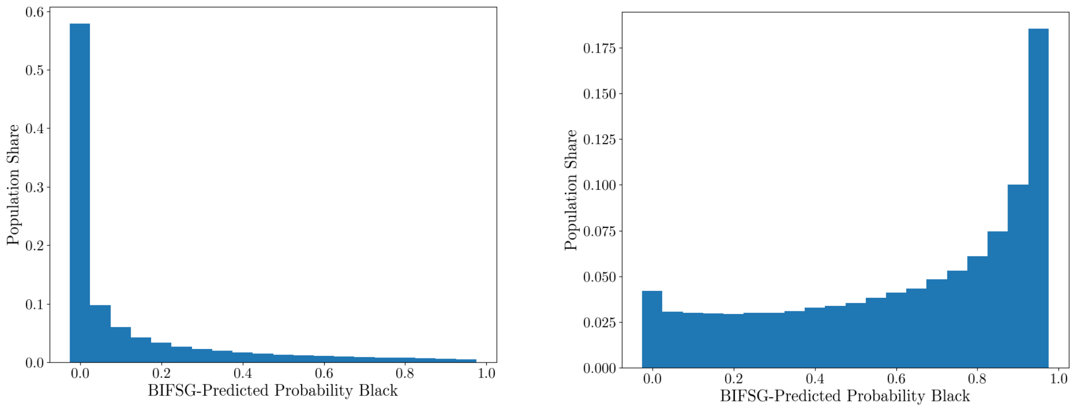
Online Appendix to Measuring and Assessing Differences in Audit Rates of Black and Non-Black Americans

Hadi Elzayn, Evelyn Smith, Thomas Hertz, Cameron Guage, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin

List of Appendices

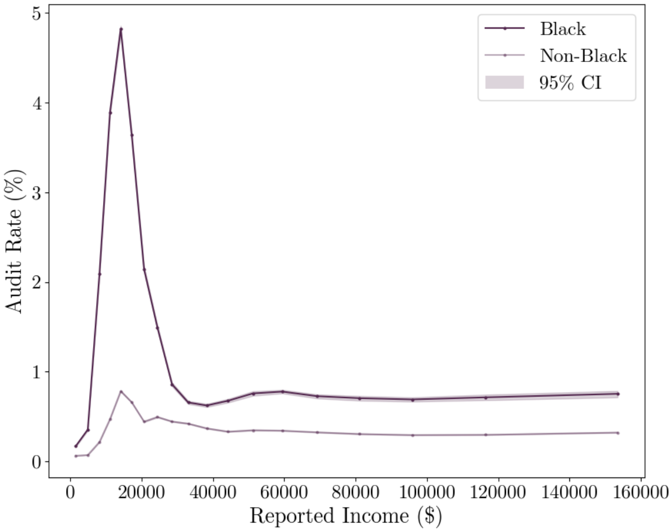
A Additional Tables and Figures

Figure A.1: Distribution of Race Imputations for Known Black and Non-Black Taxpayers



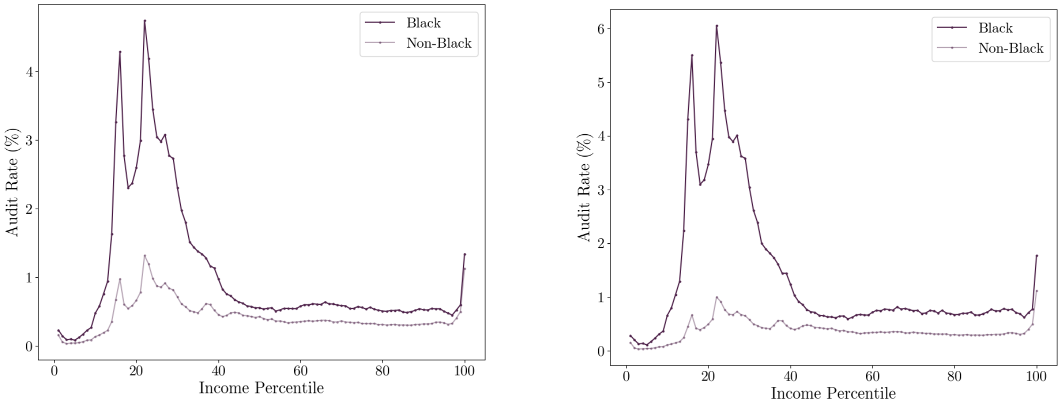
Notes: Left: The figure shows the distribution of BIFSG-predicted probabilities that a taxpayer is Black (non-Hispanic) for non-Black taxpayers in our matched North Carolina sample. Right: The figure shows the distribution of BIFSG-predicted probabilities that a taxpayer is Black (non-Hispanic) for Black (non-Hispanic) taxpayers in our matched North Carolina sample.

Figure A.2: Audit Rate Disparity by Income (Linear Estimator)



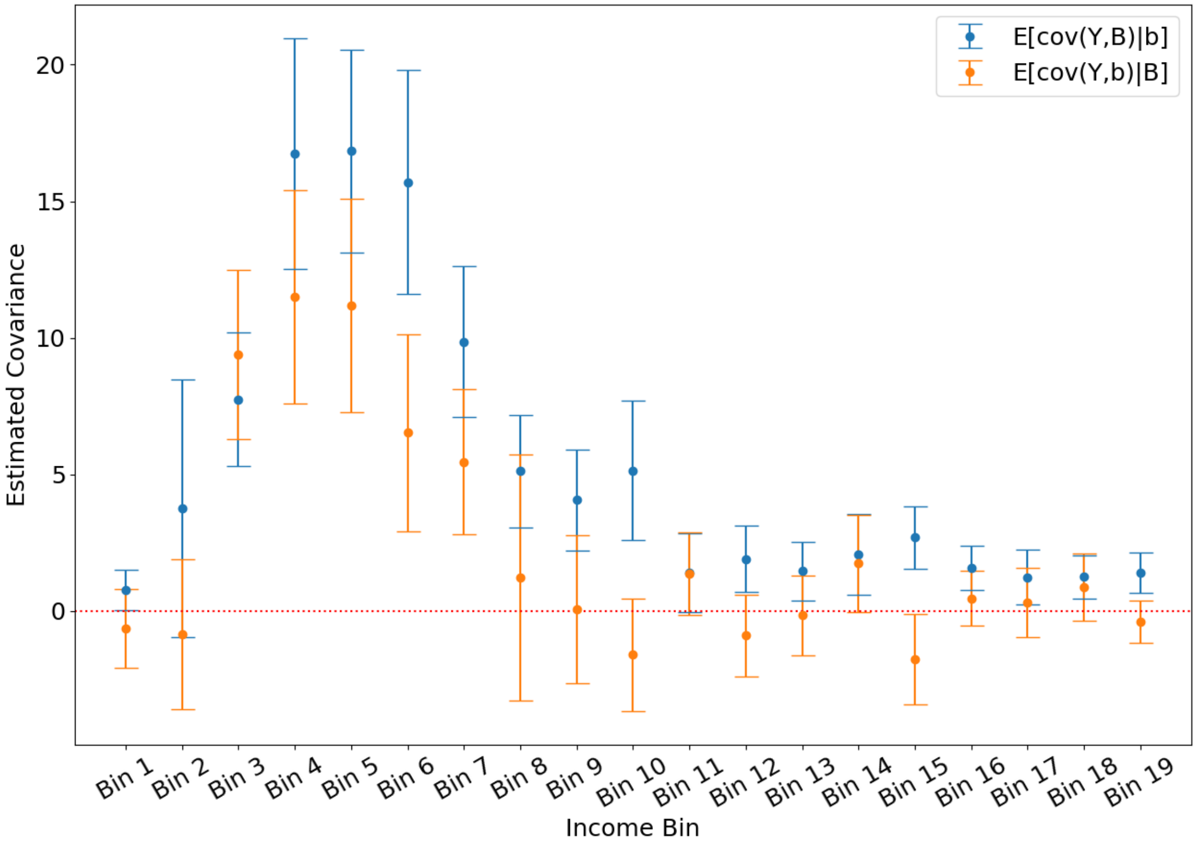
Notes: The figure shows the estimated audit rate by income among Black and non-Black taxpayers filing for tax year 2014 using the linear estimator. Income is measured according to the Adjusted Gross Income (AGI) reported on the taxpayer's return (i.e., prior to audit adjustments). The sample is limited to taxpayers reporting non-negative AGI. Taxpayers are grouped into 20 equal-sized bins. The shaded area around each line shows the 95% confidence interval based on the distribution of estimates from 100 bootstrapped samples. . To facilitate presentation, the x-axis is limited to bins with mean reported AGI under $200,000; a version of the figure with all income percentiles is presented in Appendix Figure A.3.

Figure A.3: Audit Rate Disparity by Income (All Income Percentiles)



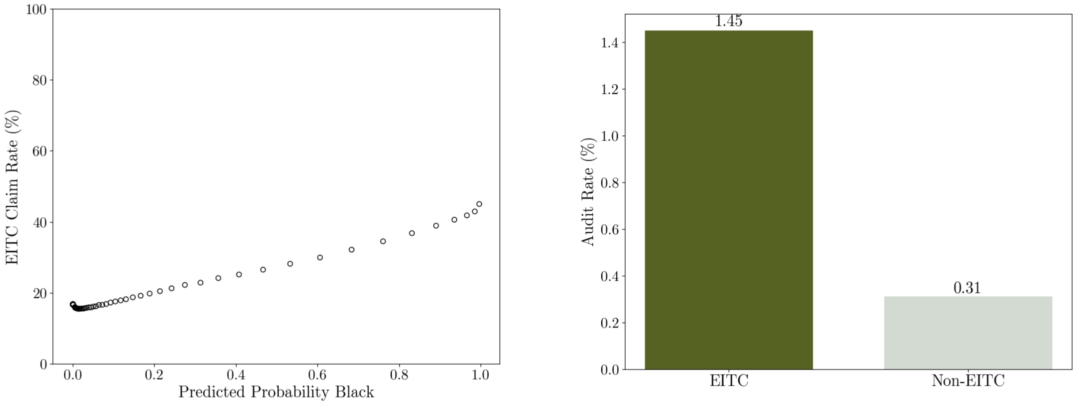
Notes: Left: The figure shows the probabilistic estimate of audit rate by income percentile among Black and non-Black taxpayers filing returns for tax year 2014. Income is measured according to the adjusted gross income (AGI) reported on the taxpayer's return (i.e., prior to audit adjustments). The sample is limited to taxpayers reporting non-negative AGI. Taxpayers have been grouped into 100 equal-sized bins, based on their AGI. Right: The figure shows the corresponding analysis to the left panel, but estimating binned audit rates using the linear estimator instead of the probabilistic estimator.

Figure A.4: Covariance Condition Estimates for Income Bins



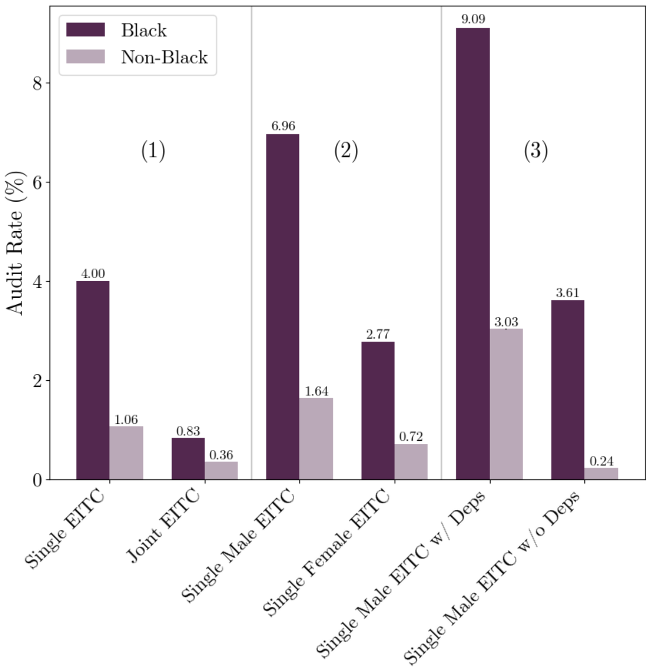
Notes: The figure displays the estimated covariance between audits and self-reported race, conditional on estimated race, as well as the estimated covariance between audits and estimated race, conditional on self-reported race, for each income bin reported in Figure 4. The estimates are calculated from the matched sample of North Carolina taxpayers, using the weights described in Appendix C. Brackets denote 95% confidence intervals.

Figure A.5: Audit Rates by EITC Claim Status



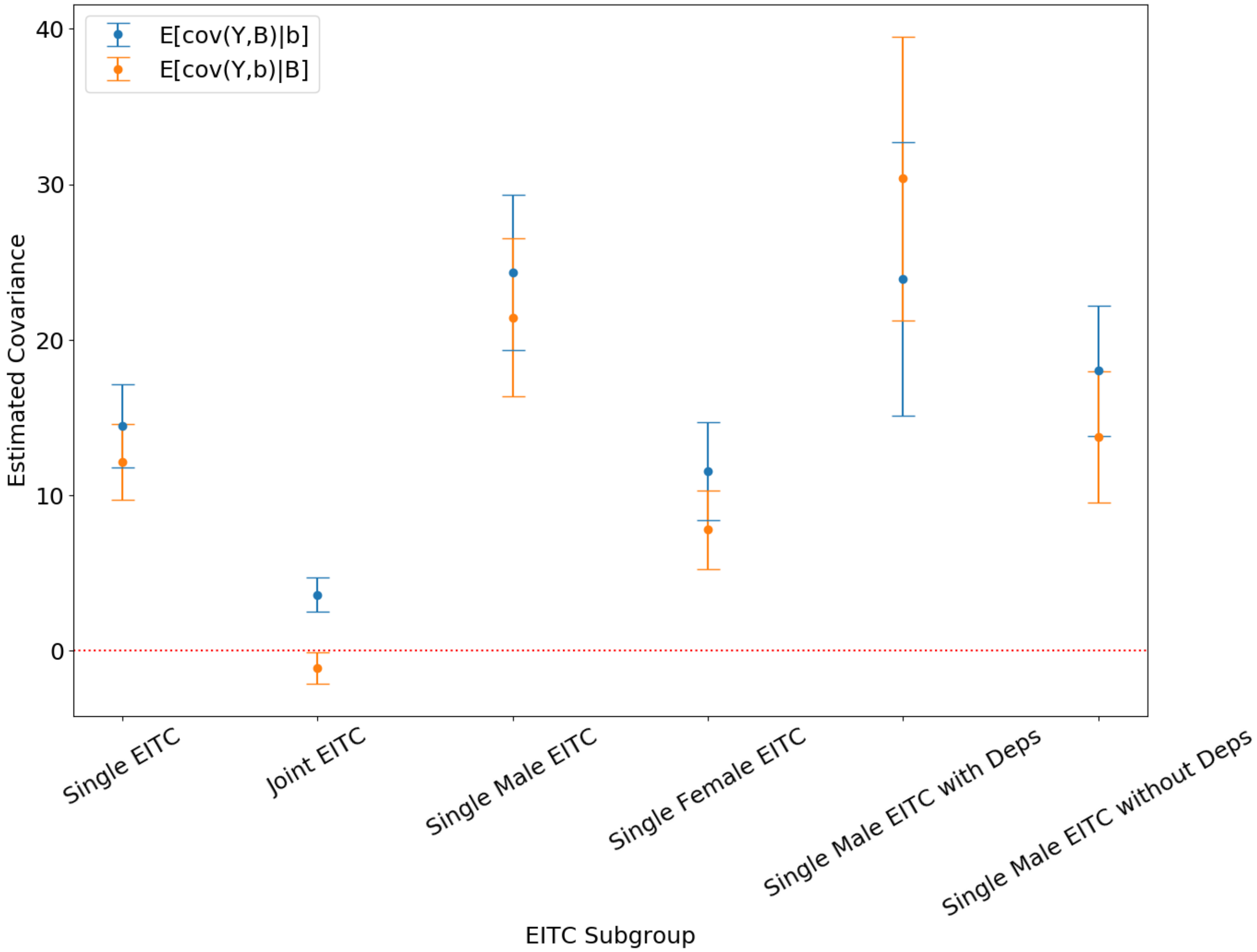
Notes: The figure shows the relationship between audits and EITC claim status among taxpayers (of any race) filing returns for tax year 2014. Left: Binned scatterplot of EITC claim rate by BIFSG-predicted probability that a taxpayer is Black. Taxpayers have been grouped into 100 equal-sized bins. Right: Audit rates among EITC claimants and non-EITC claimants.

