only examiner-determined misreporting or the DCE-adjusted misreporting. Although the misreporting of credits impacts the NMP for the EITC related activity codes, the NMPs for tax plus SE tax before credits for the EITC activity codes are still the highest.

TABLE 5. TYs 2011–2013 Individual Income Tax plus Self-Employment Tax Underreporting Gap Compliance Rates by Activity Code

Activity Code	Tax plus SE Tax After Refundable Credits (NMP)		Tax plus SE Tax Before Nonrefundable and Refundable Credits (NMP)		Total Nonrefundable and Refundable Credits (NOP)	
	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted	Examiner-Determined	DCE-Adjusted
270	66.6%	83.7%	44.4%	68.1%	34.1%	36.3%
271	96.0%	109.3%	67.2%	83.4%	33.4%	58.2%
272	6.4%	10.2%	4.5%	8.2%	17.5%	17.5%
273	13.9%	26.5%	12.4%	24.7%	11.5%	11.4%
274	18.1%	38.8%	16.3%	36.2%	8.7%	8.5%
275	27.1%	51.1%	25.5%	49.1%	4.2%	4.3%
276	30.4%	52.4%	29.1%	51.0%	7.4%	7.3%
277	44.4%	64.9%	42.8%	63.5%	12.7%	12.3%
278	23.6%	49.5%	22.1%	47.5%	4.0%	3.0%
279	2.5%	6.2%	2.5%	6.1%	-1.7%	-1.7%
280	7.5%	15.7%	7.4%	15.5%	7.1%	7.1%
281	1.0%	2.9%	0.9%	2.9%	-0.3%	-0.4%
Total	10.9%	20.5%	7.8%	17.3%	25.4%	27.7%

DCE–Detection Controlled Estimation.

NMP–Net Misreporting Percentage is the ratio of the aggregate net misreported amount over the sum of the absolute values of the amounts that should have been reported.

NOP–Net Overclaim Percentage.

Allocation of Examination Resources by Activity Code

There has often been interest in seeing the estimates of the tax gap aligned with the allocation of examination resources. The IRS has historically used activity codes for examination planning and tracking of examination results. Table 6 shows how the allocation of the individual income tax plus SE tax underreporting tax gap by activity code aligns with the number of individual income tax closed examinations, the amount of examiner time allocated to those examinations and the estimated cost associated with that examiner time. The number of examinations and the associated examination time were calculated from the Audit Information Management System (AIMS) closed case data for Fiscal Year (FY) 2019.[9] The column Calendar Year 2018 returns is taken from the 2019 IRS Data Book Table 17b and reflects returns filed in CY 2018. The examination (exam) cost is an estimate based on applying an estimate of the hourly cost according to whether the exam was conducted by an RA, a TCO or through a campus correspondence exam. The hourly costs used to create this table were taken from Table 4 of Holtzblatt and McGuire (2020). For FY 2017, they estimated the following hourly costs:

- RA/Field: $72;
- TCO/Office: $52; and
- Correspondence: $43.[10]

The total exam cost is calculated as the total RA exam time multiplied by the RA hourly cost plus the same calculation for TCOs and correspondence exams.

[9] FY 2019 examinations and CY 2018 returns were chosen as a reference point because it is the last fiscal year in which IRS published "Examination Coverage: Recommended and Average Recommended Additional Tax After Examination—IRS Data Book Table 17b." Although microdata from AIMS was used for this analysis, the author of this paper was able to replicate the examination closed and additional recommended tax amounts from Table 17b using that microdata.

[10] Although there could be some differential wage growth by examination type between FY 2017 and FY 2019, these estimates and this time period seemed reasonable for approximating the distribution of costs.

EITC Examinations

Returns that claimed EITC (activity codes 270 and 271) accounted for 17.5 percent of returns filed in CY 2018, while accounting for 44.5 percent of exams closed for FY 2019. On the surface, this might seem like examination resources are disproportionately allocated to examining returns that claimed EITC. However, those numbers are misleading for several reasons. Returns that claimed EITC accounted for 28.3 percent of the individual income tax underreporting tax gap, significantly more than the population share of those returns. Table 6 also shows that nearly 18 percent of examination time and 15 percent of the estimated cost of FY 2019 closed exams is associated with returns that claimed EITC, about the same as the share of those examinations in the CY 2018 population (17.5 percent) and significantly less than their share of the individual income tax underreporting tax gap (28.3 percent).

The share of exams closed is a poor proxy for resource allocation because of the significantly different amount of time and the cost associated with that time between face-to-face examinations and correspondence examinations. Operational risk-based examinations of returns that claimed EITC are typically conducted through correspondence. Auditing eligibility for a credit (or deduction) requires substantiating the claim, meaning eligibility for the credit or deduction can often be established through the taxpayer providing substantiating documentation. It may be more burdensome for some taxpayers to have to come into an IRS office to meet with a TCO when the taxpayer can mail in the substantiating documentation. Sending an RA to meet with the taxpayer over an issue that can be addressed through the mail would pull that RA off more complex cases. Issues that can be readily substantiated through correspondence are likely most efficiently examined that way.

It can be misleading to include correspondence examinations that focus on substantiating a limited number of issues with field examinations (whether by an RA or in an office by a TCO). The correspondence examinations take significantly less time because those examinations generally require only explaining the issues to the taxpayer and reviewing the submitted documentation. The limited scope of the audit and the limited complexity of the returns means that the much of the process can be automated, requiring less examiner time. The process efficiencies and nature of correspondence examinations mean less time on the case and less experienced employees at lower pay grades can work those examinations. Those two factors, less time on the case and less costly time, mean that a relatively small allocation of examination resources can provide greater coverage of returns suspected of misreporting that issue.

Higher TPI Examinations

In general, examiner time and exam cost estimates show that the share of examination resources allocated to returns reporting TPI of $200,000 or more (activity codes 279, 280 and 281) are higher than their share of the population and their share of the tax gap. In particular, although returns reporting TPI of $1 million or more are only 0.4 percent of the CY 2018 returns filed, examinations of those returns accounted for 2.1 percent of all FY 2019 exams closed, 13.9 percent of exam time and 15.7 percent of the estimated exam cost. Those returns are estimated to contribute only 2.6 percent to the total individual income tax plus SE tax underreporting tax gap. Although not shown in Table 6, because the examinations of these returns are predominantly through field examinations, returns with TPI of $1 million or more account for an even greater share of field examination resources.

TABLE 6. Distribution of TYs 2011–2013 Individual Income Tax plus Self-Employment Tax

Underreporting Tax Gap Compared to FY 2019 Individual Income Tax Examinations Closed by Activity Code

Activity Code	Calendar Year 2018 Returns	TYs 2011–2013 Tax Gap		FY 2019 Exams Closed [1]		
		Examiner-Determined	DCE-Adjusted	Exams Closed	Exam Time	Exam Cost
270	16.2%	28.5%	20.3%	42.1%	12.8%	10.0%
271	1.3%	7.3%	8.0%	2.4%	5.0%	5.0%
272	54.9%	15.5%	12.5%	14.6%	8.2%	7.6%
273	10.3%	12.5%	13.3%	13.3%	11.9%	11.4%
274	7.6%	8.8%	12.0%	10.5%	8.7%	8.0%
275	2.3%	5.9%	8.0%	4.4%	5.4%	5.5%
276	0.6%	3.0%	3.6%	2.2%	4.9%	5.1%
277	0.5%	4.5%	5.0%	1.5%	6.3%	6.7%
278	0.8%	1.9%	2.9%	0.4%	1.1%	1.2%
279	3.7%	4.3%	5.3%	3.0%	7.5%	8.1%
280	1.5%	6.0%	6.5%	3.5%	14.3%	15.8%
281	0.4%	1.7%	2.6%	2.1%	13.9%	15.7%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[1] Includes exams of returns that were not filed.

4. Distribution of Individual Income Tax plus SE Tax Underreporting Tax Gap by TPI

Although activity codes provide detailed classifications of returns according to various return characteristics, they do not provide complete insight by level of income because only three levels of TPI are used to classify returns (less than $200,000; $200,000 to less than $1 million; $1 million and above).

Tables 7 shows the distribution of the individual income tax plus SE tax underreporting tax gap by two alternative levels of TPI_{TG}. The first three columns present the tax gap, share of the tax gap, and the NMP when taxpayers are arrayed by their reported TPI_{TG}. The second set of three columns presents those measures when taxpayers are arrayed by their estimated DCE-Adjusted TPI_{TG}. Table 7 includes estimates for taxpayers with reported TPI_{TG} between $5 million and under $10 million and for taxpayers with reported TPI_{TG} of $10 million or more. These results are included to align Table 7 with the FY 2020 IRS Data Book Tables 17a and 18. However, these particular income levels are subject to greater sampling and estimation error due to the relatively small number of taxpayers in the NRP sample at these income levels.

When taxpayers are arrayed by reported income, tax gap shares and misreporting rates are higher at the lower end of the income distribution. When taxpayers are arrayed by estimated true income, the shares of the individual income tax underreporting tax gap shift somewhat towards higher levels of estimated true TPI_{TG}. For example, taxpayers who reported TPI_{TG} under $50,000 account for 48.8 percent of the individual income tax plus SE tax underreporting . However, when arrayed by true TPI_{TG}, taxpayers who should have reported under $50,000 TPI_{TG} accounted for only 23.6 percent. Similarly, taxpayers who reported $1 million or more of TPI_{TG} accounted for 2.4 percent of the individual income tax plus SE tax underreporting tax gap while taxpayers who *should* have reported $1 million or more of TPI_{TG} accounted for 7.6 percent.[11]

In other words, taxpayers who underreport their income typically will fall into a lower level of income when arrayed by their reported income than had they been arrayed by their true income. Therefore, one

[11] Table 4 shows that the tax gap associated with activity code 281 (reported TPI of $1 million or more) is 2.6 percent and not 2.4 percent as shown in Table 7. This small difference is due to differences in the calculation of TPI during processing (Table 4) and the calculation of TPITG used for this analysis (Table 7).

cannot simply look at reported income to assess examination coverage by true income since taxpayers who misreport their income necessarily report themselves as being lower in the income distribution. It further follows that a policy focused on examining only those who self-report as high-income would potentially miss the most egregious noncompliance. Unfortunately, the IRS does not know a taxpayer's true income without first conducting an examination.

TABLE 7. Distribution of *DCE-Adjusted* TYs 2011–2013 Individual Income Tax plus Self-Employment Tax Underreporting Tax Gap: Arrayed by Level of Reported TPI_{TG} and DCE-Adjusted TPI_{TG}

TPITG Level	Arrayed by Reported TPI_{TG}			Arrayed by DCE-Adjusted TPI_{TG}		
	$ Billions	Share	NMP	$ Billions	Share	NMP
$0 or less	$1.8	0.6%	105.2%	$0.5	0.2%	[1]
$1 under $25,000	$82.1	28.3%	83.6%	$26.6	9.2%	55.2%
$25,000 under $50,000	$57.7	19.9%	46.4%	$41.3	14.2%	41.8%
$50,000 under $75,000	$36.9	12.7%	29.0%	$35.7	12.3%	30.2%
$75,000 under $100,000	$26.3	9.0%	21.1%	$27.2	9.3%	22.4%
$100,000 under $200,000	$48.3	16.6%	15.2%	$66.5	22.9%	19.9%
$200,000 under $500,000	$24.2	8.3%	9.6%	$51.9	17.9%	17.8%
$500,000 under $1,000,000	$6.3	2.2%	5.4%	$18.7	6.4%	14.2%
$1,000,000 under $5,000,000	$5.8	2.0%	3.8%	$17.0	5.8%	10.3%
$5,000,000 under $10,000,000	$0.8	0.3%	2.3%	$3.9	1.3%	9.9%
$10,000,000 or more	$0.2	0.1%	0.3%	$1.1	0.4%	1.7%
Total	$290.4	100.0%	20.5%	$290.4	100.0%	20.5%

[1] More than 1,000 percent

DCE–Detection Controlled Estimation

NMP–Net Misreporting Percentage is the ratio of the aggregate net misreported amount over the sum of the absolute values of the amounts that should have been reported.

TPITG–Total Positive Income (Tax Gap)

Table 8 shows similar information as Table 7, however, the estimates of the individual income tax plus SE tax underreporting tax gap in Table 8 are based on just the examiner-determined misreporting, without DCE estimates of undetected income. Table 8 tells a similar story as Table 7.

Allocation of Examination Resources by Total Positive Income

Although the IRS historically has used activity codes for examination planning and tracking purposes, starting with FY 2019, the IRS began tracking compliance presence in the Data Book by tax year and level of reported total positive income in Table 17 "Examination Coverage and Recommended Additional Tax After Examination, by Type and Size of Return, TYs 2010–2018" of the Data Book. Reporting by tax year has the benefit of aligning the examinations with the actual return population from which those examinations were selected. In other words, examination coverage for TY 2017 reflects the actual filing population from TY 2017 and the actual open and closed examinations of that TY 2017 filing population. The prior Data Book Table 17b tracked compliance presence by aligning calendar year returns filed (typically the year following the tax year of the return) and fiscal year examination closures.

TABLE 8. Distribution of *Examiner Determined* TYs 2011–2013 Individual Income Tax and Self-Employment Tax Underreporting Tax Gap: Arrayed by Level of Reported TPI_{TG} and Examiner Determined TPI_{TG}

TPITG Level	Arrayed by Reported TPI_{TG}			Arrayed by Examiner-Determined TPI_{TG}		
	$ Billions	Share	NMP	$ Billions	Share	NMP
$0 or less	$0.7	0.5%	116.1%	$0.6	0.5%	[1]
$1 under $25,000	$45.0	32.3%	65.4%	$27.3	19.6%	51.3%
$25,000 under $50,000	$30.2	21.7%	30.1%	$24.9	17.9%	27.0%
$50,000 under $75,000	$16.9	12.1%	15.7%	$17.2	12.4%	16.3%
$75,000 under $100,000	$11.0	7.9%	10.0%	$12.7	9.2%	11.6%
$100,000 under $200,000	$20.9	15.0%	7.2%	$27.4	19.7%	9.2%
$200,000 under $500,000	$9.6	6.9%	4.1%	$17.0	12.2%	6.9%
$500,000 under $1,000,000	$2.8	2.0%	2.4%	$4.9	3.5%	4.2%
$1,000,000 under $5,000,000	$1.8	1.3%	1.2%	$5.7	4.1%	3.8%
$5,000,000 under $10,000,000	$0.3	0.2%	0.9%	$0.9	0.6%	2.5%
$10,000,000 or more	[2]	[2]	[2]	$0.4	0.3%	0.5%
Total	$139.1	100.0%	10.9%	$139.1	100.0%	10.9%

[1] More than 1,000 percent.

[2] Less than 0.05 percent or $0.05 billion.

DCE–Detection Controlled Estimation.

NMP–Net Misreporting Percentage is the ratio of the aggregate net misreported amount over the sum of the absolute values of the amounts that should have been reported.

TPITG–Total Positive Income (Tax Gap).

One nuance with the prior Table 17b was that the exams closed in any given fiscal year are from returns that were filed in many prior fiscal years, not necessarily the calendar year of returns filed which were included in Table 17b. For large corporation income tax returns with relatively small populations and high coverage rates, this occasionally resulted in a coverage rate greater than 100 percent. The new Table 17 solves that problem but introduces a new problem in the process. The IRS can open examinations of returns from a given tax year over multiple years into the future. So at any given point in time, the number of examinations that are open does not necessarily reflect the total number of examinations that will open. Similarly, examinations can take years to close for complicated tax returns like those reporting high income. Therefore, we will not know the total number of examinations and the actual results of those examinations for many years after the tax year of those returns. In that regard, Table 17 provides a snapshot of one point in time for the tax year and a snapshot that could be highly distorted for the most recent tax years, especially for high-income tax returns.

Beginning with the FY 2020, the IRS Data Book, Table 18, Examination Coverage: Recommended Additional Tax, and Returns with Unagreed Additional Tax, After Examination, by Type and Size of Return also changed from reporting by activity code to reporting by level of reported TPI. However, Data Book Table 18 continued to report examination results for fiscal year closed exams, instead of breaking out examinations by tax year. Table 9 of this paper aligns TY 2018 returns filed with the TYs 2011–2013 individual income tax plus SE tax underreporting tax gap estimates and measures of examination resources associated with FY 2020 examination closures by level of reported TPI. The number of examinations and associated examination time were calculated from the AIMS closed case data for FY 2020 merged with data on TPI from the Individual Returns Transaction File (IRTF).[12]

Aligning the tax gap by TPI with examination resources by TPI tells a similar story as the activity code alignment. Focusing on examination closures does not provide a reliable picture of the allocation of

[12] Although microdata from AIMS was used for this analysis, the author of this paper was able to replicate the examination closed and additional recommended tax amounts with a small margin of error from FY 2020 Data Book, Table 18, "Examination Coverage: Recommended Additional Tax, and Returns with Unagreed Additional Tax, After Examination, by Type and Size of Return" using that microdata.

examination resources and how those resources are aligned to address the tax gap that spans across all levels of reported income. A greater share of examination resources in terms of examiner time and the cost associated with examiner time, relative to the share of the tax gap, is allocated to examining higher income taxpayers. This relationship is true for all levels of taxpayers reporting at least $100,000 of TPI.

An important note concerns Table 9 (and Tables 17 and 18 of the IRS Data Book). Nonfiler examinations of taxpayers who did not file a return fall into the $0 or less TPI level. The $0 or less TPI level in these table should not necessarily be interpreted as low income. The approximately 10 percent of total individual income tax examination time and cost associated with examining taxpayers in this category (inclusive of nonfilers) is consistent with the 11.5 percent share of the tax gap attributable to taxpayers who reported $0 or less of TPI ($0.7 billion) plus the individual income tax plus SE tax nonfiling tax gap ($37 billion) as a share of the individual income tax plus SE tax underreporting tax gap ($290 billion) plus the individual income tax plus SE tax nonfiling tax gap ($37 billion).[13]

TABLE 9. Distribution of TYs 2011–2013 Individual Income Tax and Self-Employment Tax Underreporting Tax Gap Compared to FY 2020 Examinations Closed by Reported TPI

Reported TPI Level [1]	TY 2018 Returns Filed	TYs 2011–2013 Tax Gap		FY 2020 Exams Closed [2]		
		Examiner-Determined	DCE-Adjusted	Exams Closed	Exam Time [3]	Exam Cost [3]
$0 or less[4]	0.4%	0.5%	0.6%	12.5%	9.7%	10.3%
$1 under $25,000	32.1%	32.3%	28.3%	36.7%	13.8%	11.7%
$25,000 under $50,000	23.8%	21.7%	19.9%	13.3%	9.5%	8.6%
$50,000 under $75,000	14.1%	12.1%	12.7%	8.4%	7.1%	6.7%
$75,000 under $100,000	9.1%	7.9%	9.0%	6.8%	6.7%	6.4%
$100,000 under $200,000	14.4%	15.0%	16.6%	13.6%	18.5%	18.5%
$200,000 under $500,000	4.8%	6.9%	8.3%	4.6%	13.5%	14.4%
$500,000 under $1,000,000	0.8%	2.0%	2.2%	1.7%	7.0%	7.6%
$1,000,000 under $5,000,000	0.4%	1.3%	2.0%	1.8%	9.8%	10.9%
$5,000,000 under $10,000,000	[5]	0.2%	0.3%	0.3%	1.6%	1.8%
$10,000,000 or more	[5]	[5]	0.1%	0.3%	2.7%	3.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[1] TPI in this table for TY 2018 returns filed and FY 2020 Exams Closed reflects the IRS calculation of reported TPI calculated during return processing. This may differ slightly from the tax gap tax calculator calculation of reported TPITG used for the TYs 2011–2013 Tax Gap estimates.

[2] Includes nonfiler examinations.

[3] The examination time and cost does not include the time spent examining related S corporation and partnership returns associated with the examined individual income tax return. In this regard, the share of examination time and cost associated with examining higher income returns may be understated to the extent that taxpayers at these income levels are more likely to be involved in examinations with related pickups of flowthrough entities. This is a hypothesis that needs to be confirmed with additional data analysis.

[4] About 95 percent of examinations closed in the $0 or less TPI level are nonfiler examinations, in other words, exams of taxpayers who did not file a return. These exams technically should be associated with the $37 billion TYs 2011–2013 individual income tax plus SE tax nonfiler tax gap, not the underreporting tax gap.

[5] Less than 0.05 percent or $0.05 billion.

5. Data Limitations and Concerns

The individual income tax underreporting tax gap estimates are subject to various types of errors, like all estimates. There are sampling errors, measurement errors and estimation errors. Measurement error with respect to the individual income tax underreporting tax gap estimates typically refers to the correctness and completeness of the examiner's determination of what should have been reported. Tax gap estimation assumes that the recommended adjustments made by the examiners are correct and appropriate, meaning that the examiners did not make adjustments that should not have been made during the examination. Tax gap estimates

[13] ($0.7 + $37.0)/($290.4 + $37.0) = 11.5 percent.

however do assume that there may be income that examiners did not detect that impacts the completeness of the examiner's determination.

Tax Gap estimation uses the DCE methodology to address the measurement error introduced by the possibility of undetected income. The DCE methodology produces microlevel estimates that are added to the examiner recommended adjustment which then become the final data used to estimate the individual income tax underreporting tax gap. The final estimates with undetected income range from two to four times as large as estimates based solely on what the examiner determined. The fact that more than half of the estimated individual income tax underreporting tax gap is attributed to estimates of undetected income has caused some to question whether the estimates might be too high.

More recently, there have been newly raised concerns that the tax gap estimates might actually be too low. This concern is typically expressed as a question as to whether specific issues or types of noncompliance are "included" in the individual income tax underreporting tax gap estimates. There isn't a simple answer to this question for a variety of reasons. The fact that a separate tax gap estimate of a specific issue is not possible or not currently available does not mean that the tax gap related to that issue is not accounted for in the tax gap estimates. In some situations, the data are not collected at the level of detail necessary for reporting on an issue. In other situations, the issue may be rare and therefore may not be sufficient data to provide an estimate with an acceptable level of precision. In both situations, the tax gap estimates likely appropriately account for the issue, meaning it is unlikely the estimates are significantly understated in some particular way.

There specifically have been arguments that the tax gap estimates are understated with respect to flow-through income (S corporation and partnership income) and income from offshore accounts because of the difficulty in detecting this type of misreporting. Given the substantial inclusion of estimated undetected income, it is possible that the current estimates reasonably account for misreported income from these sources. However, these questions and the associated hypotheses concerning these issues warrant further research. Tax gap estimation is focused on compliance measurement and this research should be consistent with the overall approach to tax gap estimation including the principle that underreporting tax gap estimates should be grounded in examiner recommended adjustments from examined tax returns. It is not sufficient to assume the level, rate, and incidence of noncompliance; the IRS should collect the compliance data necessary to confirm or refute these hypotheses.

Another area concerns the age of the tax gap estimates and emerging issues. For example, there has been tremendous growth in digital assets, both the market capitalization and transaction volume, since the TYs 2011–2013 time period reflected in the most recent tax gap estimates. While it is often assumed that the growth in digital assets necessarily means an equivalent growth in the tax gap, there are reasons to proceed cautiously. The holding of digital assets does not necessarily indicate the generation of income. Digital assets are treated as capital assets and subject to capital gains (or losses) when sold. At this point, it is not clear how well the total global market capitalization (for example), which is highly volatile and based on daily sales at the margin, translates into the domestic taxable income of the associated digital assets. In addition, there likely is a substitution effect occurring to some extent whereby taxpayers who are noncompliant with respect to income generated through digital asset related activities were potentially already noncompliant with respect to other income generated activities and simply shifted their investments (and noncompliance) from those other activities to digital assets.

There also is the potential for overlap between the use of digital assets and illegally sourced income. Illegally sourced income is generally outside the scope of the tax gap estimates, primarily because the overall goal of the government is to stop the illegal activity, not to tax it. However, some portion of the tax gap likely includes misreporting associated with illegal activities because taxpayers who engage in illegal activities potentially comingle their illegal activities with their legal activities and/or launder the money from those illegal activities. Although income generated by illegal activities is generally outside the scope of the tax gap estimates, it is plausible that some illegal activity is reflected in the tax gap estimates.

6. Conclusions

This paper focused on several topics related to the distribution of the individual income tax plus SE tax underreporting tax gap and the allocation of examination resources. One common theme of those topics is the importance of clarity in analysis and nuance when discussing compliance estimates and examination data. For example, the tax gap associated with taxpayers who claimed EITC is significantly higher than the tax gap associated with the EITC line item. These two concepts, the total tax gap associated with taxpayers who report a given line item versus the tax gap associated with just that line item, are often conflated, which leads to incorrect conclusions.

Continuing with the EITC topic, examination measures based on the number of exams closed are a poor proxy for the allocation of examination resources. Correspondence examinations and field examinations require significantly different levels of resources to complete. EITC examinations (primarily conducted via correspondence) require a relatively small share of examination resources compared to their share of the tax gap.

Similarly, discussions of the allocation of examination resources and the distribution of the individual income tax underreporting tax gap should differentiate between the reported income of taxpayers and the true income of taxpayers. Taxpayers who misreport their income are reporting their income to be lower in the reported income distribution. Examination strategies that focus exclusively on examining taxpayers based on their self-reported high income are likely to miss some of the most egregious noncompliance of taxpayers who have high incomes, but do not report that income on their tax returns. Analyses of the distribution of the tax gap by "true income" may not be very informative for where IRS should be allocating its resources because the IRS does not observe "true income."

The final conclusion concerns the fact that more than half of the estimated individual income tax plus SE tax underreporting tax gap is associated with income which is unlikely to be detected during an examination. If the misreporting is unlikely to be detected during an examination, then the tax gap estimates do not reflect the amount of revenue that could be obtained through a major expansion of examination coverage rates. The IRS has other methods for evaluating return on investment of compliance initiatives that are more appropriate for evaluating the likely revenue impact of those initiatives.

References

Bloomquist, Kim, Ed Emblom, Drew Johns, and Patrick Langetieg (2012). "Estimates of the Tax Year 2006 Individual Income Tax Underreporting Gap," Paper Presented at the 2012 IRS and the Urban-Brookings Tax Policy Center Research Conference.

B. Erard & Associates (2005). "IRS Tax Gap Estimation: Preliminary Results of Detection Controlled Analysis," PowerPoint presentation to Internal Revenue Service Office of Research, November 1, 2005.

B. Erard & Associates (2006). "Preliminary Econometric Results," Results summary report submitted to Internal Revenue Service Office of Research, January 27, 2006.

B. Erard & Associates (2007). "Adjustment of Income Tax Underreporting Using Detection Controlled Estimation," *Final Report for IRS Contract Number TIRNO-05-D-00050 0001*, November 15, 2007.

B. Erard & Associates (2014). "Econometric Support for Developing DCE Estimates Using NRP Data for TY 2006–2008," *Final Report for IRS Contract Number TIRNO-10-D-00021-D0003*, March 21, 2014.

B. Erard and Jonathan S. Feinstein (2012). "The Individual Income Reporting Gap: What We See and What We Don't," *IRS Research Bulletin, Publication 1500* (Rev. 2012).

Feinstein, Jonathan S. (1990). "Detection Controlled Estimation," *Journal of Law and Economics*, 33(1):233–276.

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

Holtzblatt, Janet, and Jamie McGuire (2020). "Effects of Recent Reductions in the Internal Revenue Service's Appropriations on Revenues," *IRS Research Bulletin, Publication 1500*, June 2020.

Internal Revenue Service (2017). *Estimation of the Underreporting Tax Gap for Tax Years 2008–2010: Methodology*, *Research, Applied Analytics & Statistics Technical Paper*, (3–2017).

Internal Revenue Service (2016). *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2008–2010*, IRS Publication 1415 (Rev. 5–2016), Washington, D.C.

Internal Revenue Service (2019). *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011–2013* IRS Publication 1415 (Rev. 9–2019), Washington, D.C.

Internal Revenue Service (2016). *Tax Gap Estimates for Tax Years 2008–2010.* April 2016.

Internal Revenue Service (2012). *Tax Gap for Tax Year 2006: Overview.* January 2012.

Internal Revenue Service (2011). *Tax Year 2006 Tax Gap Estimate: Summary of Estimation Methods.* January 2012.

Internal Revenue Service (2011). *Tax Gap "Map": Tax Year 2006.* December 2011.

Internal Revenue Service (2012), Office of Research. *Federal Tax Compliance Research: Tax Year 2006 Tax Gap Estimation*, March 2012.

Appendix 1. DCE Implementation for the TYs 2011–2013 Tax Gap Estimates

For the first time since the TY 2001 tax gap estimates, data contemporaneous with the tax gap estimates were used to estimate undetected unreported income. TYs 2011–2013 NRP data was available for DCE estimation to estimate undetected income. Actual line item predictions for each NRP return could then be used directly for tax gap estimation. NRP data from TYs 2008–2013 were pooled to do the DCE estimation for all line items except Schedule C and F. TYs 2006–2013 NRP data were pooled for Schedules C and F.

DCE estimation requires explicit modeling of a detection equation whose arguments include the type of examiner (TCO or RA), the experience of the examiner, and binary variables that take the value of 0 or 1 to indicate which examiner conducted the exam. To differentiate the detection capabilities of different examiners, the examiners included in the detection equation must have audited a sufficient number of returns with the income item being modeled. Typically, this requirement is 15 or more returns.

The DCE methodology includes a two-part specification for modeling the noncompliance of a line item. The first noncompliance equation modeled the likelihood of noncompliance while the second equation modeled the magnitude of noncompliance conditional on the presence of noncompliance. Since some income items with significant information reporting were not routinely classified, the extension also included additional modeling conditional on whether the line item was classified and on mismatches with information documents for these items.

The data requirements for DCE meant that some income items still needed to be grouped together for purposes of estimating the detection equation, even when using NRP data pooled across multiple years. Table A1 shows the specific groupings of income items used for estimation. Income items that were routinely classified (typically because of the lack of complete information reporting) were modeled separately from items subject to significant information reporting (wages, interest income, etc.). Schedules C and F income were primarily estimated independent of each other and of other routinely classified income items. Other routinely classified income items (capital gains, rental and royalty income, partnership and S corporation income, etc.) were estimated jointly with a common detection equation. Similarly, items that were not routinely classified (typically these items are subject to significant information reporting) were also estimated jointly with a common detection equation.

TABLE A1. Grouping of Income Items for Joint Estimation

Items Not Routinely Classified	Items Routinely Classified	
Estimated Jointly	***Estimated Jointly***	***Estimated Separately***
Wages and Salaries	Short-term Capital Gains	Schedule C
Interest	Long-term Capital Gains	Schedule F
Dividends	Rents and Royalties	
State and Local Tax Refunds	Part., S corporation, Estate, Other	
Pensions and IRAs	Form 4797, Net Gains	
Gross Social Security	Other Income	
Unemployment benefits		

The joint estimation of some line items with a common detection equation meant that the expanded methodology assumed that a given examiner had similar detection capabilities across all income items within the group. Noncompliance of each income item was modeled using separate equations and parameters even though detection was modeled using a common equation. In other words, the equations and parameters that modeled the likelihood and magnitude of noncompliance were not constrained to be identical across line items within a group while the detection equation and parameters were constrained. The second extension explicitly provided separate estimates of undetected income for each income item, a marked improvement over the first extension. Additionally, because different examiners may have examined different income items, the

overall average detection rates for a given line item could still vary within the group. Although separate detection equations would be preferred to the use of a common detection equation, there were simply not enough audits in the sample to support that estimation approach.

Simulation of Undetected Income

The DCE formula underlying the return level predictions predicts a positive probability of undetected income for most returns (though this is typically very small for returns where no unreported income was detected). Simply multiplying the predicted probability of undetected income by the predicted magnitude of undetected income would result in nearly every return receiving some positive amount of undetected income for each income item, but that would not produce a realistic distribution of undetected income. A small probability of undetected income for an income item actually means that undetected income would be present on a relatively small number of returns for that item. In order to have a more realistic allocation of undetected income, a simulation approach is used to apply the DCE prediction formulas. The simulation process randomly allocates undetected income for a given income item based on the probability of undetected income for that item on each return. The simulation is repeated ten times for each income item to create ten sets of pooled NRP TYs 2011–2013 NRP data with simulated undetected income.

Tax Calculator

To estimate underreported taxes resulting from the underreported income at the line item level, a tax calculator was applied to individual observations (i.e., tax returns) from the ten simulated TYs 2011–2013 NRP datasets. The tax calculator iteratively added (or subtracted) misreported income for each income item to the reported amount of income and tentative tax was calculated. The calculated tentative tax was then compared to the reported tentative tax to calculate the marginal tax gap associated with misreporting for that line item. Then that additional income was dropped, and the process repeated for the next income item. After income a similar iterative process was applied to deductions and exemptions. The tax gap associated with credits was estimated by calculating credits based on reported income, deductions, exemptions and filing status and then recalculating credits based on the amounts of those items that should have been reported.

This process provided ten underreporting tax gap estimates for each line item that were then averaged to produce the final underreporting tax gap estimate. The marginal tax rate used to estimate the tax gap associated with a given income line item is calculated holding all other line items at their reported amounts. This calculation understates the true marginal tax rate whenever more than one line item has been underreported on the same return and the combined underreporting results in a higher marginal tax rate than when the tax on the underreported amounts is calculated separately. The total individual income tax underreporting tax gap is calculated based on the marginal tax rates associated with all misreporting for a given return. The difference between the total individual income tax underreporting tax gap and the sum of the individual line item tax gaps is characterized as "unallocated marginal effects." These unallocated marginal effects reflect a portion of the individual income tax underreporting tax gap that is not allocated to a specific line item.

Self-Employment Taxes

Self-employment taxes are required to be reported by individuals with self-employment income on individual income tax returns. The underreporting of self-employment income (primarily income reported on Schedules C and F) results in underreported self-employment taxes. Each spouse on a joint return has a separate earned income threshold above which the combined wages and self-employment income are subject to Medicare taxes but not Social Security taxes. Undetected self-employment income (Schedules C and F) was allocated to the primary taxpayer and secondary taxpayer according to each taxpayer's respective share of self-employment income as determined by the examiner. Undetected wages, salaries, and tips were allocated similarly. The tax calculator then calculated the amount of self-employment taxes that should have been reported.

3

Improving Audit Outcomes: Thinking Inside the Box

Scott ♦ Fausey ♦ Jones ♦ Warner ♦ Ortiz

Plumley ♦ Rodriguez ♦ Nicholl

Automated Discovery of Tax Schemes[1] Using Genetic Algorithms[2]

Eric O. Scott,[3] Camrynn Fausey, Karen Jones, Geoff Warner (The MITRE Corporation), and Hahnemann Ortiz (IRS)

Introduction

The tax gap, the difference between the tax imposed by law and the tax paid on time for any given tax year, was recently estimated to exceed $600 billion in the United States (Sarin (2021)). A substantial portion of this non-compliance is due to tax-planning behavior comprising structured sequences of transactions that individually satisfy the letter of the law, but collectively are designed for the sole purpose of reducing tax liability by exploiting loopholes in the Internal Revenue Code (IRC). Current efforts to combat these tax-planning schemes are almost entirely reactive: the Internal Revenue Service (IRS) uses audit filters constructed from historical data to detect anomalous filing patterns, leading to inevitable multi-year delays in its enforcement response. The lag in data availability used to design audit filters makes it challenging for the IRS to respond to tax-planning behavior in a timely manner, further exacerbating the tax gap.

The goal of this research is to provide the IRS with a proactive, simulation-based approach (not reliant on tax return data) to anticipate the emergence of tax-planning schemes in the wake of changes to the IRC or its associated regulatory enforcement regime. Such an approach will enable the IRS to mitigate tax-planning behavior as it occurs, rather than merely reacting to circumstances after the fact. By anticipating changes in taxpayer behavior, IRS tax specialists will be able to react to these changes or modify the original tax policy itself, thereby reducing the tax dollars lost, and the time allocated to audit filter development years after a tax policy is instituted.

Simulation is one strategic approach to studying policy because simulation can be implemented in the absence of expensive randomized controlled trials or real-world implementation of policy. Simulation is of particular importance for critical systems requiring high reliability, such as tax policy (because tax policy has large implications for human lives and livelihoods). Simulation reduces reliance upon historical data by leveraging the myriad of cause-and-effect assumptions that establish how taxpaying entities (entities) relate to one another within the domain. These assumptions, typically collected in close collaboration with subject-matter experts (SMEs), can be combined with data to calibrate machine-learning models (models), which in turn allow a modeling team to ask predictive questions about "what if" scenarios.

In practice, however, simulations tend to be applicable only to the specific domain for which they were designed, and the resulting models often have many free parameters, making them difficult to calibrate by hand (Carrella (2021)). This technical complexity often limits the success of simulation projects, as the work of building a model that reflects a new domain often leaves few resources for designing a generalizable system that can be adapted to new assumptions or scenarios. To combat this challenge in simulation, we approach the tax-planning problem space with a generalized structure, enabling the use of the resultant model across tax policy domains in the future.

---

[1] Sequence of transactions between entities and/or actions by entities.

[2]

[3] Corresponding author: direct questions or comments to Eric Scott at escott@mitre.org. We thank Sanith Wijesinghe and Chris Elsner for their work constructing the preliminary version of the research design and analysis that was the foundation for this paper.

Background

The Large Business & International Problem Space

The taxation of large businesses is a complex and challenging tax enforcement domain. Each year the Large Business & International division (LB&I) of the IRS develops campaigns that involve priorities for training, resource deployment, metrics, and audit filters all aiming to increase taxpayer compliance in specific ways (IRS (2022b)). Examiners and compliance officers require a high degree of specialized training to analyze these areas. However, in recent years the IRS has suffered elevated attrition of these types of employees (IRS (2021)).

An important portion of these efforts focus on the difficult task of identifying taxpaying entities that have—either through a misunderstanding of the law or through unscrupulous motives—abused or failed to comply with the tax code.

Using Simulation and Genetic and Evolutionary Algorithms

Simulation is a powerful technique for studying behavior, such as complex tax-planning behavior, especially when combined with optimization algorithms (D'Auria *et al.* (2020)). When a simulation model only contains two or three free parameters, authors often tune these parameters by hand with the help of experts (Batty *et al.* (2003)). But as models become complex it becomes necessary to search in polynomial time the often exponentially large, nondeterministic decision space that results from integrating detailed assumptions about real-world processes.

Genetic and evolutionary algorithms are one of the main classes of heuristic algorithms able to handle the complex, combinatorial, and nondifferentiable search problems that arise in simulation applications. Evolutionary algorithms (EAs)—the best-known example of which is the genetic algorithm (GA) (Mitchell (1998))—use a process of variation and selection to learn from feedback and to search large solution spaces for high-performing configurations. As such, these computational methods bear a strong (if schematic) resemblance to the process of Darwinian evolution (De Jong (2006); see Bassett (2012) for an in-depth look at how EA theory intersects with the biological discipline of quantitative genetics). Like supervised machine learning and reinforcement learning (Sutton and Barto (2018)), GAs and EAs form one of several broad approaches to artificial intelligence that rely on an iterative feedback loop which uses a series of queries to the environment to solve complex problems in a very general way.

Modeling Taxpayer Behaviors

In prior research, Warner *et al.* (2015) developed a proof-of-concept simulation and GA capable of generating a known real estate scheme called Installment Sale Bogus Optional Basis Transaction (iBOB) (Warner *et al.* (2015); Hemberg *et al.* (2016)). In this scheme, the majority owner of a network of passthrough entities arranges a sequence of transactions in which the entity makes an IRC §754 election to step up the basis of an asset subsequently sold by that network. This early research proved that a computational approach based on GAs had the power to discover transaction sequences that reduce overall tax liability.

In this research, we implemented a general approach, extending the methodology used to generate iBOB to other areas of the IRC. Specifically, we considered behaviors that may have emerged in response to significant changes in the IRC introduced under the Tax Cuts and Jobs Act (TCJA) in 2018. We explored several mechanisms large businesses might use to reduce their liability under provisions related to BEAT (Base Erosion and Anti-Abuse Tax—see IRC §59A) and IRC §163(j) (limitations on business interest expense). BEAT aims to limit efforts by multinational corporations to shift income to foreign-related parties via deductible payments such as interest and royalties. IRC §163(j) aims to limit the amount of business interest expense (BIE) that businesses can deduct when their gross receipts exceed a particular threshold. Both changes to the IRC may have spurred businesses to adjust their earlier strategies of tax avoidance to remain technically compliant while exploiting new tax "gray areas."

Using similar methods as were used in the iBOB example, we extended the use of GAs into these changing domains of tax policy—where audit filters are limited or nonexistent. Thus, we hope to anticipate novel

taxpayer avoidance behavior before it occurs, assisting the IRS in the development of audit filters or other statistical techniques. Most importantly, we will develop a single generalized framework to simulate two different tax policy scenarios, so that the simulation model can adapt properly to new, previously unseen tax policy scenarios.

Modeling Risky Behaviors

Simulating the tendency of entities to be noncompliant with the tax code requires a variety of causal assumptions. Many of these are simple, for example, the details of computing various deductions and penalties on a corporate tax return. However, a complex set of assumptions arises in considering the risk that an entity faces when it chooses to engage in noncompliant behavior.