Figure A.6: Audit Rate Disparities by EITC Subgroup (Linear Estimator)



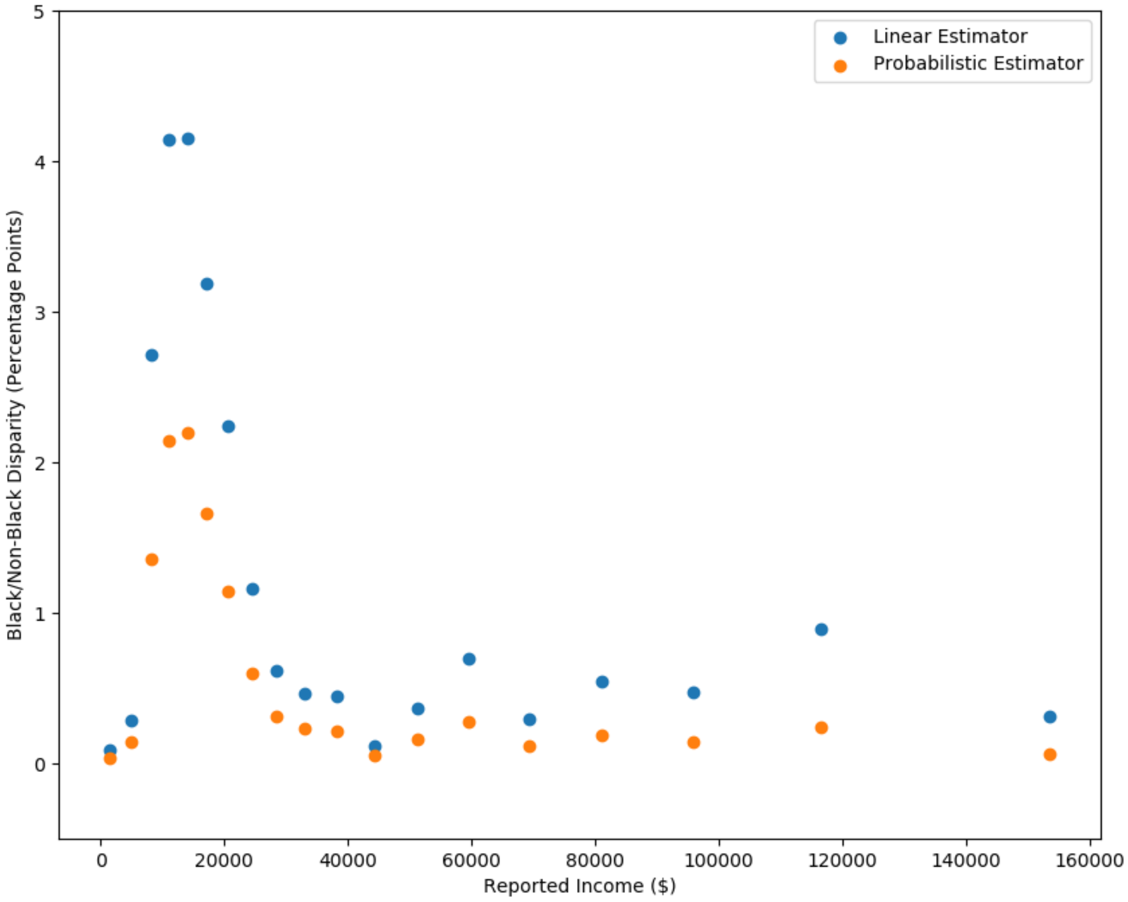
Notes: The figure shows the estimated audit rate among the specified subgroups of Black and non-Black taxpayers. Conditional audit rates by race are calculated using the linear audit rate estimator applied to BIFSG-predicted probabilities that a taxpayer is Black. Panel (1) splits EITC claimants by single vs joint filers; (2) splits single EITC claimants by taxpayer gender; and (3) splits single men claiming the EITC by whether they claim dependents.

Figure A.7: Covariance Condition Estimates for EITC Subgroups



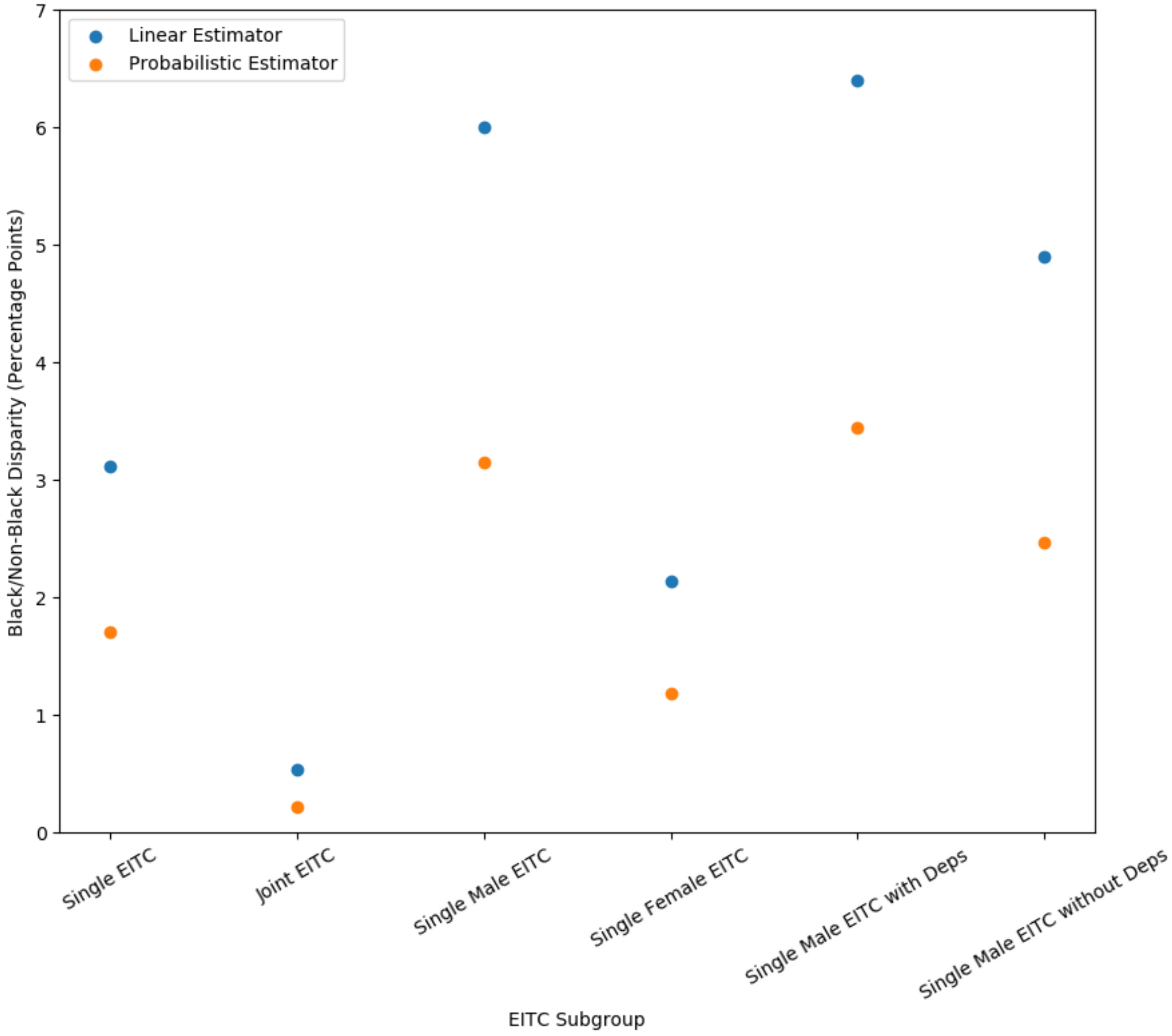
Notes: The figure displays the estimated covariance between audits and self-reported race, conditional on estimated race, as well as the estimated covariance between audits and estimated race, conditional on self-reported race, for each income bin reported in Figure 6. The estimates are calculated from the matched sample of North Carolina taxpayers, using the weights described in Appendix C. Brackets denote 95% confidence intervals.

Figure A.8: Disparity Estimates Within Income Bins Using Re-Calibrated BIFSG Probabilities



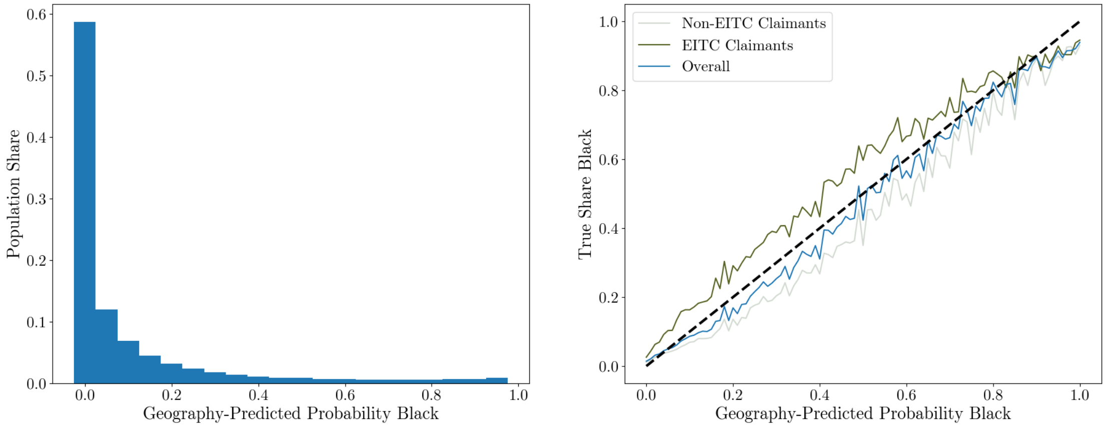
Notes: The figure reports linear and probabilistic audit rate disparity estimates in each of the income bins considered in Figure 4 using BIFSG scores recalibrated by the ground truth sample of North Carolina taxpayers in each bin. See Appendix Section B.5 for details on obtaining the recalibrated proxy. The recalibration exercise incorporates the North Carolina weights described in Appendix C.

Figure A.9: Disparity Estimates Within EITC Subgroups Using Re-Calibrated BIFSG Probabilities



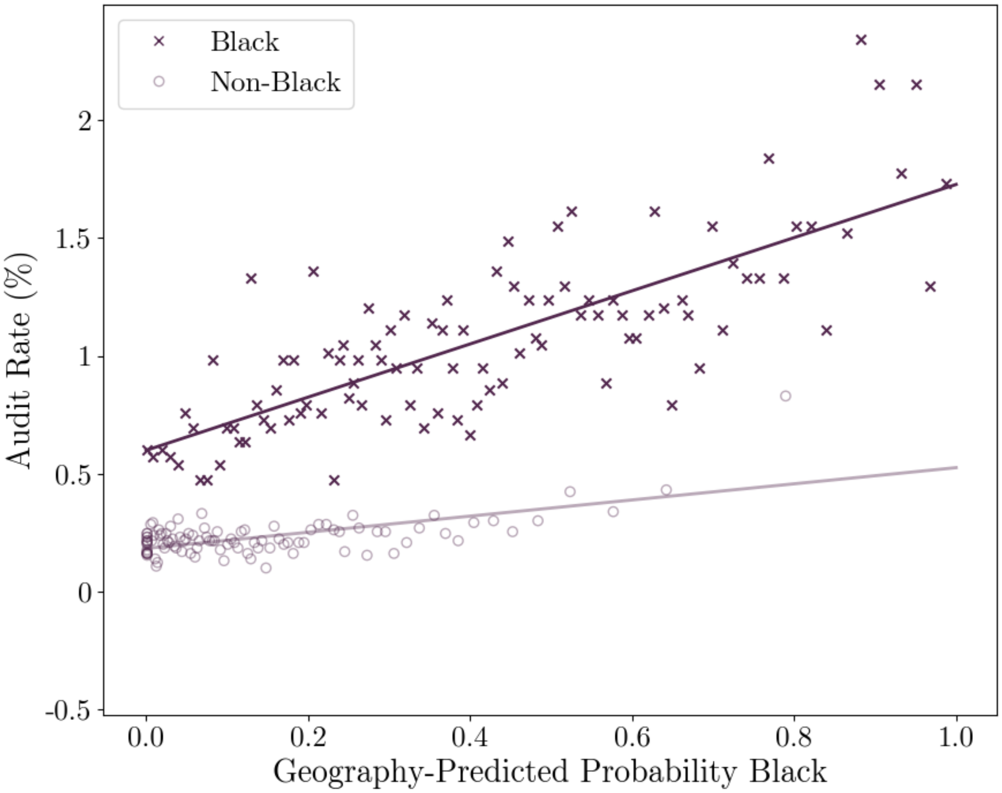
Notes: The figure shows probablistic and linear audit rate disparity estimates corresponding to each of the subgroups considered in Figure 6 using BIFSG scores recalibrated by the ground truth sample of North Carolina taxpayers in each bin. See Appendix Section B.5 for details on obtaining the recalibrated proxy. The recalibration exercise incorporates the North Carolina weights described in Appendix C.

Figure A.10: Distribution and Calibration of Geography-Based Race Imputations



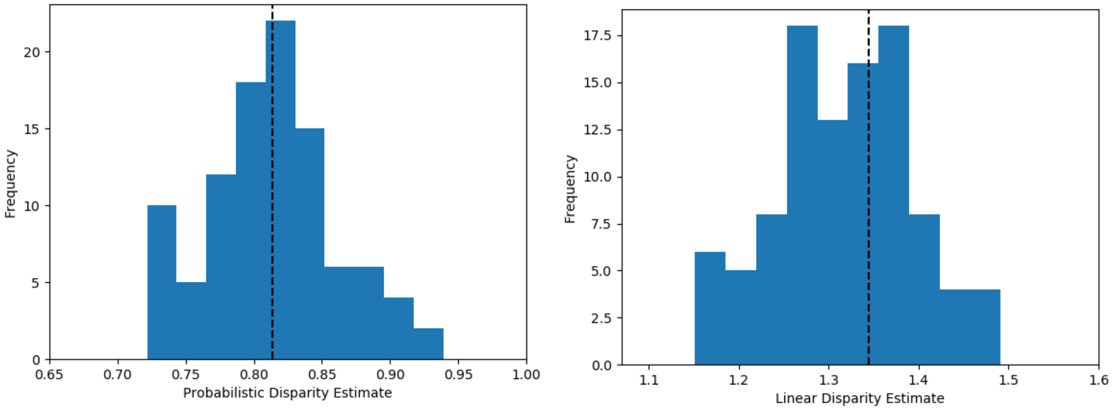
Notes: The figure replicates Figure 1 using estimated race probabilities based only on the location of a taxpayer's residence; the probability that a taxpayer is Black is set equal to the fraction of their CBG that is composed of Black residents. Left: Histogram of geography-predicted probabilities that a taxpayer is Black. The mean prediction is 11.8%. Right: The figure shows the calibration of the geographic predictions in the matched North Carolina dataset. Taxpayers are split into groups based on their predicted probability of being Black (discretized into 100 bins 1 percentage point wide). The predicted probability of being Black is on the x-axis; the y-axis represents the true proportion of each group that is Black according to the ground-truth race observed in the North Carolina matched sample, re-weighted to be representative of the overall United States (see Appendix C for details). A perfectly calibrated predictor would fall on the 45-degree line, shown as the black dotted line. The figure shows overall calibration in blue as well as calibration among EITC claimants (dark green) and non-EITC claimants (light green).