One way of quantifying the relationship between risk and decision-making is to assume that an entity will make decisions that maximize its reward in the expectation of certain outcomes. In influential studies of how corruption occurs in societies around the world, Klitgaard (1988 and 1998) expresses this assumption through the concept of corrupt gain. Under this assumption, an entity will engage in illicit behavior when the gain it expects to receive (i.e., on average) is greater than the expected value of the penalty:

$$E[\text{Corrupt Gain}] > E[\text{Penalty}] = \text{Penalty} \cdot \text{Probability of Being Detected} \quad (1)$$

This model and other expected-value models—such as Allingham and Sandmo's seminal analysis of tax evasion (Allingham and Sandmo (1972))—are useful starting points for understanding factors that go into decisions that rational but unscrupulous taxpaying entities may make. However, it is difficult to accurately model the likelihood of audit and detection in a taxpaying domain though this challenge is an active area of research in the field of behavioral economics and agent-based modeling (see Prickhardt and Prinz (2014) for a survey).

In this research, we introduce simple heuristic scoring models of risk into the scenarios, giving estimates of risk through assigning a fixed weight in the decision-making process. This simple approach circumvents the need to explicitly model or simulate the likelihood of detection at the cost of introducing strong assumptions about how entities evaluate the risk of a given decision.

Research Objectives

In this paper, the Automated Discovery of Tax Schemes simulation framework (simulation framework) informs a more detailed evolutionary behavior discovery (EBD) software framework (software framework) in the tax policy domain. The software framework focuses on representing business structures and actions they may take, and computing the tax consequences of those actions for generalization.

> **Research Objective 1**: Demonstrate the extensibility of the EBD by using it to model a variety of complex taxpayer scenarios.
>
> We illustrate, through a benchmark collection of detailed examples, that the EBD is flexible enough to model a diversity of complex scenarios. The benchmark scenarios involve recent and significant changes to the IRC, introduced under the TCJA and its ensuing regulations, particularly as applied to large business and international entities at the IRS.

Second, the hypothesis of this methodology is the claim that if we can understand the decision space available to entities in these simulations, then the resultant insights reduce the time it takes to develop—and increase the quality of—new policies, laws, and audit approaches for taxation. We therefore incorporated intelligent behaviors into the simulation framework presented here by including a heuristic search algorithm in the form of an *evolutionary behavior discovery* module. This component is based on GAs and EAs, which use a process of variation and selection to optimize an objective function defining the incentives of taxpaying entities (thereby bearing a strong resemblance to Darwinian evolution). This allows us to use a given simulation scenario to search for behaviors that one or more rational (but potentially unscrupulous) entities might take to maximize their advantages given various assumptions about the IRC and its implementation.

Research Objective 2: Validate the EBD by showing that it finds advantageous behaviors in a variety of complex taxpayer scenarios.

We performed initial validation of the EBD by showing that a simple EA that we designed for this domain can discover a sequence of actions in each scenario that a rational (but potentially noncompliant) entity might make in a similar real-world setting.

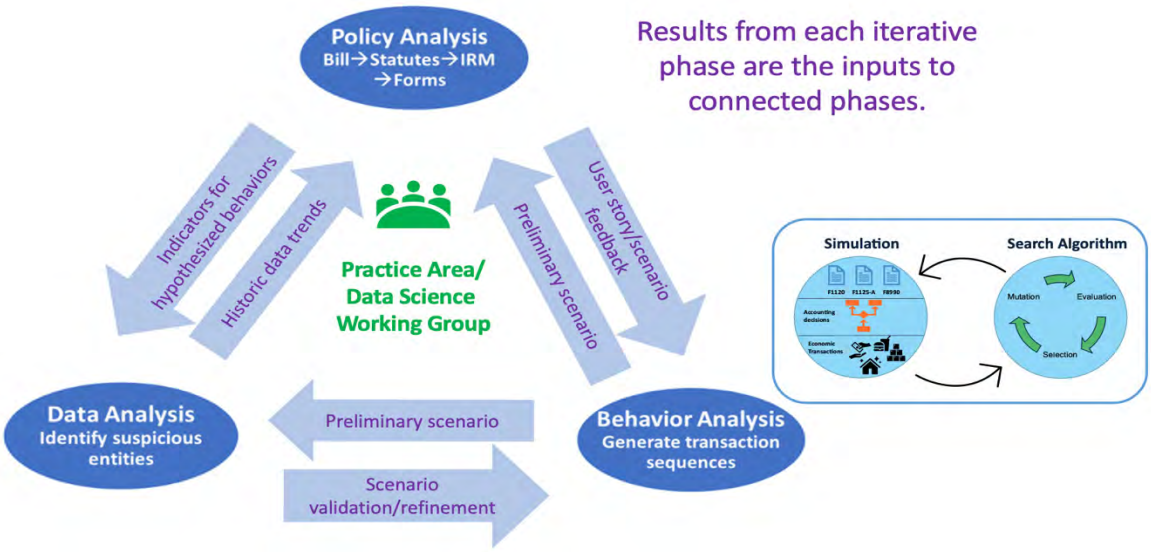
Methods

We used multiple frameworks to arrive at a generalized and extensible simulation, including an overarching simulation framework and an EBD. The overarching, three-phase simulation framework required working closely with IRS SMEs to gather model requirements and iteratively verify progress.

Simulation Framework

We implemented an overarching simulation framework consisting of three phases: Policy, Behavior, and Data Analyses (Figure 1), to gather and verify simulation requirements and drive development.

FIGURE 1. Simulation Framework: A Three-Phase Iterative Process to Elucidate Tax-Planning Behavior



Policy Analysis involves identifying a tax scenario of interest through analyzing relevant regulatory and compliance information to define a base scheme (a very simple model to exercise the framework in the target domain). The output of Policy Analysis is a set of taxpayer behavior hypotheses defined collaboratively between the SMEs and data scientists. We developed user stories to translate behavior hypotheses into model features (a capability requirement to incorporate into the simulation), written from the perspective of the end user. The user stories provided are then input for the Behavior Analysis phase to develop a model (e.g., search algorithm) that can identify and simulate the hypothesized sequence of transactions, as well as other scenario and tax-relevant decisions. The output of Behavior Analysis is a set of simulated actions and decisions of taxpayers, which is then input to Data Analysis, wherein the discovered behavior is verified and validated using any real-world data available. Additionally, SMEs iteratively verify and validate the analysis of resultant taxpayer transactions, structures, and other elements discovered in the Data Analysis phase.

While Policy Analysis and Data Analysis have been a part of our research, most of our work has been focused on the Behavior Analysis phase, and specifically the gains made in model generalization. Policy Analysis

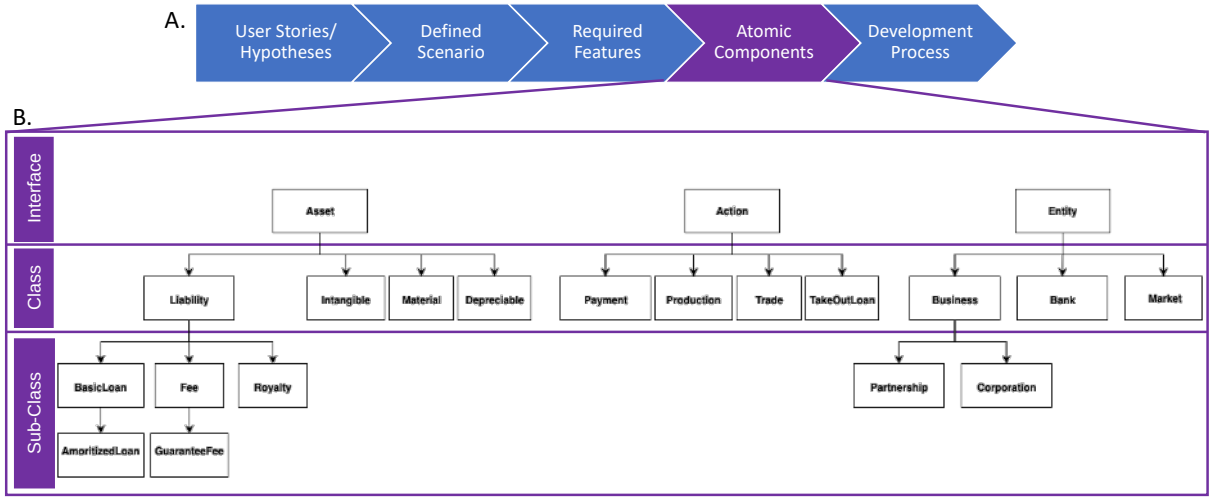
and Data Analysis will become a larger focus in future work. To verify and validate the observed behaviors from the models created in this work, we will need to complete the third and final step of the simulation framework—the Data Analysis phase.

Behavior Analysis Phase: Simulation Preparation

The Behavior Analysis phase follows five preliminary steps (Figure 2) to convert high-level hypothesized tax planning (a user story) from the Policy Analysis phase into discrete "building blocks" used as simulation input. These steps helped to generalize model development, ensuring extensibility to other tax policy scenarios of interest. First, we converted the original user story into a well-defined scenario of interest, or a detailed description of attributes and tax policy nuances involved. Next, we outlined new features, as well as the atomic components required to develop the new features (refer to Appendix B).

To simplify implementation of tax-planning behavior into any given simulation, we modularized and generalized the atomic components across scenarios of interest. These included Assets, Actions, and Entities illustrated by the class diagram in Figure 2, with any given Asset, Action, or Entity being able to be categorized into interfaces, classes, and subclasses that inherit attributes used consistently throughout the simulation. Various atomic component combinations can define new desired tax policy scenarios.

FIGURE 2. Simulation Preparation Diagram*



* A. Five preliminary steps converting high-level hypothesized tax planning into discrete building blocks used as simulation input.
B. Class Diagram of Assets, Actions, and Entities used for simulation development when converting User Stories to model mechanics.

Actions act as the "verbs" within the simulation framework, and they act upon the Assets and Entities (the "nouns"). In most cases, a fixed sequence of actions are executed on a given focus Entity during a simulation. In some cases, however, a single Action's execution may generate additional supporting actions—for example, a *TakeOutLoan* action might generate an auxiliary *Payment* action as a guarantee fee to a parent corporation to arrange a lower-interest loan.

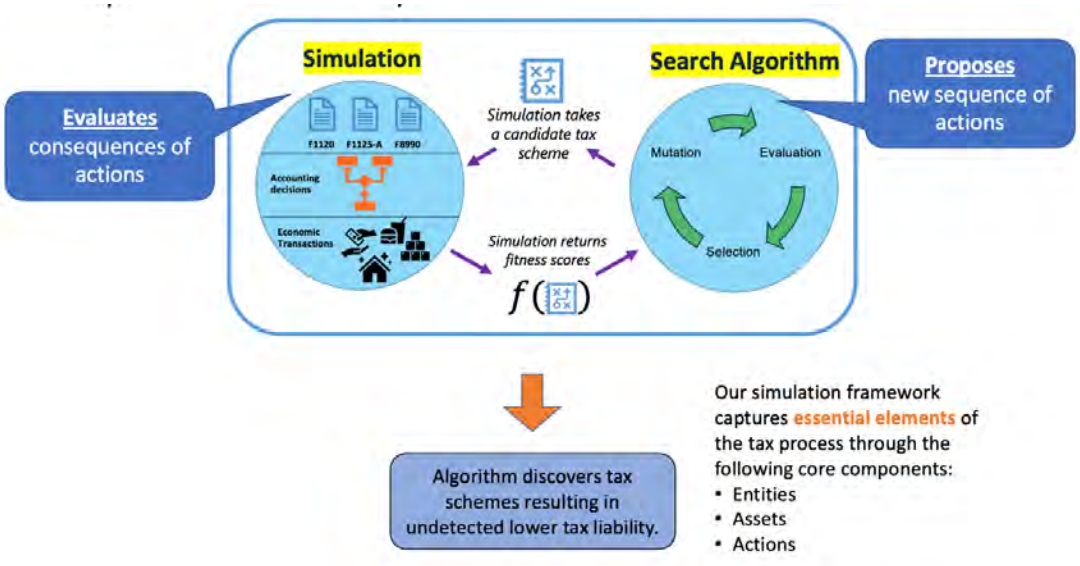
In addition to Assets, Actions, and Entities, a fourth important class of components (not shown in Figure 2) is *ProductionRules*. These govern the circumstances under which an Entity can take Production actions that create new Asset objects out of existing ones—defining the inputs and outputs of, for example, a manufacturing process.

Having discrete components enables a developer to follow a relatively straightforward process to extend or augment the simulation. With a generalized structure, new simulation feature development can scale efficiently.

Behavior Analysis Phase: Evolutionary Behavior Discovery

The Behavior Analysis Phase iteratively uses the EBD software framework to determine the most advantageous taxpayer behavior. The EBD is made up of a GA, which proposes sequences of taxpayer actions, and a simulation model, which evaluates the action sequences (see Figure 3). These two components are arranged as a nested loop: on each iteration, the GA (outer loop) calls the simulation (inner loop) a number of times to evaluate the quality of candidate solutions.

FIGURE 3. Evolutionary Behavior Discovery Software Framework*



* The search algorithm and simulation work together to discover tax schemes resulting in undetected lower tax liability.

The role of the simulation is to take a candidate tax scheme as input, and to assign it a quality score as output. This score is based on the financial consequences an entity incurs after following the given scheme. Within the simulation model, a candidate scheme is decoded into a series of transactions to be executed. Transactions include the exchange of user-defined goods and services within a network of financial entities. Possible entities include partnerships, corporations, banks, and individual taxpayers, while traded items include tangible and intangible assets, loans, contracts, and cash. When the transactions are completed, the simulation then executes applicable accounting rules, including elections or other decisions made by the taxpayer, and applies them to the computation of taxable income generated by the simulated economic activity. The taxable income and other taxpayer incentives (e.g., not being "detected by auditors") define the fitness or objective (a numeric value used to compare the success possibility of a scheme from the taxpayer's perspective, where a high score equates to a low audit risk and low tax liability, while a low score equates to a high audit risk and high tax liability) of each chromosome. Any given chromosome is deterministic, in that it will always result in the same economic outcome.

The search algorithm, which acts as the outer loop in the EBD, treats the simulation as a "fitness function": a black-box procedure that takes solutions as input and outputs a scalar quality score. The search algorithm uses a GA that begins with an initial population of randomly generated data structures, referred to as "chromosomes," each of which represents a proposed sequence of actions that a taxpayer might follow. The algorithm proceeds to evaluate the quality ("fitness") of each chromosome by executing the simulation (which returns a scalar quality score). It then uses a set of reproductive operations—mutation and recombination (wherein pairs of chromosomes are combined to produce novel offspring)—to modify the population of data structures. In this process, only those chromosomes with the highest fitness are chosen for the next iteration. As the algorithm runs, it finds solutions of better and better quality.

An essential design step in applying GAs and EAs is choosing an appropriate representation for the chromosomes. In our setting here, a solution encodes a sequence of actions. For example, a taxpayer may first take out a loan (using the *TakeOutLoan* action), then use the funds to purchase raw materials (*Trade*), engage in production (*Production*), sell the resulting product (*Trade*), and then use the proceeds to pay down the loan (*Payment*). Schemes for reducing tax liability often involve many economic actions of this type, each of which may have various arguments (the *Trade* action, for instance, requires an asset, buyer, and seller to be specified). This requires a means of representing these complex sequences of parameterized actions in a form that is easy to manipulate with mutation and recombination operators. This can be accomplished directly by modeling the action sequence as a program built from a sequence of nested functions—a route that Warner *et al*. (2015) followed in their approach to tax modeling (which is based on a kind of evolutionary program representation known as grammatical evolution).

In this work, we opt for a simple representation in which action sequences are represented by a vector of integer genes. Each integer wholly defines a single action and its parameters. This is accomplished by using the integer as a random seed to a sequence of decision values that define the action and parameters. Evolution thus proceeds by altering the vector of integral random seeds—namely by using integer-reset mutation (which randomly replaces a given integer with a newly sampled integer) to alter the seeds, and two-point crossover to combine individuals. A clear limitation of this representation is that the effects of mutation are highly random: altering a gene that represents a random seed effectively re-randomizes the decision values that it represents. The advantage, however, is that it allows defined actions whose behavior depends on the state of the simulation: the GA provides the simulation only with streams of decision values; what those values mean and how they are used to influence an entity's decision can be determined by the simulation at runtime. For example, an entity may attempt to buy a particular asset, but if no other entities are willing to sell that asset, the entity may back-track and attempt to buy a different asset instead.

The simulation and search algorithm phases continue to iterate and evolve until the fitness value reaches a steady state, upon which a final output is produced. The final output is a report that presents the most advantageous taxpayer behaviors that the simulation can discover—with an explanation of values that would appear on the taxpayer's return, as well as details of how they were able to minimize liability through financial transactions.

Our current research simulates a base tax ecosystem and verifies that the results match expected real-world taxpayer behaviors. A verifiable base tax ecosystem is a minimum requirement for later studying and discovering tax schemes. The base tax ecosystem is not limited to the tax policy scenarios presented; the modular simulation design enables simplified application of the simulation to other tax policy scenarios of interest. Additional component categories of actions, assets, and entities will enable the simulation to produce more realistic, detailed, nuanced, and informative results. The results illustrate how we used the modular components within the simulation's design to model a tax ecosystem comprising multiple tax policy scenarios.

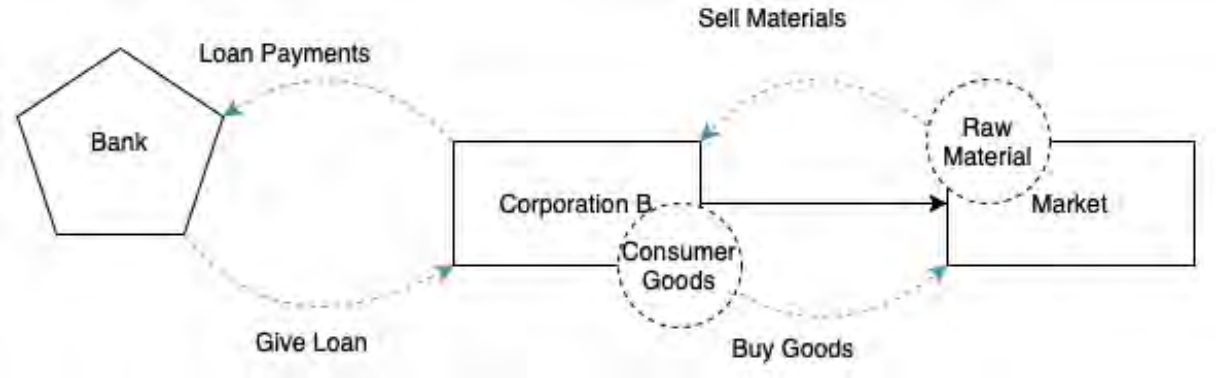
Benchmark Scenarios

The following section describes a collection of detailed examples—or benchmarks—that demonstrate the flexibility of EBD to model aspects of the two complex BEAT and IRC §163(j) scenarios. Each scenario involved recent and significant changes to the IRC introduced under the TCJA and its ensuing regulations, particularly as applied to large business and international taxpayers. Each benchmark represents important attributes of each scenario, and progress in the scenario's level of complexity.

Simple-Interest Scenario

The first and simplest scenario is a "simple interest" scenario, in which a single corporation can take out a loan from a bank to finance the production and sale of some (abstract) goods for a profit (see Figure 4).

FIGURE 4. Tax Organization Chart of the Simple-Interest Scenario



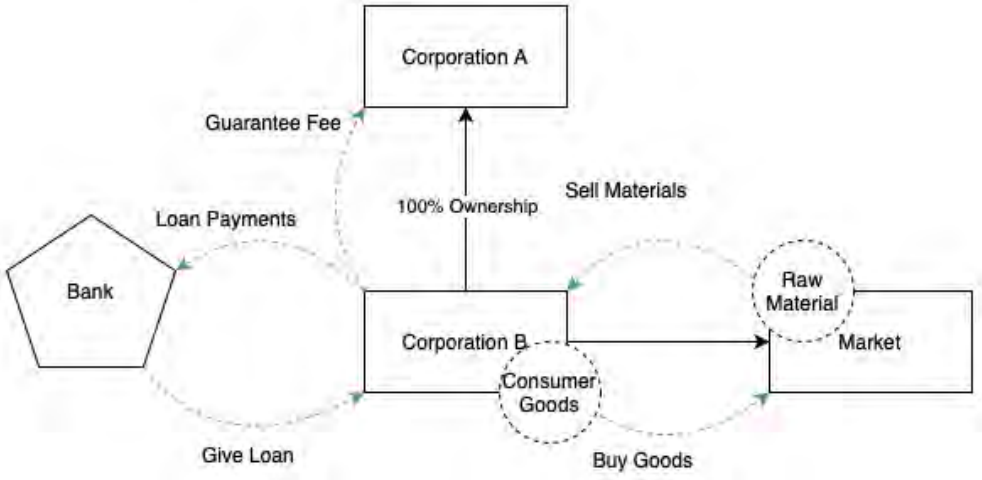
For a successful design, the simulation should model the interactions between a corporation and a market (the corporation can acquire raw materials from the market and sell processed goods to the market at some profit margin), and between a corporation and a bank (which provides funds). From a behavioral perspective, it is advantageous for the corporation to take out a loan from the bank to purchase raw materials and sell the processed goods that result in a gross profit, which it can then use to make payments on the loan. We maximize an appropriate objective function (an initial research question provided as a mathematical representation of a taxpayer's priorities) for this scenario to test the evolutionary behavior discovery approach to ensure its ability to discover this sequence of actions.

Guarantee-Fee Scenario

The second scenario in our benchmark introduces an alternative for a corporation to secure a loan. In this scenario, there are two corporations—A and B—that are related to one another in a parent-subsidiary structure as shown in Figure 5. Corporation B can acquire raw materials from the Market and sell goods to the Market that the corporation produces (just as in the simple-interest scenario above).

To fund its activities, Corporation B must take out a loan from a bank. It can do this by taking out a loan directly, in which case it will receive a loan with an interest rate of 5 percent. Alternatively, Corporation B can opt to pay a guarantee fee to Corporation A, in which case Corporation A will guarantee the loan, allowing the bank to offer a lower interest rate of 3 percent.

FIGURE 5. Tax Organization Chart of the Guarantee-Fee Scenario



This scenario is of particular interest because it introduces an opportunity for noncompliant behavior by the entities that was not available in the simple-interest scenario. Assume that Corporation B meets the criteria of IRC §163(j)—and therefore has a limit on the amount of deductible BIE. The guarantee fee effectively allows Corporation B to "buy down" the interest rate on the loan, which means it will have less overall BIE. Assuming that the deductible BIE amount was previously limited, if this new decrease in the overall business expense places Corporation B below its IRC §163(j) limitation, then it is now able to deduct other BIE that was previously disallowed or limited.

Assume that Corporation B expects to pay $5,000 interest total on the loan if it takes the 5-percent interest rate. Alternatively, it can obtain the 3 percent interest rate by paying Corporation A $2,000. In the latter case, Corporation B will still pay a total of $5,000. In this case, there is no direct financial difference between the two choices. However, by paying $2,000 as a guarantee fee rather than as BIE, Corporation B decreases its total BIE, and if Corporation B is now below its calculated IRC §163(j) limit, it can now deduct additional BIE generated through its other activities that previously would not have been deductible, thereby lowering its overall tax liability.

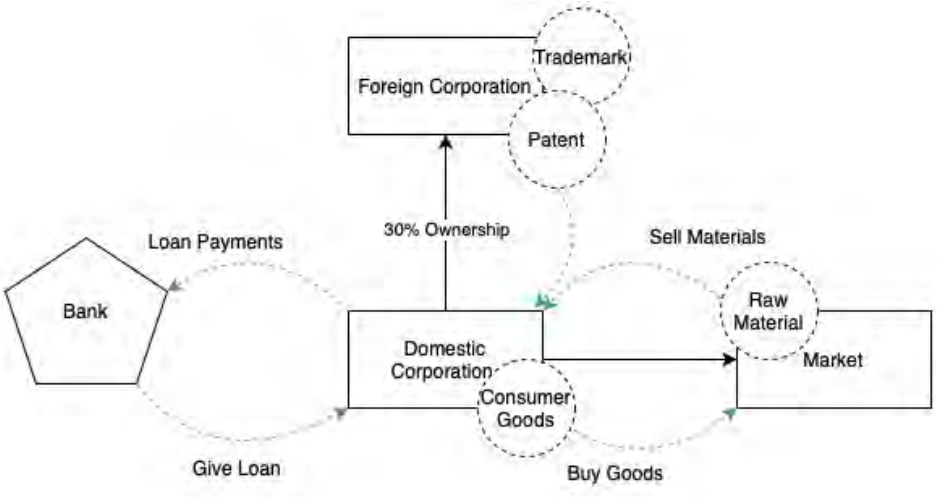
Base-Erosion Anti-Abuse Tax Scenario

The third scenario builds on the simple interest and guarantee fee scenarios by introducing royalty payments as more complex transactions. As in the guarantee-fee scenario, there are two corporations, but now with a change in nationality—A (Foreign Corporation) and B (Domestic Corporation)—that are related to one another in a parent-subsidiary structure as shown in Figure 6. The domestic corporation can trade with the Market to buy materials and sell produced goods (just as in the simple-interest and guarantee-fee scenarios above).

This scenario, however, adds dynamics and entity decisions that directly relate to the BEAT from IRC §59A. The TCJA introduced this provision to discourage large companies (>$500M of gross receipts) from avoiding tax liability by moving profits across national boundaries. In particular, BEAT is an additional tax that applies to certain payments—such as interest or royalties—that occur between domestic and international companies.

In the BEAT scenario, a domestic corporation engages in business by producing consumer goods and trading with a market. A foreign corporation partly owns the domestic corporation as shown in Figure 6 and also owns a trademark and a patent that the domestic corporation can use to create and sell more profitable consumer goods. The domestic corporation may enter into a royalty agreement with the foreign corporation, which supports the domestic corporation's production and profits, but also potentially moves some of its (otherwise taxable) profits offshore as a result.

FIGURE 6. Tax Organization Chart of the BEAT Scenario



This scenario exercises complex transaction types—namely the use of royalty payments to secure intellectual property—and elements of the tax code (computing the BEAT on the resulting transactions).

This scenario also invokes more nuanced decisions available to the domestic corporation surrounding transaction classification. The extra tax from BEAT applies only to certain tax-deductible transactions (which IRC §59A treats as base-erosion tax benefits (BETBs)), for which corporations have some latitude in determining transaction classification. In particular, under certain detailed circumstances, a corporation may choose to classify an offshore royalty payment as part of the routine cost of goods sold (COGS)—i.e., the expenses incurred in producing and distributing goods. When a corporation classifies a transaction as COGS, the transaction is no longer considered a BETB and may reduce the corporation's tax liability under BEAT. While this "reclassification" may be legitimate under well-documented circumstances, pursuing this set of actions improperly may be noncompliant and thus increase the risk of audit.

The nuanced decision-making that the taxpayer made in the BEAT scenario validates the use of EAs through the EBD. The BEAT scenario required us to model the ability of a corporation not just to engage in financially profitable transactions, but also to balance its financial goals against the risk of being found noncompliant with a complex section of the IRC. The use of EAs paired with objective functions is a strategic method to search the space of possible decisions with competing incentive structures.

Results

We accomplished the two research objectives by examining multiple benchmark scenarios. The first objective was to demonstrate the extensibility of the EBD by using it to model a variety of complex taxpayer scenarios; the second was to validate the EBD by showing that it finds advantageous behaviors in a variety of complex taxpayer scenarios.

Recall that EBD built scenarios out of three main types of objects: Actions, Assets, and Entities. In all scenarios, one run of the simulation completes two main stages:

> First, the model executed a **sequence of financial transactions** (Actions) which affect the state of the objects in the simulation—changing cash values, book entries, the ownership of certain Assets, etc.
>
> Second, the model **computed the tax implications** from the simulation state, illustrating how the various financial transactions impact line items on a simulated corporate tax return (e.g., a Form 1120).

The subsequent sections will describe each benchmark scenario, highlighting how it demonstrates the extensibility of the EBD and finds advantageous sequences of Actions for the taxpayer.

Simple-Interest Scenario

The simple-interest scenario from Figure 4 is the most basic of the benchmark scenarios, demonstrating the essential machinery of business transactions and associated tax returns. It involved three straightforward Entities: a *CCorporation* (C Corporations are taxed separately from their owners, as differentiated from S Corporations, which are not taxed separately from their owners), a *Bank*, and a *Market*. The *CCorporation* is the entity of interest (i.e., the entity driving the actions taken in the simulation).

CCorporation can perform the following *Actions* in this scenario:

- *Trade*: The entity spends cash to purchase an asset from another entity that is able to sell that asset.
- *TakeOutLoan*: The entity secures a loan of a certain amount from a bank entity that is able to loan.
- *Payment*: The entity makes a cash payment of a certain amount on a certain loan, reducing its liability.

CCorporation has access to these types of *Assets* in this scenario:

- *Loan*: represents a particular debt and its current outstanding value.

- *RawMaterial*: an asset that can be used as input to produce a *ConsumerGood*.
- *ConsumerGood*: an asset that can be produced from a *RawMaterial* and then sold with a positive profit margin.

These Entities, Actions, and Assets complete the formal description of the organizational structure depicted in Figure 4.

To meet the objective of expressing a full, executable version of the scenario in EBD, we defined rules for "producing goods," expressing the profit margins to the entity of interest from buying and selling goods. For the benchmark scenarios listed herein, the production rules outline that an entity converts raw material (*RawMaterial*, purchased from the market) into consumer goods (*ConsumerGood*, sold to the market) at a given profit margin (*ProfitMargin*: 0.8, equivalent to an 80-percent markup of the end consumer good sold in the market, relative to the raw material cost).

We tested the EBD in the simple-interest scenario by imposing a fitness function made up of a linear combination of four terms. The corporation was rewarded for maximizing its cash-on-hand at the end of the simulation (e.g., by selling consumer goods at a profit), and for minimizing its tax liability and outstanding debt (e.g., loans). A penalty term was included that discouraged the EA from creating action sequences with a large number of "dead actions,"[4] quantitatively, expressed as:

$$f(x) = c(x) - t(x) - 4l(x) - 10^6 \cdot I_{(dead \geq 85\%)}(x) \quad (2)$$

Where:

x = sequence of actions (behavior)

$c(x)$ = cash the corporation has on hand at the end of the simulation

$t(x)$ = amount of tax owed on its F1120 (U.S. Corporation Income Tax Return)

$l(x)$ = amount of outstanding liabilities (after considering all payments made)

$I_{(dead \geq 85\%)}(x)$ = a binary variable that returns 1 if more than 85 percent of the actions in x are "dead actions," otherwise zero.

Note that we weighted the liability term $l(x)$ such that outstanding debt is four times more undesirable than tax liability. This incentivized the corporation to pay off debts relatively quickly.

Table 1 shows an example of a behavior discovered by the EA using these incentives. The actions in this table show how the GA successfully incentivized the entity to use a loan to finance the purchase of raw materials, to produce consumer goods from those materials, and to sell them for a profit, all while making regular payments on the loan. Results of this kind serve to validate our Research Objective 2 by demonstrating the EBD's ability to find rational behaviors that one would expect an entity to engage in in this simple scenario.

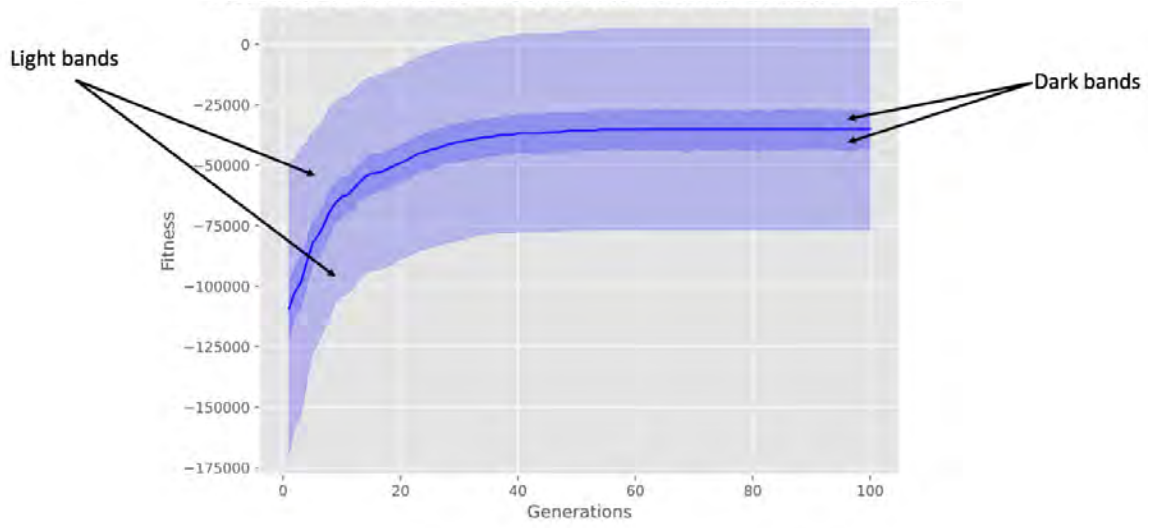
It is worth noting that because the EA is a stochastic optimization procedure, the output of the behavior discovery process may differ with each run. To demonstrate this, we ran the algorithm 100 times on the simple-interest scenario (with each run having 100 generations), each time with a different random seed but otherwise identical configuration. Figure 7 shows the results in terms of the average of the best fitness value obtained at each generation, which values are negative in large part because of the loans being taken out to drive economic activity (loans create a largely weighted negative term in Equation 2 above). There is considerable variance in the best final fitness value achieved (i.e., at generation 100).

[4] An action that cannot be completed in the context in which it is attempted—for example, attempting to sell a good that is not owned.

TABLE 1. Example of a Sequence of Actions That Was Found by Evolutionary Behavior Discovery in the Simple-Interest Scenario

Action	Target Entity	Parameters
TakeOutLoan	Bank	$100,000, interest rate 0.05
Payment	Bank	$10,500
Payment	Bank	$10,500
Payment	Bank	$10,500
Payment	Bank	$10,500
Payment	Bank	$10,500
Payment	Bank	$10,500
Trade	Market	Buy RawMaterial for $8,000
Payment	Bank	$10,500
Payment	Bank	$10,500
Produce	—	ConsumerGood
Trade	Market	Sell ConsumerGood for $14,400
Trade	Market	Buy RawMaterial for $8,000
Payment	Bank	$10,500
Produce	—	ConsumerGood
Trade	Market	Sell ConsumerGood for $14,400
Payment	Bank	$10,500

FIGURE 7. Simple Interest Mean Best-So-Far Fitness With 95% Confidence and Standard Deviation*



* Averaged over 100 independent runs of the evolutionary algorithm on the simple-interest scenario (i.e., by optimizing Equation (1)). Dark shading indicates 95-percent confidence of the mean, while the lighter shading indicates the standard deviation—showing that different runs converge to very divergent fitness values.

Guarantee-Fee Scenario

The guarantee-fee scenario is similar to the simple-interest scenario, except that the bank configuration includes an option for an entity to take out a loan in two different ways: as a standard loan or as a guaranteed loan.

The guaranteed loan posed a design challenge because it required a single decision among three Entity objects: Corporation B must request a loan from the Bank and offer a guarantee fee to Corporation A simultaneously. We resolved this by allowing the *TakeOutLoan* action itself to create a new *Payment* action when used to create a guaranteed loan. When *TakeOutLoan* creates a guaranteed loan, it takes a parameter that allows any Entity that owns the Entity requesting the loan to act as a guarantor (in this scenario, since Corporation A owns Corporation B, it will select Corporation A as the guarantor). This modification to the *TakeOutLoan* logic, and to the types of loans the Bank allows Corporation B to take out, is sufficient to express the guarantee-fee scenario.

We tested the EBD on this scenario using the following fitness function. The only difference between this fitness function and the one used in the simple-interest scenario (Equation (2)) is a lower weight on the liabilities term $l(x)$. This incentivized the algorithm to focus less on paying off debt than in the previous scenario, and more on minimizing tax liability.

$$f(x) = c(x) - t(x) - 1.5l(x) - 10^6 \cdot I_{dead \geq 85\%}(x) \quad (3)$$

From an algorithmic standpoint, de-emphasizing the liabilities term through its reduced weight in this scenario incentivizes behaviors that are more debt-reliant. This said, it is quite likely the algorithm would have eventually identified similar, if not identical, action sequences even if the same scenario would have used the fitness function given in Equation (2). In this scenario, the lower weight in Equation (3) might best be considered a "nudge" away from debt-minimization solutions that expedite the algorithm's search for more cash-generating action sequences. The process of determining weights in a fitness function is, in many ways, an art and not a science, which underscores the importance of dialogue with SMEs who have insight into the relative importance of the various financial drivers that motivate reporting behaviors.

Table 2 provides an example of a behavior discovered by the EA using Equation (3). In this scenario, Corporation B took the simultaneous actions of taking out a loan (*TakeOutLoan*: $100,000 guaranteed loan at 3-percent interest) and making a payment (*Payment*: $2,000 guarantee fee) with the Bank and Corporation A, respectively. After obtaining $100,000 in capital, Corporation B then took part in trade and production with the Market and made payments to the Bank to begin paying off Corporation B's loan. These results illustrate that, given the current simulation parameters, Corporation B chooses a guarantee-fee loan option over a simple loan on the basis of the combination of incentives expressed in Equation (3).

TABLE 2. Example of a Sequence of Actions Found by Evolutionary Behavior Discovery in the Guarantee-Fee Scenario

Action	Target Entity	Parameters
TakeOutLoan	Bank	$100,000 guaranteed loan at 3% interest
Payment	Corporation A	Pay a $2,000 guarantee fee
Trade	Market	Buy RawMaterial for $5,000
Produce	—	Produce a ConsumerGood
Trade	Market	Sell ConsumerGood for $15,000
Trade	Market	Buy RawMaterial for $5,000
Produce	—	Produce a ConsumerGood
Trade	Market	Sell ConsumerGood for $15,000
Payment	Bank	$10,000
Trade	Market	Buy RawMaterial for $5,000
Payment	Bank	$10,000

Base-Erosion Anti-Abuse Tax (BEAT) Scenario

The BEAT scenario continues the theme of *CCorporation* that interacts with a (foreign) parent corporation. There are four Entity objects in this scenario—a *Bank*, a *Market*, and two *CCorporations*. This scenario also introduced additional components to support two additional modeling goals.

First, we modeled the features necessary for the focus Entity (the domestic *CCorporation*) to increase its profit by licensing Assets from the foreign *CCorporation*. We used two specialized *IntangibleAsset* objects to achieve this, representing the intellectual property of the foreign related entity: a trademark and a patent.

The significance of the patent is that it allowed the production of the *ConsumerGood*. If the trademark was available, moreover, the value of the produced *ConsumerGood* increased. We implemented this logic using two specific *ProductionRules*: one that allows a *RawMaterial* and a *Patent* to produce a *ConsumerGood* at a profit margin of 20 percent, and one that allows *RawMaterial,* a *Patent,* and a *Trademark* to produce a *ConsumerGood* at a profit margin of 40 percent.

To allow the domestic corporation to license use of the foreign corporation's *IntangibleAsset*, we configured the *Trade* action to allow Entities to enter into a royalty contract together. This created and transferred a *RoyaltyAsset* object, which tracked the amount of royalty payments due to the owner of the intellectual property. In this scenario, we configured trademarks to be licensed for a 15-percent share in the revenue generated, while the patent is configured for a 4-percent revenue share.

The second modeling priority involved the tax consequences of BEAT. As discussed above, corporations can sometimes alter the amount of BEAT liability owed by reclassifying certain transactions that are traditionally considered deductions. In the implementation of this scenario, we gave the focus *CCorporation* the option of taking a special action that reclassifies certain deductions of royalties and license fees as COGS at tax-filing time.

The fitness function includes financial factors and a risk term. We evaluated each action sequence x's quality by running the simulation for three simulated years (treated as 2016–2018). The fitness was then calculated at the end of those three years, with the following linear combination:

$$f(x) = \Delta c(x) - t(x) - 2r(x) - g(x) - 10^7 \cdot I_{dead \geq 40\%}(x) - risk(x) \quad (4)$$

Where:

$\Delta c(x)$ = change in the domestic *CCorporation's* cash on hand

$t(x)$ = total taxes it paid

$r(x)$ = a combination of 1) a penalty value for large year-over-year deviations in corporate tax return line items and 2) total royalty fees it paid (i.e., to the foreign *CCorporation*), see Equation 5

$g(x)$ = value of its final inventory (discouraging the corporation from holding on to unsold goods)

$I_{\text{dead}\geq 40\%}(x)$ = a binary variable that penalizes the solution if it has more than 40 percent "dead" actions.

New in this scenario is an explicit inclusion of a risk model *risk(x)*. Risk is the sum of two components, each of which combines some rough aspects of the possible penalties and the probability of detection that an entity incurs (see Equation (1)) when it engages in unexpected or nontraditional reporting behaviors. The first adds a large value to the penalty if at least one line item in the *CCorporation's* corporate tax return (F1120) deviates by greater than 10 percent from the same line item in the prior year's return. The second adds a value proportional to an estimate of the amount of BEAT tax that the entity may have avoided paying—added only if a sudden change occurs from year to year in the ratio of the domestic *CCorporation's* cost of goods sold (COGS) to its gross receipts:

$$risk(x) = max\left(0, I_{\Delta item > 10\%} \cdot 10^6 + I_{\Delta \frac{COGS}{GR} > 10\%} \cdot potential\_beat\_avoided\right) \quad (5)$$

Table 3 shows an example of an action sequence the EBD found when it was run using the incentives in Equation (4). The first action to notice is that the algorithm successfully navigated the complexity of establishing royalty contracts first (which increases the value of the *ConsumerGood* assets produced and sold). Later in the simulation, this contract created a liability as the corporation made a profit from the *ConsumerGoods* it created with the licensed intellectual property. The corporation responds to this by paying the royalty fees as a fraction of its revenue: $1,050 for the trademark on each sale (15 percent for a $7,000 sale), and $280 for the patent (4 percent of $7,000).

Table 3. Example of a Sequence of Actions That Was Found by Evolutionary Behavior Discovery in the BEAT Scenario

Action	Target Entity	Parameters
Trade Payment	Foreign Corporation Foreign Corporation	Royalty Contract 1 $1,050
Trade Payment	Foreign Corporation Foreign Corporation	Royalty Contract 2 $280
Trade	Market	Buy RawMaterial for $5,000
Produce	—	Produce a ConsumerGood
Trade	Market	Sell ConsumerGood for $7,000
Trade	Market	Buy RawMaterial for $5,000
Trade	Market	Buy RawMaterial for $5,000
Produce	—	Produce a ConsumerGood
Trade	Market	Sell ConsumerGood for $7,000
Produce	—	Produce a ConsumerGood
Trade	Market	Sell ConsumerGood for $7,000
Payment	Foreign Corporation	$2,100 on Royalty 1
Payment	Foreign Corporation	$560 on Royalty 2

The final two actions in the sequence indicate that the focus corporation did choose to reclass some of its base-erosion-related transactions as COGS. The algorithm appears to have found that this reclassing was an "acceptable" risk with respect to Equation (5)—though it is worth noting that the solution returned by the EBD process only reclasses a subset of the royalty payments.

Discussion

By modeling a variety of complex taxpayer scenarios that built upon one another, *we achieved our first research objective of demonstrating the extensibility of the EBD*. The reuse of features resulted in generalization across scenarios, applicable to future tax policy. Additionally, the insights that resulted after each implementation reduced the time required to develop subsequent simulations. Iteration reduces the time needed to develop new features and scenarios to test tax-planning behavior hypotheses.

An additional and critical benefit of this research is the development of reusable common resources such as output report templates, experimental notebooks, and user story templates. These resources promote faster implementation with each future scenario developed.

By finding advantageous behaviors in a variety of complex taxpayer scenarios, we also achieved our second research objective of preliminarily validating the EBD. We demonstrated how to explore and improve decision space understanding for entities in simulations. We proved that a simple EA designed for a tax scenario can discover a sequence of actions that a rational (but potentially noncompliant) entity might make in a similar real-world setting.

We envision the EBD as a discovery method for understanding how taxpayer behavior might evolve in response to new or proposed laws—enabling examiners to concentrate their time efficiently.

Limitations

While the current research presents a flexible approach to the discovery of novel tax schemes, the transaction space is vast, and GAs tend to be computationally inefficient. This limitation could be mitigated by exploiting the native parallelism of GAs (provided that one is willing to use more computing power), and by using tax domain knowledge to prune nonviable solutions from the search space. This latter goal can be achieved by choosing representations that encode the various domain areas sparingly and accurately. For example, within the grammatical evolution paradigm developed by Warner *et al.* (2015) to generate iBOB-like schemes, the authors applied domain knowledge to tailor the context-free grammar so that it tended to produce transaction sequences that spanned the relevant search space without straying into nonviable regions of that space.

Developing the domain knowledge necessary to construct efficient GAs requires a significant investment of time and effort. Further, integrating disparate tax domains into a reasonably efficient and coherent whole poses a challenging design problem. The current simulation framework provides a high-level structure for realizing these goals, but it cannot substitute for the intricate interplay between tax domain knowledge and canny design.

We made progress toward the application of this simulation framework to the IRC §163(j) and BEAT domains; however, several key mechanics remain incomplete or do not yet accurately represent the larger tax ecosystem. For example, although we implemented partnership entities, these entities are currently unable to perform any transactions; also, any transactions involving depreciable assets, or deductions related to depreciation of assets, are not yet possible. These capabilities are central to several planned experiments whose performance awaits completion of the associated code base. Additionally, current weightings of variables in the fitness functions have not yet been validated and should be tested using sensitivity analysis during validation in the next Data Analysis phase of the simulation framework. Through sensitivity analysis (e.g., observing the changes in outcome from adjusting the weightings in singularity and combination), we will be able to quantify the relative importance of each variable's weight and thereby ensure that those weights are realistic.

Future Research

We modeled multiple scenarios within the Policy and Behavior Analysis phases of the simulation framework; however, to verify and validate the observed behaviors from these models, we will need to complete the third and final step of the simulation framework—the Data Analysis phase. While the GA may discover potential avenues of tax-planning behavior, these results will need validation and verification using real-world data.

We plan to augment the IRC §163(j) and BEAT scenarios with new features and extend the simulation framework to new tax policy scenarios. With the current two scenarios being augmented, each scenario will better represent the complex tax ecosystem and may be more easily validated in the Data Analysis phase. During extension of the simulation framework to other tax policy scenarios, current and future objects in the simulation can be refactored to arrive at greater levels of generalization. As the simulation framework is further extended, the objects will generalize better, with future extensions specified solely using configuration files.

A fundamental shortcoming of the expected-value model expressed in Equation (1) and the explicit $risk(x)$ penalty used in the BEAT scenario (Equation 5) is that they do not account for the variation in risk tolerance that different entities may have. There is an opportunity here, then, to combine simulations of the tax ecosystem with multi-objective evolutionary optimization algorithms in the future to better explore the Pareto curve of risk-reward tradeoffs inherently associated with any noncompliant decision-making process (Deb *et al.* (2002); Konak *et al.* (2006)).

Finally, to enable future development of the code base, we plan to provide documentation catered for three different roles that interact with the simulation framework: the "end user," an "intermediate user," and a "fundamental developer." The archetype of an end user is a business analyst who defines business problems and uses insight provided by the simulation framework to answer complex questions (e.g., what types of taxpayers might be most likely to be engaging in tax-planning behavior?). This end user would not make changes to the code base, but will use information from the simulations to inform business-level decisions. The archetype of

an intermediate user is a data scientist with expertise in machine learning to make small changes to the code base for sensitivity analysis or "what-if" analyses. The intermediate user should be able to make small changes to current simulations or develop new features, but would not be expected to extend the simulation framework to other tax policy scenarios of interest. The archetype of a fundamental developer is a software developer who has a deep understanding of the code base for refactoring. The fundamental developer can extend the simulation framework to other tax policy scenarios of interest. The documentation provided to each of these three users would include instructions for each role to achieve their desired outcomes.

References

Allingham, M., and Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics* 1(1972), 323–338.

Bassett, J. K. (2012). *Methods for Improving the Design and Performance of Evolutionary Algorithms* [Doctoral thesis, George Mason University]. George Mason University. ProQuest. https://www.proquest.com/openview/65a882a80e8b4eee5bc05f9b4bb01bc1

Batty, M., Desyllas, J., & Duxbury, E. (2003). Safety in Numbers? Modelling Crowds and Designing Control for the Notting Hill Carnival. *Urban Studies*, 40(8), 1573–1590.

Carrella, E. (2021). No Free Lunch When Estimating Simulation Parameters. *Journal of Artificial Societies and Social Simulation,* 24(2), 7. http://jasss.soc.surrey.ac.uk/24/2/7.html.

D'Auria, M., Scott, E. O., Lather, R. S., Hilty, J., & Luke, S. (2020, October). Assisted Parameter and Behavior Calibration in Agent-Based Models with Distributed Optimization. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (pp. 93–105). Springer, Cham.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.

De Jong, K. (2006). *Evolutionary Computation: A Unified Approach*. MIT Press.

Dodge, J. M., & Soled, J. A. (2005). Inflated Tax Basis and the Quarter-Trillion-Dollar Revenue Question. *Tax Notes*, 106, 453–462.

Hemberg, E., Rosen, J., Warner, G., Wijesinghe, S., & O'Reilly, U. M. (2016). Detecting Tax Evasion: A Co-Evolutionary Approach. *Artificial Intelligence and Law*, 24(2), 149–182.

IRS (2021). IRS Update on Audits. https://www.irs.gov/newsroom/irs-update-on-audits. Accessed April 17, 2022.

IRS (2022a). The Agency, Its Mission and Statutory Authority. https://www.irs.gov/about-irs/the-agency-its-mission-and-statutory-authority. Accessed April 16, 2022.

IRS (2022b). Large Business and International Compliance Campaigns. https://www.irs.gov/businesses/large-business-and-international-compliance-campaigns. Accessed April 17, 2022.

Klitgaard, R. (1988). Controlling Corruption. In *Controlling Corruption*. University of California Press.

Klitgaard, R. (1998). Strategies Against Corruption. Presentation at Agencia Española de Cooperación Internacional Foro Iberoamericano Sobre El Combate a la Corrupción, Santa Cruz de la Sierra, June 15–16.

Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-objective Optimization Using Genetic Algorithms: A Tutorial. *Reliability Engineering & System Safety*, 91(9), 992–1007.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press.

Prickhardt, M., & Prinz, A. (2014). Behavioral Dynamics of Tax Evasion—A Survey. *Journal of Economic Psychology* (2014), 1–19.

Sarin, N. (2021). The Case for a Robust Attack on the Tax Gap. U.S. Department of the Treasury website, https://home.treasury.gov/news/featured-stories/the-case-for-a-robust-attack-on-the-tax-gap. Accessed April 16, 2022.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Treasury (2022). 2022 Annual Audit Plan. Treasury Inspector General for Tax Administration. https://www.treasury.gov/tigta/auditplans/auditplans_fy2022.pdf. Accessed April 17, 2022.

Warner, G., Wijesinghe, S., Marques, U., Badar, O., Rosen, J., Hemberg, E., & O'Reilly, U. M. (2015). Modeling Tax Evasion with Genetic Algorithms. *Economics of Governance*, 16(2), 165–178.

Appendix A: Acronyms

Acronym	Definition
ABC	Abstract Base Class
AuDiTS	Automated Discovery of Tax Schemes simulation framework
BEAT	Base Erosion and Anti-Abuse Tax
BETB	Base-Erosion Tax Benefit
BIE	Business Interest Expense
COGS	Cost of Goods Sold
EA	Evolutionary Algorithm
EBD	Evolutionary Behavior Discovery software framework
GA	Genetic Algorithm
GR	Gross Receipts
iBOB	Installment Sale Bogus Optional Basis Transaction
IRC	Internal Revenue Code
IRS	Internal Revenue Service
LB&I	Large Business and International
SME	Subject-Matter Expert
TCJA	Tax Cuts and Jobs Act
U.S.	United States

Appendix B: Atomic Components List and Descriptions

By simulating two tax policy scenarios of interest, we developed modular components that can be restructured and specified in unique ways to represent multiple tax policy scenarios of interest. By applying these modular components across scenarios, we achieved Objective 1: Demonstrate the extensibility of the EBD by using it to model a variety of complex taxpayer scenarios.

The components used include:

Operators: Manage the chromosome state. Writing a new Operator allows for new forms of chromosomes.

Controller: Dictates actions performed in the scenario. As actions become more generalized, there may be no need for a *Controller*.

Accountant: Used to record actions into the "book" (maps result actions to a format that can be used to file simulated tax returns). We chose to represent only portions of the IRC specific to the scenario under consideration.

Tax forms: Any scenario can use the tax forms and turn off unwanted portions of the IRC using "feature flags." However, these feature flags remove broad rather than granular portions of the IRC and were added because the current implementation of some tax forms is fragile and tightly coupled with the form of the "book." Future generalization development efforts will relax these assumptions to create a more robust tax form structure.

Entities: New entities can be created easily by subclassing the *Entity* abstract base class (ABC) or one of the more specific ABCs (e.g., *Business*). All entities are initialized using an "entity config" (a JSON document) that specifies their properties. The entity implementation is responsible for validating/parsing the entity config it receives, which makes it easy to extend an entity implementation when needed. In the context of this research, we use the term "entity" in place of the term "agent" to minimize confusion between a simulation agent and an IRS revenue agent.

Assets: The asset classes are organized into a hierarchy using inheritance. New asset types can be created as an ABC that subclasses the *Assets* ABC. Specific implementations of an asset type can then be easily created (e.g., *Loan AmortizedLoan*).

Actions: The core implementation of each action is contained in a generalized class but currently, each scenario *Controller* needs to write a "set-up" method for each of the possible actions in the scenario (this method ultimately makes a call to the action class).

Operationalizing the Indirect Effect of Audits

Alan Plumley and Daniel Rodriguez (IRS, RAAS) and Leigh Nicholl (The MITRE Corporation)

The Internal Revenue Service (IRS) has long believed that after taxpayers are audited, many become more compliant in subsequent years. This effect is called the *specific indirect effect* of the audits. It is *indirect* in contrast with the direct effect of the audits—any additional tax paid or refunded for the tax year that was audited (because the auditor made adjustments to that return). The effect is *specific* in the sense that it describes the behavior of the specific taxpayer who was audited, in contrast with a possible *general* indirect effect that audits may have on those who were *not* audited. In recent years, the IRS and MITRE have estimated the specific indirect effect associated with many categories of correspondence audit and have expressed these effects as the average additional amount of tax reported by audited taxpayers in the five years[1] following the audit that they would not have reported had they not been audited.

The IRS has often drawn attention in budget discussions to the likelihood that indirect effects are both real and sizable, but without plausible empirical estimates there is no way to take these effects into account operationally.[2] This paper is a first look at how the IRS could make use of such estimates in its operational decision-making. Because of the nature of the data and the estimation methodology, these estimates of the specific indirect effect are far more appropriate for characterizing the behavior of audited taxpayers as a group (e.g., those whose audits focused on the same issues) rather than for estimating the subsequent behavior of any particular taxpayer. Therefore, these estimates can be taken into account when deciding how much of the budget to allocate to each category of audit (resource allocation), but not when deciding which return to audit within a category (case selection).

The paper provides an overview of the indirect effect estimates that we developed for all correspondence audit categories,[3] then provides a simulation to illustrate how they can be combined with the direct effect to maximize the total impact of the audits on tax revenue. Because we found that the indirect effect varies across audit categories, the budget allocation that maximizes total (direct + indirect) revenue is different from the allocation that seeks to maximize only the direct revenue. This means that when indirect effects are taken into account in the budget allocation, direct revenue will *decrease* relative to the allocation that maximizes direct revenue alone. However, the sum of direct and indirect revenue will be largest when the allocation takes both into account.

This research has been motivated in part by recommendations made by the Government Accountability Office, which has urged the IRS to estimate and account for indirect effects in its operations.[4] When indirect effects *are* taken into account for allocating budget resources, we need to be prepared for some new possibilities. For example, what if the indirect effect in a given category is negative? What if it is far larger in one category than in the others? The paper explores these and other practical considerations likely to be faced by operational decision-makers in the near future.

The paper is organized as follows. Section 1 outlines the theory underlying optimal resource allocation when considering only the direct effect of the audits, which we can observe. Section 2 shows how indirect effects (which we must estimate) can be added to this framework and how that changes the optimal allocation and outcomes. Section 3 concludes by offering practical suggestions for handling indirect effects that are negative, extremely large, or impractical to estimate.

---

[1] Our analysis shows that the effect attenuates over time.

[2] The IRS has been forced to allocate resources based on the direct effect alone, subject to subjective minimum coverage constraints. Minimum audit coverage in categories with a low direct effect is considered necessary to maintain compliance. However, if all indirect effects could be taken into account explicitly and quantitatively, minimum coverage constraints would be unnecessary.

[3] Indirect effect estimates are needed for all categories that compete against each other for resources; otherwise the categories that don't take indirect effects into account would typically be at a competitive disadvantage.

[4] GAO (2012).

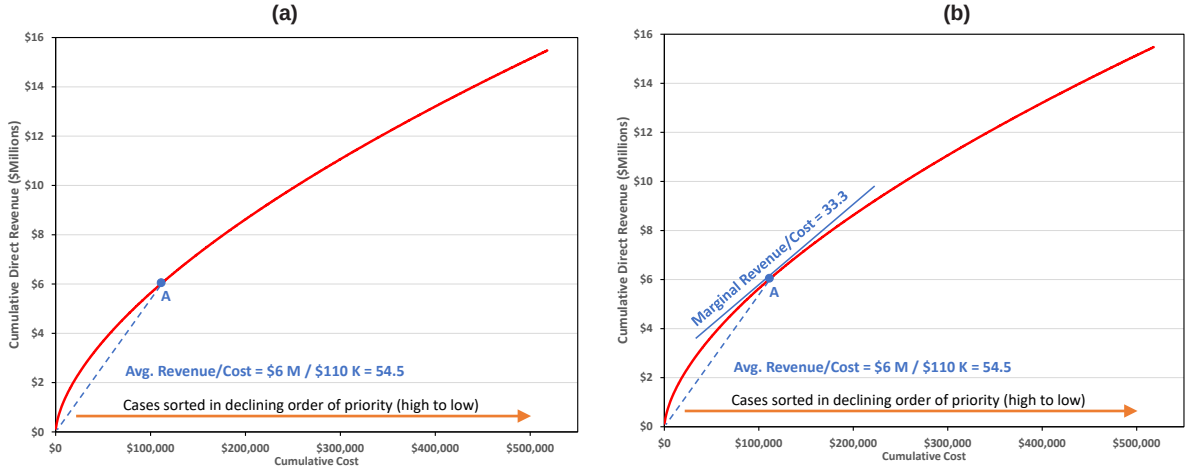
1. Optimal Resource Allocation

The IRS needs to allocate its annual budget across a wide range of service and enforcement categories. This allocation determines how much of a given activity will be conducted during the year. If the budget were unlimited, it would be simple to fund the activities as much as needed. But because the budget is limited, tough choices need to be made. To identify the "optimal" allocation of the budget, we need to be clear on what our ultimate objective is. For the purposes of this paper, the ultimate objective of the IRS is to *maximize the total tax revenue collected*, subject to the budget constraint.[5] Total tax revenue includes both enforced revenue (the direct effect of enforcement) and what taxpayers pay voluntarily (including the indirect effects of the enforcement and of other IRS activities).

Let's begin with a simple world in which there are no indirect effects; that is, the revenue to be maximized is just enforcement revenue. But even in that simple world, the revenue potential of potential cases within a given enforcement categories is not uniform; some cases have greater revenue potential than others. If the IRS selected cases at random, then the average revenue in a particular category wouldn't depend on how much of the budget were allocated to that category, and we'd allocate the budget first to the categories with the highest average revenue potential. But the IRS generally selects enforcement cases based on each case's revenue potential—the higher the revenue potential, the more likely the case will be selected for enforcement. To the extent that the IRS can successfully identify the revenue potential of candidate cases and selects them for enforcement in declining order of that revenue potential, it will collect more revenue than under a random selection strategy. However, two cases in the same category may use different amounts of the limited budget to produce the same revenue. In that case, we'd generate more revenue overall by prioritizing the case with the lesser cost over the case with the higher cost; the cost saving can be applied to the next most cost-effective case to generate more revenue.

Figure 1 illustrates the actual relationship between cumulative revenue and cumulative cost in a typical correspondence audit category. Because the axes represent cumulative concepts, the cases are sorted in declining priority order from left to right on the X-axis; the highest priority cases are on the left end of the curve and the lowest priority cases are on the right. Note that the curve is nonlinear, curving downward. That demonstrates that cases are selected in this category better than randomly (which would produce a linear relationship between revenue and cost). At any point A on the curve, the average revenue/cost is the slope of the line from the origin to point A (illustrated in panel (a)). Because the line is curved downward, the average declines as the budget expended increases.

FIGURE 1. Cumulative Revenue vs. Cumulative Cost Plot for a Typical Audit Category

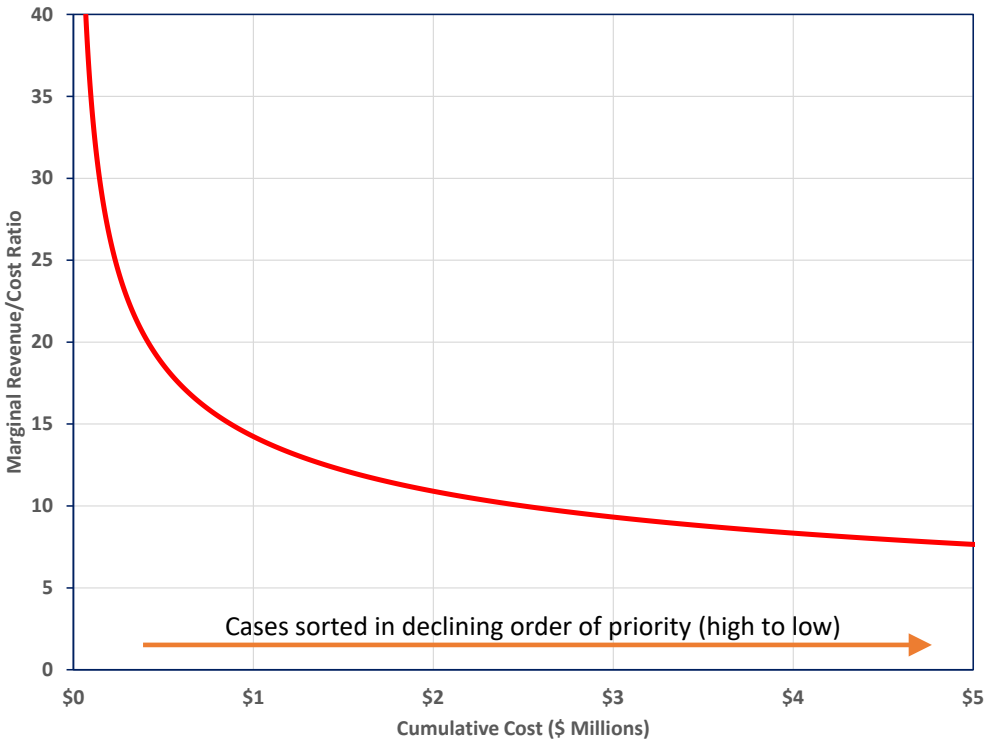


[5] The IRS exists to collect taxes. Even the services it provides are designed to help taxpayers pay their tax obligations fully and timely. Most alternative or supplemental objectives that have been proposed (e.g., minimizing taxpayer compliance burden, maximizing "fairness," or minimizing the enforcement no-change rate) are really *means* to the end of maximizing revenue or possibly even additional constraints on the maximization of revenue.

Therefore, when deciding how to allocate the budget, what's important is how the *marginal* revenue/cost ratio (the slope of the curve) varies with cumulative cost (illustrated in panel (b)). It tells us at any given point on the curve how much additional (or less) revenue to expect if one more (or less) dollar of budget were allocated to this category. Because in this category the highest priority cases (on the left) are more cost-effective than the lower priority cases (on the right), the marginal revenue/cost ratio declines as more budget is allocated, and it is smaller than the average revenue/cost at any given budget level. The marginal revenue/cost ratio can be plotted directly as a function of cumulative cost (budget), as shown in Figure 2. Remember that the revenue/cost ratio plotted at any level of cumulative cost is merely the slope of the corresponding cumulative revenue plot at that same level of cost (Figure 1(b)).

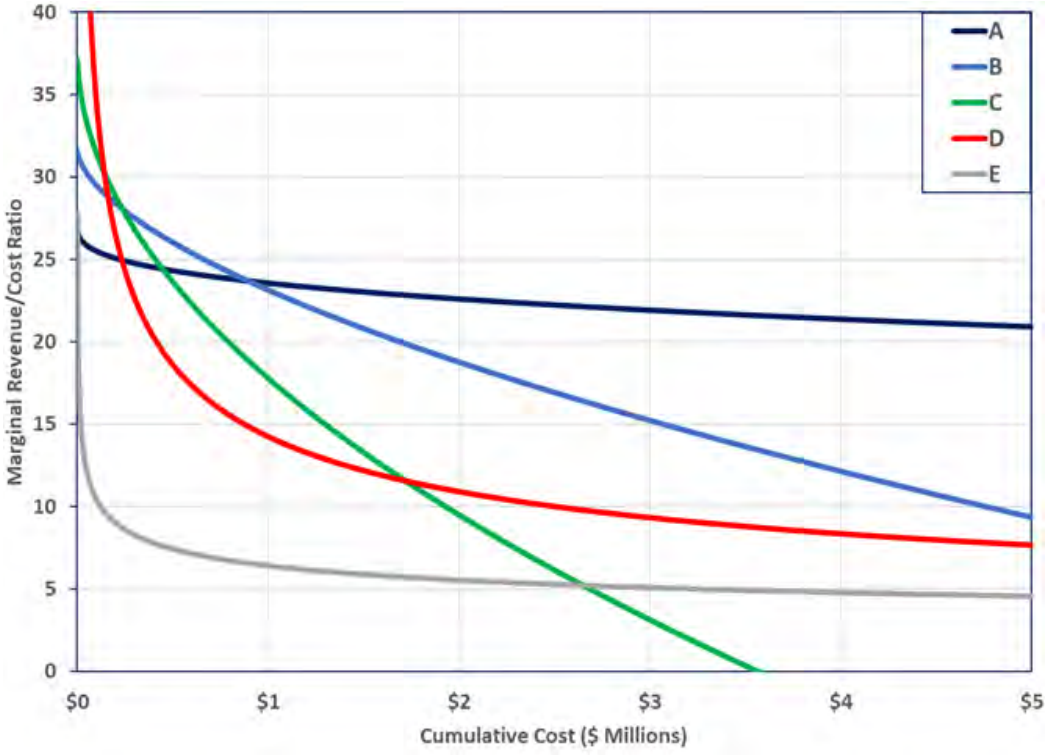
Figure 3 is the same as Figure 2, except that it adds the (actual) marginal revenue/cost curves for four other correspondence audit categories. Let's assume for illustration purposes that these five categories compete only with each other for budget allocations.[6] Let's also assume that the budget available to be allocated to these five categories is $5 million and that's the only constraint to be considered. The question is: How much of the $5 million should be allocated to each category?

FIGURE 2. Marginal Revenue/Cost vs. Cumulative Cost Plot for a Typical Audit Category



[6] In general, we'd include *all* categories that will compete for portions of the budget but limiting this illustration to just five categories will serve to illustrate the basic principles without making the illustration too complicated.

FIGURE 3. Marginal Revenue/Cost vs. Cumulative Cost Plot for Multiple Audit Categories



If our objective is to maximize direct enforcement revenue, then the optimal allocation has two key conditions:

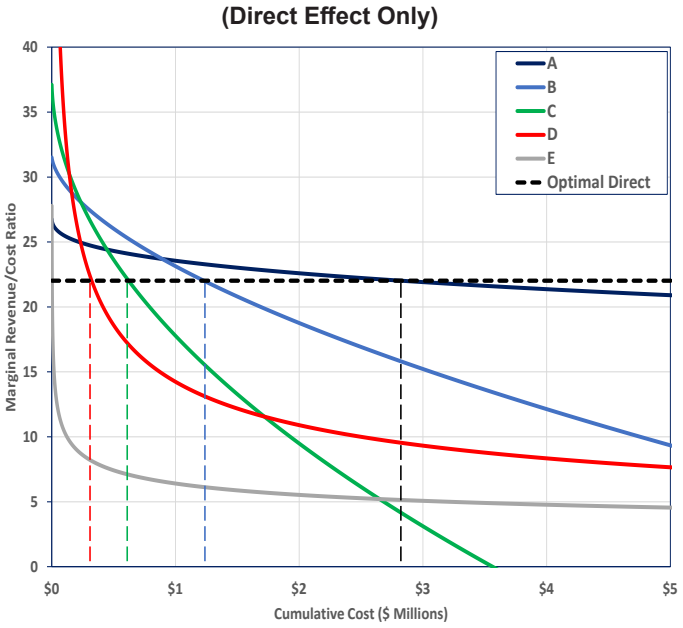
1. All of the budget is allocated; and
2. The marginal revenue/cost ratio is equalized across all unconstrained categories.

If the first condition is not very close to being met, we should be able to allocate the remaining budget somewhere to produce additional revenue. If the second condition is not met, then we should be able to increase revenue by shifting resources (budget) from categories having lower marginal revenue/cost ratios to categories with higher marginal revenue/cost ratios.[7] The optimal solution for our example is illustrated in Figure 4.

The dashed horizontal line is the optimal marginal revenue/cost ratio because it satisfies both key conditions: at that ratio, the entire $5 million budget is allocated and all five of the categories have the same marginal revenue/cost ratio (22.02). So there is no more opportunity to produce additional direct revenue from these categories. The cumulative cost where that horizontal dashed line crosses the curve for a given category indicates the optimal budget allocation for that category (which is traced down to the x-axis for each curve using vertical dashed lines). The resulting budget allocations are tabulated in Figure 4. The corresponding revenue for each category is easily derived from its cumulative revenue curve (comparable to Figure 1, but each category has a unique curve). So a budget of $5 million allocated to these five categories would produce a maximum of $126 million in revenue.

[7] This is often called the extensive margin because it addresses the extent to which each category is given resources and how many cases are worked in those categories in the optimal solution. There is also an intensive margin—addressing how intensively each case is conducted (e.g., how many line items on the return are examined) in the optimal solution. An implication of this second optimality condition is that at the optimal solution the intensive margin equals the extensive margin. That is, the marginal revenue/cost expected from examining one more line item in any given audit would be less than the marginal revenue/cost ratio that is equalized across all categories, so no more issues should be examined on that return. This means that when the budget is allocated optimally, the same optimal marginal revenue/cost ratio determines how much budget to give to each category (resource allocation), which cases to work within each category (case selection), and how intensively to examine each case (issue identification), so an increase in the budget should change how the budget is applied across categories, how many cases are worked in each category, and how many issues are audited in most cases. And because the marginal revenue/cost ratio is dimensionless, the same conditions can be applied throughout the enterprise even though the issues, methods, and costs are very different across categories.

FIGURE 4. Optimal Budget Allocation of $5 Million Across Five Audit Categories



Millions of $

Category	Optimal Allocation of $5 M	Direct Revenue ($ M)
A	$2.824	$65.9
B	$1.229	$31.5
C	$0.621	$16.9
D	$0.323	$11.6
E	$0.003	$0.1
Sum	$5.000	$126.0

2. Accounting for Indirect Effects

We noted at the outset that the ultimate objective of the IRS is not to maximize direct enforcement revenue (which we assumed for simplicity above), but rather to maximize the *total* revenue collected subject to our budget constraint, where total revenue includes both enforced tax revenue *and* the tax that is paid voluntarily by taxpayers. Some of that voluntarily paid tax revenue is due to the specific indirect of audits—the improvement in compliance among audited taxpayers in the years immediately follow their audit. Nicholl *et al.* (2020) presented estimates of the specific indirect effect in several correspondence audit categories. We have updated some of those and estimated the effect in several additional categories. At present, all of our estimates are expressed as the average effect per audit in a given category. That makes sense for the *specific* indirect effect; a taxpayer's compliance response to having been audited is not likely to depend on how many others were audited for the same year.[8] Table 1 displays our current estimates for the same five audit categories that were used in the prior example. Knowing the average cost in these categories, we have expressed the estimated indirect effect as the average indirect revenue/cost ratio for each of these categories. Note that the indirect revenue/cost ratio varies across categories, so taking these into account in the allocation of the budget is likely to change the optimal outcome. Note also that the indirect effects are not necessarily proportional to the direct effects.[9]

[8] In contrast, the general indirect effect (the extent to which unaudited taxpayers change their compliance due to the audits of others) is likely to depend on how many taxpayers are audited.

[9] Observe, for example, that the direct MR/C curve for Category E is always below (less than) the curve for Category D in Figures 3 and 4, yet in Table 1 the indirect effect is larger for Category E than for Category D.

TABLE 1. Estimated Specific Indirect Effect for Five Correspondence Audit Categories

Category	Average Indirect Revenue	Average Cost	Average Indirect Revenue/Cost
A	$4,479.57	$158.14	28.3
B	$4,282.58	$161.02	26.6
C	$1,465.31	$156.86	9.3
D	$118.50	$158.63	0.7
E	$501.66	$109.47	4.6

To include these indirect effects in the resource allocation framework, we need to plot marginal revenue/cost curves in which "revenue" includes *both* the direct and indirect effects. Because the specific indirect effect is averaged across all audited taxpayers within a given category of audits, including the indirect effect merely shifts the MR/C curve upward by the magnitude of the average indirect MR/C ratio in that category. This is illustrated in Figure 5, where the plot on the left is the same plot as in Figure 4 (direct effect only), but the vertical axis has been extended from a maximum of 40 to 70. The plot on the right shows the combined direct + indirect curve for each category, where each direct curve has been shifted upward at every point by the corresponding average indirect revenue/cost ratio given in Table 1.

FIGURE 5. Marginal Revenue/Cost Curves for Five Audit Categories

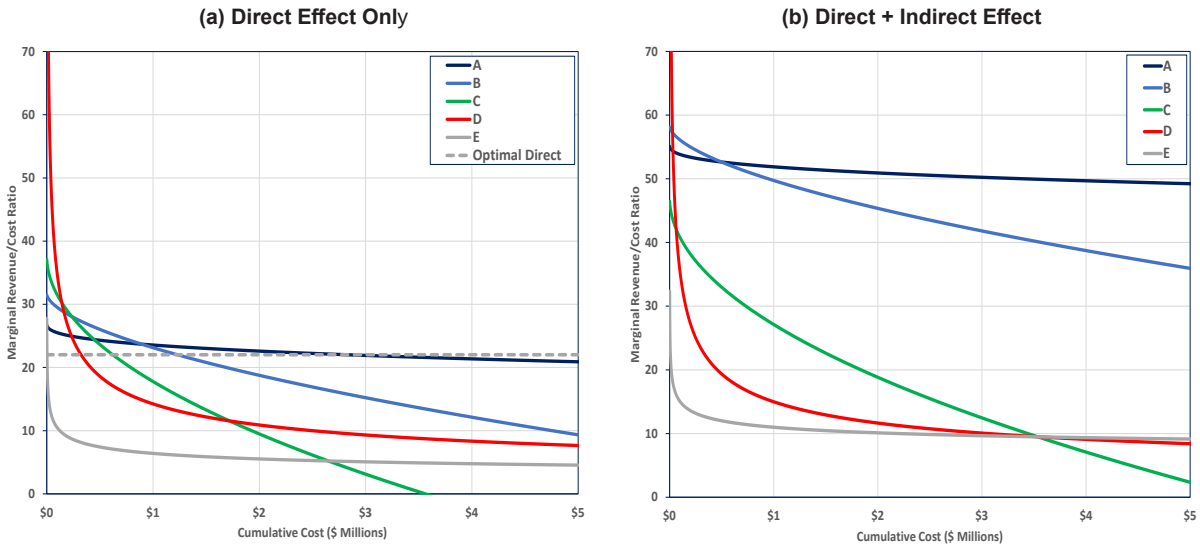
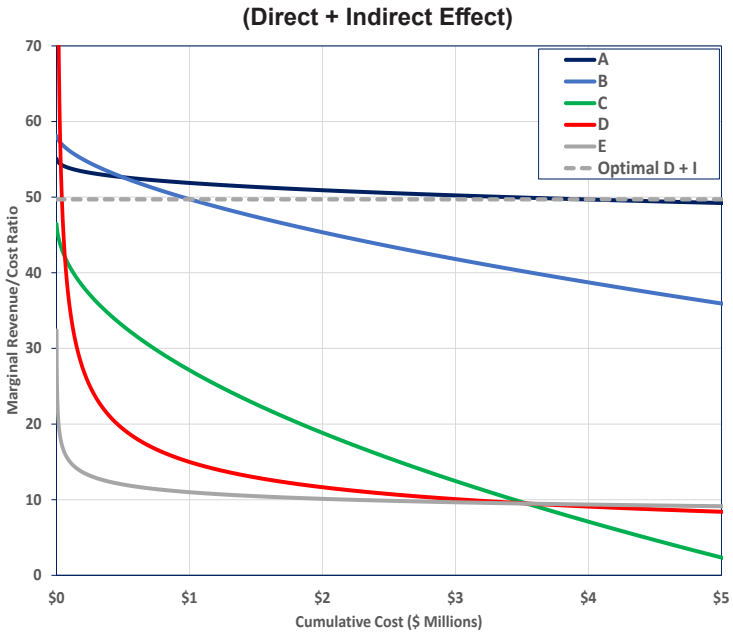


Figure 5(b) allows us to determine the optimal allocation of $5 million across these five categories accounting for both the direct and indirect effects. We do so in the same way that we did in Figure 4 when considering just the direct effect. The optimal MR/C ratio is now 49.72, as illustrated in Figure 6. This budget allocation yields $259 million in revenue.

Table 2 compares the resulting budget allocations and revenue outcomes when allocating the same $5 million of budget either to maximize the sum of direct and indirect revenue (shown on the right) or only to maximize direct revenue (shown on the left). Notice that if we count as the outcome only the direct revenue of the audits, then allocating to maximize that direct revenue produces the most direct revenue (by design); allocating to maximize the sum of direct and indirect revenue will always produce less *direct* revenue than that (in this example, around $5.9 million less than if we allocated to maximize the direct revenue alone). However, the allocation that produces the most direct revenue does not produce the most *total* revenue (direct + indirect).

FIGURE 6. Optimal Budget Allocation of $5 Million Across Five Audit Categories



Millions of $

Category	Optimal Allocation of $5 M	Direct Revenue ($ M)
A	$3.955	$202.1
B	$1.004	$53.1
C	$0	$0.0
D	$0.041	$11.6
E	$0.003	$0.1
Sum	$5.000	$126.0

TABLE 2. Impact of Accounting for Indirect Effects With a $5 Million Budget ($ Millions)

	Optimal for Direct Effect Only		Optimal for Direct + Indirect Effect		Difference in Revenue
	Optimal Budget Allocation	Revenue	Optimal Budget Allocation	Revenue	
A	$2.824	$65.870	$3.955	$90.415	$24.545
B	$1.229	$31.497	$1.004	$26.419	-$5.079
C	$0.621	$16.938	$0.000	$0.000	-$16.938
D	$0.323	$11.579	$0.041	$3.243	-$8.336
E	$0.003	$0.084	$0.0001	$0.006	-$0.078
Total Direct	$5.000	$125.968	$5.000	$120.083	-$5.885
A	$2.824	$79.984	$3.955	$112.035	$32.051
B	$1.229	$32.692	$1.004	$26.705	-$5.987
C	$0.621	$5.803	$0.000	$0.000	-$5.803
D	$0.323	$0.241	$0.041	$0.030	-$0.211
E	$.003	$0.014	$0.0001	$0.005	-$0.013
Total Indirect	$5.000	$118.734	$5.000	$138.771	$20.037
Total Direct + Indirect	$5.000	$244.702	$5.000	$258.854	$14.152

When allocating to maximize the sum of direct and indirect revenue, the smaller amount of direct revenue ($5.885 million in this example) is more than compensated for by the much larger amount of indirect revenue produced ($20.037 million in this example, for a net increase of $14.152 million in total revenue). This is completely analogous to the current practice of imposing minimum coverage constraints in categories with low direct revenue—sacrificing some greater direct revenue elsewhere to promote even greater indirect revenue (voluntary compliance) overall. However, here we are doing it with empirical estimates of indirect effects rather than by imposing subjective constraints.

The process described here can be generalized to real-life scenarios in which Examination planners need to take into account various constraints (e.g., the mix of available staffing across geographic and skill categories) in addition to the fixed budget. In such cases, the MR/C curves could be bounded by minimum or maximum constraints consistent with these nonbudgetary constraints. When such a constraint prevents us from equalizing the MR/C ratios across all categories, that constraint is a binding constraint, and the budget allocation to that category is dictated by the constraint rather than by the ideal optimization. But it's straightforward to allocate the remaining budget optimally (by equalizing the MR/C ratios of all unconstrained categories). The overall revenue result will always be less than if the budget were the only constraint; the binding constraints force that to be true.

3. Practical Considerations

It's likely to be necessary to account for various other practical outcomes when using indirect effects estimates. For example, an established estimation method could produce a negative effect in a given category. Or it could produce a very large estimate in one category compared with all the other categories. We actually encountered both of those kinds of results in our initial estimation work. How should such results be incorporated into resource allocation operationally?

A Negative Indirect Effect

A negative effect suggests that audited taxpayers become *more **non**compliant* following their audit.[10] That may be accurate, but the result might simply be due to weaknesses in the analysis; that possibility certainly needs to be explored before we blindly incorporate a negative effect in our resource allocation. That seems to have been the case in the category for which we initially estimated a negative indirect effect. Several factors seem to have influenced the initial result: (1) we had combined two similar categories into one for analysis, and they turned out to be quite different; (2) the exams focused on a particular tax credit, so an adjustment to the credit amount "should" have had a fairly direct impact on total tax (our dependent variable), but our control group turned out not to be very similar to the audited group with respect to the trend in tax reported, so the results were counter-intuitive; and (3) in both the audited group and the control group, taxpayers almost always stopped claiming the credit the very year after they claimed it inappropriately (which for the audited group was before they were audited). Subsequent analysis suggests that the indirect effect in each of the two categories we had combined is positive, but quite small.