Figure A.11: Audit Rate by Geography-Based Race Imputation and Self-Reported Race



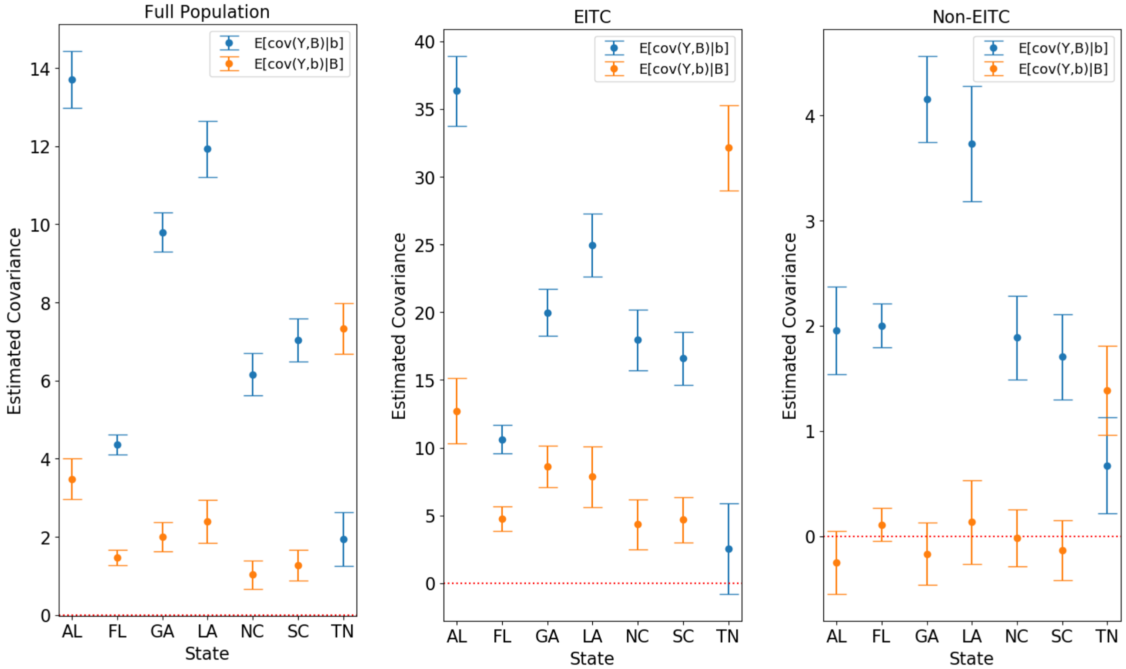
Notes: The figure replicates Figure 2 using estimated race probabilities based only on the location of a taxpayer's residence; the probability that a taxpayer is Black is set equal to the fraction of their CBG that is composed of Black residents. The figure shows the relationship between audit incidence and BIG-predicted probability that a taxpayer is Black for taxpayers filing returns for tax year 2014. Audit rates are plotted separately for Black and non-Black taxpayers in the North Carolina matched sample. Black and non-Black taxpayers are each grouped into 100 equal-sized bins, with Black taxpayers indicated by dark purple x's and non-Black taxpayers indicated by light purple circles.

Figure A.12: Distribution of Disparity Estimates from Dual-Bootstrap



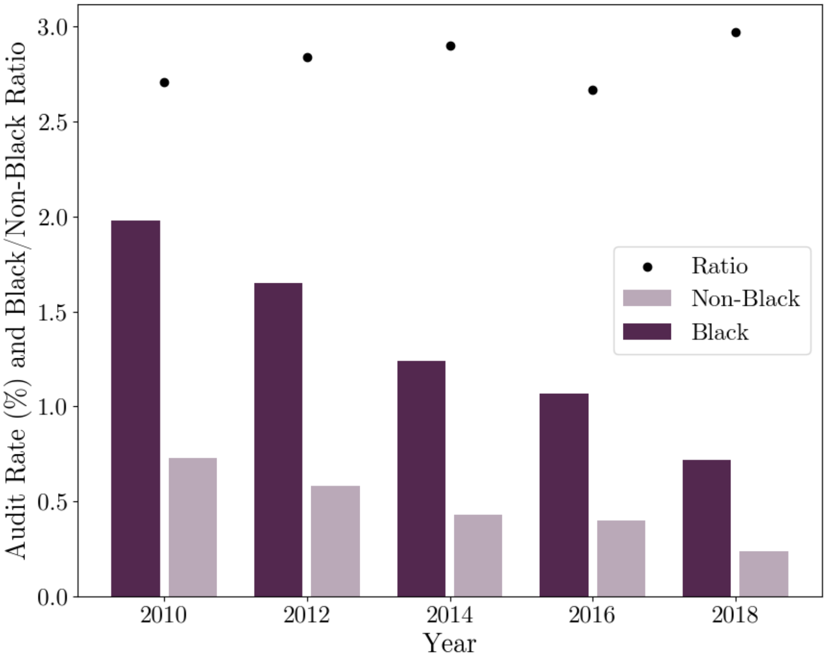
Notes: The figure shows the distribution of probabilistic (left panel) and linear (right panel) disparity estimates produced by our implementation of the Lu et al. (2024) dual-bootstrap procedure. To implement the procedure, we resampled first names and geographic information, but did not resample surnames (since the Census surname data we use corresponds to the full population rather than a sample) to generate 100 sets of BIFSG posteriors. We then estimate disparity with each set of BIFSG posteriors, resampling our taxpayer population on each iteration. Units are percentage points (0-100). The dashed line represents the disparity point estimate presented in Table 1.

Figure A.13: Residual Covariance Estimates by State



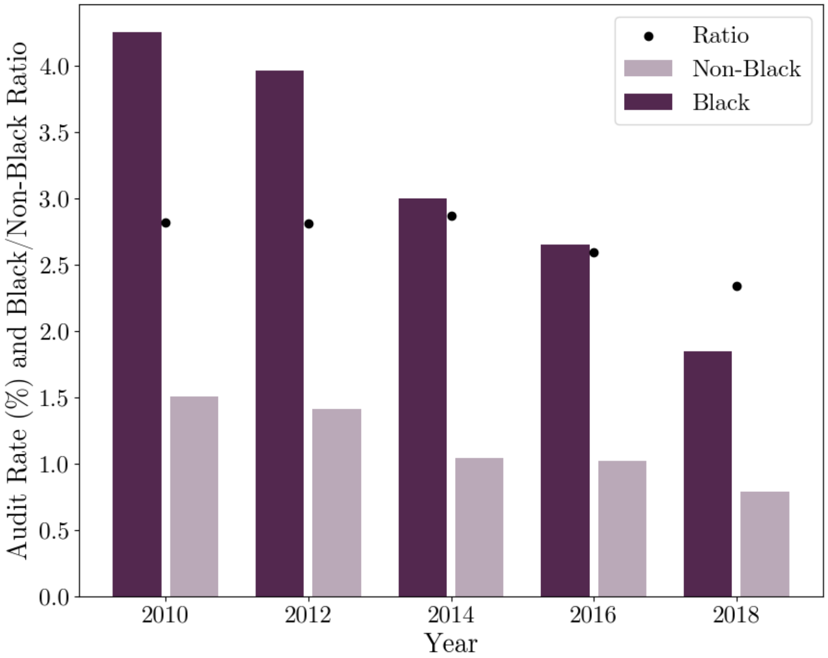
Notes: Left: The figure shows estimates of the covariance between audits (Y) and self-reported race (B), conditional on estimated race (b), as well as the estimated covariance between audits and estimated race, conditional on self-reported race (corresponding to the terms $\mathbb{E}[\text{Cov}(Y, B|b)]$ and $\mathbb{E}[\text{Cov}(Y, b|B)]$ respectively) for taxpayers that match to 2023 L2-collected voter data in seven states (match rates are displayed in parentheses): Alabama (37.18%), Florida (30.01%), Georgia (30.94%), Louisiana (38.60%), North Carolina (16.50%), South Carolina (36.21%), and Tennessee (22.63%). The match procedure we use for these states is the same as the procedure we use to match our main North Carolina data set, described in Appendix C. Error bars represent 95% confidence intervals. The displayed estimates and underlying standard errors are multiplied by 10^4. The panels show results for the full population, EITC claimants, and EITC non-claimants, respectively.

Figure A.14: Estimated Audit Rate Disparity by Year



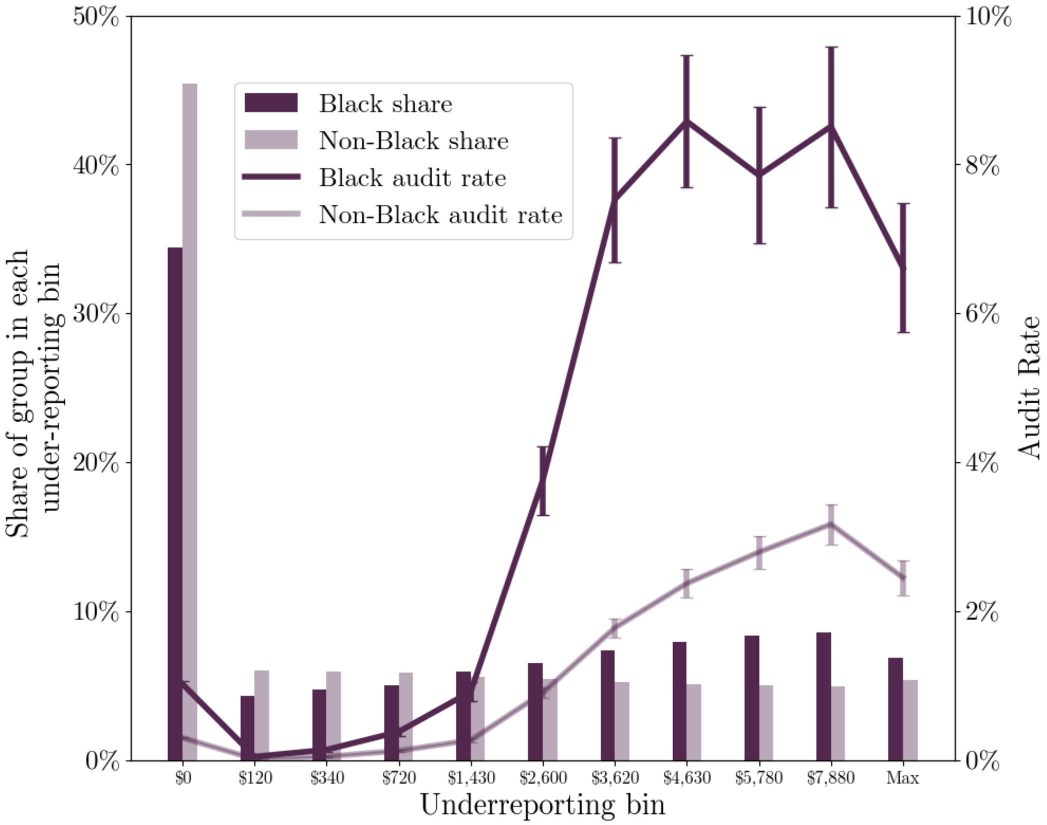
Notes: The figure reports the estimated audit rates among Black and non-Black taxpayers for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). "Ratio" refers to the ratio of the estimated Black audit rate to the estimated non-Black audit rate.

Figure A.15: Estimated Audit Rate Disparity Among EITC Claimants by Year



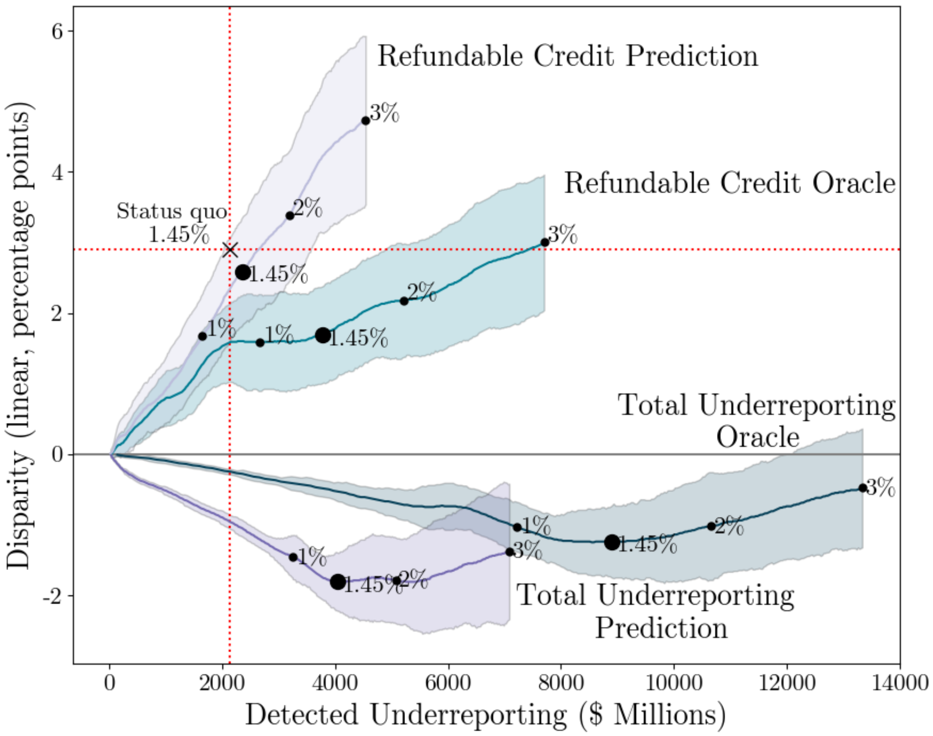
Notes: The figure reports the estimated audit rates among Black and non-Black EITC claimants for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). “Ratio” refers to the ratio of the estimated Black audit rate to the estimated non-Black audit rate.

Figure A.16: Racial Audit Disparity Among EITC Claimants by Underreported Taxes (Linear Estimator)



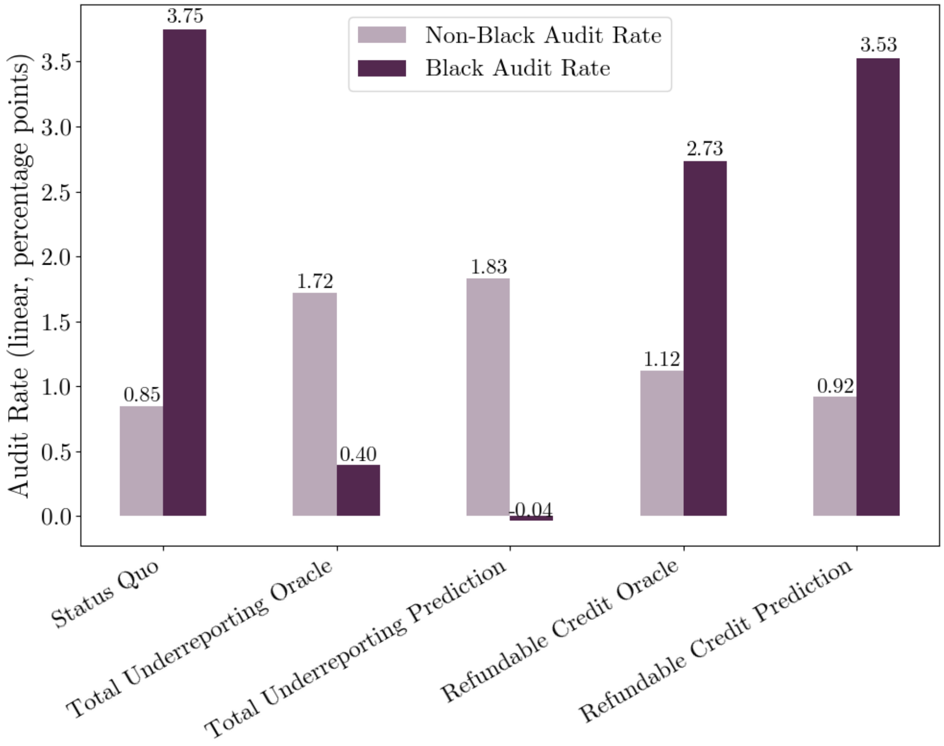
Notes: The figure shows the estimated audit rates for Black and non-Black EITC claimants, respectively, by under-reported taxes. Taxpayers are binned into 11 categories: those with less than $1 of under-reporting, and 10 equal deciles of taxpayers with positive under-reporting. Underreporting deciles are defined based on the distribution of underreporting among EITC claimants, as measured by NRP audits. Estimated audit rates by race are calculated using the linear estimator and the method described in Section 6 of the main text. All analyses account for NRP sampling weights. Brackets reflect the estimated 95% confidence interval, derived from bootstrapped standard errors (N=100). The bars show the estimated share of Black and non-Black taxpayers, respectively, that fall into each under-reporting bin.