If a negative indirect effect *does* seem to be accurate, those responsible for allocating resources have several options:

- They could include the negative effect, thus reducing the budget allocation to that category relative to the other categories;
- They could set it to zero (so that this category would compete for resources solely on the basis of its direct effect, while other categories competed on the basis of their combined direct and indirect effects), which would also direct resources away from this category;

[10] That could happen, for example, if the taxpayer is aware of noncompliance that the audit did not detect, which could embolden him to continue or even increase his noncompliance. It might also happen through revenge if the taxpayer was not treated fairly and respectfully in the audit process.

- They could allocate the budget to all categories solely on the basis of their direct effect, which would not account for the indirect effect of any of the categories, thus resulting in less overall revenue than could be achieved; or
- They could allocate the budget to the other categories based on their direct and indirect effects and allocate to this category on the basis of the direct effect alone, but subject to a *minimum* coverage constraint.

In any case, if we're fairly confident that the indirect effect is negative, we ought to do some research (e.g., interviews with audited taxpayers) to determine the underlying cause of that reaction to see if there's something that examiners should be doing differently or some other IRS action (e.g., correspondence or procedure) should be changed. And, of course, ongoing research to update our estimate the indirect effect in subsequent years should take place.

A Very Large Indirect Effect

If the indirect effect in a given category is very large relative to that in other categories, that could easily result in a large shift of resources to that category at the expense of the other categories. Again, the first step in such a situation is to double-check the analysis that produced this estimate. Are there any data issues (e.g., outlier observations that bias the estimate, missing data, too few observations in the test or control group, or data errors)? Is the control group reasonably like the audited group? Is there something unique to this category (e.g., in the population or in enforcement procedures) that is not adequately controlled for in the estimation? If further checks of this sort don't identify a need (or method) to change the estimation approach, then those responsible for allocating resources have several options:

- They could exclude this category from the resource allocation optimization, allocating a budget amount to this category somewhat subjectively up front (i.e., before the optimization, which would be subject to a smaller overall budget constraint after this category is funded); or
- They could achieve a similar, but more general result, by including this category in the optimization using the direct effect alone, but with a *maximum* coverage constraint for this category.

As in the case of a negative indirect effect, if we're fairly confident that the indirect effect is very large relative to that in other categories, we ought to do some research (e.g., interviews with audited taxpayers) to determine why that is true and whether that would suggest some ways to improve the indirect effect in other categories. And, as before, ongoing research to update our estimate of the indirect effect in subsequent years should take place anyway.

No Estimate of the General Indirect Effect

The general indirect effect is inherently more difficult to estimate; it's not as clear *who* is impacted by the enforcement, let alone by how much. And even if we *had* reliable estimates of it, incorporating it into the resource allocation optimization process would likely need to be much different. That's because the general effect would likely depend on *how many* contacts of a particular kind (e.g., the number of audits in a given category) took place. If so, the general effect would not be a simple shift upward of the direct effect MR/C curve; it would shift up more at the high-budget right end than at the low-budget left end.

While we await reliable estimates of the general indirect effect, though, we could still use estimates of the specific indirect effect as outlined in this paper. But rather than ignore the general effect entirely, budget allocators could continue to impose subjective minimum coverage constraints to account for the general effect. However, those minimum constraints should arguably be smaller than the minimum coverage constraints used currently because those presumably are intended to account for both the specific effect and the general effect combined.

References

Nicholl, Leigh C., Lucia Lykke, Max McGill, and Alan Plumley (2020). "The Specific Indirect Effect of Correspondence Audits: Moving from Research to Operational Application," *2020 IRS Research Bulletin*, Publication 1500 (Rev. 5-2021), pp. 9–32.

United States Government Accountability Office (GAO, 2012). "IRS Could Significantly Increase Revenues by Better Targeting Enforcement Resources," Report GAO-13–151.

4

Why Do Taxpayers Comply?

Erard ◆ Hertz ◆ Langetieg ◆ Payne ◆ Plumley

Grana ◆ Aw ◆ Lykke ◆ Schmitz

Angaretis ◆ Prohofsky ◆ Galle ◆ Organ

To File or Not to File? What Matters Most?

Brian Erard (B. Erard & Associates), Tom Hertz, Pat Langetieg, Mark Payne, and Alan Plumley (RAAS)

1. Introduction

Federal individual income tax returns are a key data source for understanding the drivers of many forms of taxpayer behavior. Although the information reported on these returns is not always entirely accurate, the returns provide granular details on levels and sources of the following: income; credits, deductions, and other offsets; income tax and self-employment tax obligations; and various demographic factors—all of which are potentially important for modeling and understanding taxpayer decision-making. It is the absence of the information normally reported on such returns that makes income tax nonfiling such an elusive research topic.

The relative dearth of information regarding nonfilers has largely precluded a direct comparison of filer and nonfiler characteristics in past research on the drivers of taxpayer filing behavior.[1] In our earlier work on this topic, we have attempted to overcome this information gap by comparing the characteristics of filers of federal individual income tax returns, based on tax returns and other administrative information, against the characteristics of the more general population of filers and nonfilers recorded in Census survey data (the Current Population Survey Annual Social and Economic Supplement—"CPS-ASEC"), with both data sources restricted to individuals/couples with an apparent income tax filing requirement.

Such an approach was applied to develop estimates of the drivers of individual income tax filing compliance using a novel econometric methodology that draws inferences based on differences in the characteristics of income tax filers from those of relevant overall population, rather than relying on more direct comparison of filers against nonfilers.[2] This "calibrated probit" approach (Erard (2021)) helps to overcome the lack of direct information on nonfiler characteristics and generates new insights into nonfiling behavior. Nonetheless, the reliance of the methodology on a combined data sample based on independently collected administrative and Census data sources is subject to several limitations. First, not all tax units in the population have a legal filing obligation and, given the deficiencies of income reporting in the CPS-ASEC, it sometimes can be difficult to determine whether a filing requirement exists.[3] Second, variables that are not recorded in both datasets (administrative and CPS-ASEC) must be excluded from the analysis. Among these variables are capital gains, state and local tax refunds, royalties, certain miscellaneous income sources, and various offsets to income and tax that are not present in the CPS-ASEC. Furthermore, other potentially important determinants of filing behavior are potentially unreliable, owing to systematic discrepancies in how they are defined and measured across the two data sources. In practice, several important income sources (including pensions, Social Security, and unemployment compensation) are substantially underreported in the CPS-ASEC, while other relevant taxpayer characteristics (such as earnings from self-employment, head of household filing status, and eligibility for refundable tax credits) are frequently misreported in both data sources. We have experimented with improving various CPS-ASEC income measures via imputation, but this is an imperfect and partial solution to a complex measurement error problem and results in inaccuracies at the microlevel.

Recently, however, we have been able to link detailed IRS administrative data from tax returns and third-party information reports with Census survey data for several tax years. This permits a more definitive assessment of whether a tax filing requirement is present and, if so, whether a tax return has been filed (either timely or late). Furthermore, it provides a rich set of potential determinants of filing behavior that are commonly defined and measured for both filers and nonfilers in the data sample. In this paper, we exploit this new data source to estimate several logit specifications of filing behavior to understand what factors distinguish

[1] An exception is Erard and Ho (2001), which relied on a special IRS Taxpayer Compliance Measurement Program (TCMP) study that contains line-item tax return details for both filers and *located* nonfilers of federal individual income tax returns to study the drivers of filing behavior.

[2] Erard *et al*. (2020 and 2021).

[3] To a somewhat lesser extent, this challenge extends to administrative data on filers owing to instances in which filing status, taxpayer age, dependency status, income sources, and/or income amounts are incorrectly reported.

whether a taxpayer is likely to file a timely return. Among the key factors we explore are the types and amounts of income reported to the IRS on third-party information documents, eligibility for major tax benefits (e.g., credits), prior filing behavior, filing status, changes in economic well-being, the magnitude of tax prepayments (e.g., through withholding), and demographic changes that individuals experience over time.

2. The Quest To Understand Nonfiling Behavior

a) Exclusion of individuals with no filing requirement

Since the focus of this study is on taxpayer filing compliance, it is important to restrict the analysis to individuals with a legal filing obligation. Although some individuals do file tax returns that are not required, for instance to claim a refund to which they are entitled, their motivations are outside the scope of the current analysis. As a matter of law, the majority of adults in the U.S. are required to complete a federal individual income tax return (e.g., Form 1040) each year and submit it to (i.e., "file it with") the IRS by a certain date. Whether someone is required to file a tax return depends on several economic and demographic factors. There are two main tests—either of which is sufficient to establish a filing requirement: (1) The Gross Income Test (a tax return is required if the individual's or couple's gross income exceeds the threshold established for their filing status, age category (under 65 or not), and dependency status); and (2) The Self-Employment Income Test (a tax return is required if an individual's net self-employment income exceeds $433).[4]

Because the gross income filing threshold depends on filing status (e.g., single, married filing jointly, head of household, etc.), and because the appropriate filing status is not apparent to the IRS for those who do not file a tax return (especially since information on marital status and the presence of qualifying persons associated with certain filing statuses is not generally available), the filing requirement (and tax obligation) for nonfilers whose gross income falls somewhere between the filing thresholds for singles and married couples cannot be established definitively from tax administrative data alone (e.g., third-party information returns sent to the IRS). Although reasonable aggregate estimates of the incidence of nonfiling can be made from administrative data by imputing filing statuses for those who did not file on time, evaluation of the drivers of nonfiling at the micro (individual) level (especially among those close to the filing thresholds) depends greatly on assigning them to the correct filing status. Without knowing for sure that a tax return was required, inferences about why a return was not filed are potentially misleading. One of the key advantages of linking detailed IRS administrative data to Census survey records (which contain robust demographic information for each individual surveyed) is that it permits an improved assignment of individuals and couples to their appropriate tax filing status).[5] For consistency, we use Census survey records to assign filing status for nonfilers and filers, even though for the latter we could use information from their tax return.

Having assigned individuals to their relevant filing status and age categories, we are able to determine the applicable filing thresholds for the Gross Income Test, thereby providing an improved basis for restricting our analysis to those with a legal filing obligation. However, because we rely primarily on third-party information returns for constructing income measures for the Gross Income Test (and also, in supplementary specifications, the Self-Employment Income Test), our approach will exclude certain taxpayers whose filing requirement hinges on income that goes unreported on such returns.[6] Although the third-party reports capture the vast majority of earnings from many income sources (including wages, interest, dividends, pensions, Social Security benefits, annuities, IRA distributions, state income tax refunds, and unemployment compensation), they only partially account for amounts received from certain sources, such as self-employment income, capital gains, rents, and royalties. As a consequence, our approach will exclude some taxpayers who fall below the

[4] Schedule SE indicates that self-employment tax is due (and a tax return is required) if total net self-employment income times 92.35 percent is $400 or more.

[5] However, even with the aid of Census records, there is not sufficient information to definitively identify taxpayers who qualify for head of household status. Based on National Research Program (NRP) audits, it is evident that head of household status is wrongly claimed on a significant percentage of individual returns. But, partly due to the fact that some children are underreported in the CPS-ASEC survey, our methodology tends to assign a smaller share of tax units to head of household status than even the percentage that holds up under NRP audits. Information is also lacking to reliably identify individuals who qualify for widow(er) status as well as those who would choose to report as married filing separately rather than jointly. In practice, these statuses are infrequently elected, and we have not attempted to assign them within our data sample. Instead, married taxpayers are routinely assigned married joint filing status, while unmarried individuals are assigned either to head of household or to single status, depending on whether the individual appears to have a qualifying child in the household.

[6] In the case of filers of tax returns, we could base the Gross Income Test on amounts reported on their returns for different income sources. However, we ignore such information to ensure that filers and nonfilers are assigned on the basis of consistent criteria.

gross income and self-employment income filing thresholds based on the observed third-party information returns, but not on the basis of their full earnings (including income that is not reported by third parties).

b) Terminology

We employ several terms to help describe filing (or nonfiling) behavior within our sample of taxpayers having a legal filing obligation. A key factor is whether a tax return was filed on time (either by the original, or, where applicable, extended filing deadline). We refer to returns that were filed on time as "timely filed"; all other returns are referred to as "not timely filed," including both returns filed late (not meeting the timeliness requirement) and returns not filed at all (at least within the time represented by our data). The taxpayers associated with these categories are respectively referred to as "timely filers" and "not timely filers" (or simply as "nonfilers"). Because we analyze data for two consecutive years, we can identify "stop-filers" (those who filed for the prior year but not for the current year) and "new filers" (those who filed for the current year but not for the prior year).

c) Theoretical insights

The standard economic model of tax compliance, as laid out by Allingham and Sandmo (1972) and Yitzhaki (1974), does not account for the possibility that an individual might fail to file a required income tax return. Rather, it focuses exclusively on a filer's decision with regard to how much income to report on the return. Under this framework, taxpayers approach their tax reporting decisions as they would an ordinary gamble, by weighing the potential gains from successfully underreporting taxes against the risk of audit and penalty. Specifically, a taxpayer chooses an amount X of income to report to maximize expected utility (EU):

$$EU = (1 - p)U[Y - tX] + pU[Y - tY - \theta t(Y - X)],$$

where p is the audit risk, t is the tax rate (assumed here to be proportional), Y is the taxpayer's true income that should be reported, θ represents the penalty rate on the unreported tax amount, and $U[z]$ represents the taxpayer's utility associated with having a net income of z remaining after taxes and any applicable penalties. This basic framework can be extended in various ways to account for relevant real-world considerations, such as progressive taxation, asymmetric information, nonpecuniary motivations, strategic auditing, and imperfect audit detection. However, this simple specification captures the basic insight that taxpayers have an incentive to underreport an additional dollar of income so long as the risk of audit and penalty are sufficiently low.[7]

To account for nonfilers ("ghosts"), Erard and Ho (2001) extended this framework to account for nonfiling as a strategic option. In this extended framework, a taxpayer separately considers his or her potential utility under scenarios in which (s)he does and does not file. Under the filing scenario, the taxpayer must choose not only how much income to report, but also how much tax to prepay W through withholding or estimated tax payments. The expected utility (EU_F) under this scenario is:

$$EU_F = (1 - p)U[Y - tX - \gamma(\overline{W} - W) - c] + pU[Y - tY - \theta t(Y - X) - \gamma(\overline{W} - W) - c],$$

where $\overline{W}$ is the minimum required tax prepayment amount, γ is the penalty rate on under-withholding, and c represents the burden associated with preparing and filing the return. On the basis of this scenario, the taxpayer would choose to make the minimum tax prepayment $\overline{W}$ if (s)he were to file. He or she would then choose the value of X that maximizes the above expression, conditional on $W = \overline{W}$.

Under the nonfiling scenario, the taxpayer would choose how much tax to prepay (W) after taking into account the risk of enforcement and the accompanying tax and penalty assessments. Expected utility under this scenario ($EU_{\bar{F}}$) is defined as:

$$EU_{\bar{F}} = (1 - q)U[Y - W] + qU[Y - tY - f(tY - W) - c],$$

[7] In this framework, the taxpayer will underreport by a positive amount so long as $p < 1/(1+\theta)$, although the extent of underreporting will depend on the individual's level of risk aversion and the specific values of the model's other parameters.

where q is the risk of nonfiling enforcement, f is the penalty rate on the unpaid tax balance, and represents the burden of preparing and filing a return in the event that enforcement occurs, and a return is secured. Under the nonfiling option, the taxpayer would choose the value of W that maximizes the above expression.

To decide whether to file a return, the taxpayer separately computes the corresponding expected utility values based on the optimal choices for W/X under the filing and nonfiling scenarios and chooses the option that yields the highest value. To understand what drives this decision, it is instructive to consider the base case where the filing burden c is equal to zero, the risks of enforcement under the two scenarios are equal ($p = q$), and the penalty rates are identical ($\theta = f$). In this case, Erard and Ho (2001) show that the maximum expected utility under the two scenarios is exactly the same, so that the individual is indifferent about whether to file a return. If the individual were to file, (s)he would make a tax prepayment of $W = \bar{W}$ as well as an optimal income report X^*. Alternatively, the individual could achieve the same expected utility by making a tax prepayment of $W = tX^*$ but not filing a return at all. Thus, the magnitude of the filing burden and the relative rates of risk and penalty are what drive filing behavior in this model. Individuals are more likely to be ghosts when the filing burden c is high, the risk of nonfiler enforcement q is low relative to the audit rate p, and the penalty rate facing nonfilers f is low relative to that faced by filers θ.

d) Empirical literature on drivers of nonfiling

Plumley Study

An early study of the drivers of nonfiling (Plumley (1996)), relied on an aggregate state-level panel data analysis of taxpayer behavior, IRS service and enforcement activities, and other relevant factors over the period from 1982–1991.[8] Plumley employs a two-way fixed effects model to explain the variation in the filing rate (ratio of the number of federal individual income tax returns actually filed to the estimated number of required returns) across states and over time. The state-level filing rate is found to be positively associated with several demographic factors, including the shares of potential filers that are married, under age 30, or over age 65. Average personal income within a state shows a negative association with the filing rate, while the rate of real income growth is found to have a positive impact on filing. The results also indicate that the state-level filing rate responds negatively to the share of potential filers that work as sole proprietors in the trade, finance, and service sectors; however, it is positively associated with the share of potential filers that work as sole proprietors outside of these sectors. Another significant labor market factor is the unemployment rate, which is found to be negatively associated with the state-level filing rate. A negative association is also found between the filing rate and the share of personal income that falls below the filing threshold.

The theoretical framework described above suggests that the burden associated with filing a return can be a deterrent to filing. Consistent with this perspective, the average estimated filing burden within a state is found to be strongly negatively associated with the filing rate. To the extent that IRS services reduce the filing burden experienced by taxpayers, they might play a positive role in promoting filing compliance. Consistent with this perspective, the number of tax returns per capita that are prepared with the assistance of IRS Taxpayer Service is found to be positively associated with state-level filing compliance.

The theoretical framework also suggests that the relative risk of nonfiler enforcement and penalties serves as a deterrent to nonfiling. Consistent with this perspective, the number of taxpayer delinquency investigation (TDI) notices issued per potential return in a state is found to have a significant positive impact on its filing rate. Similarly, the number of information return documents issued per potential return (a proxy for income visibility and risk of nonfiler detection) is also found to be positively associated with the state-level filing rate. On the other hand, the share of refund returns that are subject to an offset for debt collection is found to have an adverse impact on filing compliance within a state. In most states, taxpayers are required to file federal and state income tax returns. Recent state-level income tax amnesties are found to have a positive impact on the filing rate for federal returns within the state, suggesting a spillover effect of changes in the state-level enforcement environment.

[8] See Dubin *et al.* (1990) for a related analysis that applies a state-level panel to examine the determinants of income tax returns filed per capita, which serves as a proxy for filing compliance.

Update of Plumley Study

An update of the Plumley study (Erard and Morrison (2012)) was commissioned by the IRS using an improved and expanded state-level panel covering more recent years (ranging from 1992 to 2009).[9] The updated two-way fixed effects results largely corroborate Plumley's earlier findings. However, the updated results indicate that the state-level filing rate is negatively associated with the overall share of sole proprietors, not just with the share in trade, finance, and service sectors. In addition, the updated results find no significant associations of the share of the state population under age 30 with the rate of filing compliance. With regard to enforcement, the updated results no longer find the refund offset rate and state tax amnesties to be significant determinants of the state-level filing rate.

The updated analysis also experiments with some dynamic panel specifications of filing compliance that account for lagged values of the state-level filing rate. The lagged state-level filing rate is found to have a significant positive association with the current filing rate, suggesting a substantial degree of persistence in filing behavior. To account for the potential endogeneity of the TDI rate, the lagged TDI rate is employed as an instrument in this analysis using the Arellano-Bond (1991) estimation methodology. The results from this specification continue to show a positive impact of TDI investigations on filing compliance after potential endogeneity is accounted for.

Erard and Ho Study

Erard and Ho (2001) describes a micro-level analysis of the drivers of filing compliance based on data from a special TCMP study for Tax Year 1988 that investigated both filers and nonfilers of tax returns. The filer component of this study involved a stratified random sample of approximately 54,000 filers who were subjected to intensive line-by-line audits of their Tax Year 1988 returns. The nonfiler component involved a stratified random sample of 23,286 individuals for whom there was no record of a Tax Year 1988 return being filed. This sample included nonfilers as well as late filers and individuals with no legal filing obligation. An intensive effort was made to locate each individual in the sample and thoroughly investigate whether the individual had committed a filing violation. Overall, 18,689 of the individuals were successfully located, and it was determined that 3,549 of them had failed to file a required tax return for Tax Year 1988. Returns were secured from these individuals, and a random sample of 2,195 of these returns were then subjected to an intensive line-by-line audit.

Erard and Ho rely on a combined sample of the filers and the located and unlocated nonfilers from this study to investigate the determinants of filing compliance. To account for the fact that the nonfilers could not always be located, a joint model is estimated of the likelihood of filing a return and the likelihood that an individual who does not file can be successfully located. The likelihood of being located is found to be positively associated with the presence of a prior-year return, the presence of income subject to third-party information returns, being married, and being age 65 or older.

In contrast to the Plumley study, Erard and Ho find a negative association between being age 65 or older and filing a required return, while marital status is not significantly related to filing compliance. Similar to the updated Plumley study, however, filing compliance is found to be negatively associated with being a sole proprietor (regardless of business sector) or being unemployed, while it has no significant association with the level of (adjusted gross) income. Similar to the dynamic analysis that was conducted for the updated Plumley study, Erard and Ho find evidence of substantial persistence in filing behavior. All else equal, having filed a return for the prior year was associated with a 36 percent increase in the likelihood of filing for the current year. The estimated relationship between filing burden and filing compliance is more nuanced than in the Plumley study. In particular, filing burden is found to be a deterrent to filing, but only for taxpayers with income near the filing threshold. Among employees, filing compliance is found to vary across occupations. The lowest level of filing compliance is found to be among mechanics and helpers. Somewhat surprisingly, the highest level is found among those employed in construction, extraction, and production. The results indicate that individuals with a high probability of being located are much more likely to file than those with a low potential to be

[9] See also Erard (2011).

found, consistent with the theoretical model predictions with respect to the relative likelihood of enforcement and penalty.

Erard *et al.* (2020) Study

Erard *et al.* (2020) analyzes the determinants of filing compliance by combining two distinct samples of taxpayers: (1) an administrative sample of approximately 76,000 randomly selected federal income tax filers for Tax Year 2010; and (2) a combined sample of approximately 113,000 filers and nonfilers for that year. Both samples are restricted to households with a legal filing obligation.[10] The latter sample is drawn from the CPS-ASEC. Although the CPS-ASEC does not identify which respondents are filers and which are nonfilers, the authors are nonetheless able to estimate a qualitative response model using the calibrated probit methodology developed by Erard (2021). Whereas a traditional probit analysis would rely on the differences between filer and nonfiler characteristics to identify the probit model coefficients, the calibrated probit methodology relies on differences between filer characteristics and the characteristics of the overall population of filers and nonfilers.

In contrast to the earlier studies by Plumley and by Erard and Ho, this study finds that married taxpayers are relatively less likely to timely file a required federal income tax return. Consistent with Plumley, it finds that taxpayers aged 65 and older tend to be more compliant than younger taxpayers. As in previous studies, the presence of self-employment income is estimated to have a negative association with filing a timely return; however, the estimated impact is not statistically significant. Burden is found to be negatively associated with filing a timely required return, as is having income close to the filing threshold. However, the estimated impact of burden is muted for those near the filing threshold. This may reflect the fact that many low-income taxpayers are eligible for refundable tax credits. Although claiming such credits is associated with a higher filing burden, the value of the credits received in many cases justifies the burden of filing to claim the benefits. This study also finds significant regional variation in filing compliance. All else equal, compliance tends to be highest for those who reside in the Mid-Atlantic region and lowest for those who reside in the Mountain region.

Erard *et al.* (2021) Study

In Erard *et al.* (2021), the authors extend their calibrated probit analysis for Tax Year 2010 to cover a pooled sample over the period from Tax Year 2001 through Tax Year 2013. The results are qualitatively consistent with their earlier study, while contributing additional insights. For instance, the EITC benefit increase in 2009 for families with three or more children is found to have resulted in a substantial increase in timely filing among such families. The impact of the Economic Stimulus Act of 2008 on filing compliance is also investigated. Under this provision, single filers were entitled to a rebate of $600, while married filers could claim $1,200 on their Tax Year 2007 returns. An additional rebate of $300 per dependent child under the age of 17 was also available. The results show that filing compliance improved markedly in that year, but that it gradually returned to its baseline level over the subsequent three years. A widely held belief among tax administrators is that once a taxpayer is brought into the tax system, that individual will tend to remain in the system. The temporary nature of the filing response to the Economic Stimulus Act of 2008, however, calls this received wisdom into question. At the same time, however, there does seem to be substantial persistence in filing behavior. As with the studies by Plumley and Erard and Ho, the authors find that taxpayers who filed a return for the previous tax year are substantially more likely to comply with their filing requirement for the current year.

Erard *et al.* (2021) also incorporates a calibrated multinomial logit analysis to investigate what drives some taxpayers to file late, while others never file. The findings indicate that late filing is relatively more likely among self-employed taxpayers and taxpayers with high levels of income, which suggests that late filing is associated with more complex tax circumstances. Taxpayers under the age of 65 and single taxpayers are also found to be relatively more likely to file a late return. In addition, the results point to a regional pattern to late filing, with a higher incidence among taxpayers residing in the Mid-Atlantic, South-Atlantic, West South-Central, Mountain, and Pacific Census divisions and a lower incidence in the East and West North-Central divisions. On the other hand, the results indicate that taxpayers with income close to the filing threshold and those with

[10] In order to more accurately identify households with a filing requirement, additional income was imputed for certain income sources that tend to be understated in the survey.

relatively high filing burdens are substantially more likely to remain ghosts rather than file either a timely or a late return.

Stylized Facts and Ambiguities

Overall, the existing literature on filing compliance yields the following stylized facts:

- Nonfiling tends to be more prevalent when the risk of detection and penalties are relatively low, such as when taxpayers are self-employed or otherwise have only a small share of their income subject to third-party information reporting, when they are difficult to locate or contact, or when the taxpayer delinquency investigation rate is relatively low.
- Nonfiling tends to be positively associated with the magnitude of the filing burden, although this relationship seems to be rather nuanced. For instance, taxpayers who receive refundable tax credits may not be deterred from filing by the additional burden associated with claiming such benefits.
- Filing compliance can be adversely impacted by financial hardship, such as unemployment.
- Current-year filing compliance tends to be much higher among taxpayers who have recently filed a prior-year return.
- There tends to be substantial variation in filing compliance across regions, even after controlling for other relevant drivers of compliance.
- Older taxpayers tend to be more compliant with their filing obligations (at least according to most existing studies).

On the other hand, prior studies yield conflicting findings with regard to the role of certain socio-economic factors in filing compliance, such as marital status and income.

3. Data

The data used in each of the prior studies suffer from various limitations. Plumley (1996) and the subsequent updates are based on state-level aggregate data rather than data at the microlevel; the latter data source is generally better suited to explaining drivers of individual filing behavior and allows for a larger and more dynamic set of potential variables. But an advantage of the aggregate data in these studies is that it permits an examination of the impact of enforcement on filing behavior.[11] The calibrated probit method in Erard *et al.* (2020 and 2021) limits the variable set to regressors that are comparably defined and measured in both the CPS-ASEC and the IRS administrative data, and it relies on imputations of income to CPS-ASEC, which results in errors at the microlevel. Erard and Ho (2001) uses a unique dataset that includes comprehensive audits of nonfilers, but these data are now more than three decades old, had limited potential to examine dynamic factors, and were subject to potential selection bias since nonfilers who could not be located may have been different from those who were audited.[12]

The present study seeks to overcome some of these limitations. It uses microdata on filers and nonfilers, with comparably defined and measured variables for each group. We link Census survey data (in particular, the CPS-ASEC March Supplement) with detailed tax administrative data from the IRS. These datasets could be linked anonymously because the Census Bureau assigned a valid unique identifier, known as a Protected Identification Key (PIK) to most records from each dataset (see Wagner and Layne (2012)). This allows us to get the best of both worlds: relying on detailed demographic data from the Census samples that are linked at the person level with detailed micro information on income, as well as filing history from tax administrative data (including comprehensive data reported on filed Form 1040s, if filed, as well as on information documents such as Form W-2 and Form 1099 for all taxpayers). The detailed tax administrative data became available for such analysis through a special short-term IRS research project created under the authority of Internal

[11] Although most microlevel analyses have not controlled for the role of IRS enforcement activities, Erard and Ho (2001) was able to control for the likelihood that a nonfiler would be located during an enforcement action, and they found that this has a strong positive impact on the likelihood of filing a return.

[12] The estimation methodology employed by Erard and Ho attempted to account for this source of potential selection bias.

Revenue Code section 6103(n). One major advantage of this paper is that—for the first time—we have been able to: (a) establish with relatively high confidence the presence of a filing requirement at the microlevel; and (b) analyze a balanced set of data that applies equally to filers and nonfilers. The main limitation of our approach is that our sample of taxpayers with a legal filing obligation excludes taxpayers for whom the obligation hinges on income that is not fully reported by third-parties. Consequently, certain groups of taxpayers (such as the self-employed) are underrepresented in our sample. An additional challenge is that our analysis is restricted to CPS-ASEC records for which a valid PIK has been assigned and which have not been fully imputed. As a result of these restrictions, our estimation samples may not be fully representative of the underlying population of filers and nonfilers with a legal filing obligation. As described below, we employ an inverse probability weighting approach to address this potential source of selection bias.

An alternative to using these linked data would be to exclusively rely on IRS administrative data, which has been one of the approaches used in recent nonfiler tax gap estimates.[13] Under this method, the overall population of potential nonfilers is created from all individuals who appear on third-party information documents but who have not filed a federal individual income tax return. Subsequently, a subset of the potential nonfiler population is assigned as nonfilers through an imputation algorithm that seeks to achieve consistency with aggregate Census population counts by marital status, age, and numbers of dependents. For the nonfiler tax gap estimates, the errors in combining individuals into tax units and assigning filing statuses at the microlevel tend to average out and permit fairly accurate aggregate estimates. However, the relative advantages of relying on such data for analyzing the determinants of filing behavior are uncertain. On the one hand, such an approach would make it possible to identify and potentially account for a range of IRS enforcement actions experienced by members of the administrative sample. At the same time, however, the imputation errors associated with the assignment of tax units with a legal filing obligation would potentially yield misleading inferences.

We assembled the dataset for this paper as follows:

1. In order to accurately evaluate the presence of a filing requirement and have a complete set of demographic, income, and tax data for filers and nonfilers for two successive years, we take advantage of the panel component of the CPS-ASEC survey. Because most CPS-ASEC March Supplement survey respondents are included in the survey for two consecutive years (either the "previous" and current year or the current and "next" year), we first identify all CPS-ASEC records (using the unique PIK for each individual) appearing in the 2014–2015, 2015–2016, or 2016–2017 pairs of years. This excludes any records that cannot be assigned a unique PIK in one or both of the years in a pair. While theoretically half of the respondents in a given CPS-ASEC year should be matchable to a prior-year record, this is not the case because of the absence of PIKs for some records as well as survey attrition (e.g., the loss of respondents who changed their place of residence between surveys). We further restrict the sample of potential tax units to those respondents who are at least 16 years old. We treat the second year in each pair as the "current" year and refer to the first year in each pair as the "previous" year.
2. After linking the CPS-ASEC panel data to the administrative population data containing tax return and third-party information, we weight the individual person records in the two-year panels using inverse probability weights that are derived from a probit model that predicts the likelihood that a given member of the tax administration population data in the current year is present in the survey panel. This population includes all of those who were primary or secondary filers on a tax return, as well as anyone who received a third-party information document. In the probit specification, we control for factors that influence the likelihood that a member of this population can be successfully linked, including dummies for the decile position of the amount for each income line item found on a tax return (wages, interest, dividends, capital gains etc.) based on what is reported to the IRS on Forms W-2 and 1099, as well as dummies for geographic region, age group, and for whether the tax unit filed timely, late, or not at all for the current and prior tax years. The results of the probit analysis are used to predict, for each member of the linked sample, the likelihood of being present in the sample. The inverse of these predictions is then employed as a sample weight to make the linked sample approximately representative of the overall population. For instance, if a member of the linked sample is

[13] Langetieg *et al.* (2016). See also Lawrence *et al.* (2011) and Mortenson *et al.* (2009), both of which use a sample from the administrative population to estimate the extent of nonfiling.

estimated to have a 10 percent probability of inclusion, this member is assigned a weight of 10 to make this observation approximately representative of all 10 of the estimated population units with the same characteristics as the sampled unit.

3. After constructing the weights for each two-year panel, we combine the person-level records into tax units and assign children based on the information provided in the CPS-ASEC record. For married tax units, we sum the various income amounts from third-party information documents into a combined amount and then calculate the income amounts, exemptions, standard deductions, credits, total taxes, total payments, and balance due amounts that we estimate would appear on the tax returns if filed. Thus, by drawing on the third-party information documents and CPS-ASEC demographic information, we are able to compile a full set of mock tax return data for the current year and the prior year. We compare the gross income on the mock return against the filing threshold for the assigned filing status (where filing status assignment is limited to single, married-joint, or head of household) and age of the taxpayers to determine whether the tax unit was required to file for the current and/or prior year.[14] In the analysis presented in this paper, we include only the tax units that were required to file for the current year based on this gross income test.[15]

Our final base sample includes the pooled records across the three pairs of years. The full dataset has about 59,000 observations on taxpayers with an apparent filing requirement for the current year based on gross income (representing roughly 20,000 observations per panel sample). Panel A of Table 1 provides details on how the final base sample is distributed across the three pairs of years, broken down by filing status (timely or not timely). Panels B, C, and D of the table provide the corresponding counts for the three subsets considered in our analysis: those who filed timely for the prior year, those who did not file timely for the prior year, and high-income taxpayers.

TABLE 1. Sample Counts, Those Required To File for the Current† Year

Years by Category	Timely Filed in Current Year	Not Timely Filed in Current Year	Total
Panel A: Total Sample			
2014–2015	18,010	1,831	19,840
2015–2016	18,200	2,053	20,250
2016–2017	17,520	1,621	19,140
Total	53,500	5,500	59,000
Panel B: Timely Prior-Year Return Subsample			
2014–2015	17,540	597	18,140
2015–2016	17,790	741	18,530
2016–2017	17,080	569	17,650
Total	52,500	2,000	54,500
Panel C: Not-Timely Prior-Year Return Subsample			
2014–2015	667	1,027	1,694
2015–2016	618	1,094	1,711
2016–2017	636	876	1,512
Total	1,900	3,000	4,900
Panel D: High Income Subsample			
2014–2015	4,487	239	4,726
2015–2016	4,582	263	4,844
2016–2017	4,607	262	4,869
Total	13,500	1,000	14,500

† The "current" year is the second year in each pair of years.
NOTE: The totals may not equal the sum of the components due to rounding.

[14] We do not consider whether the taxpayer can be claimed as a dependent on another tax return, which would lower the filing threshold for such taxpayers.

[15] While not discussed in this paper, we have also analyzed a slightly expanded population where we include those required to file by the Self-Employment Income Test in the current year based on having Form 1099-Misc nonemployee compensation greater than $1,000. We used this larger threshold amount since the Self-Employment Income threshold of $433 applies to net self-employment income, while nonemployee compensation is a gross amount. But we are not including taxpayers who are required only because of self-employment income that is not reported on a Form 1099-Misc.

Using the final base sample, Table 2 compares the prior- and current-year filing behavior for all of the tax units that were required to file for the current year.

TABLE 2. Change and Persistence in Filing Behavior Over Two Years Among Those Required To File for the Current Year

		Current Year			
		Filed Timely	Filed Late	Did Not File	Total
Prior Year	Filed Timely	96.5%	2.0%	1.5%	100.0%
	Filed Late	57.1%	33.2%	9.7%	100.0%
	Did Not File	28.7%	3.3%	67.9%	100.0%
	Total	90.7%	3.2%	6.1%	100.0%

NOTE: Based on the 2014–2015, 2015–2016, and 2016–2017 pairs of years using normalized weights. The totals may not equal the sum of the components due to rounding.

As previous studies have shown, filing timely in one year is generally associated with timely filing in the next year, and not filing in one year is usually followed by not filing in the following year. However, it is fairly common that those who file late in one year wind up filing on time in the following year.

Presumably, the risk of detection and enforcement by the IRS increases with income, thereby resulting in a higher expected penalty for nontimely filing. Based on the theoretical framework, one would therefore expect a higher rate of nonfiling among tax units with gross income amounts near the filing threshold and lower rates among units with income further above the threshold. This pattern is confirmed in Table 3, which breaks down timely filing rates by income decile above the filing threshold.[16] In the combined population of units that did and did not file a timely prior-year return, only about 80.3 percent of all tax units in the first decile above the filing threshold filed a timely return for the current year. The incidence of timely filing for the current year increases sharply over the next two income deciles and incrementally after that, peaking at 94.7 percent for the top decile. At the same time, the incidence of not filing decreases steadily from the lower to upper deciles, a pattern that also holds within the subsets of timely and nontimely prior-year filers.

Among the subset of tax units that did not file timely for the prior year, the timely filing rate is uniformly lower across gross income deciles. For this subset, the timely filing rate is as low as 35.5 percent for those in the 4th income decile above the filing threshold and remains below 48 percent even within the top decile. These timely filing rates are illustrated in Figure 1.

Figure 2 shows that those with an apparent balance due are, on average, less likely to file than those who appear to be entitled to a refund, which matches our expectations. However, the propensity to file increases as the balance due or refund amount increases. This is likely due to the fact that those with larger balance due or refund amounts are also those with higher incomes, which makes them more likely to file. Late filing does not seem to be significantly correlated with the apparent balance due or refund amount. In the theoretical framework, the choice about the amount of taxes to pay in advance is viewed as part of (i.e., endogenous with) the filing decision so we do not include balance due or refund status as independent variables.

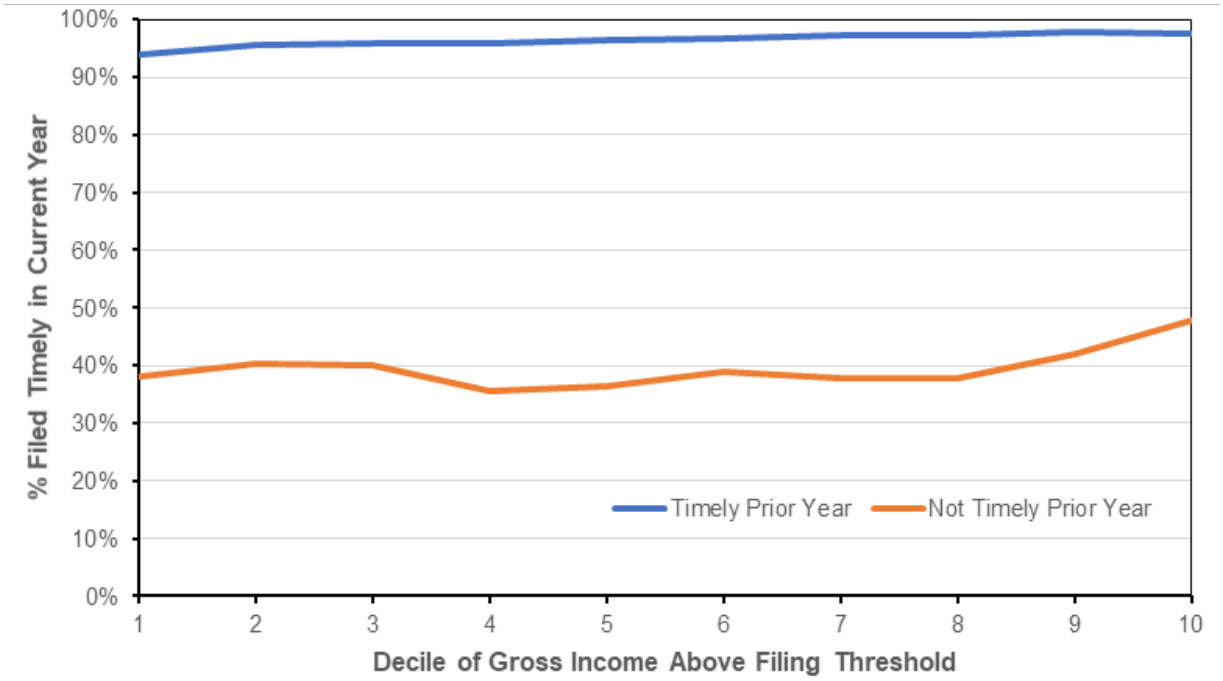
[16] This accounts for the different filing thresholds for singles, marrieds, etc.

TABLE 3. Filing Behavior for the Current Year by Decile of Gross Income Above Filing Threshold and Filing Behavior for the Prior Year

Decile	Filed Timely	Filed Late	Did Not File
Required by Gross Income in Current Year			
1	80.3%	3.3%	16.4%
2	86.9%	2.9%	10.2%
3	89.9%	3.1%	6.9%
4	90.9%	3.1%	6.0%
5	91.1%	3.6%	5.3%
6	92.1%	3.7%	4.1%
7	93.1%	3.2%	3.7%
8	93.4%	3.3%	3.3%
9	94.5%	2.9%	2.6%
10	94.7%	3.4%	1.9%
Required by Gross Income in Current Year and Timely Filed Prior-Year Return			
1	94.0%	2.4%	3.6%
2	95.6%	1.9%	2.5%
3	95.9%	1.9%	2.2%
4	96.0%	2.3%	1.8%
5	96.4%	2.2%	1.4%
6	96.7%	2.2%	1.1%
7	97.3%	1.8%	0.9%
8	97.1%	1.9%	0.9%
9	97.8%	1.5%	0.7%
10	97.4%	2.0%	0.5%
Required by Gross Income in Current Year and Prior-Year Return Not Timely Filed			
1	37.9%	6.1%	55.9%
2	40.3%	8.2%	51.4%
3	40.1%	13.1%	46.9%
4	35.5%	12.4%	52.1%
5	36.5%	17.5%	45.9%
6	39.0%	20.8%	40.2%
7	37.9%	21.8%	40.2%
8	37.8%	23.1%	39.2%
9	42.0%	24.7%	33.2%
10	47.8%	26.8%	25.5%

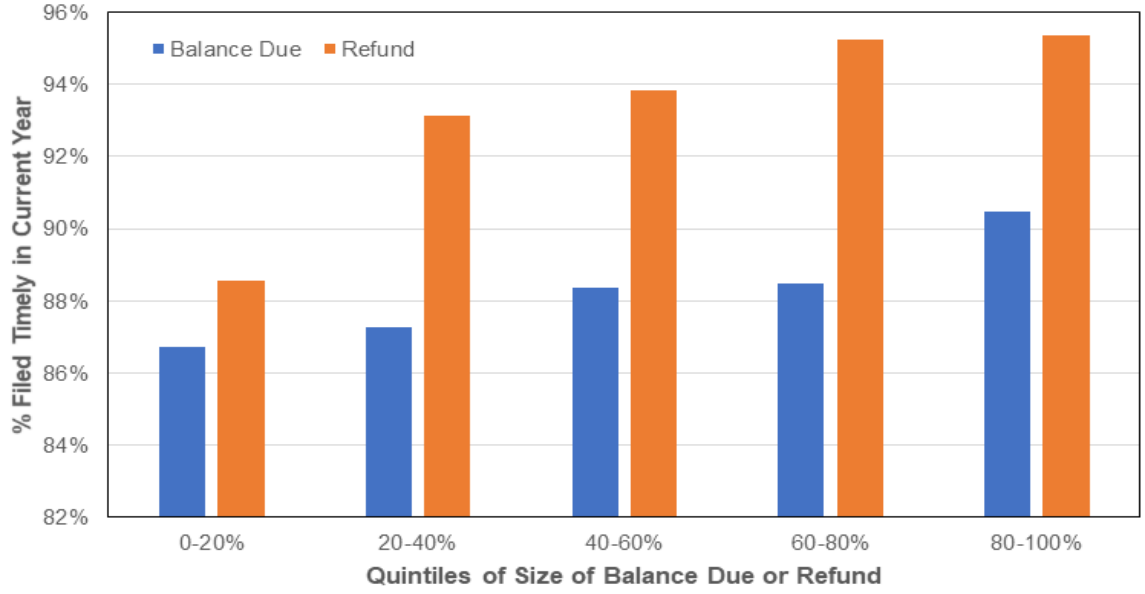
NOTE: Based on the 2014–2015, 2015–2016, and 2016–2017 pairs of years using normalized weights. The totals may not equal the sum of the components due to rounding.

FIGURE 1. Timely Filing Rate for Current Year by Decile of Gross Income Above Filing Threshold and Filing Behavior for Prior Year†



† Based on the 2014-2015, 2015-2016, and 2016-2017 pairs of years using normalized weights.

FIGURE 2. Timely Filing Rate for Current Year Among Those Required To File by Quintile of Balance Due or Refund†



† Based on the 2014-2015, 2015-2016, and 2016-2017 pairs of years using normalized weights.

4. Empirical Models and Results

We rely on a logistic regression framework to evaluate the drivers of filing compliance:

$$\Pr(TF = 1) = \frac{e^{\delta' x}}{1+e^{\delta' x}}$$

where TF is a 1/0 dummy for whether a timely tax return is filed for the current tax year, x is a set of explanatory variables, and δ is a coefficient vector to be estimated.[17] According to this framework, the log-odds of timely filing

$$\left(\ln\left[\frac{pr(TF = 1)}{pr(TF = 0)}\right]\right)$$

is a linear function ($\delta' x$) of the explanatory variables. The parameter vector δ is estimated using the method of maximum likelihood.

As reflected in prior studies, the data show a clear persistence of filing behavior from one year to the next (see Table 2). Approximately 96.5 percent of taxpayers who timely filed for the prior year filed also filed timely for the current year, compared to only 57.1 percent of those who filed late for the prior year and 28.7 percent of those who did not file at all for the prior year.[18]

Given the large impact of prior-year filing behavior and its tendency to swamp the effects of other explanatory factors, we estimate separate logit specifications for tax units that filed timely for the prior year and those that did not. And because of widespread interest in understanding the drivers of nonfiling among higher income taxpayers, we also separately examine tax units with more than $100,000 in gross income in the current year.

Specifically, we apply our basic econometric specification to three subsets of our overall sample:

1. **Timely Filed Prior Year (TPY):** Taxpayers in this subset filed a tax return on time for the prior year. Because the dependent variable in our logit analysis is an indicator for whether a return was timely filed for the current year, these regressions focus on the factors that influence a taxpayer to remain a timely filer (or the alternative, to become a "Stop-Filer").
2. **Not Timely Filed Prior Year (NTPY):** Taxpayers in this subset either did not file at all or filed late for the prior year. This is the complement to the first subset. Therefore, because the dependent variable in our logit analysis is an indicator for whether a return was timely filed for the current year, these regressions focus on the factors that influence a taxpayer who did not file timely in one year to file timely in the next year ("Start-Filers").
3. **High Income (HI):** This subset includes all taxpayers with gross income over $100,000 in the current year (which by definition makes them all required to file for this year), regardless of whether they filed a return for the prior year.[19] The analyses for this sample are focused on developing an understanding of what drives some high-income taxpayers to become nonfilers.

Using the linked data, we are able to examine many determinants of filing behavior that have been examined in past studies, as well as some potential determinants that have not previously been investigated. Table 4 describes the variables that are included in the logit specifications for each of the three samples (TPY, NTPY, and HI).

[17] A return is designated as timely filed in the IRS administrative data if it is filed before the April filing deadline, before a valid extension when this is requested, or within a grace period granted in the case of natural disaster.

[18] The percentages for timely filing in Tables 2 and 3 differ because of a data error. In Table 2 the requirement to file is determined based on a gross income amount that includes at least one half of the gross Social Security income amount. However, in Table 3, the gross income amounts over the threshold are based on a correct calculation of gross income in which only the taxable amount of Social Security is included.

[19] This definition includes a somewhat low threshold, which we chose in order to maintain an adequate sample size.

TABLE 4. List of Variables and Definitions

Variable	Description
timely_cy	1/0 indicator for a timely filed return for the current year
late_py	1/0 indicator for a late-filed return for the prior year
timely_py	1/0 indicator for a timely filed return for the prior year
year2016	Record from CPS-ASEC panel corresponding to Tax Years 2016 and 2015
year2017	Record from CPS-ASEC panel corresponding to Tax Years 2017 and 2016
gi_req_py	1/0 indicator for a prior-year filing requirement based on the Gross Income Test
singlefemale	1/0 indicator for a female taxpayer with single filing status
married	1/0 indicator for married joint filing status
oneeicchild twoeicchild	1/0 indicator for 1 EITC qualifying child based on age 1/0 indicator for 2 EITC qualifying children based on age
threeeicchild gtthreeicchild	1/0 indicator for 3 EITC qualifying children based on age 1/0 indicator for more than 3 EITC qualifying children based on age
logrefundcred	Natural logarithm of one plus the sum of EITC and Additional Child Tax Credits for which the tax unit is estimated to be eligible
age	Age of primary taxpayer
agesq	Age of primary taxpayer squared
newengland	1/0 indicator for residence in Maine, New Hampshire, Vermont, Rhode Island, Massachusetts, or Connecticut
midatlantic	1/0 indicator for residence in New York, New Jersey, or Pennsylvania
esouthcentral	1/0 indicator for residence in Kentucky, Tennessee, Mississippi, or Alabama
wsouthcentral	1/0 indicator for residence in Texas, Oklahoma, Arkansas, or Louisiana
enorthcentral	1/0 indicator for residence in Wisconsin, Illinois, Michigan, Indiana, or Ohio
wnorthcentral	1/0 indicator for residence in North Dakota, South Dakota, Minnesota, Nebraska, Iowa, Kansas, or Missouri
mountain	1/0 indicator for residence in Montana, Idaho, Wyoming, Nevada, Utah, Colorado, Arizona, or New Mexico
pacific	1/0 indicator for residence in California, Oregon, Washington, Hawaii, or Alaska
notaxstate	1/0 indicator for residence in a state with no income tax (Alaska, Tennessee, Wyoming, Florida, South Dakota, Texas, Nevada, or Washington)
gi_thresh_dec1	1/0 indicator for gross income within the first (lowest) decile of all required tax units
gi_thresh_dec2	1/0 indicator for gross income within the second decile of all required tax units
gi_thresh_dec34	1/0 indicator for gross income within the third or fourth decile of all required tax units
gi_thresh_dec56	1/0 indicator for gross income within the fifth or sixth decile of all required tax units
gi_thresh_dec78	1/0 indicator for gross income within the seventh or eighth decile of all required tax units
plur_retire_cy	1/0 indicator for retirement income (taxable Social Security and pension income) as largest income source
plur_seinc_cy	1/0 indicator for Form 1099 self-employment earnings (from Forms1099-Misc nonemployee compensation and Form 1099-G farm subsidies) as largest income source
plur_investinc_cy	1/0 indicator for investment income (dividends, interest, rents, and capital gains) as largest income source
plur_other_cy	1/0 indicator for other income (unemployment compensation, other income) as largest income source
burden_tpi	Ratio of the estimated monetary burden of filing to total positive income reported on third-party information returns
unemployed_cy	1/0 indicators for Received a presence of a Form 1099-G with a positive amount reported in Box 1 (unemployment compensation)
gotmarried	1/0 indicator for single prior-year and married-joint current-year filing status
gotdivorced	1/0 indicator for married-joint prior year and single current-year filing status. (Note: widow(er)s will potentially be misclassified as recent divorcees)
addkids	1/0 indicator for year-to-year increase in the number of EITC-qualifying children based on age
dropkids	1/0 indicator for year-to-year decrease in the number of EITC-qualifying children based on age

NOTE: The omitted (reference) category is single males with no children from the mid-Atlantic in gross income threshold decile 9 or 10 with wages as the primary source of income, no change in marital status, no change in number of children, and no unemployment compensation.

In each of the models, the dependent variable is an indicator for whether a return was timely filed for the current year (*timely_cy*), thereby grouping together taxpayers who file late and those who do not file at all into the nonfiling category. In future work, we plan to explore what factors determine whether a taxpayer is likely to file late rather than not at all.

Whereas members of the TPY sample exclusively filed a timely prior-year return, the NTPY sample includes both members that filed their prior-year returns late and members that never filed them at all. To control for these differences in prior-year filing behavior, the dummy variable *late_py* is included as a regressor in the NTPY logit analysis. We expect that a taxpayer who filed late for the prior year would be more likely to file a timely return for the current year than one who did not file the priory-year return at all.

For the HI sample, the dummy variables *timely_py* and *late_py* are respectively included as regressors to control for the role of prior-year filing behavior on current-year compliance outcomes. Given the persistence of filing behavior, we would expect that, all else equal, the likelihood of filing timely for the current year would be highest among timely filers for the prior year, followed by those who filed their prior-year returns late, followed by those who did not file them at all.

The logit specifications for each of the populations include the indicators *year2016* and *year2017* to distinguish observations from the different two-year panels in order to control for differences in behavior across different time periods. In addition, the TPY and NTPY models also control for whether a return was required for the tax unit for the prior year (*gi_req_py*). For the TPY sample, a requirement to file for the prior year might be associated with a longer history of filing and make timely filing for the current year more likely. Within the NTPY sample, taxpayers who had a filing requirement for the prior year are expected to be less likely to timely file for the current year, since they have already failed to comply with their prior-year filing requirement. We do not include this variable in the logit specifications for the HI sample since very few of these taxpayers were not required to file for the prior year.

Dummy variables are included as regressors to explore the roles of marital status and gender (*singlefemale* and *married*) in filing decisions. Prior studies reported conflicting findings on the effect of marriage on filing compliance. While Plumley (1996) found a positive association between the marriage rate in a state and its filing rate, the micro-level analyses by Erard *et al.* (2020 and 2021) found a negative relationship between marriage and filing. The role of gender has not been explored in prior studies.

We also include indicators for the number of children in the tax unit who meet the age criteria to qualify for EITC (*oneeicchild, twoeicchild, threeeicchild, gtthreeeicchild*). Given that the size of the credit for any given level of earned income and filing status increases with the number of qualifying children up to three, one would expect that timely filing would be more likely the larger the number of qualifying children (up to 3). However, in our specifications, we also explicitly control for the natural logarithm of the estimated amount of refundable credits (*logrefundcred*) for which a taxpayer is eligible. Consequently, the indicators should capture any differences in the filing rate that are attributable to the number of dependent children in the household, irrespective of the value of refundable credits.

We control for age using the continuous variables *age* and *agesq*. Prior research reported conflicting findings for the relationship between particular age categories and filing. While Plumley (1996) and Erard *et al.* (2020 and 2021) found that older persons are more likely to file, Erard and Ho (2001) found the opposite. Plumley (1996) also found that those under 30 were more likely to file, but the follow-up studies (Erard (2011) and Erard and Morrison (2012)) found this association was not significant. We choose the quadratic form for age to allow for the possibility that the conditional impact of age on filing compliance is nonlinear.

We also include dummy variables for the Census regional divisions to control for any variation in filing behavior across geographical regions. Erard *et al.* (2020 and 2021) found that taxpayers in the Mid-Atlantic division tended to exhibit more filing compliance while those in the Mountain and Pacific divisions were less compliant.

Apart from the geographic divisions, we include an indicator for whether the taxpayer resides in a state where there is no state income tax. In principle, taxpayers may be more likely to file a federal return when they also have to file a state return, both because one usually needs information from the federal return to file the

state return and because the presence of a state income tax may imply a higher likelihood of detection and higher expected penalties for not filing. However, prior studies (Erard and Ho (2001) and Erard *et al.* (2020 and 2021)) have not found such an indicator to be a statistically significant determinant of filing compliance.

Given the patterns observed in Table 3 on the relationship between the amount of gross income above the filing threshold and the filing rate, we also include dummy variables for the decile position of these amounts (*gi_thresh_dec1*, *gi_thresh_dec2*, *gi_thresh_dec34*, *gi_thresh_dec56,* and *gi_thresh_dec78*). A complicating factor is that, depending on filing status and number of qualifying children, taxpayers with gross income close to the filing threshold may have an added incentive to file given that they may be eligible for the EITC and/or the Additional Child Tax Credit (ACTC). For this reason, in some specifications we interact the dummies for deciles of gross income above the threshold with the log of the amount of refundable credits for which the taxpayer appears to be eligible.[20]

To investigate how filing compliance varies with the primary source of taxpayer income, the "plurality of income source" indicators *plur_retire_cy*, *plur_seinc_cy*, *plur_other_cy,* and *plur_investinc_cy* are included as regressors. A fairly consistent finding across prior studies is that sole proprietors are relatively less likely to file. This may be because the risk of detection is lower given that third-party reporting is incomplete and also because prepayments are not automatic as they are with wage withholding. This specification allows this finding to be tested, but also to check whether filing behavior varies in accordance with other primary income sources.

Consistent with the theoretical framework and several prior studies, filing burden is also expected to be negatively associated with filing compliance (particularly among those whose gross income is close to the filing threshold). To account for this possibility, we include the variable *burden_tpi* as an explanatory variable, which is the ratio of total monetary burden to total positive income. In addition, since some studies have found that the role of burden depends on how far gross income is above the filing threshold, in some specifications we also interact burden with the dummies for deciles of gross income above the threshold.

Economic hardship, as proxied by unemployment, was found to be negatively associated with filing compliance in several earlier studies. To account for this possibility, the list of regressors includes the dummy variable *unemployed_cy* for the presence of a positive reported amount in Box 1 of Form 1099-G (signifying the receipt of unemployment compensation in the current year).

Finally, we include two new sets of dummy variables that are feasible owing to the panel aspect of the CPS-ASEC portion of the data. The first captures whether the tax unit changed its filing status from single or head of household to married joint or whether it changed from married joint to single or head of household (*gotmarried*, *gotdivorced*). One might expect that both types of changes to the tax unit structure might reduce the likelihood of filing. The second set of dummy variables captures whether there was a change in the number of children (*addkids*, *dropkids*), which might also affect the likelihood of filing.

A. Timely Filers for the Prior Year (TPY, "Stop-Filers")

In the TPY sample, we focus on tax units that filed a tax return on time for the prior year and were required to file for the current year. Although, as previously shown in Table 1, most taxpayers who file timely in one year continue to file timely in subsequent years, we seek to understand why some taxpayers with a history of filing on time choose to stop filing.

Where possible, we compare the current results for the TPY sample against the findings of the earlier studies surveyed in Section 2. However, it is important to keep in mind that the earlier studies focused on the determinants of filing compliance within the general population of taxpayers with a legal filing obligation, not with respect to the large, but nonetheless distinct, subgroup who filed a timely prior-year return.

[20] We consider earned income (in our data, earned income is calculated as the sum of wages, Form 1099-Misc nonemployee compensation and farm subsidies on Form 1099-G), filing status, and the number of children to estimate the amount of EITC for which a taxpayer is eligible. If a child is under 19 or under 24 and a student, he/she is considered as a qualifying child for EITC. We cannot verify that the child is a full-time student or that he/she satisfies the formal relationship or residency tests. Children are first assigned to parents in the household, if present, then to relatives and then to adults with income.

Panel A of Table 5 presents the logistic regression results for the TPY sample. The logit specification includes the panel period indicators *year2016* and *year2017* to control for differences in behavior across time periods. The negative estimated effect associated with the *year2016* dummy indicates that otherwise similar taxpayers were significantly less likely to file timely for Tax Year 2016 than for either Tax Year 2015 or Tax Year 2017.

Consistent with Plumley (1996), but in contrast to Erard *et al.* (2020 and 2021) and Erard and Ho (2001), we find that married taxpayers are significantly more likely than single taxpayers to timely file a tax return. Furthermore, among single taxpayers, females are relatively more likely to timely file than males.

Our logit specification also includes dummy variables for the number of children in the tax unit of EITC qualifying age (under age 19). Erard *et al.* (2021) found that taxpayers with three or more EITC-qualifying children were significantly more likely to file following the expansion of EITC benefits for such taxpayers in 2009. However, this added financial incentive for large families to file is captured in our specification by the logarithm of refundable credits (which has a predictably positive and significant coefficient). The coefficient estimates associated with the EITC-qualifying children dummies are all insignificant, suggesting that the number of children in the household impacts filing compliance only insofar as they entitle the household to additional refundable credits.

As indicated earlier, previous studies had conflicting findings regarding the impact of the age of the primary taxpayer on filing compliance. Instead of relying on dummies for one or more age ranges, we employ a quadratic specification for age. The estimated coefficients of *age* and *agesq* indicate that filing compliance follows a U-shaped pattern with age. All else equal, the rate of filing compliance among those who filed timely in the prior year tends to decline until taxpayers reach their mid-40s, followed by a steady increase as they continue to age.

We also control for the geographic region in which the taxpayer resides. Consistent with Erard *et al.* (2020 and 2021), we find that those living in the Mid-Atlantic region are significantly more likely to timely file. In contrast to their findings, however, the Mountain and Pacific regions do not stand out as regions where taxpayers are particularly less likely to timely file. In fact, none of the other estimated regional dummy coefficients are statistically significant.

As in prior studies that have examined the issue, no significant impact was found for residence in a state that imposes its own income tax.

Consistent with the descriptive statistics in the previous section and prior studies (Erard *et al.* (2020 and 2021)), the current results indicate that taxpayers who have gross income near the filing threshold are relatively less likely to timely file. Specifically, the declining negative coefficient estimates for the dummy variables for the decile position of the amount of gross income above the threshold suggest that as a taxpayer's gross income increases relative to the filing threshold, the likelihood of filing becomes closer to that of the top two income deciles (the reference category). All else equal, taxpayers with income in the first (lowest) decile are least likely to timely file, followed by those in the second decile.

The presence of self-employment earnings was found in many prior studies to be a significant determinant of filing compliance. Plumley (1996) found that sole proprietors in certain specific sectors were relatively less likely to file a timely return, while the follow-up studies of Erard (2011) and Erard and Morrison (2012), as well as the TCMP study of Erard and Ho (2001), found that sole proprietors in general were less likely to file. From a theoretical perspective, one might expect that many self-employed taxpayers will perceive a lower likelihood that the IRS would recognize that a return was required and be able to establish that a balance is due (given that the amount reported by third parties on Forms 1099-Misc may greatly understate a sole proprietor's true income). Moreover, self-employed taxpayers tend to have a greater opportunity to avoid making required tax prepayments, which makes undetected nonfiling potentially more rewarding. The estimation results indicate that taxpayers with wages (the reference category) or interest as their primary source of income are relatively more likely to timely file than those for whom retirement income is the primary source, while those who receive the largest share of their income from self-employment or another source are least likely to file on time. It is important to recognize, however, that self-employment earnings and certain other income sources are

Table 5. Logit Model Results Predicting Timely Filed in Current Year Among Returns Required in Current Year*

Predictors	A. Timely Prior Year		B. Not Timely Prior Year		C. High Income	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
timely_py					5.183	0.1793
late_py			1.709	0.0896	1.685	0.1906
year2016	-0.190	0.0568	-0.101	0.0841	0.018	0.1147
year2017	0.042	0.0604	-0.056	0.0870	0.064	0.1153
gi_req_py	0.078	0.0919	-1.565	0.1143		
singlefemale	0.416	0.0630	0.163	0.0920	0.115	0.2229
married	0.630	0.0675	0.282	0.1030	0.260	0.1652
oneeicchild	-0.114	0.0826	-0.439	0.1271	-0.024	0.1402
twoeicchild	-0.102	0.0905	-0.315	0.1410	0.297	0.1448
threeeicchild	-0.192	0.1250	-0.323	0.2098	-0.154	0.1830
gtthreeeicchild	0.170	0.1905	-0.810	0.3206	-0.066	0.2821
age	-0.053	0.0097	-0.098	0.0128	-0.121	0.0309
agesq	0.0006	0.0001	0.0007	0.0001	0.0012	0.0000
newengland	0.014	0.0732	0.103	0.1081	-0.096	0.1438
midatlantic	0.538	0.1298	0.252	0.1643	0.317	0.2577
esouthcentral	-0.008	0.1051	0.045	0.1488	-0.225	0.1941
enorthcentral	0.194	0.1056	-0.189	0.1647	0.130	0.2148
wsouthcentral	-0.027	0.0786	0.102	0.1125	-0.003	0.1534
wnorthcentral	0.016	0.0860	-0.035	0.1290	-0.111	0.1706
mountain	0.031	0.0869	-0.045	0.1372	-0.103	0.1808
pacific	0.053	0.1192	-0.050	0.1764	-0.140	0.2454
notaxstate	-0.056	0.0855	-0.024	0.1285	-0.286	0.1564
logrefundcred	0.063	0.0111	0.059	0.0165	0.077	0.0451
githresh_dec1	-1.096	0.1144	-0.658	0.1708		
githresh_dec2	-0.658	0.1057	-0.298	0.1555		
githresh_dec34	-0.570	0.0905	-0.224	0.1390		
githresh_dec56	-0.353	0.0859	-0.188	0.1288		
githresh_dec78	-0.139	0.0858	-0.246	0.1276		
plur_retire_cy	-0.482	0.0936	-0.296	0.1275	-0.538	0.1949
plur_seinc_cy	-0.965	0.0989	-0.973	0.1567	-1.055	0.1781
plur_other_cy	-0.971	0.2367	-1.226	0.2968	-0.652	0.4212
plur_investnc_cy	-0.114	0.1440	-0.328	0.2141	-0.469	0.1869
gotmarried	-0.195	0.2082	0.200	0.3526	0.193	0.4103
gotdivorced	-0.288	0.1760	0.150	0.2660	-1.063	0.4074
addkids	-0.215	0.1214	0.309	0.1979	-0.214	0.2222
dropkids	-0.383	0.1257	-0.138	0.2049	-0.203	0.2209
burden_tpi	-0.647	0.3624	-2.301	2.6790	-0.479	13.6700
unemployed_cy	-0.340	0.0919			-0.602	0.2086
Intercept	4.430	0.2427	3.198	0.3297	1.460	0.796
Observations	54,500		4,900		14,500	
Pseudo-R^2	0.031		0.2133		0.3495	

Coefficients in plain bold are significant at the 5% level and coefficients in bold italics are significant at the 10% level.

* These estimation results rely on a somewhat flawed algorithm for determining whether a taxpayer had a legal filing obligation, which results in a somewhat inflated number of retired taxpayers. Except in a few cases, the removal of these excess observations does not alter the conclusions regarding the direction and significance of the explanatory variables considered.

understated in our data, because the measures only include the portions of income that are reported on third-party information returns.

Consistent with Plumley (1996) and Erard and Ho (2001), we find that those who are unemployed during the tax year are significantly less likely to file. This might be because taxpayers are not aware that unemployment compensation is subject to income tax or because those experiencing economic hardship may be relatively more likely to have a balance due and lack the means to pay their obligations.[21]

The linked data with two-year panels of the CPS-ASEC survey allow us to examine the impact of changes in a taxpayer's marital status and family size. These changes might be expected to cause life disruptions that complicate the tasks of recordkeeping and filing a return. At the same time, such changes can impact one's balance due or refund status (marriage tax or penalty, tax offsets associated with dependent children, etc.). The results in Column A of Table 5 indicate that getting married or divorced/widowed do not have a significant impact on the likelihood of filing a timely return for those who filed timely for the previous year. However, when the number of EITC-qualifying children decreases (as identified by the *dropkids indicator*), the likelihood of filing decreases significantly. More generally, a reduction in the number of qualifying dependent children may lessen the incentive to file because it reduces the number of exemptions, deductions, and credits for which a taxpayer may be eligible. In this regard, however, it is rather surprising that the estimation results also indicate that an increase in the number of EITC-qualifying children (as identified by the *addkids* indicator) has a negative and marginally significant impact on timely filing.

We also explore the role of the ratio of the estimated burden of filing to the taxpayer's total positive gross income. Similar to the results reported in Plumley (1996) and Erard *et al.* (2020 and 2021), a higher burden of filing is found to be associated with a lower likelihood of timely filing (a result that is marginally significant).

Erard and Ho (2001) found that filing burden is negatively associated with timely filing only among taxpayers who have gross income near the filing threshold, while Erard *et al.* (2020 and 2021) found that the negative association of filing burden with timely filing is attenuated for those near the threshold. The latter finding might be due to the fact that those near the threshold are now more likely to be eligible for refundable credits, which may more than offset the burden that filing entails. To explore this issue, we have estimated an augmented specification that includes interactions between our burden measure and the gross income decile indicators, as well as between the gross income deciles and the natural log of refundable credits. With these controls in place, the results indicate that there is no statistically significant variation in the impact of burden at different gross income deciles.

B. Not Timely for the Prior Year (NTPY, "Start-Filers")

The second sample we analyze is restricted to taxpayers with a current-year filing requirement who did not file a timely return for the prior year. This sample is considerably smaller (approximately 4,900 observations) than the one focused on those who timely filed for the prior year. The explanatory variables included in the specification for this sample are mostly the same as those used in the specification for the TPY sample, but the estimation results are quite different.

The logistic regression results for the NTPY sample are presented in Panel B of Table 5. The estimated coefficients of the panel period indicators (*year2016 and year2017*) are statistically insignificant, suggesting an absence of unexplained period-specific differences in timely filing rates within the sample.

As expected, the results indicate that taxpayers who filed a late return for the prior year are relatively more likely to file a timely return for the current year than taxpayers who failed to file at all for the prior year.

The indicator for the presence of a prior-year filing requirement (*gi_req_py*) distinguishes taxpayers in the NTPY sample who have a recent history of filing noncompliance from those who do not. The significant negative estimated coefficient of this variable indicates that the former group is relatively less likely to comply with their current-year filing requirement.

[21] Although taxpayers are encouraged to file even when they cannot currently pay their outstanding tax balance, the presence of a large balance can serve as a deterrent to filing.

The NTPY estimation results indicate that many of the regressors have a qualitatively similar impact on current-year filing compliance to what was found for the TPY sample. Specifically, the NTPY findings indicate that (all else equal):

- Single females and married taxpayers are somewhat more likely to timely file than single male filers;
- The propensity to file a timely return increases with the magnitude of refundable credits that the taxpayer is eligible to receive (and, based on an augmented specification that includes interaction terms between the natural log of refundable credits and the gross income decile indicators, it does not appear that the magnitude of the impact differs across the income deciles);
- Filing compliance tends to follow a U-shaped pattern with age (although the compliance rate tends to bottom out at a much higher age within the NTPY sample (around age 70 rather than in the mid-40s), so that relatively young and relatively old taxpayers have similar rates of filing compliance, all else equal;[22]
- Residents of the Mid-Atlantic region are relatively more likely to timely file than residents in other jurisdictions (although the estimated difference is not statistically significant);
- Residence in a state with no income tax has no significant impact on filing compliance;
- Those who have gross income close to the filing threshold are significantly less likely to timely file; and
- Taxpayers with wages as their primary income source tend to be the most likely to file on time, while those for whom self-employment or "other" income is the primary source tend to be the least likely to do so.

The NTPY estimation results for certain regressors are qualitatively different, however. For instance, the estimated impact of filing burden on the propensity to file a timely return is negative (as expected), but it is very imprecisely estimated. In an augmented specification, the interactions of the burden measure with the indicators for gross income deciles were also found to be statistically insignificant. Rather surprisingly, the presence of EITC-qualifying children is found to have a statistically significant adverse impact on current-year filing compliance after controlling for magnitude of refundable credits, with a particularly large impact for taxpayers with more than three qualifying children. On the other hand, a year-to-year change in the number of EITC qualifying children (in either direction) is found to have no statistically significant impact on filing compliance.[23]

Similar to the results for the NTPY sample, the estimated coefficient of the filing burden measure is extremely imprecise for the HI sample, so that the impact of filing burden on filing compliance cannot be inferred with any degree of confidence.

C. High-Income (HI) Taxpayers

In this section we restrict our analysis to taxpayers with gross income over $100,000 (HI sample), comprising about 14,500 observations. This group is of particular interest because at this level of income there is no potential ambiguity about the requirement to file and little chance to escape detection by the IRS. Nonetheless, a nontrivial number of taxpayers within this group fail to file a tax return and pay their outstanding tax liabilities in a timely manner. Moreover, their outstanding tax liabilities can be quite substantial. In Tax Year 2010, for instance, the top income decile of nonfilers is estimated to have been responsible for about 60 percent of the overall individual income tax nonfiler tax gap.[24] Therefore, obtaining a better understanding of the factors contributing to nonfiling within the HI sample would be valuable for tax administration. Panel C of Table 5 contains the logistic regression results for the HI sample.

---

[22] This finding was probably influenced greatly by the data error mentioned in footnote 18. The error was that we included all Social Security income (rather than just the taxable portion) as gross income when determining who had a filing requirement, which resulted in too many older people being classified as required to file. It makes sense that this would affect the NTPY sample much more than the TPY sample. Preliminary analysis indicates that this error generally has little impact on most of our results. Indeed, we suspect that the U-shaped relationship we find between the filing rate and age will still hold for the NTPY sample, but the turning point probably takes place at an age significantly younger than 70.

[23] Though not shown in Table 5, the indicator for unemployment benefits also was not found to be significantly associated with timely filing within the NTPY sample.

[24] Hertz et al. (2021).

As with the NTPY sample, the estimated coefficients of the panel period indicators (*year2016 and year2017*) for the HI sample are statistically insignificant, suggesting an absence of unexplained period-specific differences in timely filing rates. The HI sample includes subsamples of taxpayers who filed their prior-year returns timely, late, or not at all. As expected, the estimated incidence of timely filing in the current year is highest within the first subsample and lowest within the third subsample.

Similar to the findings for the TPY and NTPY samples, filing compliance is positively associated with the magnitude of refundable credits within the HI sample (which mostly involve additional child tax credit payments within this sample) and tends to be most prevalent among taxpayers whose primary source of income is wages and least prevalent among those whose primary source is self-employment. In contrast to the findings for those other samples, however, residence in a state with no income tax is found to have a negative and marginally significant impact on filing compliance within the HI sample, while marital status and gender are found to have no significant impact.

Similar to the results for the TPY sample, filing compliance seems largely unaffected by the presence of EITC-qualifying children (who would also qualify for the refundable and nonrefundable portions of the child tax credit) in the HI sample; for reasons that are unclear, however, having precisely two qualifying children is found to have a significant positive impact on the propensity to file a timely return within this sample. Also consistent with the TPY sample findings, filing compliance within the HI sample is found to be negatively associated with the receipt of unemployment benefits.

In contrast to the TPY and NTPY sample results, the HI sample results indicate that the propensity to file a timely return is adversely impacted by a recent divorce (or loss of a spouse).

5. Conclusions and Future Research

The unique dataset constructed for this study has provided an opportunity to deepen our understanding of the drivers of tax filing behavior. Like Erard *et al.* (2020 and 2021), we conduct a microanalysis of filing behavior. Unlike those studies, however, the current research is able to rely on a single data source that contains a commonly defined and measured set of variables for both filers and nonfilers of income tax returns. These variables include detailed information on income sources and levels, as well as demographic characteristics, which permit a more reliable assessment of whether a given tax unit has a legal filing obligation. In addition, the longitudinal nature of this dataset makes it possible to conduct separate analyses of current-year filing behavior of those for taxpayers who did and did not file a timely return for the prior year and to account for the role of changes in income and demographic characteristics in filing compliance. By separately analyzing the drivers of filing compliance among those with different filing histories, the study yields new insights regarding "stop-filers" and "start-filers." The current study is also the first to focus specifically on the potential drivers of nonfiling behavior among high-income taxpayers.

"Stop-filing" (among the TPY sample) seems to be much more likely among single males, those with lower income, those with a primary income source other than wages or investment, those who can no longer claim as many children as dependents, and those who received some unemployment compensation.

"Start-filing" (among the NTPY sample) seems to be less likely among those who failed to comply with a filing requirement for the prior year, single males, most groups of EITC-eligible taxpayers with children, taxpayers with income close to the filing threshold, and those with self-employment or "other" income as their primary income source.

High-income nonfiling seems to be more likely among those with self-employment or retirement as their primary income source, those who lost a spouse or got divorced, and those who received some unemployment compensation.

These findings could potentially be useful in supporting efforts to improve filing compliance. Understanding the factors associated with "stop-filing," "start-filing," and high-income nonfiling could help narrow the population of filers most at risk of not filing who could be contacted with a reminder of the importance of filing required tax returns every year. This would be most effective if the relevant third-party information documents

for the current year were available for analysis before the end of the filing season (April 15 of the next year), although this would require legislative changes.

Although the dataset used for this analysis has many desirable features, it also has several limitations. For example, the linked Census-IRS data does not include any direct information about enforcement actions taken against individuals in the Census surveys; we expect that enforcement activities would influence filing behavior, but we will have to explore that using tax administrative data alone (which has its own disadvantages). In addition, our current data may not fully reflect all instances of late filing, because we have considered only two years of filed returns for each tax year in the analysis. We hope to expand the data soon to be able to better distinguish between late filing and not filing, particularly to see if there are factors that tend to lead to filing late rather than not filing at all.

Expanding the number of years in the analysis may also shed some light on the impact of key tax law changes as well as one's longer-term filing history on current filing behavior. A limitation of the current study design is that it relies solely on third-party information reports to measure income, which means taxpayers with certain income sources (such as self-employment earnings) are underrepresented in our sample of taxpayers with a legal filing obligation.[25] Furthermore, relying on the pairs of records from selected years restricted the number of observations available to us for analysis. We plan to estimate the same models using only administrative data to assess whether the access to much larger samples and additional potential determinants of filing compliance (such as IRS enforcement actions) outweighs the benefits of having supplementary demographic information to incorporate as controls and more reliably assess whether a filing requirement is present.

[25] We considered imputing self-employment income to the linked records in a manner similar to Hertz *et al*. (2021), but we opted not to do that for this first use of these data to study the drivers of nonfiling because such imputations would presumably be far less reliable for this microanalysis than they are for estimating aggregate totals. We may consider alternative approaches in future research.

References

Allingham, M.G., and A. Sandmo (1972). "Income Tax Evasion: A Theoretical Analysis," *Journal of Public Economics* 1(3/4), 323–338.

Arellano, M., and S. Bond (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and Application to Employment Equations," *Review of Economic Studies* 58, 277–297.

Dubin, J., M. Graetz, and L. Wilde (1990). "The Effect of Audit Rates on the Federal Individual Income Tax, 1977–1986," *National Tax Journal* 43(4), 395–409.

Erard, B. (2011). "Research Focus Area 4 Documentation Report, Predicting Taxpayer Behavior," produced for IRS by IBM and B. Erard & Associates under Contract TIRNO-09-Z-00021, December 22.

Erard, B. (2021). "Modeling Qualitative Outcomes by Supplementing Participation Data with General Population Data: A New and More Versatile Approach," *Journal of Econometric Methods* 11(1), 35–53.

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

Erard, B., and L. Morrison (2012). "Predicting Taxpayer Behavior, Final Report," produced for IRS by IBM and B. Erard & Associates under Contract TIRNO-09-Z-00021, July 31.

Erard, B., P. Langetieg, M. Payne, and A. Plumley (2020). "Flying Under the Radar: Ghosts and the Income Tax," *CESIFO Economic Studies* 66(3), 185–197.

Erard, B., P. Langetieg, M. Payne, and A. Plumley (2021). "Ghosts in the Income Tax Machinery," Working Paper.

Hertz, T., P. Langetieg, M. Payne, A. Plumley, and M. Jones (2021). "Estimating the Extent of Individual Income Tax Nonfiling," *2021 IRS Research Bulletin*, Internal Revenue Service, Publication 1500, Washington, DC, pp. 93–124.

Langetieg, P., M. Payne, and A. Plumley (2016). "The Individual Income Tax and Self-Employment Tax Nonfiling Gaps for Tax Years 2008–2010," *2016 IRS Research Bulletin*, Internal Revenue Service, Publication 1500, Washington, DC, pp. 39–60.

Langetieg, P., M. Payne, and A. Plumley (2017). "Counting Elusive Nonfilers Using IRS Rather Than Census Data," *2017 IRS Research Bulletin*, Internal Revenue Service, Publication 1500, Washington, DC, pp. 38–59.

Lawrence, J., M. Udell, and T. Young (2011). "The Federal Tax Position of Persons Who Were Not Reported on Filed Tax Returns in 2005," *2011 IRS Research Bulletin*, Internal Revenue Service, Publication 1500, Washington, DC, pp. 143–155.

Mortenson, J, J. Cilke, M. Udell, and J. Zytnick (2009). "Attaching the Left Tail: A New Profile of Income for Persons Who Do Not Appear on Federal Income Tax Returns," *Proceedings of the NTA 102nd Annual Conference on Taxation*. NTA, Washington, DC, pp. 88–102.

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

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

Yitzhaki, S. (1974). "A Note on Income Tax Evasion: A Theoretical Analysis," *Journal of Public Economics* 3(2), 201–202.

Economic Influencers of Total Enforcement Revenue Collected and Operational Implications[1]

Jess Grana[2], Astou Aw, Lucia Lykke, and Sam Schmitz (The MITRE Corporation), and Ron Hodge (IRS, RAAS)

1. Introduction

Internal Revenue Service (IRS) resource planning considers the average revenue per case, among other metrics, at each step of the enforcement process. Revenue per case is estimated from historical data. However, changing economic conditions, IRS resources, and taxpayer population may cause future revenue per case to deviate from past values. Resource allocation models that use outdated revenue per case inputs will not generate staffing allocations that optimize total enforcement revenue.[3]

We build two proof-of-concept models to forecast enforcement revenue. First, we develop a prototype "macro level" model to estimate the effect of economic conditions, IRS resources, and taxpayer attributes on broad measures of enforcement revenue. This model helps illuminate the potential drivers of enforcement revenue at a high level. For example, our results suggest that Collection revenue is positively associated with two-year prior business bankruptcies and negatively associated with the two-year prior consumer price index (CPI).