Figure A.17: Detected Underreporting and Disparity by Algorithm (Linear Estimator)



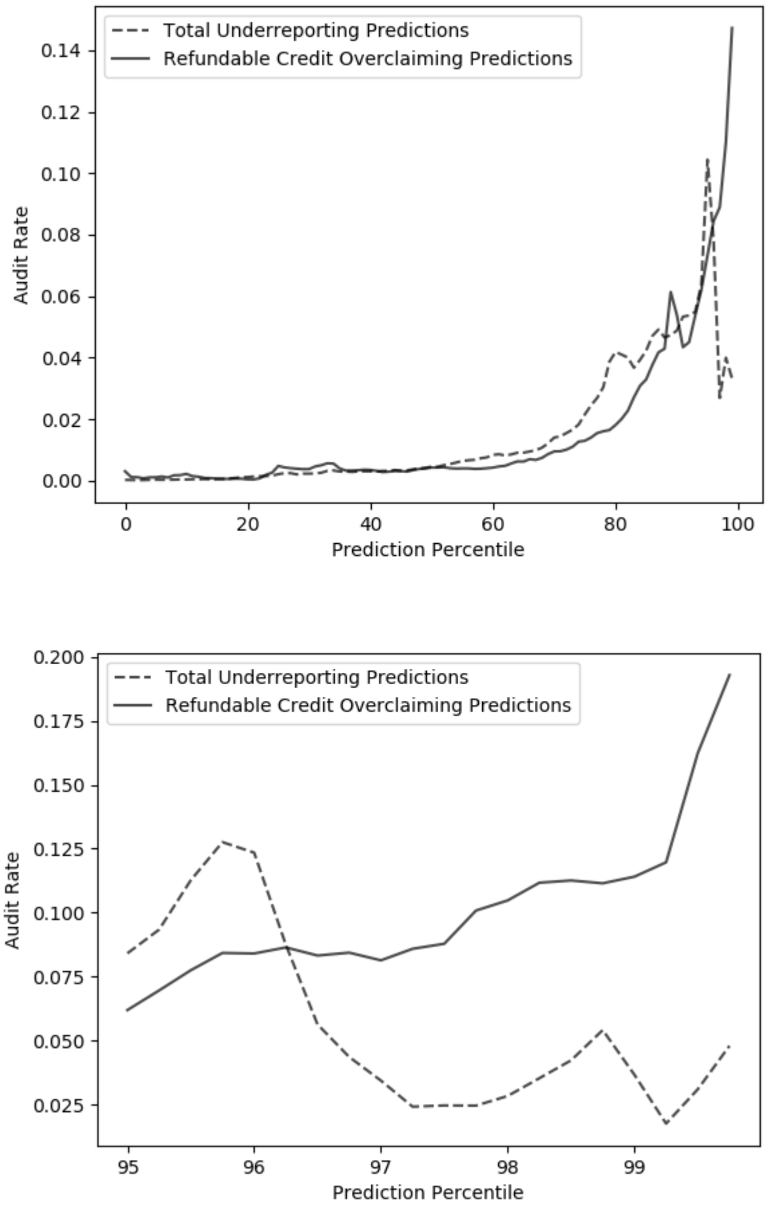
Notes: The figure replicates Figure 8, but with disparity calculated based on the linear disparity estimator. The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates. The trajectories correspond to the total underreporting oracle (dark blue), total underreporting prediction (dark purple), refundable credit oracle (light blue), and refundable credit prediction (light purple) algorithms. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified algorithm at the audit rate specified in the label. For each algorithm, the audit rates considered range from 0.1% to 3%; the audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The total underreporting prediction algorithm is based on a random forest regressor trained to predict total underreporting. The refundable credit prediction algorithm is based on a random forest regressor trained to predict total adjustments to EITC, CTC, and AOTC amounts. The total underreporting oracle selects returns in descending order of true underreporting. The refundable credit oracle selects returns in descending order of true EITC, CTC, and AOTC overclaiming. Annualized detected underreporting is calculated as the total amount of adjustments (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm. Detected underreporting and disparity estimates are constructed using the full set NRP EITC claimants from 2010-14. Details relating to the predictive models and associated estimates are contained in Appendix E. The point labeled "Status quo" shows the estimated disparity in audit rates between Black and non-Black taxpayers and detected underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around disparity estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Figure A.18: Group-Specific Audit Rates by Algorithm (Linear)



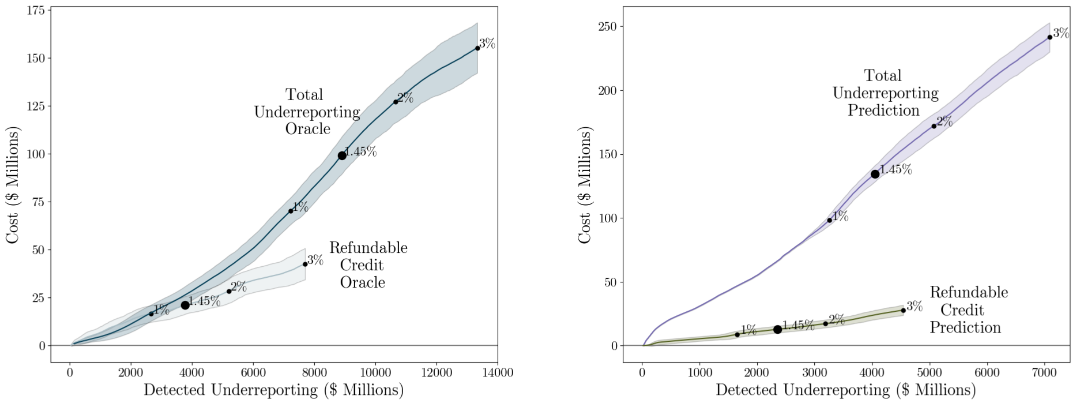
Notes: The figure replicates Figure 9, but with audit rates calculated using the linear estimator. It reports estimated audit rates for Black and non-Black EITC claimants that would be induced by the algorithms considered in Figure A.17 under an assumed audit rate of 1.45%. The population of tax returns available for audit is based on the NRP sample of taxpayers claiming the EITC; see notes to Figure A.17 for additional detail. Status quo refers to the estimated audit rates by race for tax year 2014 returns claiming the EITC, reported in Figure 5.

Figure A.19: Audit Rates by Prediction Percentiles



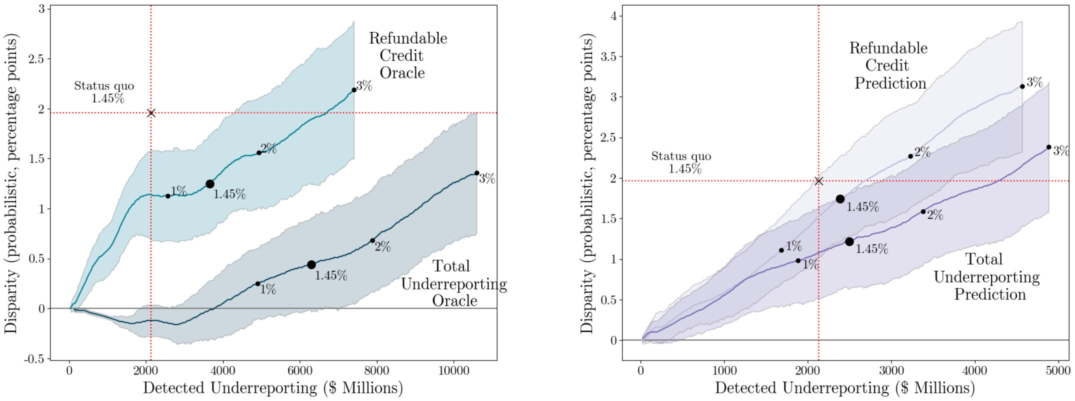
Notes: This figure displays the relationship between operational audit rates and predictions from the Total Underreporting and Refundable Credit Overclaiming Prediction algorithms, applied to the population of EITC claimants in 2014. For tractability, the versions of the Total Underreporting and Refundable Credit Overclaiming algorithms applied for the analysis in this figure are simplified in the following sense: We apply our NRP-based models to non-NRP taxpayers by training models with the forty most important features from both the total underreporting prediction and refundable credit overclaiming prediction models respectively and using them to obtain predictions. We validate that these simplified models yield similar results to the full-featured models in the NRP data set. In the top panel, we divide taxpayers into 100 equally-sized bins separately for each model and compute the audit rate separately for each bin. In the bottom panel, we divide taxpayers into 400 equally-sized bins separately for each model and filter to the top 20, computing audit rates separately for each bin. In this panel we divide the bin numbers on the x-axis by 4 to report results for each quarter of a percentile from 95 through 100.

Figure A.20: Detected Underreporting and Cost by Algorithm



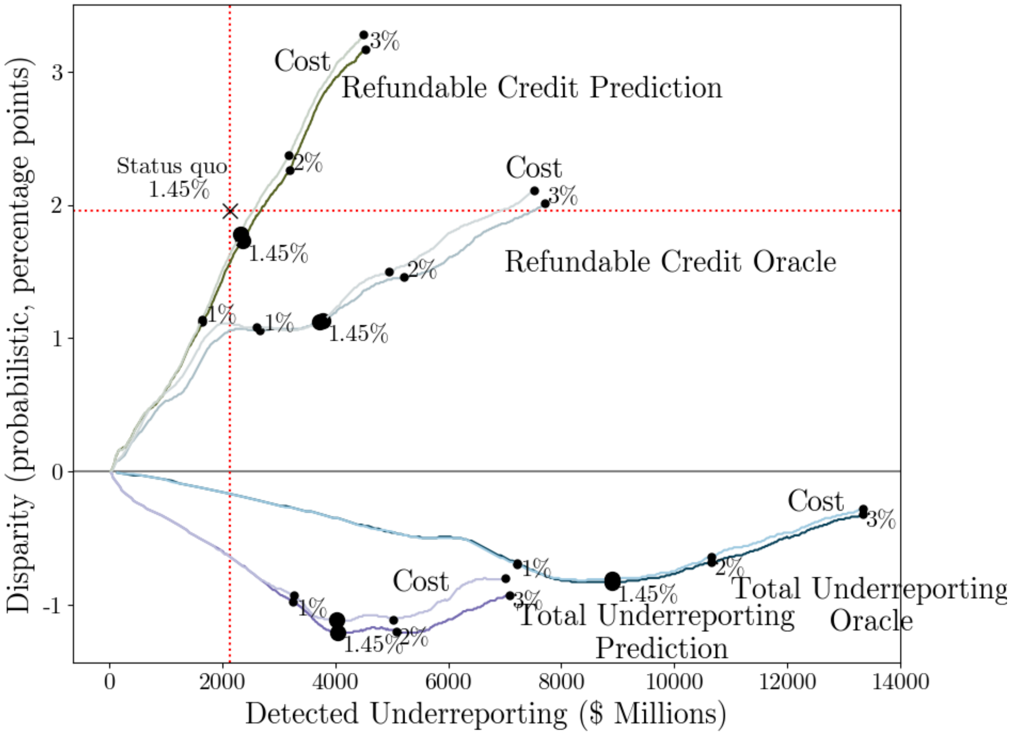
Notes: The figure shows the annualized cost (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates. Predictive models are trained and evaluated on the set of NRP EITC claimants from 2010-14; see Appendix E for details. The left panel displays trajectories for the total underreporting oracle (dark blue) and the refundable credit oracle (light blue) algorithms. The right panel displays trajectories for the total underreporting prediction (dark purple) and the refundable credit prediction (light purple) algorithms. The labeled points along each trajectory represent estimated detected underreporting and cost for the specified algorithm at the audit rate specified in the label. The audit rates considered range from 0.1% to 3%. The audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The total underreporting prediction algorithm is based on a random forest regressor trained to predict total underreporting. The refundable credit prediction algorithm is based on a random forest regressor trained to predict total adjustments to EITC, CTC, and AOTC amounts. The total underreporting oracle selects returns in descending order of true underreporting. The refundable credit oracle selects returns in descending order of true EITC, CTC, and AOTC overclaiming. Annualized cost is calculated as the total cost incurred to audit the returns selected under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. The cost estimates are calculated from operational audits based on the average time logged by IRS employees dealing directly with the case, multiplied by the applicable General Schedule payscale based on the employee level. Returns without substantial business income (not reporting gross business receipts in excess of $25,000) are classified into activity code 270 and are assigned a cost of $23.09, the winsorized average cost of tax returns in activity code 271 in our sample. Returns with substantial business income (reporting gross business receipts in excess of $25,000) are classified into activity code 271 and are assigned a cost of $369.70, the winsorized average cost of tax returns in activity code 271 in our sample. Annualized detected underreporting is calculated as the total detected underreporting (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around cost estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Figure A.21: Detected Underreporting and Disparity by Algorithm (Constrained Models)



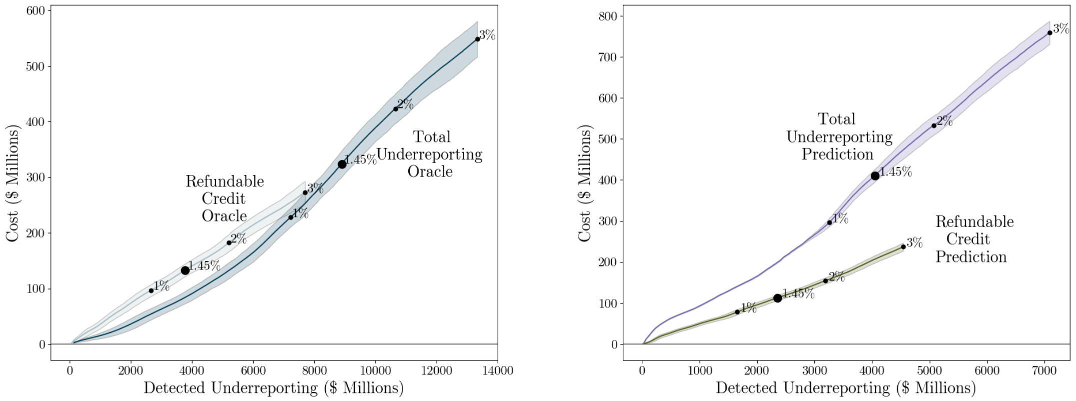
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates, where each algorithm's allocation of audits between EITC business and non-business activity codes is constrained to match the status quo allocation. The left panel shows trajectories corresponding to the total underreporting oracle (dark blue) and the refundable credit oracle (light blue) algorithms. The right panel shows trajectories corresponding to the total underreporting prediction (dark purple) and refundable credit prediction (light purple) algorithms. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified algorithm at the audit rate specified in the label. For each algorithm, the audit rates considered range from 0.1% to 3%; the audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The total underreporting prediction algorithm is based on a random forest regressor trained to predict total underreporting. The refundable credit prediction algorithm is based on a random forest regressor trained to predict total adjustments to EITC, CTC, and AOTC amounts. The total underreporting oracle selects returns in descending order of true underreporting. The refundable credit oracle selects returns in descending order of true EITC, CTC, and AOTC overclaiming. Disparity is calculated from the probabilistic disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. Annualized detected underreporting is calculated as the total amount of adjustments (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm. Detected underreporting and disparity estimates are constructed using the full set NRP EITC claimants from 2010-14. Details relating to the predictive models and associated estimates are contained in Appendix E. The point labeled "Status quo" shows the estimated disparity in audit rates between Black and non-Black taxpayers and detected underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around disparity estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Figure A.22: Allocating Audits Based on Reward Net of Audit Costs