Second, we develop a "micro" model that forecasts revenue at the level of steps in the enforcement process (e.g., Automated Collection System (ACS) cases and Field Collection cases derived from correspondence audits). Using taxpayer characteristics, we predict taxpayers' progression through the enforcement process as well as their expected revenue once we have arrived at each step. We use the "micro" forecasts of average revenue at these steps in a workforce allocation model, which optimizes resource allocation based on expected revenue per case at each step of the enforcement process. We demonstrate that using forecasted revenue per case rather than historical estimates leads to workforce allocations that can generate higher total enforcement revenue.

2. Macro Influencers of Total Enforcement Revenue Collected

Total enforcement revenue collected (TERC) is the amount of revenue collected through enforcement activities by the IRS. These enforcement activities are broadly categorized as Automated Underreporter (AUR)[4], Examination (Exam), Collection, and Appeals. Economic conditions affect TERC by influencing not only total tax but also voluntary compliance and enforced and late payments. We develop a prototype "macro level" model to estimate the effect of economic conditions, IRS resources, and taxpayer attributes on broad measures of TERC.

---

[2] Corresponding author. Direct questions or comments to Jess Grana at cheny@mitre.org.

[3] To optimize total enforcement revenue, resource allocation should be based on the marginal revenue to cost of each enforcement activity. Marginal revenue to cost is typically computed from historical data. In this paper, our focus is on forecasting revenue per case rather than estimating it from historical data. As such, estimating marginal revenue to cost is outside the scope of this work.

[4] This is the document-matching program that compares what is reported to the IRS on third-party information documents with what taxpayers report on their tax returns.

The goal of this study is two-fold: 1) to use correlation analysis to understand the macro-level drivers of TERC, and 2) to use those drivers in forecasting analysis to forecast TERC. This report seeks to understand the drivers of total enforcement revenue for Appeals, Collection, Exam, and AUR. We show the percentage of variation in total revenue explainable by certain influencer variables, and we use this model to produce a two-year forecast of total revenue. This will help IRS anticipate enforcement revenue in a changing economic environment and evaluate revenue implications of high-level IRS policies such as total workforce levels.

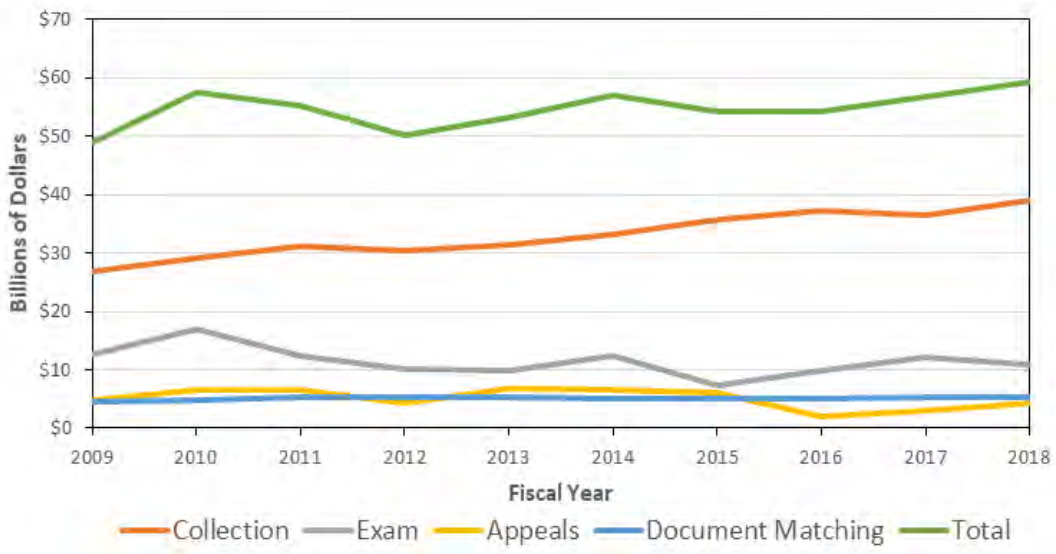
2.1 Background

IRS enforcement spans four functions: Document Matching, Examination (Exam), Appeals, and Collection (IRS (2019a)). Document Matching includes the IRS's AUR Program, which is a highly automated program comparing tax returns with third-party reporting (such as Forms W-2 and 1099). AUR cases are not audits (examinations); they follow different rules by law. Exam investigates certain tax returns that are selected for correspondence (mail) or field (face-to-face) audits of one or more entries on the tax return. Appeals are initiated by taxpayers who disagree with an IRS examination, collection, or penalty judgment. Finally, Collection handles unpaid taxes and unfiled returns through both automated and human systems.

Total enforcement revenue collected (TERC) is the combined revenue, including taxes, penalties, and interest, from all four enforcement functions. Figure 1 plots enforcement revenue by function and in total. Collection makes up the largest portion of enforcement revenue, and its share has been increasing. In Fiscal Year 2018 (FY2018), it brought in a record $39 billion. It also accounted for 80 percent of enforcement cases over the past 10 years (Macheret *et al.* (2020)). Exam is consistently the second largest portion of enforcement revenue, bringing in almost $11 billion in FY2018. Note that enforcement cases that progress through other enforcement functions often end up in Collection.[5] The Program Assessment Model (PAM) Optimizer[6] details these enforcement workflows, including the likelihood of an enforcement case progressing from one enforcement activity to another.

FIGURE 1. Total Enforcement Revenue Collected by Function

SOURCE: IRS (2019a)

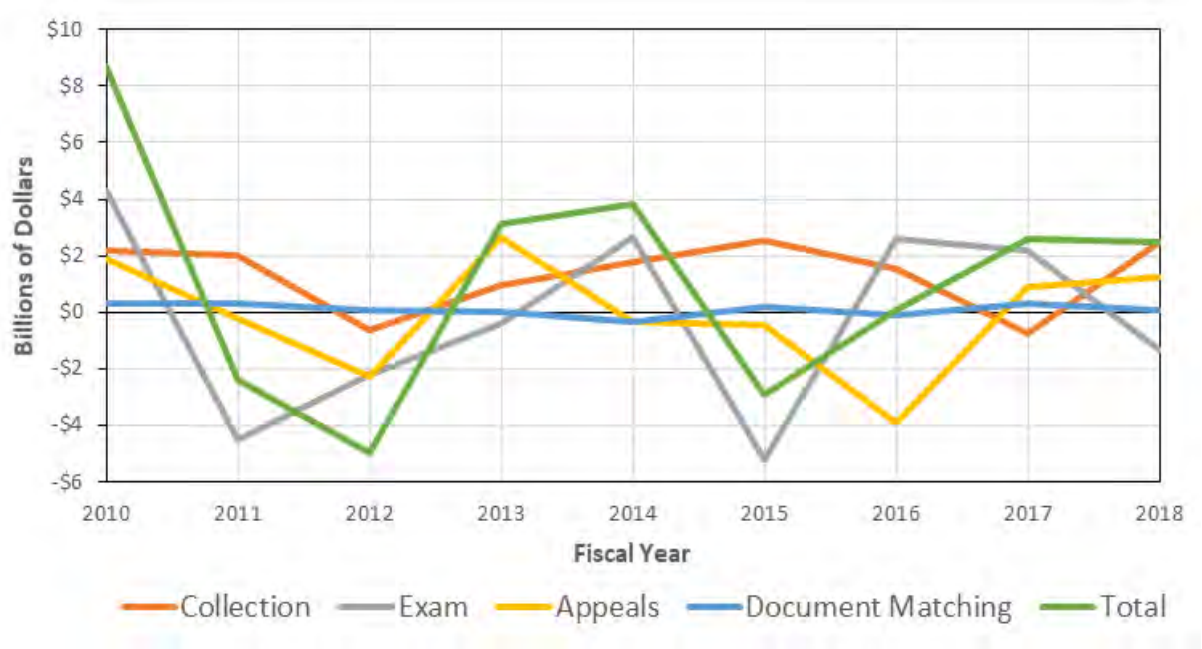


[5] Revenue collected is credited to the origination of the enforcement case. For example, cases that begin in Exam and progress to Collection will have any revenue collected in Collection credited to Exam.

[6] Developed by MITRE for SB/SE.

TERC fluctuated over FY2009–FY2018, with peaks in FY2010, FY2014, and FY2018 and troughs in FY2009 and FY2012. Figure 2 shows the change in enforcement revenue from the prior year. Collection has mostly seen revenue growth in those 10 years. Exam has shown the largest swings over prior year revenue. Document Matching has the lowest variability, but it also brings in the least revenue of the four enforcement areas.

FIGURE 2. Change in Total Enforcement Revenue from Prior Year



SOURCE: IRS (2019a)

The drivers of fluctuations in TERC are not completely understood. A major area of scrutiny is IRS resources. The Treasury Inspector General for Tax Administration (TIGTA) finds that recent increases in enforcement revenue were driven by automated notices and processes (such as ACS), while decreases in staffing reduced labor-intensive enforcement activities (such as field exams) (TIGTA (2019)). In fact, the Congressional Budget Office (CBO) predicts that a $20 billion increase in IRS enforcement funding over 10 years could increase enforcement revenue by $61 billion and a $40 billion funding increase could lead to $103 billion higher revenue (Congressional Budget Office (2020)). CBO notes other factors that affect the size of the tax gap (and thus enforcement revenue), such as tax policies related to third-party reporting, tax code complexity, and IRS taxpayer assistance such as educational materials and Taxpayer Assistance Centers. Previous studies of TERC have identified factors related to economic outcomes and taxpayer attributes that may also affect enforcement revenue (Macheret *et al.* (2020)).

2.2 Research Questions

This study addresses the following research questions:

1. What is the correlation between macro-level variables and TERC outcomes?
2. What are the forecasted values of TERC based on these macro-level influencers?

This study seeks to understand the drivers of total enforcement revenue for Appeals, Collection, Exam, and AUR on the basis of economic variables, taxpayer attributes, and IRS resources. We show the percentage of variation in total revenue explainable by influencer variables, and we use this model to produce a two-year forecast of total revenue. This will help IRS anticipate enforcement revenue in a changing economic environment and evaluate revenue implications of high-level IRS policies such as total workforce levels.

2.3 Data and Methods

2.3.1 Methods

Conceptual Framework

IRS enforcement revenue can change on the basis of "demand-side" factors related to total tax and the tax gap, such as the number of people required to file a tax return or the level of noncompliance. It is also dependent on "supply-side" factors that affect the ability of IRS to enforce compliance, such as IRS enforcement staffing levels. We develop a conceptual framework for predicting enforcement revenue as a function of the total true tax, voluntary compliance rate, and recovery rate:

$$\textit{IRS Enforcement Revenue} = f(\textit{total true tax}, \textit{voluntary compliance rate}, \textit{recovery rate})$$

The **total true tax** is the estimated total true tax liability in the U.S. It is estimated to be around $2.683 trillion on average during Tax Years 2011 through 2013 (TY2011–TY2013) (IRS (2019b)). Total true tax depends on various demographic factors (such as birth and death rates and immigration), economic factors (such as gross domestic product (GDP) growth, labor force statistics, and inflation), and changes in tax code.

The voluntary compliance rate reflects the amount of total true tax that is paid on time and voluntarily. This rate is estimated to be around 82-84 percent for TY2008–TY2013 (IRS (2019b)). It is calculated on the basis of the gross tax gap, which includes the nonfiling gap, the underreporting gap, and the underpayment gap.[7] The net tax gap is the gross tax gap minus taxes collected through IRS enforcement activities and other late payments.

We define the recovery rate as the percent of the gross tax gap recovered through enforcement activities (e.g., accounting for the recovery rate, what is left is the net tax gap). This is not directly measured, since enforcement revenue credited to a fiscal year may cover tax returns filed in multiple tax years. However, a related measure is "Enforced and Other Late Payments," which includes tax owed for the tax year that is paid late or due to enforcement. Enforced and other late payments made up 13.6 percent of the gross tax gap in TY2011–TY2013 (IRS (2019b)).

A change in any of these three factors—total true tax, voluntary compliance rate, and recovery rate—can change IRS enforcement revenue. An increase in total true tax, keeping constant the voluntary compliance rate and recovery rate, will increase enforcement revenues by increasing the gross tax gap. This may occur with stronger economic growth or population growth, for instance. An increase in the voluntary compliance rate, holding total true tax and the recovery rate constant, will decrease enforcement revenues by lowering the gross tax gap.[8] This could occur through greater tax education, for example. Finally, an increase in the recovery rate, holding the other two factors constant, can also increase enforcement revenues. This could occur through hiring of more enforcement staff or increased enforcement productivity.

Econometric Framework

We do not model total true tax, voluntary compliance rate, or recovery rate directly.[9] Instead, we predict enforcement revenue on the basis of variables that influence these three factors. Our influencers cover economic variables, IRS resource variables, and taxpayer attributes. Some influencers may affect enforcement revenue through more than one channel. For example, a stronger economy can increase the total true tax by raising employee wages and increasing corporate revenues. It can also raise the voluntary compliance rate by making taxpayers more able to pay their taxes on time (or conversely, decrease the voluntary compliance rate by motivating some taxpayers to seek more avenues for hiding income). A stronger economy can also affect the recovery rate by influencing which tax returns are selected for examination and how collectable assessed changes are.

[7] For IRS estimates of the tax gap, see IRS (2019b).

[8] To the extent that the increase in voluntary compliance arises from taxpayers who would not be subject to enforcement in the first place, enforcement revenues could stay constant.

[9] These concepts are related to the tax gap, which is not estimated frequently and recently enough for our purposes.

We estimate a separate equation for each of our four monthly revenue variables (Appeals, Collection, Exam, and AUR). Since each enforcement activity involves inherently different returns, estimating separate equations allows for the relationships between predictors and revenue to vary as well. Our basic regression equation, using Collection as an example, is:

$$\begin{aligned}\text{Collection monthly revenue}_t &= \beta_o + \beta_1 CPI_{t-24} + \beta_2 nasdaq_{t-24} + \beta_3 pbank_{t-24} + \beta_4 bbank_{t-24} \\ &+ \beta_5 unemp_{t-24} + \beta_6 housessold_{t-24} + \beta_7 cci_{t-24} + \beta_8 gdp_{t-24} \\ &+ \beta_9 buildpermits_{t-24} + \beta_{10} rent_{t-24} \\ &+ \beta_{11} Emonth.business.systems.modernization_{t-24} \\ &+ \beta_{12} FTEmonth.examinations.and.collections_{t-24} \\ &+ \beta_{13} FTEmonth.filing.and.account.services_{t-24} \\ &+ \beta_{14} FTEmonth.investigations_{t-24} \\ &+ \beta_{15} FTEmonth.prefiling.taxpayer.assistance.and.education_{t-24} \\ &+ \beta_{16} FTEmonth.regulatory_{t-24} \\ &+ \beta_{17} FTEmonth.shared.services.and.support_{t-24} + \beta_{18} perc65_{t-24} \\ &+ \beta_{19} realwage_{t-24} + \beta_{20} top1share_{t-24} + \alpha_m\, month_m + \pi cyear_y + \varepsilon_t\end{aligned}$$

Monthly revenue in month t is modeled as a function of predictor variables. The main predictor variables are the economic, IRS resource, and taxpayer variables. These variables are lagged by two years (24 months) in our main model to allow for forecasting. We also include a month fixed effect (*month*) to capture seasonal variation in revenue collected. Calendar year (*cyear*) is included to control for long-term trends in revenue. ε_t is the residual variation in monthly revenue that is not related to the predictor variables.

2.3.2 Data

Outcome Variable

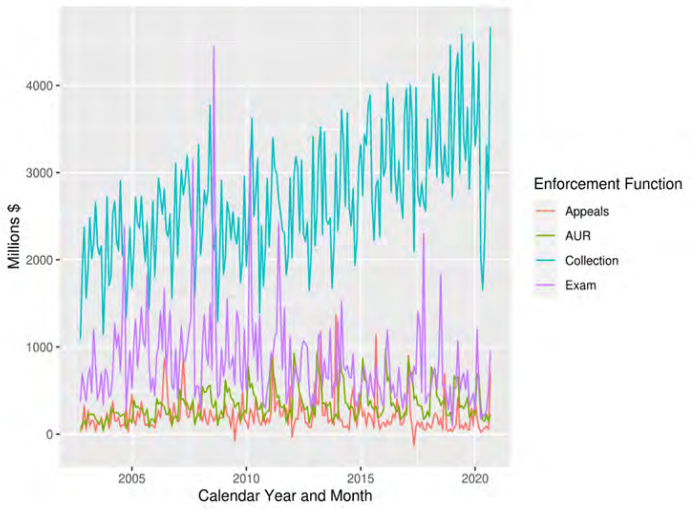
We focus on total enforcement revenue collected (TERC) as the primary outcome of interest. We aggregate TERC up to the fiscal year, collection month, and enforcement function level. The monthly data are expressed, for example, in terms of revenue collected in March 2018 for all Exam activities. Our revenue variable includes taxes, interest, and penalties collected. Revenue is also aggregated across IRS business operating divisions[10] and across tax return groups.[11]

Figure 3 plots revenue for each collection month in calendar years 2003–2020. There are seasonal trends in revenue for all for enforcement functions and a long-term upward trend in Collection revenue. Current month revenue is mostly positive, except for a few months within the Appeals function.

[10] These are Wage & Investment (W&I), Small Business/Self Employed (SB/SE), Large Business & International (LB&I), and Tax Exempt/Government Entities (TEGE).

[11] There are 12 tax return groups based on the Master File Tax Account Code (MFT).

FIGURE 3. Enforcement Revenue by Report Case Area and Collection Month (Nominal Dollars)



NOTE: AUR is Automated Underreporter Program (document matching)

Economic Variables

Economic variables can affect enforcement revenue by influencing total true tax, the voluntary compliance rate, or the recovery rate. The economic variables we include are:

- Consumer price index (CPI), Bureau of Labor Statistics
- NASDAQ Composite index close level, NASDAQ OMX Group
- Personal and business bankruptcies, American Bankruptcy Institute
- Unemployment rate, Bureau of Labor Statistics
- Housing purchases, U.S. Census Bureau
- Consumer confidence index, University of Michigan Survey of Consumers
- Gross domestic product, Bureau of Economic Analysis
- Construction building permits, Census Bureau Building Permits Survey
- Rental expenditures, Bureau of Labor Statistics [12]

The CPI can affect the nominal dollars received through enforcement due to inflation. Other economic variables, such as the NASDAQ close level, housing purchases, and consumer confidence index, are leading economic indicators (i.e., predicting the direction of economic growth) (The Conference Board (2020)). Variables such as the unemployment rate and bankruptcies are also important indicators of economic health. We acquire monthly data for these indicators and, when necessary, we interpolate monthly values from quarterly or annual values.[13]

IRS Resource Variables

We take a general approach to measuring IRS resources in order to capture the drivers of TERC across all business categories. Some budget activities are directly linked to enforcement, such as examinations and collections. However, nonenforcement activities also influence taxpayer compliance. For example, more resources dedicated to prefiling taxpayer assistance and education may improve the voluntary compliance rate. More resources dedicated to criminal investigations may improve deterrence—TIGTA found that the small number

[12] Rents, royalties, and proprietor income are subject to little or no third-party reporting and therefore affect the tax gap. See Congressional Budget Office (2020).

[13] We calculate the compound monthly growth rate (CMGR) from quarterly or annual values and apply the CMGR to interpolate monthly values. In other words, monthly values are assumed to grow at a constant rate within a quarter or year.

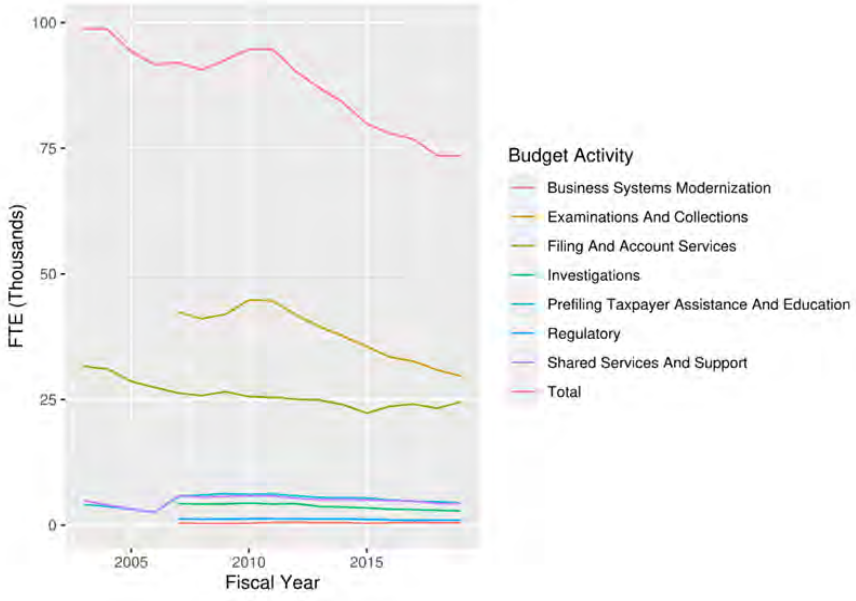
of criminal convictions each year likely does little to deter employers from willfully avoiding employment tax payments (TIGTA (2017)).

Our primary source of IRS resource data is the IRS Data Book (IRS (2020)). The Data Book is published annually and contains statistical tables covering the entire Service. Table 32 of the Data Book presents the IRS personnel summary by employment status, by budget activity, and for selected type of personnel.[14] In particular, we use the number of full-time equivalent positions (FTEs) by budget activity. Definitions of budget activities vary somewhat over the years but largely cover:

- Examinations and collections;
- Filing and account services;
- Prefiling taxpayer assistance and education;
- Shared services and support;
- Investigations;
- Regulatory; and
- Business systems modernization.

Figure 4 summarizes the number of IRS FTEs by budget activity and fiscal year. Some budget activities, such as business system modernization and examinations and collections, were not defined until 2007. The total number of FTEs across the IRS has fallen, from almost 100,000 in FY2003 to about 74,000 in FY2019. In fact, all budget activities except for business systems modernization employed fewer FTEs in FY2019 than in FY2007.

FIGURE 4. IRS FTEs by Budget Activity and Fiscal Year



IRS policies not captured by FTEs can also affect TERC. For example, the Tax Cuts and Jobs Act of 2017 contained 119 tax provisions and required IRS to produce more than 500 new tax products (TIGTA (2019)). This puts a strain on IRS resources and requires taxpayers to become educated on the new laws. Policies such as these can affect TERC through several channels (total tax, voluntary compliance rate, or recovery rate), but are not easily measurable. In our regressions, we include dummy variables (fixed effects) for each calendar year to capture when new policies take effect.

[14] The format of the IRS Data Book has changed over the years. The personnel summary is Table 32 in FY2019, Table 30 in FY2006–FY2018, and Table 32 in FY2003–FY2005.

Taxpayer Attributes

Previous MITRE research uses a set of taxpayer attributes that cover demographic, asset, and expense information (Macheret *et al.* (2020)). These variables are drawn from the individual returns transaction file database in the Compliance Data Warehouse (CDW), an internal research database within the IRS. The attributes include taxpayer age, net worth, and adjusted gross income (AGI), among other characteristics. While microdata provide information at a highly granular level, we take a broader approach by using aggregate characteristics of the taxpayer population at the monthly level. Aggregate taxpayer characteristics are measured at the same level as our TERC variable (i.e., across the entirety of IRS enforcement) rather than on a by-taxpayer basis.

We include these demographic and income variables:

- Population 65 and above, World Bank;
- Inflation-adjusted income, Bureau of Labor Statistics; and
- Share of total net worth held by top 1 percent, Board of Governors of the Federal Reserve System.

The share of the population 65 and above proxies the retirement population, which has unique tax circumstances compared to wage earners. Inflation-adjusted income is expressed as the median weekly wage and captures wage growth representative of the general population. Share of total net worth held by the top 1 percent indicates the growth of wealth at the high end of the distribution. Changing wealth and income distributions have implications for total tax and also affect enforcement revenue. For example, the closing of a high-income enforcement case can yield substantially more revenue than do typical cases.

Table 5 in the Appendix summarizes the outcome and influencer variables. Dollar-denominated variables are adjusted for inflation. Some variables have fewer observations than do others due to data availability, such as a lag in data reporting or a change in variable definitions.

2.4 Results

2.4.1 Correlation Analysis

Table 1 presents the results of our main regression on the four enforcement revenue variables. A total of 144 observations (i.e., by month and year) are used in each model. These are the observations for which all revenue and predictor variables are nonmissing. The adjusted R^2 value in the bottom panel is a measure of model performance. It indicates the percent of variation in monthly revenue that is explained by the predictor variables. The model does best with AUR revenue, explaining 92 percent of its monthly variation. Sixty percent of Collection revenue variance is explainable by predictor variables, while 30 percent of Exam revenue variance is explainable by predictor variables. The adjusted R^2 value for Appeals is slightly negative, indicating that the model provides a very poor fit of that revenue.[15]

Table 1 also displays individual coefficients and their standard errors in parentheses. Only the main predictor variables are displayed. (Month fixed effects are included in the model for seasonality trends but are not displayed.) Across all four revenue variables, certain predictor variables have a consistently positive or consistently negative association with revenue (i.e., consistent sign in at least three of four models). Although many coefficients are not statistically significant, these patterns appear:

- **Positive association with revenue**: Predictors that consistently have a positive association with revenue variables are two-year lags of the NASDAQ Composite level, business bankruptcies, business systems modernization FTEs, Exam/Collections FTEs, and the share of wealth held by the top 1 percent.
- **Negative association with revenue**: Predictors that consistently have a negative association with revenue variables are two-year lags of personal bankruptcies, unemployment rate, houses sold, the consumer confidence index, GDP, building permits, filing and account services FTEs, investigations FTEs, prefiling taxpayer assistance/education FTEs, and the percent of population 65 and older.

[15] Both Appeals and Exam involve diverse tax returns filed over a range of years, thereby complicating the association between revenue and economic conditions at any one point in time.

Note these patterns do not necessarily indicate a causal interpretation and could be due to random correlation.

TABLE 1. Main Regression Results

Independent Variables	Dependent Variable			
	Monthly Appeals Revenue (1)	Monthly Collection Revenue (2)	Monthly Exam Revenue (3)	Monthly AUR Revenue (4)
cyear	-229.04	512.44	-209.87	181.35***
	(185.52)	(334.56)	(326.70)	(49.99)
CPI	0.10	-52.99**	2.04	-17.46***
	(14.66)	(26.44)	(25.82)	(3.95)
nasdaq	30.17	61.33	31.90	55.02***
	(63.88)	(115.20)	(112.49)	(17.21)
pbank	-0.85	-17.69**	2.27	-3.68***
	(4.83)	(8.71)	(8.51)	(1.30)
bbank	-2.69	484.18**	200.88	56.84*
	(109.10)	(196.74)	(192.12)	(29.40)
unemp	46.26	-112.62	-15.77	-12.55
	(49.49)	(89.25)	(87.15)	(13.34)
housessold	-3.17	-4.34	-4.79	-0.46
	(3.07)	(5.54)	(5.41)	(0.83)
cci	-1.97	-3.77	-2.67	-1.02
	(2.37)	(4.28)	(4.18)	(0.64)
gdp	-93.86	-233.79	-23.28	-197.25**
	(336.54)	(606.90)	(592.64)	(90.68)
buildpermits	-65.47	-252.87	-23.76	27.36
	(171.60)	(309.45)	(302.18)	(46.24)
rent	72.20	-252.27**	90.72	-28.48
	(64.08)	(115.57)	(112.85)	(17.27)
FTEmonth.business.systems.modernization	99.89	729.17	825.86	-50.43
	(291.08)	(524.92)	(512.59)	(78.43)
FTEmonth.examinations.and.collections	6.41	18.43	-58.86	5.52
	(27.67)	(49.90)	(48.73)	(7.46)
FTEmonth.filing.and.account.services	-16.78	-44.21	-88.45	13.21
	(43.52)	(78.48)	(76.64)	(11.73)
FTEmonth.investigations	-102.29	-70.17	-231.64	54.25
	(183.78)	(331.43)	(323.64)	(49.52)
FTEmonth.prefiling.taxpayer.assistance.and.education	-69.91	-66.52	-333.27	56.09
	(206.62)	(372.60)	(363.85)	(55.67)
FTEmonth.regulatory	432.71	-581.53	1,337.06	-519.72***
	(625.88)	(1,128.69)	(1,102.17)	(168.64)
FTEmonth.shared.services.and.support	-129.21	18.34	-108.16	14.77
	(120.07)	(216.53)	(211.44)	(32.35)
perc65	515.37	-678.80	-202.42	-292.79**
	(511.01)	(921.54)	(899.89)	(137.69)
realwage	-3.37	8.98	8.70	-3.45**
	(5.77)	(10.41)	(10.16)	(1.55)
top1share	17.30	269.65**	159.39	22.84
	(60.95)	(109.91)	(107.33)	(16.42)
Constant	455,651.30	-1,012,708.00	421,851.00	-355,293.50***
	(365,465.10)	(659,064.50)	(643,580.10)	(98,471.98)
Observations	144	144	144	144
Adjusted R^2	-0.003	0.60	0.30	0.92

NOTES: Standard errors in parentheses. All models include month fixed effects and year trend.
*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

Sensitivity Analysis of Correlation Results

The fact that only a few predictor variables have statistically significant associations in Table 1 may be due to the fact that some variables act on revenue with a deeper (or shallower) lag than two years. We conduct sensitivity analysis using two alternate approaches: one with contemporaneous (same period) values of the predictors and one with the "optimal" lags corresponding to lags with the strongest correlation with the TERC variables.[16] The primary takeaways from the sensitivity analysis are these:

- **There are tradeoffs in optimizing model fit**: The model with optimal lags of predictors has the highest adjusted R2 among all models for Appeals, Exam, and AUR revenue. The model with contemporaneous values has the highest adjusted R2 among all models for Collection revenue. Optimal lags should theoretically improve model fit, since lags are chosen on the basis of the highest degree of correlation with revenue. However, using long lags reduces the number of complete observations (sample size), thereby reducing a model's power.
- **Selection of lags matters for the sign of the association**: The association between revenue and some predictor variables switches signs when different lags are applied. For example, CPI has a negative association when its two-year lag is used, but a positive association when its contemporaneous value is used. This likely reflects the delays between economic conditions, tax filing, enforcement activity, and revenue collection.
- **Selection of lags matters for statistical significance**: The statistical significance of some coefficients changes depending on which lag is used. For example, the unemployment rate has a positive association with Appeals revenue in all three models, but this association is only statistically significant in the optimal lag model.

Overall, the sensitivity analysis indicates the difficulty in discerning the individual relationships between monthly enforcement revenue and economic, IRS resource, and taxpayer data. The selection of lags is an important consideration, although this is limited by the need to *forecast* revenue using this model. There are also likely other predictor variables that could be included. Regardless of whether a variable has a statistically significant coefficient for *association* purposes, it can be useful for *forecasting* purposes.[17]

2.4.2 Forecasting Analysis

The regressions presented in Table 1 contain two-year lags of the predictor variables, which accommodate a forecast of revenue up to two years ahead. The forecast window is limited by how recently predictor variables are updated. Although many variables, including monthly revenue and economic data, are updated through September 2020, the IRS FTE data end in September 2019.[18] Thus, we are able to forecast two years ahead of September 2019 (through September 2021). Predicted revenues for FY2020 (October 2019–September 2020) can be compared against actual revenues as validation.

Figure 5 through Figure 8 show our forecasts of Appeals, Collection, Exam, and AUR monthly revenue. Note that all revenues are in real (inflation-adjusted) terms, not nominal terms. The black dots in each figure represent actual monthly revenues. These are observed for October 2002–September 2020. The blue line is the predicted value of revenue from the regression model. The predicted values begin 24 months after the first month of "complete data" (i.e., first month for which all variables are nonmissing) and end two years ahead of the last month of complete data. The gray ribbon shows the 95th percent confidence interval of the predicted value ("error bars").

Figure 5 shows the forecast of Appeals revenue. The model is a fairly good fit of Appeals revenue historically, with the exception of some high-value outliers during the middle of the period. It is also a fairly good fit

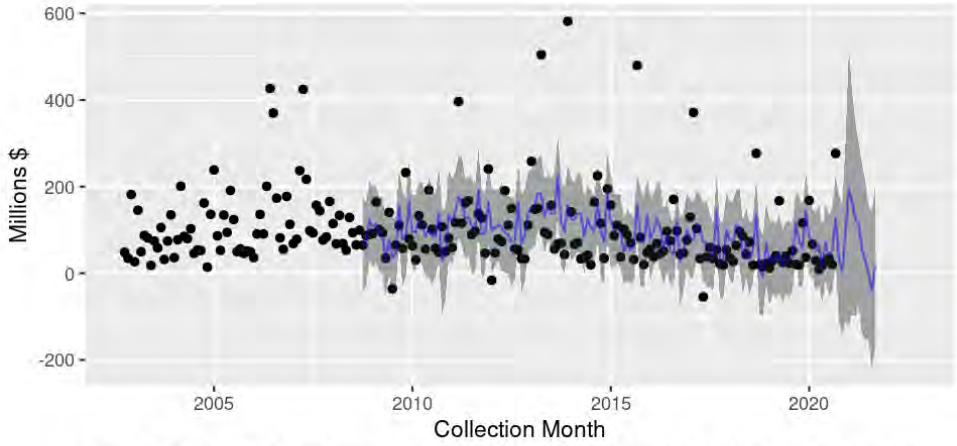
[16] We estimate correlation coefficients for each influencer variable and outcome variable, using one- to 36-month lags. For length, we have excluded the regression tables for the sensitivity analyses.

[17] For example, a model could suffer from "omitted variable bias" whereby the association between a predictor and the outcome is actually due to a third, unmeasured variable. This would invalidate any conclusions about the causal effect of that predictor. However, the model could still prove useful for forecasting if the predictor is closely associated with the third unmeasured variable.

[18] At the time of this study.

in FY2020 except for the last month. The error bars around our forecasts are wide, indicating a high degree of uncertainty in the estimates. This is unsurprising given the poor fit of the model as indicated in Column 1 of Table 1. Appeals revenue was negative for three months historically, and the model predicts that it will again dip into negative territory near the end of 2021.

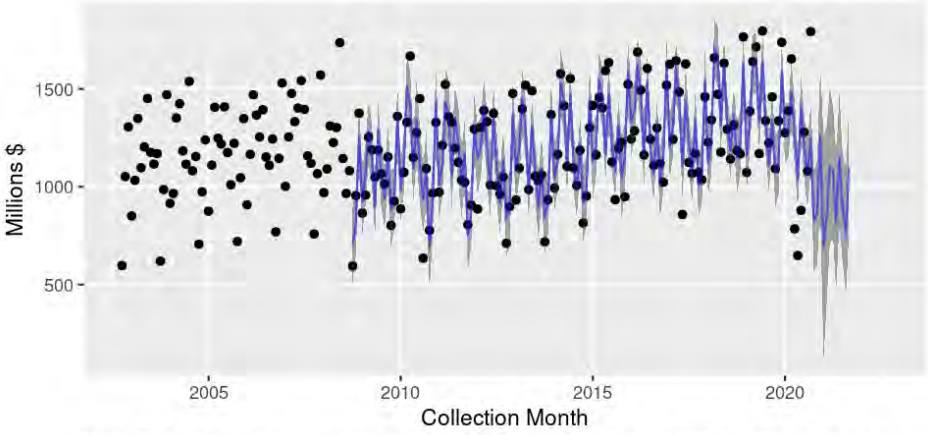
FIGURE 5. Forecast of Monthly Appeals Revenue (Inflation-Adjusted)



Notes: Black dots indicate actual monthly revenues, which are observed from Oct 2002-Sep 2020. The blue line indicates the point estimate of our predicted monthly revenue, and the gray ribbon indicates the 95% confidence interval of that estimate.

Figure 6 shows the forecast of Collection revenue. The model predicts an overall downward trend in Collection revenue for the next two years, with some cyclical variation. For FY2020, predicted revenue matches actual revenue for more moderate values but does not capture the high and low outliers. Collection revenue appears to exhibit larger cyclical swings in the last year (punctuated by a few low months) than in previous years. Nevertheless, the model predicts an average of around $1 billion (in real terms) in monthly Collection revenue in the coming months.

FIGURE 6. Forecast of Monthly Collection Revenue (Inflation-Adjusted)

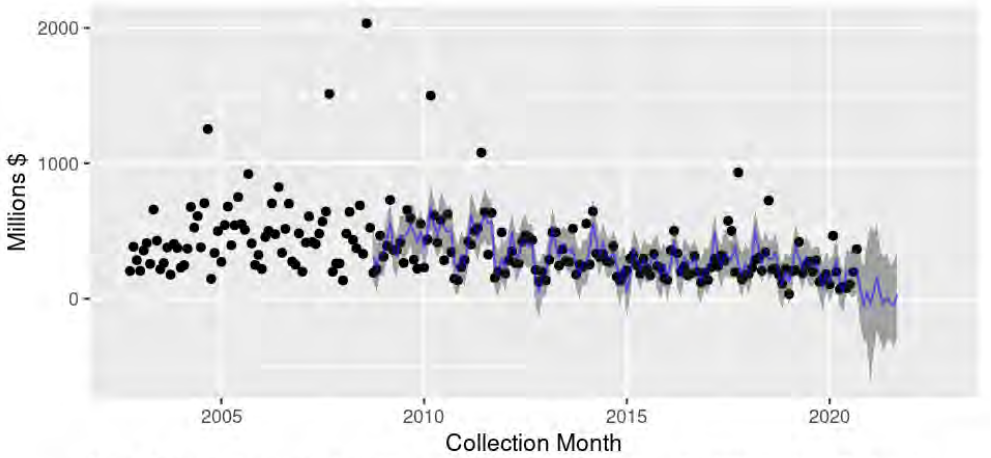


Notes: Black dots indicate actual monthly revenues, which are observed from Oct 2002-Sep 2020. The blue line indicates the point estimate of our predicted monthly revenue, and the gray ribbon indicates the 95% confidence interval of that estimate.

Figure 7 shows the forecast of Exam revenue. Predicted values of revenue track fairly closely with actual revenues in FY2020. However, the model appears to pick up a long-term declining trend, driving forecasts down to slightly negative for a few months in FY2021. This should be interpreted with caution, as Exam revenue has not been negative in any month historically. Recent declines in economic conditions may be driving this downward trend, whereby historical relationships between economic variables and Exam revenue do not

hold. We conduct sensitivity analysis in the next section to assess whether alternate models also forecast these negative values.

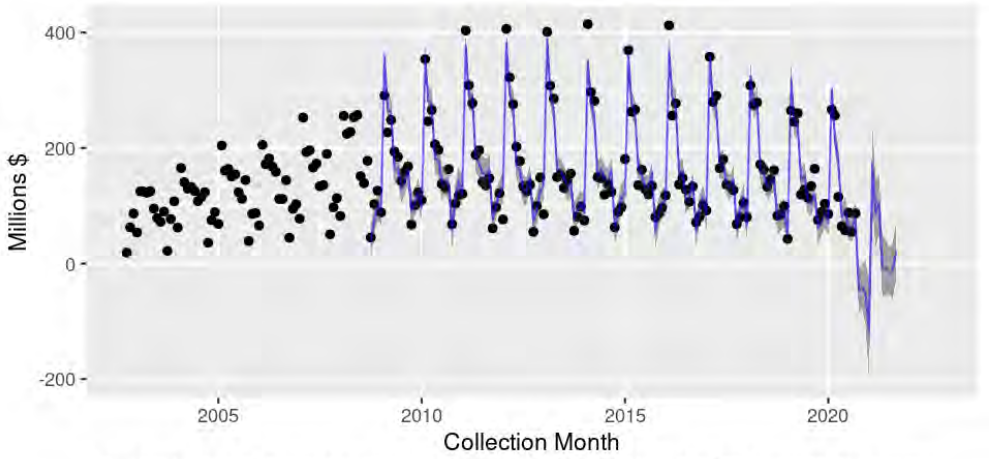
FIGURE 7. Forecast of Monthly Exam Revenue (Inflation-Adjusted)



Notes: Black dots indicate actual monthly revenues, which are observed from Oct 2002-Sep 2020. The blue line indicates the point estimate of our predicted monthly revenue, and the gray ribbon indicates the 95% confidence interval of that estimate.

Figure 8 shows the forecast of AUR revenue. Predicted values of revenue track closely with actual revenue historically as well as in FY2020. As with Exam, the model forecasts negative values of AUR revenue in the coming months. This should also be interpreted with caution, as AUR revenue has not dipped below zero historically. The negative forecasts could be due to the influence of a few dominant predictors. Table 1 shows that regulatory FTEs and GDP have large negative effects on AUR revenue. Both measures have declined in recent years, which may drive down the forecasted values of revenue. We conduct sensitivity analysis in the next section to see if negative forecasts persist under an alternate model specification.

FIGURE 8. Forecast of Monthly AUR Revenue (Inflation-Adjusted)



Notes: Black dots indicate actual monthly revenues, which are observed from Oct 2002-Sep 2020. The blue line indicates the point estimate of our predicted monthly revenue, and the gray ribbon indicates the 95% confidence interval of that estimate.