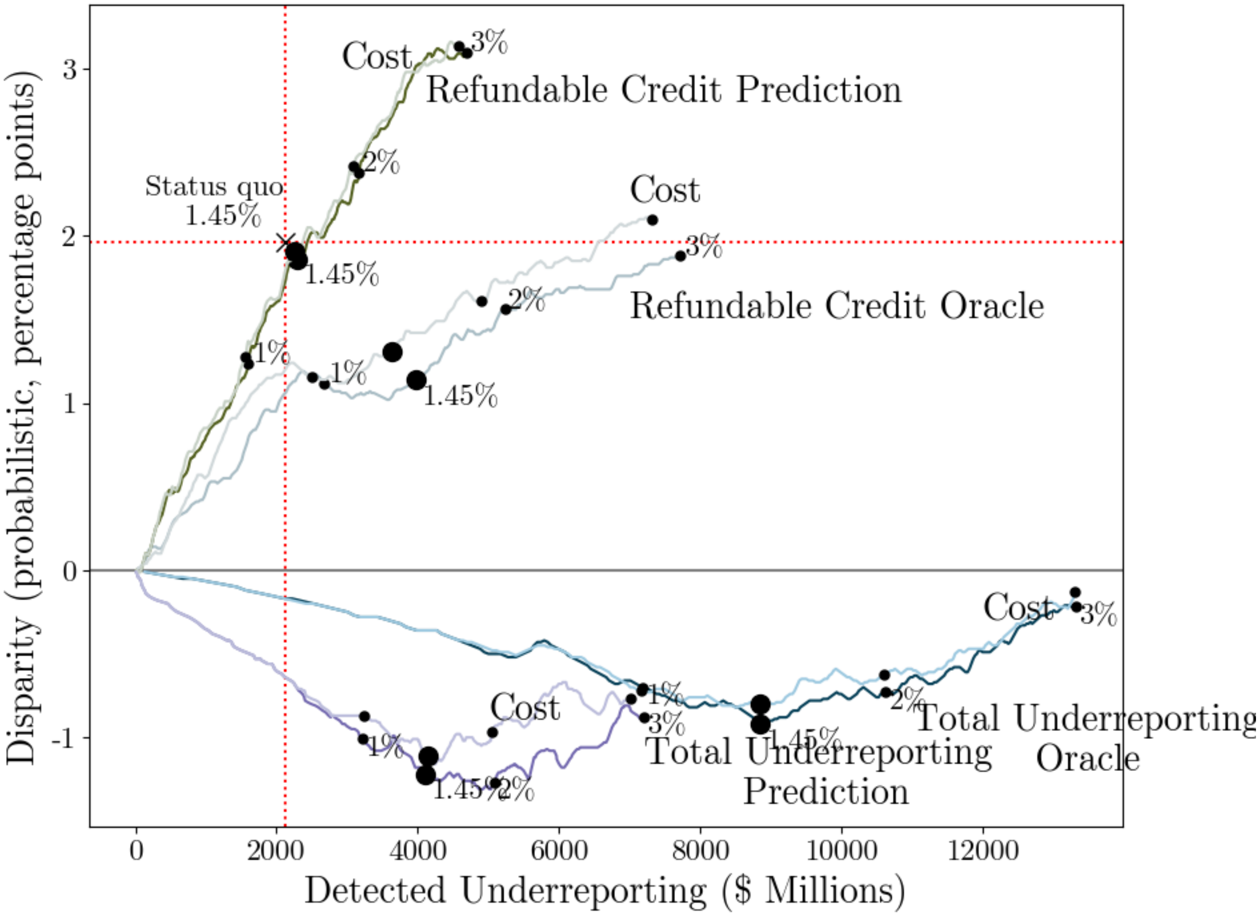
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative assumptions about whether returns are selected for audit based on detected reward, whether total underreporting or refundable credit overclaiming (gross of audit costs, as in our other analyses), or based on detected reward minus expected audit costs. Underreporting/refundable credit overclaiming is based on either the oracle or the random forest regressor prediction algorithm, as specified. Audit costs are measured at the activity code level, using data on the time spent on audit examination and the salary grade of the examiner, and abstracting from non-salary costs associated with the enforcement process, such as appeals, litigation, and collections, or fixed costs, such as overhead. Using this approach, the average cost of auditing an EITC non-business return (activity code 270) is $23.09, whereas the average cost of auditing an EITC business return (activity code 271) is $369.70. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified algorithm at the audit rate specified in the label. Disparity is calculated from the probabilistic disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. For each algorithm, the audit rates considered range from 0.1% to 3%; the audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. Annualized detected underreporting is calculated as the total amount of adjustments (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm. Detected underreporting and disparity estimates are constructed using the full set NRP EITC claimants from 2010-14. Details relating to the predictive models and associated estimates are contained in Appendix E. The point labeled "Status quo" shows the estimated disparity in audit rates between Black and non-Black taxpayers and detected underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights.

Figure A.23: Detected Underreporting and Cost by Algorithm (NRP-Derived Audit Costs)



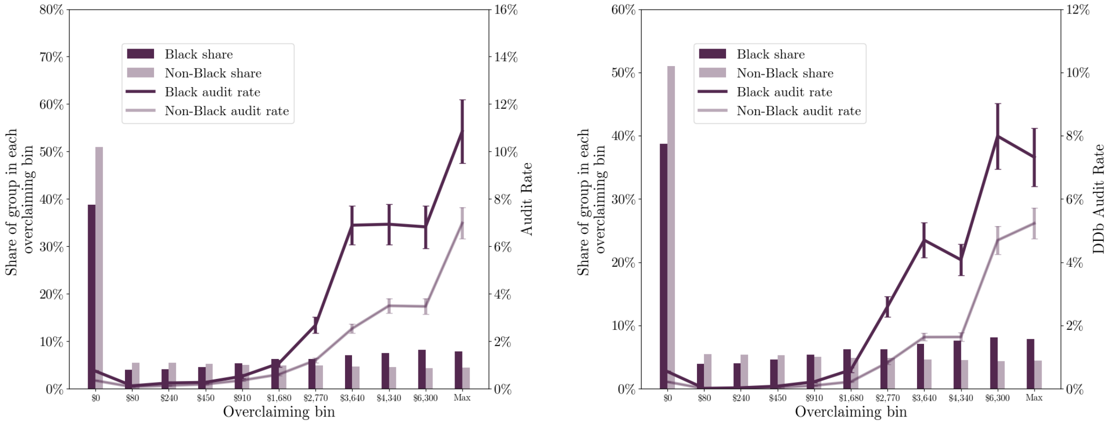
Notes: The figure replicates Appendix Figure A.20 using an alternative measure of costs derived from the number of hours reported by examiners conducting NRP examinations. Audit costs are measured at the activity code level, with the average cost of auditing an EITC non-business return (activity code 270) being $263.51 (compared to $23.09 under our original cost measure), and the average cost of auditing an EITC business return (activity code 271) being $1,110.90 (compared to $369.70 under our original cost measure). The figure shows the annualized cost (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates. Predictive models are trained and evaluated on the set of NRP EITC claimants from 2010-14; see Appendix E for details. The left panel displays trajectories for the total underreporting oracle (dark blue) and the refundable credit oracle (light blue). The right panel displays trajectories for the total underreporting prediction model (dark purple), and the refundable credit prediction model (light purple). The labeled points along each trajectory represent estimated detected underreporting and cost for the specified algorithm at the audit rate specified in the label. The audit rates considered range from 0.1% to 3%. The audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The total underreporting prediction algorithm is based on a random forest regressor trained to predict total underreporting. The refundable credit prediction algorithm is based on a random forest regressor trained to predict total adjustments to EITC, CTC, and AOTC amounts. The total underreporting oracle selects returns in descending order of true underreporting. The refundable credit oracle selects returns in descending order of true EITC, CTC, and AOTC overclaiming. Annualized cost is calculated as the total cost incurred to audit the returns selected under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. Annualized detected underreporting is calculated as the total detected underreporting (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around cost estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Figure A.24: Allocating Audits Based on Reward Net of Audit Costs (NRP-Derived Audit Costs)



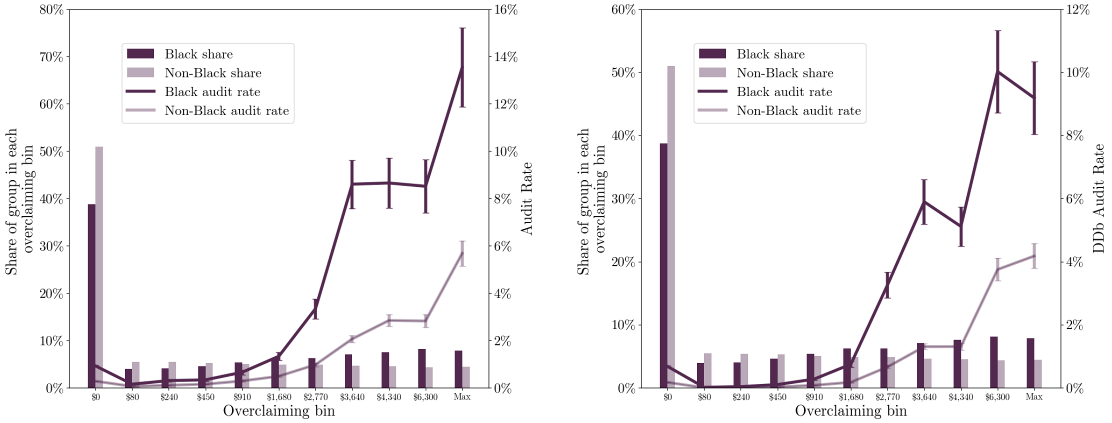
Notes: The figure replicates Appendix Figure A.22 using an alternative measure of costs derived from the number of hours reported by examiners conducting NRP examinations. Audit costs are measured at the activity code level, with the average cost of auditing an EITC non-business return (activity code 270) being $263.51 (compared to $23.09 under our original cost measure), and the average cost of auditing an EITC business return (activity code 271) being $1,110.90 (compared to $369.70 under our original cost measure). The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative assumptions about whether returns are selected for audit based on detected reward, whether total underreporting or refundable credit overclaiming (gross of audit costs, as in our other analyses), or based on detected reward minus expected audit costs. Underreporting/refundable credit overclaiming is based on either the oracle or the random forest regressor prediction algorithm, as specified. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified algorithm at the audit rate specified in the label. Disparity is calculated from the probabilistic disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. For each algorithm, the audit rates considered range from 0.1% to 3%; the audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. Annualized detected underreporting is calculated as the total amount of adjustments (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm. Detected underreporting and disparity estimates are constructed using the full set NRP EITC claimants from 2010-14. Details relating to the predictive models and associated estimates are contained in Appendix E. The point labeled “Status quo” shows the estimated disparity in audit rates between Black and non-Black taxpayers and detected underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights.

Figure A.25: Racial Audit Disparity Among EITC Claimants by Overclaimed Refundable Credits



Notes: The figure shows two analogs of Figure 7. Each panel group taxpayers into bins along the x-axis based on refundable credit overclaiming. For the left panel, the y-axis reflects any type of audit; for the right panel, the y-axis reflects DDb audits only. For the left panel, audit-imposed adjustments to refundable credits are used to measure a taxpayer's overclaimed refundable credits, conditional on being selected for operational audit, by race. For the right panel, total audit adjustments from DDb audits are used as a proxy for a taxpayer's overclaimed refundable credits, conditional on being selected for operational audit, by race. In both panels, taxpayers are binned into 11 categories: those with less than $1 of overclaimed refundable credits, and 10 equal deciles of taxpayers with positive overclaimed refundable credits. Overclaiming deciles are defined based on the distribution of overclaiming among EITC claimants, as measured by NRP audits. Bin labels on the x-axis reflect the upper dollar limit of each overclaiming bin (rounded for confidentiality). Estimated audit rates by race are calculated using the probabilistic estimator. All analyses account for NRP sampling weights. Brackets reflect the estimated 95% confidence interval, derived from bootstrapped standard errors (N=100). The bars show the estimated share of Black and non-Black taxpayers, respectively, that fall into each overclaiming bin. A similar analysis, corresponding to the linear estimator, is presented in Appendix Figure A.26.

Figure A.26: Racial Audit Disparity Among EITC Claimants by Overclaimed Refundable Credits (Linear Estimator)



Notes: The figure replicates the analysis in Figure A.25, using the linear estimator to calculate audit rates by race.

Table A.1: EITC Audit Frequency by Timing and Type

	Correspondence	Field/Office	All Audit Types
Pre-Refund	270,940	0	270,940
	(66.4%)	(0%)	(66.4%)
Post-Refund	112,689	24,361	137,050
	(27.6%)	(6.0%)	(33.6%)
All Audit Times	383,629	24,361	407,990
	(94.0%)	(6.0%)	(100%)

Notes: The table reports the frequency of audits of 2014 tax returns claiming the EITC by audit timing (whether the audit occurred pre- or post-refund) and by audit type (whether the audit was conducted by correspondence or as a field or office examination). Percentages (reported in parentheses) reflect the share of all audits of the specified taxpayer population that fall into the specified audit category.

Table A.2: Coverage of BIFSG Features Among 2014 Taxpayers

Case	Count	First Name	Last Name	CBG	Share of Total
1	107,624,714	X	X	X	72.6%
2	10,087,515	X	X		6.8%
3	10,455,708	X		X	7.1%
4	14,981,324		X	X	10.1%
5	2,572,849			X	1.7%
6	1,431,541		X		1.0%
7	903,311	X			0.6%
8	248,356				0.2%
Total	148,305,318				100%

Notes: The table shows the availability of the data used to calculate race probabilities for primary filers on tax year 2014 returns. The distribution of race by first name is tabulated from mortgage applications, following Tzioumis (2018); it is missing for names not among the 4,250 most common names in that data. The distribution of race by last names is tabulated from 2010 Census data and includes the 162,253 most common surnames. The distribution of race by Census Block Group (CBG) is tabulated from the Census 2014 5-Year American Community Survey and covers all CBGs. CBG data is missing for taxpayers who cannot be reliably geo-coded to a specific CBG. In our main analysis, taxpayers in row 8 are assigned a BIFSG-predicted probability Black based on the national average share of the population that is Black.

Table A.3: Calibration Metrics for BIFSG Predictions.

	Full Population		EITC Population	
Metric	Imputed	Recalibrated	Imputed	Recalibrated
	(1)	(2)	(3)	(4)
Area Under ROC Cruve	0.9048	0.9048	0.9038	0.9038
Panel A: 50% Threshold				
False Positive	0.0880	0.0632	0.1119	0.1181
True Positive	0.6804	0.6066	0.7237	0.7377
False Negative	0.3196	0.3934	0.2763	0.2623
True Negative	0.9120	0.9368	0.8881	0.8819
Precision	0.6529	0.7002	0.8002	0.7946
Recall	0.6804	0.6066	0.7237	0.7377
Accuracy	0.8667	0.8722	0.8252	0.8268
Panel B: 75% Threshold				
False Positive	0.0338	0.0112	0.0504	0.0504
True Positive	0.4740	0.2822	0.5169	0.5170
False Negative	0.5260	0.7178	0.4831	0.4830
True Negative	0.9662	0.9888	0.9496	0.9496
Precision	0.7733	0.8598	0.8641	0.8641
Recall	0.4740	0.2822	0.5169	0.5170
Accuracy	0.8699	0.8506	0.7842	0.7842
Panel C: 90% Threshold				
False Positive	0.0114	-	0.0216	0.0191
True Positive	0.2855	-	0.3180	0.2946
False Negative	0.7145	-	0.6820	0.7054
True Negative	0.9886	-	0.9784	0.9809
Precision	0.8588	-	0.9013	0.9053
Recall	0.2855	-	0.3180	0.2946
Accuracy	0.8511	-	0.7259	0.7184

Notes: The table characterizes the predictive power of BIFSG in the task of predicting whether a taxpayer self-reports as Black as measured in the North Carolina data. A positive label refers to the individual self-reporting as Black and a negative label refers to the individual self-reporting as non-Black. We can then characterize performance using standard metrics from the machine learning literature. We use columns to demarcate different versions of the final predictions and comparing against different populations: Columns 1 and 3 correspond to the standard BIFSG score, while Columns 2 and 4 correspond to the re-calibrated BIFSG score (described in Section B.5); and Columns 1 and 2 are evaluations against the full population, while in Columns 3 and 4 are evaluations against only the EITC claimant population. We use rows to demarcate each error metric. The first error metric we consider, the Area Under the Receiver Operator Characteristic (ROC) curve, requires as input only the probabilistic predictions and labels; the other metrics, which are classification-based, require discrete label choices. We thus convert BIFSG scores into predicted labels of Black/non-Black via thresholding, i.e. labeling all observations with predicted probability Black of t or greater as Black and all others as non-Black. We consider thresholds at 50%, 75%, and 90%, demarcated by Panels A-C.

Table A.4: Residual Covariance Estimates

	Full Population		EITC		Non-EITC	
E[cov(Y,B)|b]	5.76^{***}	5.05^{***}	14.68^{***}	13.39^{***}	1.65^{***}	1.20^{***}
	(0.20)	(0.23)	(0.81)	(1.03)	(0.14)	(0.14)
E[cov(Y,b)|B]	2.03^{***}	2.28^{***}	8.76^{***}	9.62^{***}	0.00	-0.28
	(0.14)	(0.24)	(0.65)	(0.97)	(0.10)	(0.18)
Weighted	No	Yes	No	Yes	No	Yes
N	1,613,130		277,064		1,336,062	

Notes: The table displays the estimated covariance between audits and self-reported race, conditional on estimated race, as well as the estimated covariance between audits and estimated race, conditional on self-reported race. The estimates are for the matched sample of North Carolina taxpayers for the full population (columns 1 and 2), EITC claimants (columns 3 and 4), and non-EITC claimants (columns 5 and 6). Columns 2, 4, and 6 are re-weighted to be representative of the U.S. population, using the weights described in Appendix C. Standard errors are displayed in parentheses. The displayed estimates and standard errors are multiplied by 10^4. Stars correspond to p-values derived from two-sided hypothesis tests. $^{*}: P < .10; ^{**}: P < .05; ^{***}: P < .01$.