Sensitivity Analysis of Forecasting Results

Forecasts of enforcement revenue using the main regression model may not be robust. Historical relationships between revenue and variables such as IRS FTEs may not be generalizable if, for example, FTEs dip to unprecedented levels. We conduct an alternate forecast using a regression model that only uses seasonal (month fixed effects) and long-term (calendar year) trends as predictors.[19] Economic, IRS resource, and taxpayer attributes are not included in this alternate model. The primary takeaways from this sensitivity analysis are these:

- **Simple models perform better in some cases**: For Appeals and Exam, the alternate model using just seasonal and long-term trends yields forecasts that are more certain (with smaller error bars) and more reasonable (with no negative values) than our main regression using economic, IRS FTE, and taxpayer variables.
- **Simple models are not restricted by dated predictor variables**: Since only month and year are used to forecast revenue, the alternate model is not limited by how recently predictor variables were updated (e.g., in the case of the IRS FTE variables). We are able to forecast two years ahead from September 2020 (the most recent monthly revenue data).
- **Simple models are unable to account for idiosyncrasies**: The alternate model assumes a constant seasonal cycle and a steady long-term trend. It is unable to account for idiosyncratic movements in revenue. For example, there are high-value outliers in Appeals and AUR revenue that fall outside of the main cyclical movements. Also, there is a recent declining trend in Collection revenue that does not match long-term trends.
- **Simple models do not explain the "why"**: The alternate model simply assumes cyclical movements in revenue and a long-term trend. Cyclical movements could arise from logistical patterns in tax collection during the fiscal year. However, the reasons for any long-term trends are not immediately clear.

Ultimately, a simpler forecasting model may perform best in producing precise forecasts that can yield meaningful insights for workforce allocation. This would suggest choosing a model with just seasonal and long-term trends or a model with these trends and only a handful of influencer variables.

2.5 Discussion

In this report, we have built regression models that estimate associations between IRS enforcement revenue and economic, IRS resource, and taxpayer attributes. We also use the models to forecast enforcement revenue one to two years forward. Through sensitivity analysis, we show the benefits and drawbacks of using alternate specifications, such as different lags of predictor variables or a simplified forecasting model excluding external predictors. This study produced these key insights:

- **Different enforcement activities require different models**: We estimated separate equations for each enforcement activity (Appeals, Collection, Exam, and AUR) and found associations to differ by activity. This confirms the intuition that these activities are inherently different, acting on different types of taxpayers/tax returns and are subject to different timetables.
- **A simple model has advantages in forecasting**: A simplified model using only seasonal and long-term trends produces more precise estimates and more realistic ones (such as only positive predicted values for revenues that historically did not dip below zero). It is also not limited by delays in data availability for predictor variables. These advantages are useful for workforce allocation purposes.
- **A simple model cannot explain all revenue fluctuations**: A simplified model using only seasonal and long-term trends failed to account for time periods in which cyclical patterns became more dramatic or shifts in the long-term trend.
- **Causal stories are difficult to tell, especially for some enforcement activities**: The individual effects of the predictor variables often depend on the lag chosen for that variable. On the whole, fluctuations

[19] Recall that all influencer variables (except for preparer assistance) are measured in thousands of dollars (see Table 6).

in Appeals revenue are very difficult to explain, while fluctuations in AUR revenue are easier to explain (based on adjusted R^2 values).

- **IRS resources have some association with revenue, but other factors come into play**: Sensitivity analysis shows that the number of Collection and Exam FTEs six months prior has a positive, statistically significant effect on both Collection and Exam monthly revenue. However, many of the other FTE variables were not statistically significant in most models.

Future Work

This study builds a prototype for understanding TERC influencers and forecasting. In addition to early results, the prototype includes a conceptual framework, econometric specification, and framework for interpretation of findings. We envision various areas of potential future work to refine the prototype, which areas include but are not limited to:

- **Alternate outcome variables**: Future work can forecast TERC at a more granular level than the overall enforcement function, such as breaking out Collection into individual and business case revenue. Previous MITRE work also presented some entirely different outcomes of interest (Macheret *et al.* (2020)). For example, "enforcement effectiveness" is defined as TERC divided by enforcement net assessment.[20] Another possible measure is return on investment, which is TERC divided by enforcement investment. Enforcement investment can be measured in terms of FTEs or some other measure.
- **Alternate model specifications**: We can improve model fit through a variety of techniques, including exploring different transformations of the variables (such as log-transforming revenues), adding or removing predictor variables, or imputing missing values of the data. Akaike information criterion and Bayesian information criterion are model statistics that can be compared to see what model provides the best predictive fit of TERC.
- **Alternate regression methods**: Advanced machine learning methods, such as artificial neural networks (ANN), are more flexible than linear regression and thus may offer better predictive power. However, linear regression produces coefficients that are easily interpretable for causation, while highly nonlinear methods such as ANN do not.

3. Micro Influencers of IRS Enforcement Revenue and Implications for Workforce Allocation

While the "macro" model provides a big-picture view of TERC, it does not model revenue at a sufficiently detailed level for operational purposes. In this study, we develop a "micro" model to forecast revenue at the enforcement-step level. The forecasts are used in MITRE's Program Assessment Model (PAM) for workforce allocation, which models the IRS enforcement chain and optimizes resource allocation on the basis of expected revenue per case at each step of the chain.[21] We demonstrate that using *forecasted* revenue per case rather than *historical* estimates leads to workforce allocations that can generate higher revenue across the enforcement system (i.e., higher TERC).

3.1 Background

The IRS enforcement workflow involves many possible routings for an enforcement case. Figure 9 illustrates the enforcement chain modeled within PAM. PAM was developed to optimize workforce allocation for the IRS Small Business/Self-Employed (SB/SE) Division. As such, PAM does not cover workforce allocations for other IRS enforcement divisions.[22] The Appendix provides a review of PAM.

Within PAM, we focus on taxpayers who enter the enforcement chain by first experiencing a correspondence audit (Group A in Figure 9). Correspondence exams are organized around a specific line item or

[20] 0.990 – 1 = -0.01 or -1 percent.
[21] 0.940 – 1 = -0.06 or -6 percent.
[22] 0.729 – 1 = -0.271 or -27 percent.

schedule. These examinations are conducted by letter and are limited in scope to the line item in question. The exam may end with an assessment requiring the taxpayer to pay a balance. The taxpayer would be issued a series of notices alerting them to the balance. If the taxpayer pays the assessed amount, their enforcement case will be subsequently closed.

If the taxpayer fails to pay the full assessed amount, their case enters the balance due status (Group B in Figure 9). The "Collection_Bal_Due" step of PAM mimics the function of the Inventory Delivery System as a clearinghouse for balance-due tax modules and assigns those modules to one of the various collection programs.

For the purposes of this study, we focus on two collection programs: Automated Collection System (ACS) (Group C1 in Figure 9) and Field Collection (Group C2 in Figure 9). These steps are used for demonstration because they collect revenues that are large but not so large as to be subdivided into further steps within PAM (which would complicate our analysis).

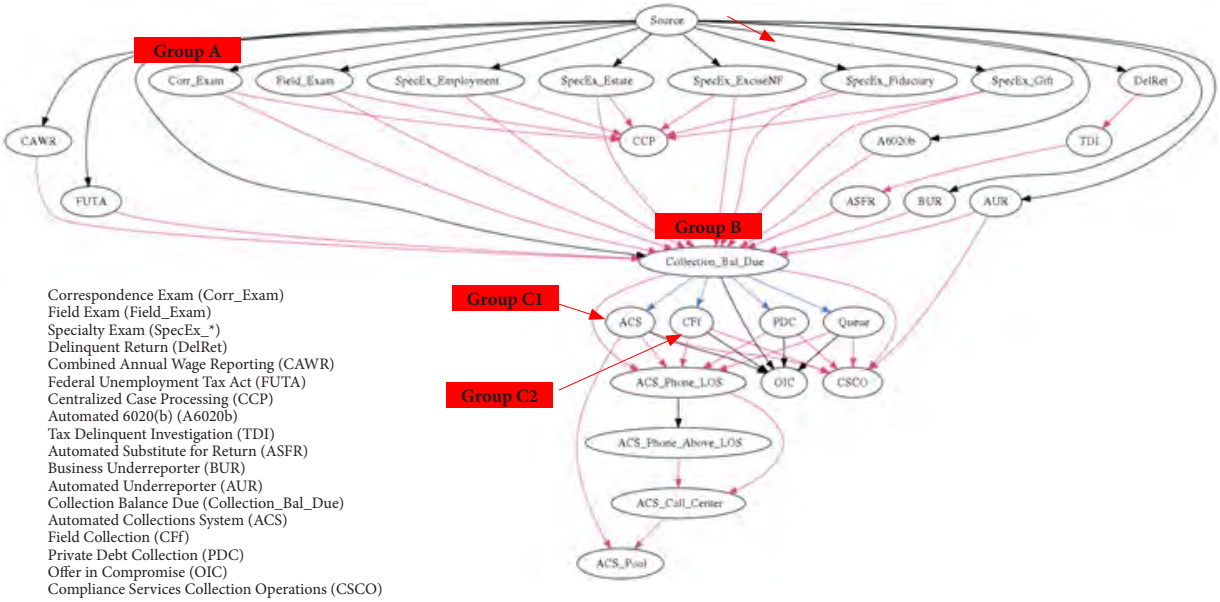
ACS

ACS works balance-due and nonfiler cases requiring telephone contact. In addition to phone calls, ACS issues notices, liens, and levies. The type of cases being worked by ACS are not the highest-value cases (in terms of balance due), which are generally worked by Field Collection. They are also not the lowest-value cases, which are not actively worked, nor the least likely to result in collection, which are given to Private Debt Collection (PDC).

Field Collection

Field Collection (or Collection Field Function (CFf)) works cases that tend to be older, more complex, and with higher balances. Most CFf cases proceed directly from balance due, although in rare cases, some can proceed to CFf from ACS. In CFf, revenue officers make in-person visits to taxpayers to better understand their financial circumstances and determine their ability to pay taxes owed.

FIGURE 9. PAM Compliance Chain for SB/SE



In summary, the two enforcement steps we model in this report deal with enforcement cases that progress in the following manner:

1. Correspondence exam (Group A) → Collection balance due (Group B) → ACS (Group C1); and
2. Correspondence exam (Group A) → Collection balance due (Group B) → Field collection (Group C2).

Although we demonstrate our prototype on these two enforcement steps, our model can be expanded to other steps in future work.

Timing of Compliance Activities

Our forecasting approach takes advantage of natural time lags between enforcement steps. As an example, Figure 10 outlines a typical enforcement timeline for a taxpayer who is selected for a correspondence audit and who then progresses into Field Collection. The taxpayer may file a return in spring 2020 for income earned during TY2019. In December 2021, a correspondence exam may be opened for this tax return. The exam is conducted over the course of a year and is closed in December 2022. The taxpayer is given a year to pay the assessed balance. Suppose the taxpayer fails to pay, and the case ends up in balance due status. The case is then sent to Field Collection in December 2023. The case progresses through Field Collection, and the taxpayer pays the adjustment in December 2024. The enforcement case is then closed.

FIGURE 10. Example of Enforcement Case Timeline

2019	2020	2021	2022	2023	2024
January – December	Spring	December	December	December	December
Taxpayer earns income	Taxpayer files return for TY2019	Correspondence exam starts	Correspondence exam closes	Field collection begins	Adjustment collected; enforcement case closed

The timing of these events is important for workforce allocation across downstream enforcement steps. Suppose the IRS is evaluating workforce allocation in January 2022, with new hires to start in January 2023. In January 2022, the IRS is aware of which taxpayers have an open correspondence exam (as in the case of the taxpayer represented in Figure 10). Some of these taxpayers (but not all) will progress to Field Collection in 2023, when the new hires begin their assignments. Thus, the IRS can view the *potential* pool of Field Collection cases when it is deciding how to allocate workforce into Field Collection. In constructing our forecasting model of revenue per case, we exploit the time lag between active correspondence exams and collection activities for these cases.

3.2 Research Questions

This study addresses the following research questions:

1. What taxpayer characteristics influence expected revenue for certain steps in the enforcement chain?
2. What is the forecasted expected revenue for each enforcement step?
3. Does integrating forecasts of expected revenue into resource allocation models improve overall enforcement revenue?

3.3 Data and Methods

3.3.1 Methods

Two-Part Forecasting Model

We develop a two-part model to forecast revenue per case for ACS and CFf. First, we forecast which taxpayers with an active correspondence exam (Group A in Figure 9) will progress into ACS (Group C1 in Figure 9),

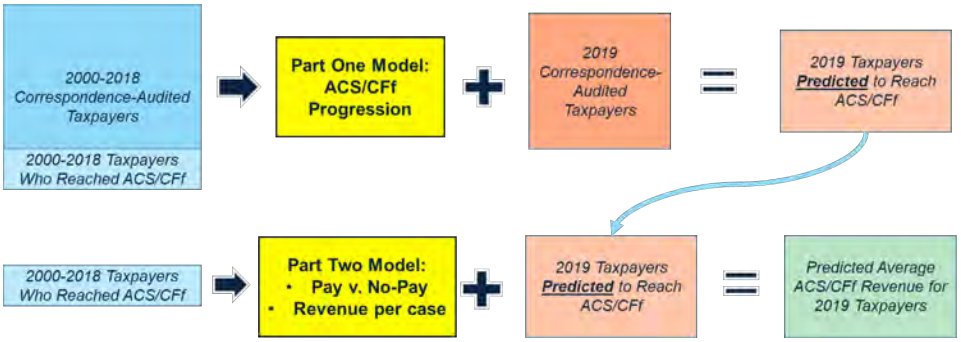
Field Collection (Group C2 in Figure 9), or another status. Second, on the basis of this forecasted group of ACS and CFf cases, we forecast the likelihood of making a payment and the average collection revenue per case.[23] These revenue parameters will be used to determine workforce allocation in PAM.

As outlined in Figure 9, allocation of new FTEs for ACS and CFf is affected by the rising cohort of ACS and CFf cases. This rising cohort is derived from the current cohort of active correspondence exams. However, we cannot yet observe who will progress from current correspondence exams into ACS or CFf, nor their future collection revenue upon case closure. However, we can observe these outcomes for *historical* enforcement cases. Our forecasting model is trained on historical data relating the attributes of correspondence-audited taxpayers to the outcomes observed for them (e.g., progression into ACS/CFf and collection revenue at those steps). The forecasting model applies these historical relationships to the current correspondence exam cohort to forecast revenue per case.

Figure 11 illustrates the two-part forecasting approach whereby historical data (TY2000-TY2018) is used to train the models, which are then applied to current data (TY2019) to generate expected revenue per case. The two-part forecasting model is summarized as follows:

- Part One Model: ACS/CFf progression
  - We forecast which taxpayers with an active correspondence exam will have a balance due and progress to ACS or Field Collection.
  - We employ a discrete choice model of whether a taxpayer advances to ACS, CFf, or another outcome. The model is trained on historical data from 2000-2018 and is applied to 2019 correspondence-audited taxpayers to forecast which ones progress to ACS or CFf.
  - The discrete choice model we use is called a multinomial logistic (logit) regression, which models multiple discrete (categorical) outcomes.
- Part Two Model: ACS/CFf revenue per case
  - We forecast the probability of payment and collection revenue for each taxpayer expected to progress into ACS or CFf (based on results from the Part One Model).
  - We employ a model used for continuous outcomes—the collection revenue for each ACS or CFf case. The model is trained on historical data from 2000–2018, based on the taxpayers who reached ACS or CFf during that time. The model is applied to 2019 taxpayers expected to reach ACS or CFf to forecast revenue per case for this rising cohort.
  - The model we use is called a zero-inflated model, which models continuous outcomes that have many zeros (i.e., cases with no revenue). A zero-inflated model contains two components: One predicts the likelihood of payment (versus paying nothing), and the other predicts the amount paid (if the taxpayer pays something).
  - Average revenue per case is calculated as the average across the rising ACS cohort (for ACS) and across the rising CFf cohort (for CFf).

FIGURE 11. Forecasting Approach



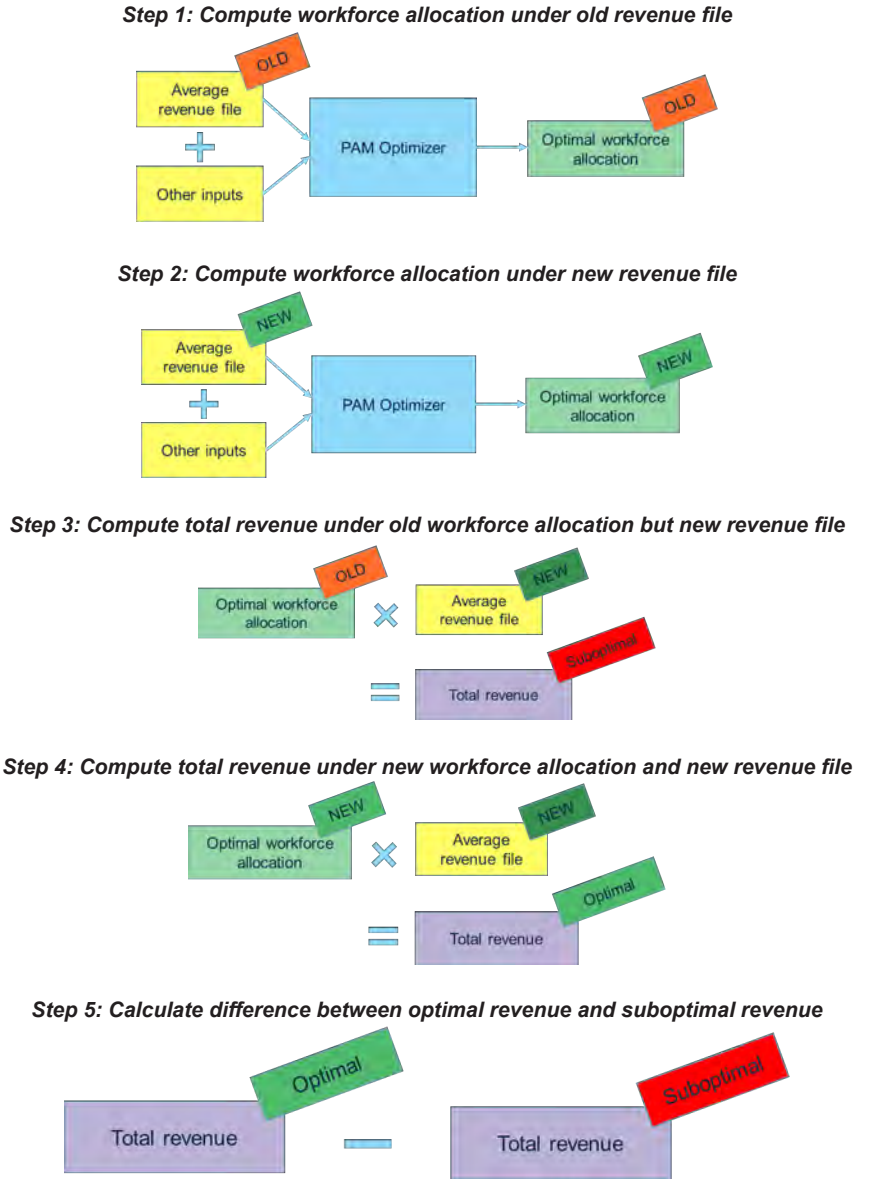
[23] 1.162 – 1 = 0.162 or 16%.

Incorporating Revenue Forecasts into PAM

We incorporate the forecasted revenue per case in the PAM model for workforce allocation. PAM uses historical revenue parameters to optimize workforce allocation. However, to the extent that future revenues do not reflect past revenues, the model yields suboptimal allocations. In particular, we quantify the amount of revenue lost if historical revenue parameters are not replaced with more updated ones.

Figure 12 shows the process by which we quantify the effect on TERC. First, we compute the "optimal" workforce allocation assuming the old revenue file (Step 1) and then again under the new revenue file (Step 2). Second, we calculate total revenue under the "old" approach (Step 3) and under the "new" approach (Step 4). In both cases, we multiply the allocated workforce by the "new" average revenue parameters, which yields the more accurate representation of FTE productivity. Finally, Step 5 calculates the difference in total revenue between the two cases to show the effect on TERC of using outdated revenue parameters.

FIGURE 12. Calculating the Effect on TERC



3.3.2 Data

As Figure 11 shows, our population of interest for the Part One Model is taxpayers who are issued a correspondence audit (some of whom then advance to ACS or CFf). For the Part Two Model, we are interested only in taxpayers who advance to ACS or CFf. As such, we select a 10-percent random sample of correspondence-audited taxpayers for returns filed in each tax year from 2000 to 2019. The sample is limited to individual taxpayers with a valid social security number. We collect information on their Master File Collection Status Code (to indicate their progression within the enforcement chain) and ACS and CFf revenue (if they progressed to these steps).

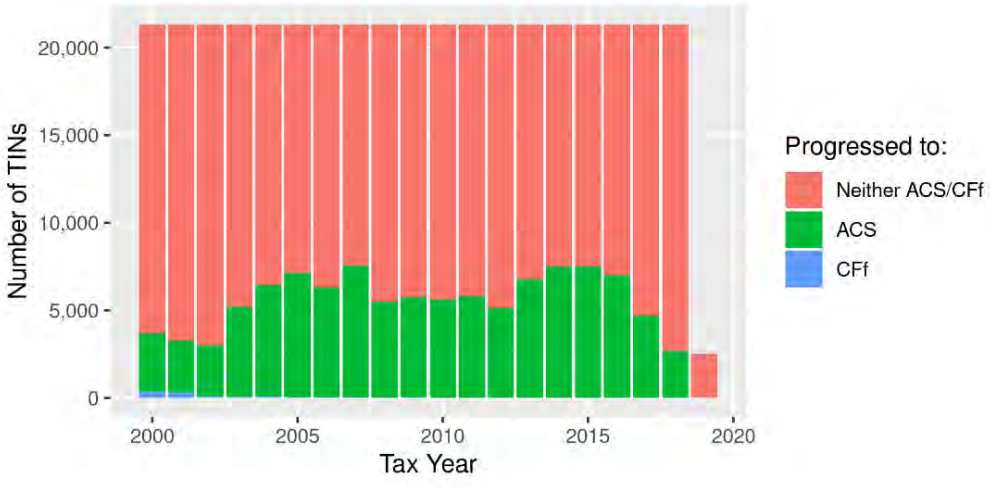
We also collect data on their tax characteristics to use as influencer variables. In particular, we select characteristics that are likely to affect progression into ACS/CFf, as well as revenue once we have arrived at these steps. Taxpayers progress into ACS or CFf only if they enter balance-due status (i.e., they do not pay the assessed amount from the correspondence exam). Thus, we select predictors that affect tax compliance or the ability to pay: AGI, total tax, deductions and credits (medical deductions and child tax credit amount), and profits and losses (casualty and theft loss, business net profit or loss, rents and royalties loss, and farm net profit or loss). We also include an indicator for whether the taxpayer received preparer assistance.

Part One Model: ACS/CFf Progression

Table 6 in the Appendix summarizes the sample data for the Part One Model (predicting taxpayer progression from correspondence audit to ACS, CFf, or neither). Dollar-denominated variables are adjusted for inflation. The sample for the Part One Model has 407,869 observations across the 2000–2019 period. The outcome variable is *status*, defined as 2 if the taxpayer reached ACS, 3 if the taxpayer reached CFf, and 1 otherwise. Many of the influencer variables are highly skewed, with large positive or negative outliers.

Figure 13 summarizes the number of taxpayers (by taxpayer identification number (TIN)) in our sample by tax year. We create a repeated cross-sectional dataset with the same number of taxpayers every year.[24] There are 21,333 observations for each year during 2000–2018. In addition to sample size, Figure 13 indicates the number of taxpayers in each year who progressed to ACS, CFf, or neither status. Depending on the year, 13-35 percent of sampled TINs were selected for ACS during the period 2000–2018. Only a small sample of TINs were selected for CFf, peaking at 1.6 percent in 2000.

FIGURE 13. Number of TINs by Tax Year (Entire Sample)



[24] For each year, we include a different sample of taxpayers to ensure independence between years (to satisfy assumptions of our statistical models). The sample size for each year is restricted to the smallest size sampled for any given year (21,333), not including 2019. This sampling approach causes modeling results to differ slightly depending on what sample is selected. For reproducibility, we set a sampling seed in our code to ensure that the same sample is drawn each time the model is run.

Part Two Model: ACS/CFf Revenue per Case

The Part Two Model is applied to taxpayers who advanced to ACS (CFf) to predict their likelihood of payment as well as ACS (CFf) revenue. We summarize the taxpayers in our 2000–2019 sample who advanced to these steps.

Table 7 of the Appendix summarizes ACS revenue and tax characteristics of taxpayers who advanced to ACS. Some 105,356 taxpayers in our 2000–2019 sample advanced to ACS. *acs_rev* describes the total payments collected from the taxpayer while in ACS. Across our sample of correspondence-audited taxpayers *who advanced to ACS*, the average payment collected per taxpayer while in ACS was $623 (in inflation-adjusted terms). Individual ACS payments ranged from $0 to over $100,000. Note that ACS revenue has not yet been resolved for 2019 taxpayers; our model will predict revenue for this cohort of taxpayers.

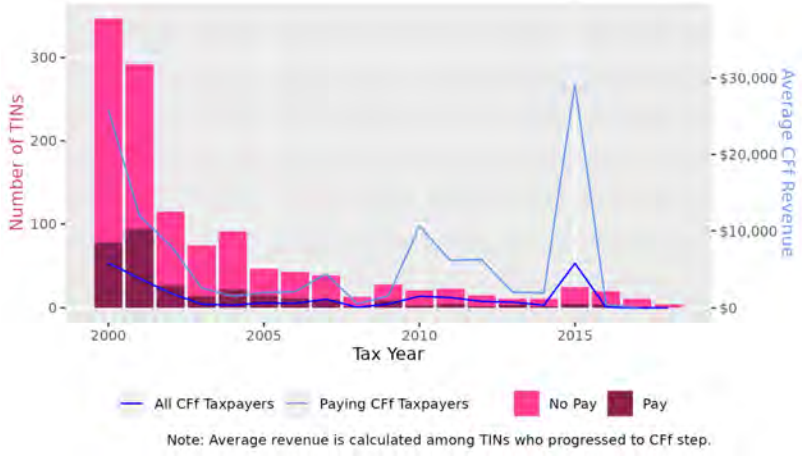
Table 8 of the Appendix summarizes CFf revenue and tax characteristics of taxpayers who advanced to CFf. Some 1,231 taxpayers in our 2000–2019 sample advanced to CFf. *cff_rev* describes the total payments collected from the taxpayer while in CFf. Across our sample of correspondence-audited taxpayers *who advanced to CFf*, the average payment collected per taxpayer while in CFf was $3,097 (in inflation-adjusted terms). Individual CFf payments ranged from $0 to over $1.3 million. The higher payments compared to those for ACS reflect the nature of CFf enforcement in focusing on high-value, complex tax cases. Average *agi* and *tax* are also higher in Table 8 than in Table 7. Since no 2019 taxpayer in our sample has yet advanced to CFf, CFf revenue has not yet been resolved for that year.

Figure 14 shows the sample size and average ACS revenue for our ACS sample, by tax year. The bars indicate the number of TINs in our sample for each year (corresponding to the primary y-axis on the left). The number of TINs is broken down into taxpayers who paid ACS revenue (pink) versus those who paid nothing while in ACS (purple). The lines indicate the average ACS revenue per case for each year (corresponding to the secondary y-axis on the right). Average ACS revenue is calculated for all taxpayers who advanced to ACS (dark blue) versus taxpayers who paid some amount while in ACS (light blue). Average ACS revenue (both lines) peaks in 2001 and gradually declines thereafter. Conversations with the sponsor indicate that this recent declining trend is likely driven by taxpayers' increased ability to pay. More taxpayers paying during the notice stream (while in balance-due status) leaves less revenue to be collected once we are at the ACS step.

Figure 15 shows the sample size and average CFf revenue for our CFf sample, by tax year. As the bars indicate, the number of CFf cases within our sample has declined over time from almost 350 cases in 2000 to under 25 cases in 2015. Average CFf revenue per case has fluctuated over time, with notable peaks in 2000 and 2015. The large swings in average CFf revenue by year reflect the sensitivity of revenue to case selection: If only a small number of cases are selected each year, the average revenue among them is highly sensitive to outliers.

FIGURE 14. ACS Sample Size and Average Revenue by Tax Year (Among TINs Progressed to ACS)

FIGURE 15. CFf Sample Size and Average Revenue by Tax Year (Among TINs Progressed to CFf)



3.4 Results

3.4.1 Part One Model: ACS/CFf Progression

The Part One Model predicts progression from correspondence audit to ACS or CFf. We estimate the model on 2000–2018 data and show which tax characteristics are most influential in ACS/CFf progression. We then apply the model to the 2019 cohort of taxpayers to forecast which ones will progress to ACS/CFf.

To estimate the Part One Model, we use the sample of correspondence-audited taxpayers for TY 2000-TY2018. We employ a discrete choice model called a multinomial logit regression, in which the outcome variable is the status of the taxpayer. We model three levels of the status variable: Each taxpayer reaches either ACS, CFf, or some other status (such as installment agreement or balance paid). The three status levels are mutually exclusive and collectively exhaustive.[25] Future work can extend the status variable to explicitly address other outcomes as well.

The multinomial logit regression estimates the odds of reaching a status relative to a base comparison group. In this case, the base comparison group is the "other" status (neither ACS nor CFf). A taxpayer's likelihood of reaching ACS or CFf (compared to the "other" status) is modeled as a function of taxpayer characteristics such as AGI and total tax. In addition to these influencers, we also include as control variables indicators for year to account for year-specific factors that affect ACS and CFf progression.

Table 2 presents the results of the multinomial logit regression. This table reports the odds ratios associated with each influencer. An odds ratio *less than one* means that an increase in the influencer variable *decreases* the odds of advancing to ACS or CFf compared to the base category, while an odds ratio *more than one* means that an increase in the influencer variable *increases* the odds of advancing to ACS or CFf compared to the base category.

Column 1 shows how the influencers affect the odds of reaching ACS, while Column 2 shows their effects on the odds of reaching CFf. Except for total medical deductions, all influencers are statistically significant (at the 10-percent level or lower) in at least one column of the table. A greater number of influencers have both statistically significant and economically meaningful effects for ACS advancement than have such effects for CFf advancement. This likely reflects the small number of CFf cases to begin with and therefore a lack of statistical power to estimate its influencers.

Higher AGI and higher total tax are associated with decreased odds of advancing to ACS relative to the base category. For example, increasing total tax by $1,000[26] decreases the odds of advancing to ACS by 1

[25] A very small number of taxpayers are processed through both ACS and CFf—we drop these taxpayers in our analysis.

[26] Recall that all influencer variables (except for preparer assistance) are measured in thousands of dollars (see Table 6).

percent.[27] AGI and total tax have negligible effects on the odds of advancing to CFf. Increases in casualty and theft loss, rents and royalties loss, and child tax credit amount are also associated with decreased odds of ACS advancement. These influencers are also associated with decreased odds of CFf advancement (although the rents/royalties loss effect is not statistically significant for CFf). The effect of the child tax credit is especially large: A $1,000 increase in child tax credit amount decreases the odds of ACS advancement by 6 percent[28] and decreases the odds of CFf advancement by 27 percent.[29]

On the other hand, increases in net capital gain/loss and net farm profit/loss are associated with *increased* odds of ACS advancement compared to the base category. Receiving preparer assistance increases the odds of ACS advancement by 16 percent.[30] Some of these influencers have the opposite effect on CFf advancement. Both net business profit/loss and receiving preparer assistance are associated with *decreased* odds of CFf advancement. In fact, receiving preparer assistance decreases the odds of CFf advancement by 77 percent.[31]

TABLE 2. Odds Ratios from Part One Model (TY2000–TY2018)

Influencers	Explained Variable: Status	
	(1) ACS	(2) CFf
agi	0.999***	1.000***
	(0.0001)	(0.00001)
tax	0.990***	1.000
	(0.0004)	(0.00003)
meddeduc	1.000	0.996
	(0.0003)	(0.011)
casualtyloss	0.993***	0.985***
	(0.002)	(0.001)
busprofloss	1.000	0.999**
	(0.0002)	(0.0003)
capgainloss	1.002***	1.000*
	(0.0002)	(0.00003)
rentroyloss	0.946***	0.988
	(0.003)	(0.013)
farmprofloss	1.006***	0.999
	(0.001)	(0.002)
prepasst	1.162***	0.231***
	(0.008)	(0.008)
childcr	0.940***	0.729***
	(0.009)	(0.004)
Constant	0.188***	0.036***
	(0.006)	(0.025)
Akaike Information Criterion	462,067.800	462,067.800

Standard errors in parentheses. Subsumed in the intercept are taxpayers who did not progress to ACS or CFf, those who did not use preparer assistance, and TY2000.
*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

[27] 0.990 – 1 = -0.01 or -1 percent.
[28] 0.940 – 1 = -0.06 or -6%.
[29] 0.729 – 1 = -0.271 or -27%.
[30] 1.162 – 1 = 0.162 or 16%.
[31] 0.231 – 1 = -0.769 or -77%.

Forecasts of ACS/CFf Progression

We use the multinomial logit model estimated on 2000-2018 taxpayers to forecast ACS and CFf progression for the most recent cohort of taxpayers filing in 2019. In essence, we use historical relationships between influencers and ACS/CFf progression to predict which of today's correspondence-audited taxpayers will progress, based on their tax characteristics.

To maximize the usefulness of this forecast, the model should be run once all (or almost all) 2019 taxpayers have been selected for correspondence audit (i.e., all correspondence exams on 2019 returns have been started). At the time of this analysis, only 2,542 taxpayers from 2019 have been selected for correspondence audit. As Figure 13 shows, the 2019 count is far less than even the truncated samples selected for prior years. This suggests that a good number of taxpayers are still "in the queue" to be selected for correspondence exam. Since our Part One Model is applied to 2019 correspondence audits to forecast ACS/CFf progression, we will only be able to predict which of the *existing* correspondence audits will likely progress. *Our model can be applied again once all or almost all correspondence audits have been started for 2019 returns.*

We show forecasts for ACS and CFf progression on the small number of 2019 correspondence exams that have already been started. Figure 16 summarizes the predicted probabilities of ACS progression for the 2,542 taxpayers in the 2019 sample. Figure 17 shows the predicted probabilities of CFf progression for these taxpayers. Most predicted probabilities for ACS fall under 20 percent, indicating that the likelihood for any taxpayer to be selected for ACS is not particularly high, given their tax characteristics. Likewise, the predicted probabilities for CFf progression are well under 1 percent. This reflects the rarity of being selected for CFf.

One approach to "advance" some of these taxpayers to ACS and CFf is to use historical progression rates as a benchmark. Across our 2000-2018 sample, 26 percent of correspondence-audited taxpayers progressed to ACS and 0.3 percent of taxpayers progressed to CFf (corresponding to counts in Figure 13). We apply these progression rates to 2019 taxpayers, selecting the top 26 percent of predicted ACS probabilities to advance to ACS and the top 0.3 percent of predicted CFf probabilities to advance to CFf. The cutoff for advancement is shown in the vertical lines in Figure 16 and Figure 17. In Figure 16, the top 26 percent of predicted ACS probabilities corresponds to a probability of 0.17 or higher. In Figure 17, the top 0.3 percent of predicted CFf probabilities corresponds to a probability of 0.03 or higher. Taxpayers with predicted probabilities above these thresholds are forecasted to advance to ACS or CFf.

Our forecast of ACS and CFf progression among 2019 taxpayers can be used in the Part Two Model to forecast ACS and CFf revenue.

FIGURE 16. Predicted Probability of ACS Progression for TY2019 Sample of Taxpayers

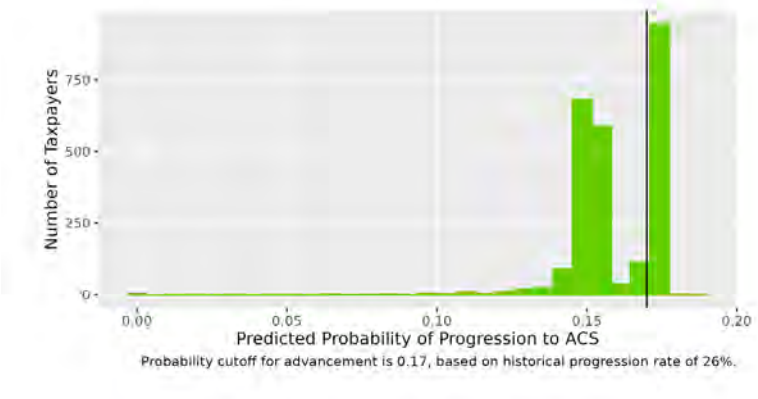
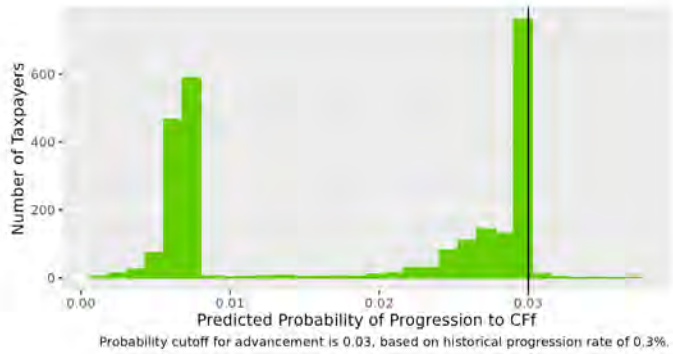


FIGURE 17. Predicted Probability of CFf Progression for TY2019 Sample of Taxpayers



3.4.2 Part Two Model: ACS/CFf Revenue per Case

The Part Two Model predicts the probability of payment as well as ACS and CFf revenue per case, for taxpayers who progressed to these enforcement steps. We estimate the model on TY2000–TY2018 data and show which tax characteristics are most influential in ACS and CFf revenue. We then apply the model to more recent years to forecast revenue per case.

To estimate the Part Two Model, we use the sample of taxpayers for TY2000–TY2018 who reached ACS (for the ACS model) or who reached CFf (for the CFf model). A unique feature of ACS and CFf revenue is an abundance of zeros—these are taxpayers who progress through these enforcement steps but end up paying none of the amount assessed. On the other hand, there are several outliers of extremely high revenue collected. Modeling ACS and CFf revenue while ignoring these two statistical features (abundance of zeros and large outliers) will lead to model estimates that lack precision and accuracy (i.e., biased and inconsistent estimates). Further, from a conceptual point of view, the drivers of a "pay or no pay" outcome can be very different from the drivers of the actual amount someone pays.

We use a zero-inflated model[33] that addresses both processes: why someone pays (or does not pay) due to the enforcement action, and the amount they pay. This model segregates the outcome's probability distribution into two processes: one that generates the zero versus nonzero revenue values and another that predicts positive revenue values. Alternatively stated, one can think of revenue collection as an event. If the event happens, we observe a positive revenue, and if it does not happen, we observe a zero value. We model both processes as the outcome of "influencer" variables related to taxpayer characteristics such as AGI and total tax. In addition to these influencers, we also include as control variables indicators for year to account for year-specific factors that affect ACS and CFf revenue.

Modeling Pay v. No-Pay

The portion of the zero-inflated model that predicts a "pay or no pay" outcome is carried out through a logit regression. A logit regression is similar to the multinomial logit regression used for the Part One Model, but the outcome variable has only two levels. We include all taxpayers in our sample who advanced to ACS or CFf during 2000–2018. The outcome variable is an indicator for whether the taxpayer paid nothing (i.e., zero ACS revenue) or a positive amount. A set of taxpayer characteristics, as well as year effects, is used to estimate the odds of observing a positive revenue value.

Table 3 presents the results of the logit regression. This table reports the "odds ratios" associated with each influencer. An odds ratio *less than one* means that an increase in the influencer variable increases the odds of a "no pay" outcome, while an odds ratio *more than one* means that an increase in the influencer variable increases the odds of a "pay" outcome.

[32] Zero-inflated models are technically referred to as "two-part models." We use the term "zero-inflated model" to avoid confusion with the Part One vs. Part Two models outlined in this report with a Gamma. After we tried several link functions, an identity link with a Gaussian distribution fitted ACS and CFf revenue data the best.

Column 1 shows how the influencers affect the probability of a "pay" outcome for ACS, while Column 2 shows how the influencers affect the probability of a "pay" outcome for CFf. Note that the regressions for the two columns are estimated separately, each using the taxpayers who advanced to that particular enforcement step. Most of the influencers are statistically significant at least at the 10-percent level.