Table A.5: Estimated Audit Rate by Race (Probabilistic)

		Audit Timing		Audit Type	
	Any	Pre	Post	Correspondence	Field/
	Audit	Refund	Refund		Office
	(1)	(2)	(3)	(4)	(5)
Panel A: Full Population					
Black	1.241	0.743	0.497	1.139	0.103
	(0.003)	(0.002)	(0.002)	(0.003)	(0.001)
Non-Black	0.427	0.174	0.253	0.335	0.093
	(0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
N (Millions)	148.3	148.3	148.3	148.3	148.3
Panel B: EITC Population					
Black	2.989	2.126	0.863	2.897	0.092
	(0.006)	(0.005)	(0.003)	(0.006)	(0.001)
Non-Black	1.039	0.651	0.389	0.954	0.085
	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
N (Millions)	28.3	28.3	28.3	28.3	28.3
Panel C: Non-EITC Population					
Black	0.401	0.080	0.322	0.294	0.108
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Non-Black	0.300	0.074	0.225	0.206	0.094
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
N (Millions)	120.0	120.0	120.0	120.0	120.0

Notes: The table reports estimates of the audit rate (in group-specific levels, not differences between groups) for Black and non-Black taxpayers filing income tax returns for tax year 2014. Units are percentage points (0-100). All estimates are based on the probabilistic audit rate estimator. The category of audit considered varies across columns; for example, the results in column (4) show the estimated rate at which Black and non-Black taxpayers are selected for correspondence audit. Panel A includes all taxpayers, whereas Panels B and C restrict the analysis to EITC claimants and non-claimants, respectively. Standard errors, reported in parentheses, correspond to the standard deviation of the distribution of estimates from 100 bootstrapped samples.

Table A.6: Estimated Audit Rate by Race (Linear)

		Audit Timing		Audit Type	
	Any	Pre	Post	Correspondence	Field/
	Audit	Refund	Refund		Office
	(1)	(2)	(3)	(4)	(5)
Panel A: Full Population					
Black	1.707	1.070	0.637	1.599	0.108
	(0.004)	(0.003)	(0.002)	(0.004)	(0.001)
Non-Black	0.363	0.129	0.234	0.271	0.092
	(0.001)	(< 0.001)	(< 0.001)	(0.001)	(< 0.001)
N (Millions)	148.3	148.3	148.3	148.3	148.3
Panel B: EITC Population					
Black	3.732	2.688	1.044	3.637	0.095
	(0.009)	(0.007)	(0.005)	(0.009)	(0.001)
Non-Black	0.846	0.505	0.342	0.762	0.085
	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
N (Millions)	28.3	28.3	28.3	28.3	28.3
Panel C: Non-EITC Population					
Black	0.471	0.083	0.388	0.355	0.117
	(0.003)	(0.001)	(0.002)	(0.002)	(0.001)
Non-Black	0.292	0.074	0.218	0.199	0.093
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
N (Millions)	120.0	120.0	120.0	120.0	120.0

Notes: The table replicates Appendix Table A.5, but using the linear audit rate estimator instead of the probabilistic audit rate estimator.

Table A.7: Ground-Truth Disparities in Matched North Carolina Sample

Estimator	Full Population	EITC	Non-EITC
	(1)	(2)	(3)
Unweighted	0.835	1.849	0.204
Re-weighted	1.284	2.393	0.265
N	1,613,124	277,064	1,336,060

Notes: The table shows audit rate disparities within the matched North Carolina sample, where we observe self-reported race. Units are percentage points (0-100). The Black/non-Black audit disparity is shown for the full population (column 1), EITC claimants (column 2), and non-EITC claimants (column 3). The first row computes audit rate disparities directly using the data, while the second row re-weights the data to be representative of the full population using the North Carolina weights described in Appendix C.

Table A.8: Audit Burden Disparity Estimates (Probabilistic)

	Correspondence	Office	Field	Overall
Audit Rate (pp)	0.43	0.06	0.03	0.53
Share of Total Audits	0.82	0.12	0.06	1.00
Taxpayer Hours	30	38	34	31.18
Taxpayer Compliance Cost	643	1,717	4,431	1,002.98
Penalties and Interest	320.71	1,580.77	6,434.52	846.77
Assessed Taxes	5,252.56	7,130.46	24,960.56	6,694.85
Disparity (Audit Rate, pp)	0.80	0.02	-0.01	0.66
Disparity (Hours)	0.241	0.008	-0.004	0.199
Disparity (Compliance Cost)	5.17	0.36	-0.50	4.26
Disparity (Penalties and Interest)	2.58	0.34	-0.73	2.11
Disparity (Assessed Taxes)	42.20	1.51	-2.83	34.68

Notes: The table reports quantities for estimating differences by race in the consequences to taxpayers of being selected for an audit. All results are presented separately by type of audit (Columns 1-3), as well as aggregated based on the share of 2014 audits in each category (Column 4). The first six rows report summary audit statistics. The audit rate (row 1) is the percent of taxpayers that receive the type of audit in question. Taxpayer hours (row 3) and compliance costs (row 4) are taken from estimates reported in Guyton & Hodge (2014). Hours refers to the total hours the taxpayer spends responding to the audit. Compliance cost refers to the sum of out-of-pocket compliance costs and monetized hours spent responding to the audit (accounting for income differences in the cost of taxpayer time), inflation-adjusted from 2009 dollars to 2014 dollars. Penalties and interest (row 5) are net of IRS abatements. The bottom five rows report the estimated disparity in each outcome in rows two through six. The audit rate disparity is our main outcome of interest in the rest of the paper, here calculated using the probabilistic disparity estimator. For columns 1-3, the estimated disparity for other variables is calculated by multiplying the audit rate disparity by the average value of that outcome for the corresponding type of audit. For example, the estimated disparity in hours from a correspondence audit is the product of the correspondence audit disparity (0.80) and average taxpayer hours per correspondence audit (30). The disparity results in column 4 are obtained by taking the weighted average of the disparity results in columns 1-3, using the weights in row 2. These estimates assume that within-audit category, the components of audit burden do not vary by race. In addition, these estimates do not incorporate factors known to be important to taxpayers such as the non-financial stress of experiencing an audit, delays in receiving anticipated refunds, or reductions in refundable credits claims in future years. For context, taxpayers (whether audited or not) spend an average of 0.16 hours complying with audits

Table A.9: Audit Burden Disparity Estimates (Linear)

	Correspondence	Office	Field	Overall
Audit Rate (pp)	0.43	0.06	0.03	0.53
Share of Total Audits	0.82	0.12	0.06	1.00
Taxpayer Hours	30	38	34	31.18
Taxpayer Compliance Cost	643	1,717	4,431	1,002.98
Penalties and Interest	320.71	1,580.77	6,434.52	846.77
Assessed Taxes	5,252.56	7,130.46	24,960.56	6,694.85
Disparity (Audit Rate, pp)	1.33	0.04	-0.02	1.09
Disparity (Hours)	0.398	0.013	-0.006	0.329
Disparity (Compliance Cost)	8.54	0.60	-0.83	7.04
Disparity (Penalties and Interest)	4.26	0.55	-1.20	3.49
Disparity (Assessed Taxes)	69.77	2.50	-4.67	57.34

Notes: The table replicates Table A.8, but using the linear disparity estimator instead of the probabilistic disparity estimator to calculate the audit rate disparity.

Table A.10: Dual-Bootstrap Confidence Intervals

Estimator	Full Population	EITC	Non-EITC
	(1)	(2)	(3)
Linear	1.345	2.885	0.180
95% CI	(1.169, 1.461)	(2.633, 2.991)	(0.138, 0.230)
Probabilistic	0.813	1.950	0.102
95% CI	(0.729, 0.909)	(1.846, 2.031)	(0.080, 0.135)
N	148,305,318	28,338,472	119,966,846

Notes: The table reports the linear and probabilistic disparity estimates along with 95% confidence intervals obtained via our implementation of the Lu et al. (2024) dual-bootstrap procedure. To implement the procedure, we resampled first names and geographic information, but did not resample surnames (since the Census surname data we use corresponds to the full population rather than a sample) to generate 100 sets of BIFSG posteriors. We then compute our outcomes of interest with each set of BIFSG posteriors, resampling our taxpayer population on each iteration. Units are percentage points (0-100). Confidence intervals were obtained based on the distribution of estimates obtained by this procedure (see Appendix Figure A.12).

Table A.11: DDb vs. Non-DDb Audit Disparity Estimates Among EITC Claimants

Estimator	DDb Audit Disparity	Non-DDb Audit Disparity	Share of Disparity Attributable to DDb
	(1)	(2)	(3)
Linear	2.265	0.620	78.5
	(0.008)	(0.005)	
Probabilistic	1.531	0.419	78.5
	(0.007)	(0.004)	
Mean	1.071	0.372	
N	28,338,472	28,338,472	

Notes: Columns 1 and 2 report the estimated audit rate disparities for DDb and non-DDb audits among EITC claimants. Column 3 is calculated as the ratio of the DDb audit disparity in Column 1 to the total audit disparity (the sum of Columns 1 and 2), multiplied by 100. Units are percentage points (0-100). The "Mean" row is the share of the EITC claimant population that is selected for the specified category of audit. In our data, 74.31% of the audits of EITC claimants are selected through the DDb program.

Table A.12: Audit Disparity Robustness Checks

	BIFSG	Gibbs	Re-calibrated Unweighted	Re-calibrated Weighted	Geography- Only
	(1)	(2)	(3)	(4)	(5)
Panel A: Full Population					
Linear	1.27	1.55	1.623	1.543	1.738
	(0.005)	(0.04)	(0.005)	(0.004)	(0.005)
Probabilistic	0.811	1.02	0.774	0.735	0.696
	(0.004)	(0.03)	(0.003)	(0.003)	(0.003)
N	107,624,714	1,390,219	148,305,318	148,305,318	134,671,876
Panel B: EITC Population					
Linear	3.00	2.82	3.48	3.31	3.500
	(0.01)	(0.08)	(0.01)	(0.01)	(0.01)
Probabilistic	2.14	2.03	1.842	1.816	1.623
	(0.01)	(0.07)	(0.008)	(0.008)	(0.008)
N	19,357,514	283,055	28,338,472	28,338,472	25,768,606
Panel C: Non-EITC Population					
Linear	0.172	0.165	0.217	0.206	0.352
	(0.003)	(0.03)	(0.003)	(0.003)	(0.004)
Probabilistic	0.104	0.100	0.097	0.091	0.128
	(0.002)	(0.02)	(0.002)	(0.002)	(0.002)
N	88,267,200	1,107,164	119,966,846	119,966,846	108,903,270

Notes: The table shows the estimated audit rate disparity from the linear and probabilistic disparity estimators, under various modifications to our baseline approach. Units are percentage points (0-100). standard errors, reported in parentheses, are calculated from the asymptotic distributions described in Appendix B.3. Column 1 restricts the analysis to the subset of taxpayers for which each of first name, last name, and census block group are available. Column 2 predicts taxpayer race using the Gibbs sampling approach described in Appendix B.6. Column 3 predicts taxpayer race after re-calibrating the race probability estimates using the North Carolina data, as described in Appendix C. Column 4 replicates Column 3, but re-weights the data to be representative of the full population using the North Carolina weights described in Appendix C. Column 5 predicts taxpayer race using only taxpayers' geographic location. Panel A shows results for the full population; Panel B for the EITC population; and Panel C for the non-EITC population. Each displayed disparity estimate is in terms of percentage points and is statistically different from zero ($p < .01$).

Table A.13: Residual Covariance Estimates (Geography-Only)

	Full Population		EITC		Non-EITC	
E[cov(Y,B)|b]	7.52^{***}	8.51^{***}	22.58^{***}	24.87^{***}	1.98^{***}	1.64^{***}
	(0.23)	(0.38)	(0.94)	(1.41)	(0.16)	(0.28)
E[cov(Y,b)|B]	2.16^{***}	2.73^{***}	8.49^{***}	7.79^{***}	0.08	0.53^{**}
	(0.12)	(0.24)	(0.57)	(0.79)	(0.08)	(0.24)
Weighted	No	Yes	No	Yes	No	Yes
N	1,612,713		277,005		1,335,708	

Notes: The table displays the estimated covariance between audits and self-reported race, conditional on geography-only estimated race, as well as the estimated covariance between audits and estimated race, conditional on self- reported race. The estimates are for the matched sample of North Carolina taxpayers for the full population (columns 1 and 2), EITC claimants (columns 3 and 4), and non-EITC claimants (columns 5 and 6). Columns 2, 4, and 6 are re-weighted to be representative of the U.S. population, using the weights described in Appendix C. Standard errors are displayed in parentheses. The displayed estimates and standard errors are multiplied by 10^4. Stars correspond to p-values derived from two-sided hypothesis tests. $^* : P < .10;^{**} : P < .05;^{***} : P < .01$.

Table A.14: Audits of EITC Claimants by Presence of Business Income

	No Substantial Business Income	Substantial Business Income
Share of EITC Claimants	0.94	0.07
Share of Audits of EITC Claimants	0.96	0.04
Audit Rate (pp)	1.46	0.96
Disparity (Probabilistic, pp)	1.98	0.70
Disparity (Linear, pp)	2.93	1.15
Black Audit Rate (Probabilistic, pp)	3.02	1.58
Black Audit Rate (Linear, pp)	3.76	1.99
Non-Black Audit Rate (Probabilistic, pp)	1.04	0.89
Non-Black Audit Rate (Linear, pp)	0.84	0.84
N	26,330,090	1,853,037

Notes: The table reports descriptive statistics and estimated audit disparity for EITC claimants, based on whether they report substantial business income on their return. Substantial business income is defined based on the IRS's definitions for activity codes 270 and 271, which partition the set of EITC claimants with total positive income below \$200,000. Activity code 271 includes those taxpayers who report substantial business income (i.e., schedule C or schedule F gross receipts in excess of \$25,000); activity code 270 includes the remainder. A very small share of EITC claimants are classified into activity codes other than 270 or 271 because they report total positive income above \$200,000; we exclude them for purposes of this analysis. Disparity estimates are presented for both the probabilistic disparity estimator and linear disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. Audit rates by race present group-specific levels (rather than differences between groups) for Black and non-Black taxpayers filing income tax returns for tax year 2014.

Table A.15: Audit Cost and Revenue by Model

	Prediction Models		Oracles	
	Refundable Credit Overclaiming	Total Underreporting	Refundable Credit Overclaiming	Total Underreporting
Share with Substantial Business Income	0.028	0.927	0.091	0.655
Cost (Operational, $ Millions)	13	134	23	99
Cost (NRP, $ Millions)	113	408	138	324
Detected Underreporting ($ Millions)	2,299	4,109	3,976	8,855

Notes: The table reports information about the performance of the alternative audit selection algorithms considered in Figure 8 at the status quo (1.45%) audit rate for EITC returns. The first row refers to the share of selected returns that fall into activity code 271 (gross business receipts above $25,000). Cost refers to the estimated cost to the IRS of conducting the selected audits; it is calculated based on the share of selected returns in activity codes 270 and 271, and the average audit cost per return in those respective categories. The second row reports costs calculated from the average hours reported by IRS examiners dealing directly with the case multiplied by the applicable General Schedule payscale given the employee level. The third row instead calculates costs similarly but based on the number of hours reported by examiners in the much more intensive and comprehensive NRP exams. Costs are annualized to reflect our use of five years of NRP data. Detected underreporting refers to the total amount of underreporting (positive or negative) discovered on returns selected for audit under the specified audit selection algorithm, and is annualized to reflect our use of five years of NRP data.

B Additional Results Relating to Disparity Estimation

In this section, we provide additional theoretical results relating to our estimation of disparity from BIFSG-derived race probability estimates.

Contents

B.1 Results Relating to BIFSG Estimator

To derive Equation (1), use Bayes rule to write:

$$\begin{aligned}\Pr[B|F,S,G] &= \frac{\Pr[F,S,G|B]\ \Pr[B]}{\Pr[F,S,G]} \\ &= \frac{\Pr[F|B]\ \Pr[S|B]\ \Pr[G|B]\ \Pr[B]}{\Pr[F,S,G]}\end{aligned}$$

where the second equation follows from the "naive" conditional independence assumption underlying the approach. Equation (1) then follows by dividing $\Pr[B=1|F,S,G]$ by $\Pr[B=0|F,S,G]$, and using the fact that $\Pr[B=1|F,S,G]+\Pr[B=0|F,S,G]=1$.