In Column 1, higher AGI, total medical deductions, and child tax credit amount are associated with a higher probability of paying ACS revenue. For example, a $1,000 increase in child tax credit amount raises the odds of an ACS pay outcome by 66.1 percent. Additionally, taxpayers who received preparer assistance have a higher probability of paying ACS revenue. In contrast, higher total tax liability, total casualty and theft loss, Schedule C net profit or loss, and rents and royalties loss are associated with a greater "no pay" probability for ACS. For example, a $1,000 increase in Schedule C net profit/loss decreases the odds of an ACS "pay" outcome by 3.7 percent.[33]

Column 2 shows how the influencers affect a "pay" outcome for CFf. There are far fewer observations for CFf than for ACS, which results in less statistical precision in the estimated effects. This result is to be expected given that CFf works on fewer cases and is highly sensitive to case selection. The only statistically significant influencer is preparer assistance: Taxpayers who received preparer assistance have a 31.8-percent lower probability of making a payment in CFf.[34]

TABLE 3. Odds Ratios from Part Two Model: Pay vs. No-Pay (2000–2018)

Influencers	Explained Variable: Pay	
	(1) ACS	(2) CFf
agi	1.013***	1.000
	(0.0005)	(0.0001)
tax	0.956***	1.000
	(0.002)	(0.0001)
meddeduc	1.011***	0.990
	(0.002)	(0.025)
casualtyloss	0.990***	0.725
	(0.004)	(1,830.39)
busprofloss	0.963***	0.999
	(0.001)	(0.002)
capgainloss	1.000	0.999
	(0.001)	(0.001)
rentroyloss	0.989***	1.011
	(0.004)	(0.012)
farmprofloss	0.992	2.96x1089
	(0.005)	(6,247.56)
prepasst	1.075***	0.682**
	(0.014)	(0.183)
childcr	1.661***	1.137
	(0.022)	(0.208)
Constant	0.820***	0.302***
	(0.036)	(0.129)
Observations	105,332	1,231
Log Likelihood	-67,334.710	-674.636
Akaike Information Criterion	134,727.400	1,407.27

Standard errors in parentheses. Subsumed in the intercept are taxpayers who reached the enforcement step but did not make any payments (i.e., ACS or CFf revenue = 0), those who did not use a preparer's assistance, and TY2000.

*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

[33] 0.963 – 1 = -0.037 or -3.7 percent.

[34] 0.682 – 1 = -0.318 or -31.8 percent.

Modeling Amount Paid

The previous section estimated the drivers of a "pay or no pay" outcome. In this section, we model the drivers of the amount paid itself. The portion of the zero-inflated model that predicts the amount paid (given that someone is a "pay" outcome) is carried out through a generalized linear model.[35] We include TY2000-TY2018 taxpayers in our sample who 1) advanced to ACS (CFf), *and* 2) paid some amount while in ACS (CFf). Taxpayers who advance to the step but pay nothing are not included in this section. The outcome variable is ACS (CFf) amount paid, and the influencer variables are the same set used in the "pay or no pay" regression from the previous section.

Table 4 presents the results of the generalized linear model. The coefficients should be interpreted as the change in revenue from a one-unit increase in the influencer variable (note that a one-unit increase corresponds to a thousand-dollar increase in many of the influencers). Column 1 shows how the influencers affect ACS revenue, while Column 2 shows how the influencers impact CFf revenue. Note that the regressions for the two columns are estimated separately, each using the taxpayers who advanced to that particular enforcement step. Most of the influencers are statistically significant at least at the 10-percent level.

In Column 1, higher AGI, casualty and theft loss, rents and royalties loss, and child tax credit amount are associated with higher ACS amount paid. For example, a $1,000 increase in AGI is associated with a $28 increase in ACS amount paid, and a $1,000 increase in child tax credit amount is associated with a $154 increase in ACS amount paid. Higher total tax, Schedule C net profit or loss, net capital gain or loss, and net farm profit or loss are associated with lower ACS amount paid. In addition, receiving preparer assistance is associated with lower ACS amount paid. In fact, using a preparer decreases the average ACS amount paid by $161. It is interesting to note that casualty and theft loss and rents and royalties loss negatively impact the probability of an ACS pay outcome but are positively associated with the ACS amount paid. In other words, taxpayers with higher casualty/theft loss and rents/royalties loss have a lower probability of making an ACS payment, but if they do pay, the amount paid tends to be higher.

Column 2 shows how influencers affect the amount paid in CFf. Compared to that in the "pay vs. no-pay" part of the model, the sample size is even smaller here. Only 310 taxpayers during 2000-2018 reached CFf *and* paid some amount while in CFf. Almost all influencers are statistically significant at least at the 10-percent level. Higher AGI and net capital gain or loss are associated with increases in CFf amount paid. On the other hand, higher total tax, Schedule C net profit or loss, rents and royalties loss, and receiving preparer assistance are associated with lower CFf amount paid. It is interesting to note that AGI has no impact on the probability of making a payment but is a relevant predictor of what amount is paid. This illustrates how the drivers of a pay/no-pay outcome could be different from what predicts the actual amount paid. Preparer assistance decreases both the probability of a CFf pay outcome and the amount paid in CFf.

[35] We use a generalized linear model with a Gaussian distribution and identity link function. A generalized linear model allows the response variable's error to follow any distribution in the exponential family. A Gamma distribution is often well suited for right-skewed data. However, in our case, the parameters did not converge with a Gamma. After we tried several link functions, an identity link with a Gaussian distribution fitted ACS and CFf revenue data the best.

TABLE 4. Coefficients from Part Two Model: Amount Paid (2000–2018)

Influencers	Explained Variable: Amount Paid	
	(1) ACS	(2) CFf
agi	28.103***	281.898***
	(0.363)	(8.448)
tax	-87.674***	-530.826***
	(1.159)	(33.341)
meddeduc	0.114	-773.619
	(0.135)	(781.244)
casualtyloss	14.316***	--
	(4.525)	--
busprofloss	-11.393***	-77.263**
	(0.921)	(35.013)
capgainloss	-22.756***	137.273***
	(1.277)	(21.681)
rentroyloss	59.257***	-613.578**
	(3.877)	(307.469)
farmprofloss	-9.786	393.333
	(7.959)	(1,773.22)
prepasst	-161.411***	-8,864.463**
	(15.136)	(3,707.14)
childcr	153.596***	-5,808.50
	(16.918)	(4,488.53)
Constant	1,338.690***	6,017.389**
	(40.374)	(2,647.44)
Observations	50,825	310
Log Likelihood	-447,197.200	-3,539.46
Akaike Information Criterion	894,452.400	7,132.92

Standard errors in parentheses. Subsumed in the intercept are taxpayers who reached ACS and did not use a preparer's assistance as well as TY2000.
*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

Forecasts of ACS and CFf Revenue per Case

We use the zero-inflated model estimated on TY2000-TY2018 taxpayers to forecast ACS and CFf revenue per case. In essence, we use historical relationships between influencers and ACS/CFf revenue to predict revenue for recent ACS/CFf taxpayers, based on their tax characteristics.

Since not all correspondence audits have been issued for TY2019 at the time of this analysis, we have a small sample size for this year that may lead to unstable results. Our approach can be applied to TY2019 taxpayers once all or almost all correspondence audits have been issued.[36]

In the meantime, we apply the zero-inflated model to 2018 taxpayers for demonstration purposes. The TY2018 sample includes 365 taxpayers who 1) progressed to ACS *and* 2) paid some amount while in ACS. Figure 18 shows the distribution of predicted ACS revenue across the 365 taxpayers in our prediction sample. The vertical line shows the average revenue per case: $569.58. This estimate is close to the historical average

[36] If applied to 2019 correspondence-audited taxpayers, the Part One Model would first predict who advances to ACS or CFf. Given the group that advances, the Part Two Model then predicts 1) the probability of each taxpayer making a payment, and then 2) if predicted to pay, how much each taxpayer would pay in ACS or CFf.variation in the draws within each simulation that affects the TERC calculation.

from TY2013–TY2016 of $603, the revenue parameter currently being used in PAM. In the following section, we replace the historical PAM parameter of $603 with our new estimate of $569.58 to update workforce allocation and assess the impact on TERC.

For the CFf prediction sample, we use taxpayers in our sample from TY 2013, TY2014, TY2016, TY2017, and TY2018 for demonstration purposes.[37] We incorporate more years for CFf in order to generate a larger sample size. Figure 19 shows the distribution of predicted CFf revenue among the 11 taxpayers who 1) progressed to CFf *and* 2) paid some amount while in CFf. The average revenue per case is $1,265.49. This is close to the historical average from TY2013-TY2016 of $1,394. In the next section, we replace the historical PAM parameter of $1,394 with our new estimate of $1,265.

FIGURE 18. Predicted ACS Revenue (TY2018 Sample)

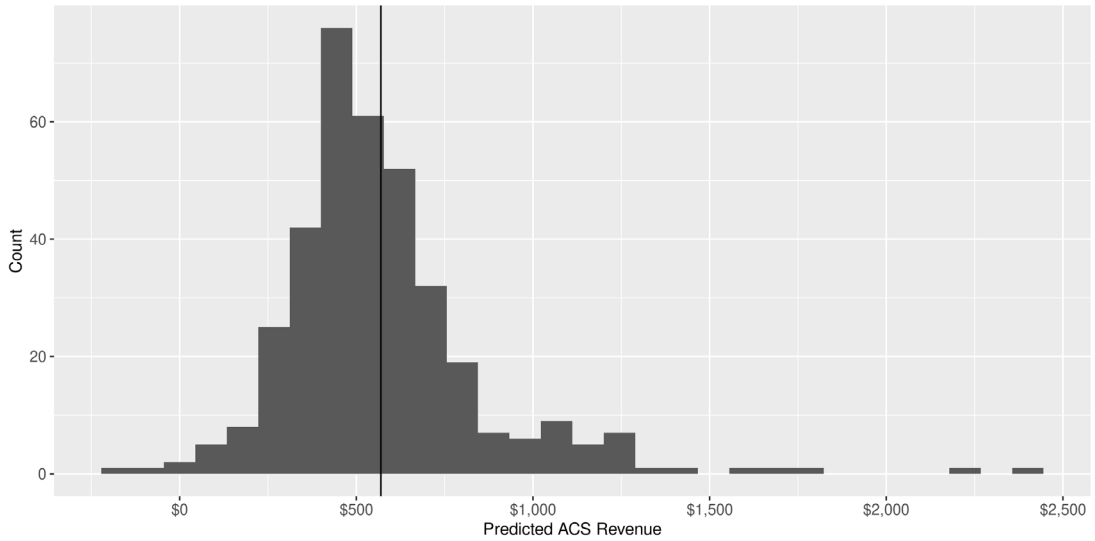
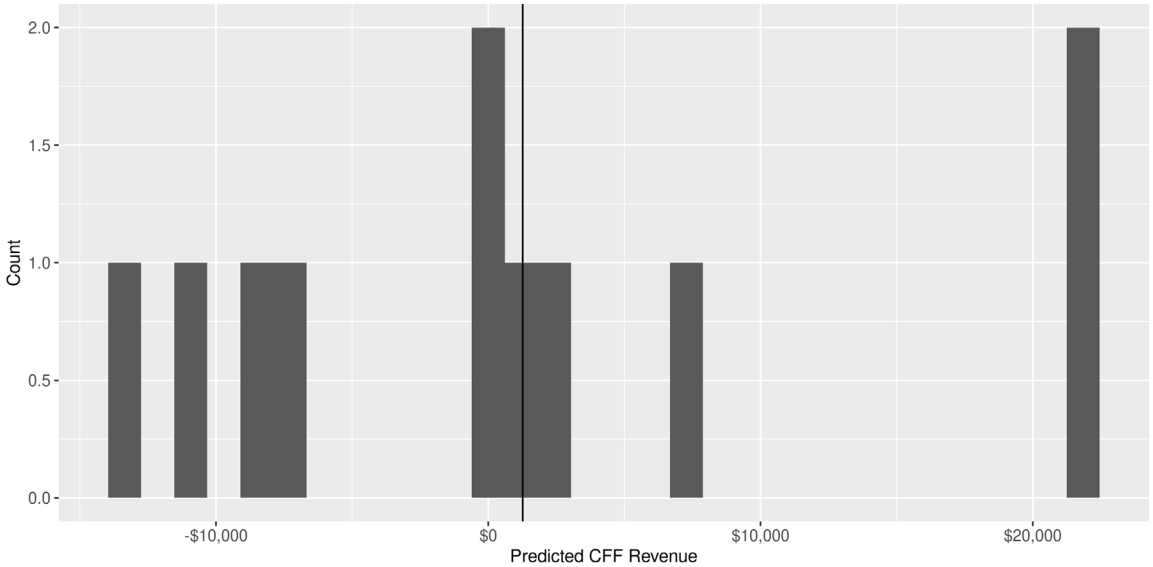


FIGURE 19. Predicted CFf Revenue (TY2013, TY2014, TY2016, TY2017, and TY2018 Sample)



[37] We exclude TY2015, which is an outlier year in terms of CFf revenue.

3.4.3 Using Revenue Forecasts in PAM

The micro model forecasts revenue at the enforcement-step level in order to improve workforce allocation with the goal of maximizing TERC. When historical revenue parameters no longer reflect future conditions, workforce allocation decisions based on these parameters are suboptimal. This leads to lower overall TERC. We use the forecasted revenue parameters from the previous section to allocate workforce in PAM. We also model the bottom-line effect on TERC using the approach summarized in Figure 12.

Decreasing the ACS revenue parameter from $603 to $570 results in four FTEs being removed from the ACS step. Decreasing the CFf parameter from $1,394 to $1,265 does not affect the CFf FTE allocation. In either case, changing the ACS or CFf parameter does not affect FTEs allocated to other steps. Using the new ACS revenue parameter leads to a very small increase in TERC of 0.001 percent.[38] This increase is not substantially different from zero—indicating that changing the ACS parameter from $603 to $570 is not enough of a change to affect TERC. Changing the CFf parameter from $1,394 to $1,265 also does not affect TERC.

Since these changes are not very large, we explore sensitivity analysis by changing all revenue parameters within PAM by 50 percent (a 50-percent increase and a 50-percent decrease). We find that the biggest increases in TERC come from a 50-percent increase in the average revenue per case for specialty exam employment tax (SpecEx_Employment). Increasing this parameter from $7,740 to $11,610 increases TERC by 0.094 percent or $26 million relative to the baseline approach. This parameter change is associated with an additional 67 FTEs assigned to this enforcement step. Decreasing this parameter from $7,740 to $3,870 increases TERC by 0.18 percent or $49 million. This parameter change is associated with 163 fewer FTEs being assigned to specialty employment exam.

Whether increasing or decreasing the revenue parameter by 50 percent, other enforcement steps with notable effects on TERC are the medium automated underreporter (AUR) commodity (aurMed) at the AUR, the medium AUR commodity (aurMed) at collection balance due (Collection_Bal_Due), and the specialty exam gift commodity (exam_gift) at specialty exam gift (SpecEx_Gift). Overall, the enforcement steps with the largest effect on revenue tend to be ones with large revenue per case (such as specialty exam employment tax) or ones with very large caseloads (such as AUR). Future work can apply the micro model to these types of steps.

3.5 Discussion

The micro model presented in this report achieves the objectives of 1) understanding the drivers of enforcement revenue at the enforcement-step level; and 2) operationalizing this knowledge for workforce allocation in order to maximize TERC. We demonstrate the modeling framework on two enforcement steps: ACS and CFf cases proceeding from correspondence audits. Our analysis sheds light on the largest drivers of revenue for these two enforcement steps:

- What increases the probability of proceeding to the enforcement step (i.e., progression):
  - For ACS, receiving preparer assistance increases the odds of progression, while higher child tax credit amount and rents and royalties loss decrease the odds of progression.
  - For CFf, receiving preparer assistance, higher child tax credit amount, and casualty and theft loss all decrease the odds of progression.
- What increases the probability of a "pay" outcome in the given step:
  - For ACS, higher child tax credit amount and receiving preparer assistance increase the probability of paying. On the other hand, higher total tax and Schedule C net profit/loss decrease the probability of paying.
  - For CFf, receiving preparer assistance decreases the probability of paying.

[38] In some runs of PAM, a small change in the revenue parameter actually leads to a decrease in TERC. While this seems counterintuitive from an optimization standpoint, the drop in TERC does not result from a different workforce allocation (due to the swapping of revenue parameters) but rather from random variation within the PAM simulation framework. PAM iterates across a range of parameters, each represented with a distribution, and there is random variation in the draws within each simulation that affects the TERC calculation.

- What increases amount paid in the given step:
  - For ACS, higher child tax credit and rents and royalties loss increase the amount paid, while receiving preparer assistance and higher total tax decrease the amount paid.
  - For CFf, higher AGI increases the amount paid, while receiving preparer assistance, higher rents and royalties loss, and higher total tax decrease the amount paid.

This work also provides a basis for greater understanding of taxpayer compliance and is generalizable beyond the immediate scope of ACS and CFf revenue. The key takeaways of this research include the following:

We provide a modeling framework for understanding the "micro" drivers of TERC.

While the macro model presents a framework for understanding macroeconomic and other aggregate drivers of TERC, the micro model approach shows how individual tax characteristics can help explain enforcement-step revenue. This framework sheds light on why enforcement-step revenue fluctuates over time and can serve as a basis for adding other tax characteristics in future work.

We show how to forecast revenue per case for new taxpayer cohorts.

In addition to *explaining* fluctuations in enforcement step revenue, the micro model can also be used to *forecast* revenue. The model estimates historical relationships between tax characteristics and revenue per case, and these relationships can be extrapolated to new cohorts of correspondence-audited taxpayers. This provides a forward-looking rather than backward-looking view of revenue.

We provide a framework that is generalizable to other enforcement steps.

Although we demonstrate the framework on ACS and CFf revenue, the model (and accompanying codebase) can be applied to other enforcement steps, as well. The modeling approach is particularly fruitful for steps with large caseloads (i.e., sample sizes) and less precise for steps with small caseloads and high sensitivity to workload selection.

We show that operational and revenue implications can be large.

Updating revenue parameters for some enforcement steps can have a particularly large effect on workforce allocation (and thus TERC). Our PAM sensitivity analysis showed which revenue parameters, if changed, would have the largest impact on TERC. These steps tend to be the ones with the largest revenue per case or the largest caseloads.

Future Work

This task builds a prototype for understanding and forecasting the microlevel drivers of enforcement revenue. Potential future work can include the following:

- **Inclusion of additional influencers**: We explored a limited set of taxpayer characteristics thought to influence ACS and CFf progression and revenue. Future work can explore other characteristics known or theorized to affect enforcement revenue.
- **Application of model to other enforcement steps**: ACS and CFf revenue are not necessarily the most impactful parameters in PAM—other enforcement steps may have a greater impact on workforce allocation and TERC. The micro model framework can be used to explore fluctuations within these other revenue parameters.
- **Alternate modeling specifications**: Future work can explore alternate specifications to improve model fit, such as transforming influencer variables, conducting out-of-sample tests to select the most important influencers, and subsample analysis of different time periods. We could also focus on estimating causal relationships between influencers and outcomes (as opposed to primarily aiming for predictive accuracy). To understand predictive accuracy, future work should include model validation and calibration on holdout taxpayer data.

References

Congressional Budget Office (2020). "Trends in the Internal Revenue Service's Funding and Enforcement," Washington, DC.

Internal Revenue Service (IRS) (2019a). "IRS Enforcement and Service Results—Fiscal Year 2018." Online. Available at https://www.irs.gov/pub/irs-news/fy-2018-enforcement-and-service-results-final.pdf. Accessed 18 December 2020.

IRS (2019b). "Tax Gap Estimates for Tax Years 2011–2013," Washington, DC.

IRS (2020). "SOI Tax Stats—Personnel Summary, by Employment Status, Budget Activity, and Selected Type of Personnel—Databook Table 32." Online. Available at https://www.irs.gov/statistics/soi-tax-stats-personnel-summary-by-employment-status-budget-activity-and-selected-type-of-personnel-databook-table-32. Accessed 18 December 2020.

Macheret, C., S. Michel, and S. Rosen (2020). "Data and Analytical Methodologies Useful for Simulation-Based Analysis of TERC," The MITRE Corporation, McLean, VA.

The Conference Board (2020). "U.S. Leading Indicators." Online. Available at https://conference-board.org/data/bcicountry.cfm?cid=1. Accessed 18 December 2020.

Treasury Inspector General for Tax Administration (TIGTA) (2017). "A More Focused Strategy Is Needed to Effectively Address Egregious Employment Tax Crimes," Washington, DC.

TIGTA (2019). "Trends in Compliance Activities Through Fiscal Year 2018," Washington, DC.

Appendix

1. Supplementary Tables

TABLE 5. Summary Statistics for Macro Model (Calendar Years 2003–2020)

Category	Description	Variable Name	Obs	Mean	Std. Dev	Min	Max
Outcome	Current month revenue, Appeals ($ Million)	monthrev.Appeals	216	99.5	89.9	-54.5	581.6
Outcome	Current month revenue, Collection ($ Million)	monthrev.Collection	216	1,201.9	256.7	595.0	1,796.1
Outcome	Current month revenue, Exam ($ Million)	monthrev.Exam	216	371.7	247.4	34.8	2,033.6
Outcome	Current month revenue, AUR ($ Million)	monthrev.AUR	216	151.7	81.5	19.0	414.5
Econ	Consumer price index, seasonally adjusted (1982-1984=100)	CPI	216	223.2	22.3	181.2	260.2
Econ	NASDAQ Composite index close level (1,000s)	nasdaq	216	3.9	2.3	1.2	11.2
Econ	Personal bankruptcies (1,000s)	totalpbank	213	92.5	36.7	34.0	350.9
Econ	Business bankruptcies (1,000s)	bbank	213	2.8	1.0	1.3	6.0
Econ	Unemployment rate (%)	unemp	216	6.2	2.1	3.5	14.7
Econ	New single-family houses sold (1,000s)	housessold	216	55.3	26.5	20.0	127.0
Econ	Consumer confidence index (1966Q1=100)	cci	216	84.0	11.8	55.3	103.8
Econ	Gross domestic product ($ Trillion)	gdp	216	7.2	0.6	6.1	8.6
Econ	New housing authorized by building permits (Million)	buildpermits	216	1.3	0.5	0.5	2.3
Econ	Rental expenditures ($ Billion)	rent	207	15.1	3.1	11.2	20.5
FTE	FTE, business systems modernization (1,000s)	FTEmonth.business.systems.modernization	156	0.4	0.1	0.3	0.7
FTE	FTE, exams and collections (1,000s)	FTEmonth.examinations.and.collections	156	38.1	5.1	29.7	46.1
FTE	FTE, filing and account services (1,000s)	FTEmonth.filing.and.account.services	204	25.9	2.5	21.7	32.3
FTE	FTE, investigations (1,000s)	FTEmonth.investigations	156	3.7	0.6	2.8	4.5
FTE	FTE, prefiling taxpayer assistance and education (1,000s)	FTEmonth.prefiling.taxpayer.assistance.and.education	204	5.0	1.1	1.8	6.4
FTE	FTE, regulatory (1,000s)	FTEmonth.regulatory	156	1.1	0.1	1.0	1.3
FTE	FTE, shared services and support (1,000s)	FTEmonth.shared.services.and.support	204	4.8	0.9	1.7	6.1
Attr	Percent of US population age 65 and above (%)	perc65	207	13.6	1.3	12.3	16.2
Attr	Weekly median earnings ($)	realwage	213	341.4	10.3	325*	390*
Attr	Share of net worth held by top 1% (%)	top1share	213	29.5	1.5	25.1	31.6

NOTE: Figures marked with * have been rounded for taxpayer privacy. Dollar-denominated variables are adjusted for inflation and expressed in terms of 1982–1984 dollars.

TABLE 6. Summary Statistics for Part One Forecasting Model, Micro Model (TY2000–TY2019)

Category	Description	Variable Name	Obs	Mean	Std. Dev	Min	Max
Outcome	Indicator for progression to ACS (1), CFf (2), or neither (0)	*status*	407,869	1.26	0.45	1	3
Influencer	AGI ($1,000s)	*agi*	407,869	$66.85	$1,777.18	**	**
Influencer	Total tax liability ($1,000s)	*tax*	407,869	$13.16	$1,042.26	**	**
Influencer	Total medical deductions ($1,000s)	*meddeduc*	407,869	$0.55	$19.30	**	**
Influencer	Total casualty and theft loss ($1,000s)	*casualtyloss*	407,869	$0.05	$2.43	**	**
Influencer	Schedule C net profit or loss ($1,000s)	*busprofloss*	407,869	$10.03	$5,474.90	**	**
Influencer	Net short-term capital gain or loss ($1,000s)	*capgainloss*	407,869	-$1.16	$188.97	**	**
Influencer	Rents and royalties loss ($1,000s)	*rentroyloss*	407,869	$0.31	$6.01	**	**
Influencer	Net farm profit or loss ($1,000s)	*farmprofloss*	407,869	-$0.09	$8.92	**	**
Influencer	Indicator for preparer assistance on return[40]	*prepasst*	407,869	0.59	0.49	0	1
Influencer	Child tax credit amount[41] ($1,000s)	*childcr*	407,869	$0.16	$0.62	**	**

NOTE: Figures marked with ** have been redacted for taxpayer privacy. We predict whether a taxpayer advances to the ACS or Field Collection enforcement steps (outcome variables), given their tax characteristics (influencer variables). Dollar-denominated variables are adjusted for inflation and are expressed in terms of year 2000 dollars.

TABLE 7. Summary Statistics for Part Two Forecasting Model: ACS Step, Micro Model (TY2000–TY2019)

Category	Description	Variable Name	Obs	Mean	Std. Dev	Min	Max
Outcome	Total payments received while taxpayer is in ACS enforcement step, given taxpayer reaches that step ($)	*acs_rev*	105,356	$623.10	$1,376.14	**	**
Influencer	AGI ($1,000s)	*agi*	105,356	$28.26	$380.19	**	**
Influencer	Total tax liability ($1,000s)	*tax*	105,356	$2.77	$122.24	**	**
Influencer	Total medical deductions ($1,000s)	*meddeduc*	105,356	$0.55	$36.87	**	**
Influencer	Total casualty and theft loss ($1,000s)	*casualtyloss*	105,356	$0.04	$1.76	**	**
Influencer	Schedule C net profit or loss ($1,000s)	*busprofloss*	105,356	$1.01	$22.99	**	**
Influencer	Net short-term capital gain or loss ($1,000s)	*capgainloss*	105,356	-$0.10	$10.62	**	**
Influencer	Rents and royalties loss ($1,000s)	*rentroyloss*	105,356	$0.10	$2.12	**	**
Influencer	Net farm profit or loss ($1,000s)	*farmprofloss*	105,356	-$0.01	$1.69	**	**
Influencer	Indicator for preparer assistance on return[42]	*prepasst*	105,356	0.62	0.48	0	1
Influencer	Child tax credit amount[43] ($1,000s)	*childcr*	105,356	$0.15	$0.38	**	**

NOTE: Figures marked with ** have been redacted for taxpayer privacy. This dataset only includes taxpayers who advanced to the ACS step. We predict a taxpayer's total payments while in ACS (outcome variable), given their tax characteristics (influencer variables). Dollar-denominated variables are adjusted for inflation and are expressed in terms of year 2000 dollars.

[40] Equals zero if self-prepared. Equals one if return is preparer-assisted (such as by IRS, IRS Volunteer Income Tax Assistance, H&R Block or other paid preparer, tax counseling for the elderly, etc.).

[41] "Child tax credit amount" was replaced by "Child and other dependent credit amount" in 2019.

[42] Equals zero if self-prepared. Equals one if return is preparer-assisted (such as by IRS, IRS Volunteer Income Tax Assistance, H&R Block or other paid preparer, tax counseling for the elderly, etc.).

[43] "Child tax credit amount" was replaced by "Child and other dependent credit amount" in 2019.

TABLE 8. Summary Statistics for Part Two Forecasting Model: CFf Step, Micro Model (TY2000–TY2019)

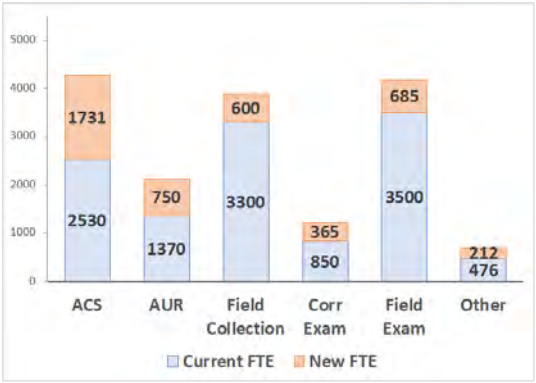
Category	Description	Variable Name	Obs	Mean	Std. Dev	Min	Max
Outcome	Total payments received while taxpayer is in Field Collection enforcement step, given taxpayer reaches that step ($)	*cff_rev*	1,231	$3,096.95	$38,747.16	**	**
Influencer	(AGI) ($1,000s)	*agi*	1,231	$1,206.96	$24,804.99	**	**
Influencer	Total tax liability ($1,000s)	*tax*	1,231	$464.46	$15,237.89	**	**
Influencer	Total medical deductions ($1,000s)	*meddeduc*	1,231	$0.40	$5.53	**	**
Influencer	Total casualty and theft loss ($1,000s)	*casualtyloss*	1,231	$0.06	$2.04	**	**
Influencer	Schedule C net profit or loss ($1,000s)	*busprofloss*	1,231	$8.46	$65.20	**	**
Influencer	Net short-term capital gain or loss ($1,000s)	*capgainloss*	1,231	-$36.06	$1,203.94	**	**
Influencer	Rents and royalties loss ($1,000s)	*rentroyloss*	1,231	$0.44	$5.94	**	**
Influencer	Net farm profit or loss ($1,000s)	*farmprofloss*	1,231	-$0.43	$12.33	**	**
Influencer	Indicator for preparer assistance on return[44]	prepasst	1,231	0.31	0.46	0	1
Influencer	Child tax credit amount[45] ($1,000s)	childcr	1,231	$0.14	$1.77	**	**

NOTE: Figures marked with ** have been redacted for taxpayer privacy. This dataset only includes taxpayers who advanced to the Field Collection step. We predict a taxpayer's total payments while in Field Collection (outcome variable), given their tax characteristics (influencer variables). Dollar-denominated variables are adjusted for inflation and are expressed in terms of year 2000 dollars.

2. Review of MITRE Program Assessment Model

PAM is a linear optimization program developed by MITRE to help optimize workforce allocation within SB/SE. Given a total number of new FTEs to hire, PAM allocates FTEs across the enforcement chain in order to maximize total enforcement revenue. The model acknowledges the effect of upstream steps on downstream enforcement activity and considers these interdependences in the allocation optimization. Further, it incorporates operational constraints such as level of service. Figure 20 is a hypothetical example of optimal workforce allocation under a scenario in which 4,500 new FTEs are allocated and a 40-percent level of service must be achieved. PAM's recommended allocation of new FTEs is shown in the pink sections of each bar.

FIGURE 20. Sample PAM Workforce Allocation (4,500 New FTEs and 40-Percent Level of Service)

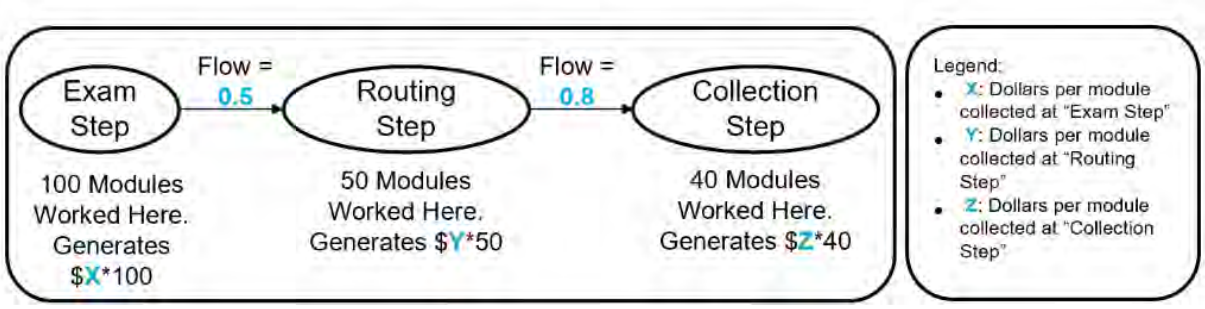


[44] Equals zero if self-prepared. Equals one if return is preparer-assisted (such as by IRS, IRS Volunteer Income Tax Assistance, H&R Block or other paid preparer, tax counseling for the elderly, etc.).

[45] "Child tax credit amount" was replaced by "Child and other dependent credit amount" in 2019.

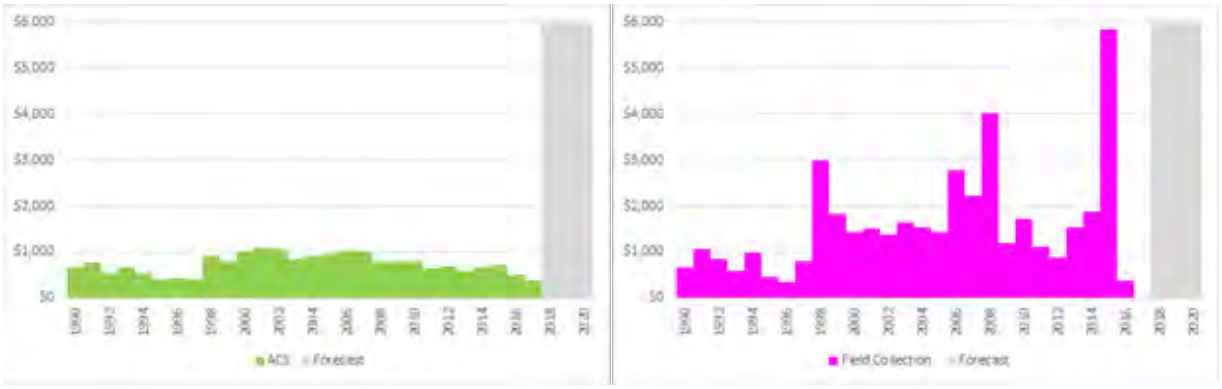
Certain input parameters affect PAM's optimization procedure. We focus on the revenue parameters, which define the expected revenue per case at each step of the enforcement chain. Generally speaking, PAM allocates more FTEs to higher-revenue enforcement steps. Figure 21 illustrates how the revenue parameters are tracked within PAM. "X" represents the revenue per case at the Exam step, "Y" the revenue per case at the Routing step, and "Z" the revenue per case at the Collection step. The flow parameters (0.5 and 0.8) dictate the percentage of cases expected to move from Exam to Routing and from Routing to Collection, respectively. For this study, we focus on revenue parameters such as X, Y, and Z.

FIGURE 21. Use of Revenue Parameters in PAM (X, Y, and Z)



PAM currently uses historical averages of revenue per case for the revenue parameters. Figure 22 shows the average revenue per case by tax year, for ACS (on left) and CFf (on right). The revenue parameters used in PAM are taken as the average of the most recent years of data (e.g., last five years) with some adjustments for outliers. A drawback of this approach is that historical trends may not reflect future conditions. Certain enforcement steps have relatively steady average revenue (such as for ACS below), while others (such as for CFf) see average revenue fluctuate significantly from year to year. For activities such as CFf, in which only a small number of cases are worked each year, average revenue in a given year is highly sensitive to case selection. For activities such as ACS, there is a recent declining trend in average revenue due to taxpayers' increased ability to pay before the account reaches the ACS step. If past trends do not reflect future conditions, then PAM's assumed inputs will produce suboptimal workforce allocations. Our forecasting model produces forecasted revenue parameters that can address this issue.

FIGURE 22. Historical Average Revenue per Case



Nonmonetary Sanctions as Tax Enforcement Tools: Evaluating California's Top 500 Program[1]

Chad Angaretis and Allen Prohofsky (California Franchise Tax Board), Brian Galle (Georgetown University Law Center), and Paul R. Organ (University of Michigan)

Many U.S. states and countries around the world use nonmonetary sanctions to encourage tax compliance, including public disclosure, license suspension, and withholding of other government-provided benefits or privileges. Little is known about the effectiveness of these programs. Using administrative tax microdata from California's "Top 500" program, we study whether notices warning of the imminent publication of a taxpayer's personal information and potential license suspension affect payment and other compliance outcomes, as well as whether these notices affect subsequent reported earnings. Exploiting variation over time in the cutoff balance for program eligibility, we find evidence of strong positive compliance responses to the program, with no evidence of an impact on subsequent reported earnings. We also develop estimates of the deadweight loss caused by publication of noncompliers and conclude that the program generates positive net social welfare. Together, these results suggest that nonmonetary sanctions can be efficient tax enforcement tools, at least among the relatively high-income population we study.

[1] We thank Jeff Hoopes, Joel Slemrod, and participants at the 2020 National Tax Association annual conference, ComplianceNet 2021, and 2021 IIPF annual congress for many helpful comments and suggestions. We are especially grateful to Nghia-Nhan Duong and You Zhan for exceptional research assistance, and to Ken Kulhavy, Alaina Andrews, Jeff Geisler, Cesar Ramos, and the rest of the CART team at the Franchise Tax Board. Any views expressed in this paper are those of the authors and not official positions of the California Franchise Tax Board. Corresponding author: Paul R. Organ, prorgan@umich.edu.

5

Appendix

Conference Program

12th Annual IRS-TPC Joint Research Conference on Tax Administration
Virtual Conference
June 16, 2022

Program

9:00–9:30 Opening

Eric Toder (Co-Director, Urban-Brookings Tax Policy Center) and

Melanie Krause (Chief Data and Analytics Officer, Research, Applied Analytics and Statistics, IRS)

9:30–11:00 Session 1: Balancing Audits: Enforcement vs. Measuring Noncompliance

Moderator: *Aaron Katch (IRS, RAAS)*

» Improving Risk Models by Supplementing Random NRP Audits with Non-Random Operational Audits Using Statistical Controls for Bias
Ishani Roy, Brett Collins, Alex Turk, Mark Payne (IRS, RAAS)

» Augmenting National Research Program Tax Change Estimates by Incorporating Operational Audit Information: A New RAAS Research Initiative
Lou Rizzo, John Riddles, Xiaoshu Zhu, Richard Valliant (Westat); Kimberly Henry (IRS, RAAS)

» Integrating Reward Maximization and Population Estimation: Sequential Decision-Making for Internal Revenue Service Audit Selection
Peter Henderson, Ben Chugg, Kristen Altenburger, Daniel E. Ho (Stanford University); Brandon Anderson (Stanford University/IRS); John Guyton, Alex Turk (IRS, RAAS); Jacob Goldin (Stanford University/U.S. Department of the Treasury)

Discussant: *Alan Plumley (IRS, RAAS)*

11:00–12:30 p.m. – Session 2: Burden vs. Opportunity

Moderator: *Alex Ruda (IRS, RAAS)*

» The Spiderweb of Pass-Through Tax Planning
Jacob Goldin, Ryan Hess, Daniel E. Ho, Rebecca Lester, Mansheej Paul (Stanford University)

» Automatic Tax Filing: Simulating a Pre-Populated Form 1040 Automatic Tax Form
Lucas Goodman, Andrew Whitten (U.S. Department of the Treasury, Office of Tax Analysis); Katherine Lim (Federal Reserve Bank of Minneapolis); Bruce Sacerdote (Dartmouth College)

» The Distribution of the Individual Income Tax Underreporting Tax Gap
Drew Johns (IRS, RAAS)

Discussant: *Steve Rosenthal (Tax Policy Center)*

12:30–12:40 p.m. – Break

12:40–1:15 p.m. – Keynote Speaker

John M. Abowd (Associate Director and Chief Scientist, Research and Methodology Directorate, U.S. Census Bureau)

1:15–2:45 p.m. – Session 3: Improving Audit Outcomes: Thinking Inside the Box

Moderator: *Evan Schulz (IRS, RAAS)*

» Graph-Based Machine Learning Methods for Case Selection and Population Segmentation
Matt Olson, Ben Howard, Devika Mahoney-Nair (MITRE); Annette Portz (IRS, RAAS)

» Automated Discovery of Tax Schemes Using Genetic Algorithms
Karen Jones, Camrynn Fausey, Eric O. Scott, Geoff Warner, Sanith Wijesinghe (MITRE); Hahnemann Ortiz (IRS, LB&I)

» Incorporating the Specific Indirect Effect of Correspondence Audits Into IRS Resource Allocation Decisions
Alan Plumley, Daniel Rodriguez (IRS, RAAS); Leigh Nichol (MITRE)

Discussant: *Mike Stavrianos (ASR Analytics)*

2:45–4:15 p.m. – Session 4: Why Do Taxpayers Comply?

Moderator: *Robert McClelland (Tax Policy Center)*

» To File or Not to File? What Matters Most?
Brian Erard (B. Erard & Associates); Tom Hertz, Pat Langetieg, Mark Payne, Alan Plumley (IRS, RAAS)

» Economic Influencers of Total Enforcement Revenue Collected and Operational Implications
Jess Grana, Lucia Lykke, Sam Schmitz (MITRE); Ron Hodge (IRS, RAAS)

» Non-Monetary Sanctions as Tax Enforcement Tools: Evaluating California's Top 500 Program
Chad Angaretis, Allen Prohofsky (California Franchise Tax Board); Brian Galle (Georgetown University Law Center); Paul R. Organ (University of Michigan)

Discussant: *Alex Yuskavage (U.S. Department of the Treasury)*

4:15–4:20 p.m. – Wrap-up

Barry Johnson *(Deputy Chief Data and Analytics Officer, Research, Applied Analytics, and Statistics (IRS))*