In the Census data we use to estimate BIFSG scores, we observe $\Pr[B|S]$ rather than $\Pr[S|B]$, and we cannot back out $\Pr[B|S]$ due to censoring of uncommon surnames. Hence, the actual BIFSG scores we estimate are derived from

$$\Pr[B|F,S,G] = \frac{\Pr[F|B]\ \Pr[B|S]\ \Pr[G|B]\ \Pr[S]}{\Pr[F,S,G]}$$

Dividing $\Pr[B=1|F,S,G]$ by $\Pr[B=0|F,S,G]$ leads the (unobserved) $\Pr[S]$ terms to cancel, and following the same procedure as above we obtain:

$$\Pr[B=1|F,S,G] = \frac{\Pr[F|B=1]\ \Pr[B=1|S]\ \Pr[G|B=1]}{\sum_{j=0}^{1}\Pr[F|B=j]\ \Pr[B=j|S]\ \Pr[G|B=j]}$$

which we use to estimate taxpayer-level race probabilities.

B.2 Proof of Proposition 1

Recall Proposition 1:

Proposition 1. Suppose that b is a taxpayer's probability of being Black given some observable characteristics Z, so that $b = \Pr[B = 1|Z]$. Define D_p as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_p$, and D_l as the asymptotic limit of the linear disparity estimator, $\widehat{D}_l$. Then:

1.
$$D_p = D - \frac{\mathbb{E}[\text{Cov}(Y, B|b)]}{\text{Var}(B)} \tag{1.1}$$

2.
$$D_l = D + \frac{\mathbb{E}[\text{Cov}(Y, b|B)]}{\text{Var}(b)} \tag{1.2}$$

3. Suppose $\mathbb{E}[\text{Cov}(Y, B|b)] \geq 0$ and $\mathbb{E}[\text{Cov}(Y, b|B)] \geq 0$. Then
$$D_p \leq D \leq D_l \tag{1.3}$$

4. Suppose $\mathbb{E}[\text{Cov}(Y, B|b)] \leq 0$ and $\mathbb{E}[\text{Cov}(Y, b|B)] \leq 0$. Then
$$D_l \leq D \leq D_p \tag{1.4}$$

Proposition 1 follows from the more general Proposition 2, given below. Before stating and proving it, we state and prove a lemma showing that $D_p = D_l \cdot \frac{\text{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}$ (under the mild condition that b be almost surely nontrivial; in practice, observations for which b is 0 or 1, i.e. ground truth is available, can be analyzed separately).

Lemma 1. Suppose that $0 < b < 1$ almost surely, and that $\mathbb{E}|Y|$ is finite. Then as sample size grows, the probabilistic estimator converges almost surely to:

$$D_p = D_l \cdot \frac{\text{Var}(b)}{\mathbb{E}[b](1 - \mathbb{E}[b])}$$

Proof. We can write D_p as:

$$D_p = \frac{\sum_i b_i Y_i}{\sum_i b_i} - \frac{\sum_i (1 - b_i) Y_i}{\sum_i 1 - b_i} = \frac{\frac{1}{n}\sum_i b_i Y_i}{\frac{1}{n}\sum_i b_i} - \frac{\frac{1}{n}\sum_i (1 - b_i) Y_i}{\frac{1}{n}\sum_i (1 - b_i)}$$

For both the numerator and denominator, the strong law of large numbers holds (since $\mathbb{E}|Y|$ is finite and, since $0 < b < 1$, $\mathbb{E}|bY|$ also is also finite), so the numerator and denominator of each of the two terms converge almost surely to their expectations. Since $0 < b < 1$ almost surely, the continuous mapping theorem gives that the ratio of the terms converges to the ratio of their limits. That is:

$$\left[\frac{\frac{1}{n}\sum_i b_i Y_i}{\frac{1}{n}\sum_i b_i} - \frac{\frac{1}{n}\sum_i (1-b_i) Y_i}{\frac{1}{n}\sum_i (1-b_i)}\right] \xrightarrow[n\to\infty]{\text{a.s.}} \left[\frac{\mathbb{E}[bY]}{\mathbb{E}[b]} - \frac{\mathbb{E}[(1-b)Y]}{\mathbb{E}[1-b]}\right]$$

Now, simply combining fractions, we note:

$$\begin{aligned}\frac{\mathbb{E}[bY]}{\mathbb{E}[b]} - \frac{\mathbb{E}[(1-b)Y]}{\mathbb{E}[1-b]} &= \frac{\mathbb{E}[bY] - \mathbb{E}[b]\mathbb{E}[bY] - \mathbb{E}[b]\mathbb{E}[Y] + \mathbb{E}[b]\mathbb{E}[bY]}{\mathbb{E}[b](1-\mathbb{E}[b])} \\ &= \frac{\mathbb{E}[bY] - \mathbb{E}[b]\mathbb{E}[Y]}{\mathbb{E}[b](1-\mathbb{E}[b])} \\ &= \frac{\text{Cov}(Y,b)}{\mathbb{E}[b](1-\mathbb{E}[b])}\end{aligned}$$

Finally, we recall that $D_l = \frac{\text{Cov}(Y,b)}{\text{Var}(b)}$ by construction; substituting in $\text{Cov}(Y,b) = D_l\text{Var}(b)$ yields the result. □

Proposition 2. Suppose that b is a (potentially imperfectly calibrated) estimate of the probability that a taxpayer is Black, based on some observable characteristics Z. Let $\varepsilon = B - b$ denote the error in a taxpayer's predicted race. Define D_p as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_p$, and D_l as the asymptotic limit of the linear disparity estimator, $\widehat{D}_l$. Define $\mu = \text{Cov}(\mathbb{E}[\eta|b], \mathbb{E}[\varepsilon|b])$, where η denotes the residual from the linear projection of Y on b.

Then:

1.
$$D_l = D\left(1 + \frac{\text{Cov}(b,\varepsilon)}{\text{Var}(b)}\right) + \frac{\mathbb{E}[\text{Cov}(Y,b|B)]}{\text{Var}(b)}$$

2.
$$D_p = \frac{D\cdot\text{Var}(B) - D_l\cdot\text{Cov}(b,\varepsilon)}{\mathbb{E}[b](1-\mathbb{E}[b])} - \frac{\mathbb{E}[\text{Cov}(Y,B|b)] + \mu}{\mathbb{E}[b](1-\mathbb{E}[b])}$$

Proof of Proposition 2. Consider the linear projections of Y on b and of Y on B:

$$Y = \alpha + \beta b + \eta$$

$$Y = \alpha' + \gamma B + \nu$$

By construction, $\text{Cov}(b,\eta) = \text{Cov}(B,\nu) = 0$. In addition, $E[\nu] = 0$, so

$$\gamma = E[Y|B=1] - E[Y|B=0] = D$$

Also, by construction:

$$\gamma \,\mathrm{Var}(B) = \mathrm{Cov}(Y, B)$$

and similarly,

$$\beta \,\mathrm{Var}(b) = \mathrm{Cov}(Y, b)$$

Using the law of total covariance, we can write:

$$\mathrm{Cov}(Y, b) = E[\mathrm{Cov}(Y, b|B)] + \mathrm{Cov}(E[Y|B], E[b|B])$$

The latter term can be expanded as:

$$\begin{aligned}\mathrm{Cov}(E[Y|B], E[b|B]) &= \mathrm{Cov}(E[\alpha' + \gamma B + \nu|B], E[B - \varepsilon|B]) \\ &= \mathrm{Cov}(\alpha' + \gamma B + E[\nu|B], B - E[\varepsilon|B]) \\ &= \gamma\mathrm{Var}(B) - \gamma\mathrm{Cov}(B, E[\varepsilon|B]) + \mathrm{Cov}(E[\nu|B], B) - \mathrm{Cov}(E[\nu|B], E[\varepsilon|B]) \\ &= \gamma\mathrm{Var}(B) - \gamma\mathrm{Cov}(B, E[\varepsilon|B])\end{aligned}$$

where the last equality follows from the fact that since B is binary, $\mathrm{Cov}(B, \nu) = 0 \implies E[\nu|B] = 0$ for all B.

Next, note that

$$\begin{aligned}\mathrm{Cov}(B, E[\varepsilon|B]) &= E[B\, E[\varepsilon|B]] - E[B]\, E[E[\varepsilon|B]] \\ &= E[E[B\,\varepsilon|B]] - E[B]\, E[E[\varepsilon|B]] \\ &= E[B\,\varepsilon] - E[B]\, E[\varepsilon] \\ &= \mathrm{Cov}(B, \varepsilon) \\ &= \mathrm{Cov}(b + \varepsilon, \varepsilon) \\ &= \mathrm{Cov}(b, \varepsilon) + \mathrm{Var}(\varepsilon)\end{aligned}$$

Combining these results, we have:

$$\begin{aligned}\beta\,\mathrm{Var}(b) &= \mathrm{Cov}(Y, b) \\ &= E[\mathrm{Cov}(Y, b|B)] + \mathrm{Cov}(E[Y|B], E[b|B]) \\ &= E[\mathrm{Cov}(Y, b|B)] + \gamma\,\mathrm{Var}(B) - \gamma\,\mathrm{Var}(\varepsilon) - \gamma\mathrm{Cov}(b, \varepsilon)\end{aligned}$$

From the definition of ε, we have:

$$\mathrm{Var}(B) = \mathrm{Var}(b) + \mathrm{Var}(\varepsilon) + 2\mathrm{Cov}(b, \varepsilon) \implies \mathrm{Var}(B) - \mathrm{Cov}(b, \varepsilon) - \mathrm{Var}(\varepsilon) = \mathrm{Var}(b) + \mathrm{Cov}(b, \varepsilon)$$

Thus

$$\beta\,\mathrm{Var}(b) = \gamma[\mathrm{Var}(b) + \mathrm{Cov}(b, \varepsilon)] + \mathbb{E}[\mathrm{Cov}(Y, b|B)]$$

and dividing through by $\mathrm{Var}(b)$ yields part 1 of the proposition.

To prove part 2 of the proposition, again use the law of total covariance:

$$\mathrm{Cov}(Y, B) = E[\mathrm{Cov}(Y, B|b)] + \mathrm{Cov}\left(E[Y|b]\,,\ E[B|b]\right)$$

Expanding the second term of the right-hand side of the equation, we have

$$\begin{aligned}\mathrm{Cov}\left(E[Y|b]\,,\ E[B|b]\right) &= \mathrm{Cov}\left(E[\alpha + \beta\, b + \eta|b]\,,\ E[b + \varepsilon|b]\right)\\ &= \mathrm{Cov}\left(\alpha + \beta\, b + E[\eta|b]\,,\ b + E[\varepsilon|b]\right)\\ &= \beta\,\mathrm{Var}(b) + \beta\mathrm{Cov}(b, E[\varepsilon|b]) + \mathrm{Cov}(E[\eta|b], b) + \mathrm{Cov}(E[\eta|b]\,,\ E[\varepsilon|b])\end{aligned}$$

Note that:

$$\begin{aligned}\mathrm{Cov}(b, E[\varepsilon|b]) &= E[b\, E[\varepsilon|b]] - E[b]E[E[\varepsilon|b]]\\ &= E[E[b\,\varepsilon|b]] - E[b]E[E[\varepsilon|b]]\\ &= E[b\,\varepsilon] - E[b]E[\varepsilon]\\ &= \mathrm{Cov}(b, \varepsilon)\end{aligned}$$

By the same logic:

$$\mathrm{Cov}(E[\eta|b], b) = \mathrm{Cov}(\eta, b) = 0$$

Define $\mu := \mathrm{Cov}(E[\eta|b]\,,\ E[\varepsilon|b])$. Then collecting results, we have

$$\begin{aligned}\gamma\,\mathrm{Var}(B) &= \mathrm{Cov}(Y, B)\\ &= E[\mathrm{Cov}(Y, B|b)] + \mathrm{Cov}\left(E[Y|b]\,,\ E[B|b]\right)\\ &= E[\mathrm{Cov}(Y, B|b)] + \beta\,\mathrm{Var}(b) + \beta\mathrm{Cov}(b, \varepsilon) + \mu\end{aligned}$$

Rearranging, and recalling that $D_p = \beta\frac{\mathrm{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}$ from Lemma 1 yields the result.

□

Now, we prove Proposition 1 as a consequence of Proposition 2.

Proof of Proposition 1. If $b = \Pr[B = 1|Z] = \mathbb{E}[B|Z]$, it follows from the definition of ε that

$$\begin{aligned}E[\varepsilon|Z] &= E[B|Z] - E[b|Z]\\ &= E[B|Z] - E[E[B|Z]|Z]\\ &= E[B|Z] - E[B|Z]\\ &= 0\end{aligned}$$

Hence, we can write

$$\begin{aligned}\mathrm{Cov}(b,\varepsilon) &= E[b\,\varepsilon] - E[b]\,E[\varepsilon]\\ &= E[b\,\varepsilon]\\ &= E[E[b\,\varepsilon|Z]]\\ &= E[b\,E[\varepsilon|Z]]\\ &= E[b\,0]\\ &= 0,\end{aligned}$$

where the third equality follows from the law of iterated expectations, and the fourth from the fact that b is a function of Z.

Substituting the fact that $\mathrm{Cov}(b,\varepsilon) = 0$ into Proposition 2.1, and noting that since $\mathbb{E}[b] = \mathbb{E}[\mathbb{E}[B|Z]] = \mathbb{E}[B]$,

$$\mathbb{E}[b](1-\mathbb{E}[b]) = \mathbb{E}[B](1-\mathbb{E}[B]) = \mathrm{Var}(B).$$

yields Proposition 1.2.

Proposition 1.1 follows by again substituting in $\mathrm{Cov}(b,\varepsilon) = 0$ and noting that $\mathbb{E}[\varepsilon|b] = 0$ because:

$$\begin{aligned}E[\varepsilon|b] &= E[E[\varepsilon|b,Z]|b]\\ &= E[E[\varepsilon|Z]|b]\\ &= E[0|b]\\ &= 0,\end{aligned}$$

where the second equality follows from the fact that b is a function of Z.

Finally, once the forms of D_l and D_p are established, Proposition 1.3 and 1.4 follow directly when the respective assumptions on the signs of $\mathbb{E}[\mathrm{Cov}(Y,b|B)]$ and $\mathbb{E}[\mathrm{Cov}(Y,B|b)]$ are met. □

The following Proposition extends Proposition 1 to the case in which the estimand of interest is the level of the outcome by group, rather than the difference in the levels of the outcome across groups. In particular, Proposition 3 characterizes the bias of the linear and probabilistic approaches for estimating the audit rate by race (not the audit rate disparity).

Proposition 3. (Statistical Bias of Level Estimators). Suppose $b = \Pr[B|Z]$. Consider the following estimators:

$$\begin{aligned}\widehat{Y}_p^B &:= \frac{\sum b_i Y_i}{\sum b_i} &\quad\text{and}\quad&& \widehat{Y}_p^{NB} &:= \frac{\sum(1-b_i)Y_i}{\sum(1-b_i)}\\ \widehat{Y}_l^B &:= \widehat{\alpha}+\widehat{\beta} &\quad\text{and}\quad&& \widehat{Y}_l^{NB} &:= \widehat{\alpha},\end{aligned}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the intercept and slope, respectively, from the regression of Y on b. Let Y_p^B,Y_p^{NB}, Y_l^B, Y_l^{NB} be the respective limits the estimators described above converge to. Then:

1. Y_l^B and Y_l^{NB} have the following biases relative to the true audit rates Y^B and Y^{NB}:

$$Y_l^B - Y^B = \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{E(B)} \qquad \text{and} \qquad Y_l^{NB} - Y^{NB} = -\frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{1 - E(B)}$$

2. Y_p^B and Y_p^{NB} have the following biases relative to the true audit rates Y^B and Y^{NB}:

$$Y_p^B - Y^B = -\frac{\mathbb{E}[\mathrm{Cov}(Y, B|b)]}{\mathbb{E}[B]} \qquad \text{and} \qquad Y_p^{NB} - Y^{NB} = \frac{\mathbb{E}[\mathrm{Cov}(Y, B|b)]}{1 - \mathbb{E}[B]}$$

3. Suppose $\mathbb{E}[\mathrm{Cov}(Y, b|B)] = 0$. Then:

$$Y_l^B = Y^B \qquad \text{and} \qquad Y_l^{NB} = Y^{NB}$$

4. Suppose $\mathbb{E}[\mathrm{Cov}(Y, B|b)] = 0$. Then:

$$Y_p^B = Y^B \qquad \text{and} \qquad Y_p^{NB} = Y^{NB}$$

Proof. Notice that 3) and 4) follow directly from 1) and 2). For 1): By construction, we have

$$Y = \alpha + \gamma B + \nu$$

From this, we know $Y^{NB} = \alpha$ and $Y^B = \alpha + \gamma$.
Taking expectations, and rearranging:

$$Y^{NB} = \alpha = E[Y] - \gamma\, E[B]$$

In contrast, our sample estimate of Y^{NB} from the linear estimator, $\widehat{Y}_l^{NB}$, is given by:

$$\widehat{Y}_l^{NB} = \widehat{\alpha} = \overline{Y} - \widehat{\beta}\, \overline{b}$$

which converges to

$$Y_l^{NB} = E[Y] - D_l\, E[b] = \mathbb{E}[Y] - D_l\, \mathbb{E}[B],$$

since $\mathbb{E}[b] = \mathbb{E}[B]$ (because $\mathbb{E}[b] = \mathbb{E}_Z[\Pr[B = 1|Z]] = \mathbb{E}[B]$).
From Proposition 1, we know $D_l = \gamma + \frac{\mathbb{E}[\mathrm{Cov}(Y,b|B)]}{\mathrm{Var}(B)}$
Substituting this into the above, we have:

$$\begin{aligned} Y_l^{NB} &= E[Y] - \gamma\, E[B] - E[B]\, \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{\mathrm{Var}(B)} \\ &= Y^{NB} - \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{1 - E(B)} \end{aligned}$$

Turning to Y_l^B, we have

$$\widehat{Y}_l^B = \widehat{\alpha} + \widehat{\beta}$$

which converges to

$$\begin{aligned}
Y_l^B &= Y_l^{NB} + D_l \\
&= \left(\alpha - E[B]\,\frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{\mathrm{Var}(B)}\right) + D_l \\
&= \alpha - E[B]\,\frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{\mathrm{Var}(B)} + \gamma + \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{\mathrm{Var}(B)} \\
&= \alpha + \gamma + (1 - E[B])\,\frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{\mathrm{Var}(B)} \\
&= Y^B + \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B)]}{E(B)}
\end{aligned}$$

We prove 2) in a very similar manner as the related statement is in Chen et al. (2019): Note that:

$$Y^B = \mathbb{E}[Y|B = 1] = \frac{\mathbb{E}[YB]}{\mathbb{E}[B]} = \frac{\mathbb{E}[\mathbb{E}[YB|b]]}{\mathbb{E}[B]}$$

On the other hand,

$$\widehat{Y}_p^B = \frac{\frac{1}{n}\sum b_i Y_i}{\frac{1}{n}\sum b_i} \longrightarrow \frac{\mathbb{E}[Yb]}{\mathbb{E}[b]} := Y_p^B$$

since the law of large numbers applies to the numerator and the denominator separately (and the boundedness away from the end of the interval guarantees that the limits of the ratio converges to the ratio of the limits).

But $\mathbb{E}[b] = \mathbb{E}_Z[\Pr[B = 1|Z]] = \mathbb{E}[B]$, and $\mathbb{E}[Yb] = \mathbb{E}[\mathbb{E}[Yb|b]] = \mathbb{E}[b\mathbb{E}[Y|b]] = \mathbb{E}[\mathbb{E}[B|b]\mathbb{E}[Y|b]]$, so:

$$Y_p^B - Y^B = \frac{\mathbb{E}\left[\mathbb{E}[Y|b]\mathbb{E}[B|b]\right]}{\mathbb{E}[B]} - \frac{\mathbb{E}[\mathbb{E}[YB|b]]}{\mathbb{E}[B]} = -\frac{\mathbb{E}[\mathrm{Cov}(Y, B|b)]}{\mathbb{E}[B]}$$

where the second equality follows from the definition of conditional covariance. This establishes the result for Y_p^B. To see the analogous result for Y_p^{NB}, let $A = 1 - B$ and $a = b$, and observe that $-\mathbb{E}[\mathrm{Cov}(Y, A|a)] = \mathbb{E}[\mathrm{Cov}(Y, B|b)]$. The result then follows in the same manner as above.

□

B.3 Inference in Finite Samples

This section characterizes the asymptotic distributions of the linear and probabilistic estimators.

B.3.1 Standard Errors of Disparity Estimators

Call $\widehat{D}_l^n$ and $\widehat{D}_p^n$ the empirically-constructed linear and probabilistic estimators using a sample size of n observations. (D_l and D_p as written above are what $\widehat{D}_l^n$ and $\widehat{D}_p^n$ converge to as $n \to \infty$.)

Lemma 2. For any fixed dataset, we relate $\widehat{D}_p^n$ and $\widehat{D}_l^n$ as:

$$\widehat{D}_p^n = \widehat{D}_l^n \cdot \frac{\frac{1}{n}\sum_i b_i^2 - \bar{b}^2}{\bar{b}(1-\bar{b})}$$

And asymptotically,

$$\widehat{D}_p^n \to \widehat{D}_l^n \cdot \frac{\mathrm{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}.$$

Proof. Notice that:

$$\begin{aligned}\widehat{D}_p^n = \frac{\sum b_i Y_i}{\sum b_i} - \frac{\sum (1-b_i)Y_i}{\sum 1-b_i} &= \frac{\frac{1}{n}\sum b_i Y_i}{\frac{1}{n}\sum b_i} - \frac{\frac{1}{n}\sum(1-b_i)Y_i}{\frac{1}{n}\sum(1-b_i)} \\ &= \frac{\frac{1}{n}\sum b_i Y_i}{\bar{b}} - \frac{\frac{1}{n}\sum(1-b_i)Y_i}{1-\bar{b}}\end{aligned}$$

where we use $\bar{\cdot}$ to indicate the sample average. We can then write:

$$\begin{aligned}\frac{\frac{1}{n}\sum b_i Y_i}{\frac{1}{n}\sum b_i} - \frac{\frac{1}{n}\sum(1-b_i)Y_i}{\frac{1}{n}\sum(1-b_i)} - \frac{\frac{1}{n}\sum b_i Y_i}{\bar{b}} - \frac{\frac{1}{n}\sum(1-b_i)Y_i}{1-\bar{b}} \\ = \frac{\frac{1}{n}\sum b_i Y_i - \frac{\bar{b}}{n}\sum b_i Y_i - \frac{\bar{b}}{n}\sum Y_i + \frac{\bar{b}}{n}\sum b_i Y_i}{\bar{b}(1-\bar{b})} \\ = \frac{\frac{1}{n}\sum b_i Y_i - \bar{b}\bar{y}}{\bar{b}(1-\bar{b})}\end{aligned}$$

Now consider the regression estimator. By definition:

$$D_l^n = \frac{\sum(b_i-\bar{b})(y_i-\bar{y})}{\sum(b_i-\bar{b})^2} = \frac{\sum b_i y_i - \bar{b}\sum y_i - \bar{y}\sum b_i + n\bar{b}\bar{y}}{\sum(b_i-\bar{b})^2} = \frac{\frac{1}{n}\sum b_i Y_i - \bar{b}\bar{y}}{\frac{1}{n}\sum(b_i-\bar{b})^2}$$

But notice the numerator in both terms are the same. That is:

$$\widehat{D}_p^n = \frac{\frac{1}{n}\sum(b_i-\bar{b})^2}{\bar{b}(1-\bar{b})}\widehat{D}_l^n = C_n\widehat{D}_l^n$$

where $C_n = \frac{\frac{1}{n}\sum(b_i-\bar{b})^2}{\bar{b}(1-\bar{b})}$.

But now recall Slutsky's theorem, which says that if A_n, B_n are random variables and $B_n \to c$ for some constant c, then $A_n B_n \to A_n c$. In particular,

$$C_n \longrightarrow \frac{\mathrm{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}.$$

The second half of the lemma follows. □

The asymptotic distribution of $\widehat{D}_l^n$ is well understood, as it is the OLS estimator.

Given the relationship between $\widehat{D}_p^n$ and $\widehat{D}_l^n$ shown above, it is mechanically true that $\widehat{D}_p^n$ will, under the same conditions, be distributed normally as well. Formally:

Proposition 4. The asymptotic distribution of D_n^p is given by:

$$\frac{\widehat{D}_p^n - D_p}{\sqrt{V_l^n \frac{\mathrm{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}}} \longrightarrow \mathcal{N}(0,1)$$

where $D^p = \frac{\mathrm{Cov}(Y,b)}{\mathbb{E}[b](1-\mathbb{E}[b])}$ and V_l^n is the variance of $\widehat{D}_l^n$.

B.4 Incorporating Sampling Weights into Disparity Estimation

In some of our analyses, we use data which is re-weighted to be representative of the full population of U.S. taxpayers. $\widehat{D}^l$ can be naturally extended to incorporate sample weights via weighted regression. How to extend the probabilistic estimator, however, may be less obvious. We propose the following as the weighted probabilistic estimator $\widehat{D}_{p,w}$:

$$\widehat{Y}_{p,w}^B = \frac{\sum_i \omega_i b_i Y_i}{\sum_i \omega_i b_i}, \qquad \widehat{Y}_p^{NB} = \frac{\sum_i \omega_i (1-b_i) Y_i}{\sum_i \omega_i (1-b_i)}, \qquad \widehat{D}_{p,w} := \widehat{Y}_{p,w}^B - \widehat{Y}_{p,w}^{NB}$$

where ω_i is a sample weight for observation i. (Notice that as with D_p, replacing Y_i with any other random variable gives an estimator for disparity in said random variable.) This estimator is closely related to the Horwitz-Thompson family of estimators; see Robinson (1982); Berger (1998); Delevoye & Sävje (2020) for prior results regarding convergence and consistency.

What is the purpose of the weighted estimator? The intention behind it is to use the data we have to estimate what D_p *would* be given a different dataset or distribution. If $\widehat{D}_{p,w}$ provides this fidelity, then it is the 'correct' weighted analogue. We show here that $\widehat{D}_{p,w}$ is the 'correct' weighted analogue in a sense we make formal below.

We will distinguish between two cases. In the first, we have access to a subset of a finite population of individuals, and are given *replicate weights*. The replicate weight for observation i corresponds to the number of individuals in the full population that i represents. In other words, we have some dataset of observations $\mathscr{D} := \{X_i\}_{i=1}^n$, but the full dataset which we do not have access to has observations $\mathscr{D}' := \bigcup_i \{X_i\}_{j=1}^{w_i}$. The hope is that $\widehat{D}_{p,w}$ estimated on $\mathscr{D}$ corresponds to $\widehat{D}_p$ estimated on $\mathscr{D}'$.

Proposition 5. Suppose we are in the case of replicate weights and $\mathscr{D}$ and $\mathscr{D}'$ are as above. Let $\widehat{D}_{p,w}|\mathscr{D}$ be estimated over $\mathscr{D}$ and $\widehat{D}_p|\mathscr{D}'$ be what would be estimated over $\mathscr{D}'$. Then:

$$\widehat{D}_{p,w}|\mathscr{D} = \widehat{D}_p|\mathscr{D}'$$

Proof. This follows simply from the linearity of the numerator and denominator of $\widehat{Y}^B_{p,w}$ and $\widehat{Y}^{NB}_{p,w}$. Take $\widehat{Y}^B_{p,w}$:

$$\widehat{Y}^B_{p,w}|\mathscr{D} = \frac{\sum_i w_i b_i Y_i}{\sum_i w_i b_i} = \frac{\sum_i \sum_{j=1}^{w_i} b_i Y_i}{\sum_i \sum_{j=1}^{w_i} b_i} = \widehat{Y}^B_p|\mathscr{D}'.$$

$\widehat{Y}^{NB}_{p,w}$ follows similarly and thus too $\widehat{D}_{p,w}$. □

Notice that the case of replicate weights corresponds to our analyses in which we use NRP to estimate quantities over the population.

The second case is more general: weights may be not be integers corresponding the number of people represented in some larger dataset, but rather changes of measure intended to capture some other distribution. (For instance, weighting for non-response attempts to map the data from responders to the overall population.) In this setting, we are agnostic to how the weights are generated; instead, we merely assume that they successfully accomplish re-weighting at the level of the sample mean. We make this precise in the following proposition:

Proposition 6. Suppose we have data drawn from a distribution $\mathcal{D}$; this data includes both a quantity of interest, Y_i, as well as sample weights ω_i that map $\mathcal{D}$ to some other distribution $\mathcal{D}'$ in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n} \omega_i Q_i \overset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{D}'}[Q]],$$

for any random variable Q. Then:

$$\widehat{D}^n_{p,w}|\mathcal{D} \overset{n\to\infty}{\longrightarrow} D_p|\mathcal{D}'.$$

Proof. Consider $\widehat{Y}^B_{p,w}$. Let $Q := bY$. Then by assumption:

$$\frac{1}{n}\left[\sum_{i=1}^{n} \omega_i b_i Y_i\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{D}'}[bY]$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} \omega_i b_i \overset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{D}'}[b]$$

But we have that:

$$\widehat{Y}_{p,w}^{B,n} = \frac{\sum_i \omega_i b_i Y_i}{\sum_i \omega_i b_i} = \frac{\frac{1}{n}\sum_i \omega_i b_i Y_i}{\frac{1}{n}\sum_i \omega_i b_i} \overset{n\to\infty}{\longrightarrow} \frac{\mathbb{E}_{\mathcal{D}'}[bY]}{\mathbb{E}_{\mathcal{D}'}[b]}$$

Proceeding similarly with $\widehat{Y}_{p,w}^{NB,n}$ and taking the difference, we obtain:

$$\widehat{D}_{p,w}^{n} \overset{n\to\infty}{\longrightarrow} \frac{\mathbb{E}_{\mathcal{D}'}[bY]}{\mathbb{E}_{\mathcal{D}'}[b]} - \frac{\mathbb{E}_{\mathcal{D}'}[(1-b)Y]}{\mathbb{E}_{\mathcal{D}'}[1-b]} = D_p|\mathcal{D}'$$

□

Notice that the choice of unit weights, i.e. $\omega_i = 1$ satisfies the assumption of the theorem and recovers the original convergence results. For another example, suppose we have groups A and B in equal number throughout the population, but in our data we obtain twice as many observations from group B as group A. Then it is easy to verify that the choice of weights $\omega_i = \begin{cases} 2/3 & i \in A \\ 3/4 & i \in B \end{cases}$ would satisfy the assumptions, and thus this choice of weights would allow us to recover D_p in the population from our data.

B.5 Estimating Disparity from a Recalibrated Race Probability Estimate

In general, our estimate of $\Pr[B_i = 1|Z_i]$ may be drawn from a population that differs from the population of interest; for example, among EITC claimants, $\Pr[B_i = 1|Z_i]$ may be different than the equivalent quantity for the population as a whole. (This is reflected in the imperfect calibration visible in Figure 2.) Proposition 2 shows that systematic deviation like this can bias our estimates. However, given access to a *recalibrated* b^*, (i.e. the linear projection of B on to the space of b and a constant, based on a subset of data with ground truth race labels), we can use this re-calibrated proxy to obtain similar results as those in Proposition 1 by applying Proposition 2.

To see this, suppose access to such a b^* and note that by construction, $\mathrm{Cov}(b^*, \varepsilon^*) = 0$, so Proposition 2 applies. Moreover, $\mathbb{E}[b^*] = \mathbb{E}[B]$, so $\mathbb{E}[b^*](1 - \mathbb{E}[b^*]) = \mathbb{E}[B](1 - \mathbb{E}[B]) = \mathrm{Var}(B)$. So designating D_l^* and D_p^* as the linear and probabilistic estimators, respectively, applied to b^*, as well as η^* and ε^* for the analogues of η and ε, Proposition 2 indicates that:

$$D_l^* = D + \frac{\mathbb{E}[\mathrm{Cov}(Y, b^*|B)]}{\sigma_{b^*}^2} \tag{2}$$

and

$$D_p^* = D - \frac{\mathbb{E}[\mathrm{Cov}(Y, B|b^*)] + \mathrm{Cov}(\mathbb{E}[\eta^*|b^*], \mathbb{E}[\varepsilon^*|b^*]}{\mathrm{Var}(B)}. \tag{3}$$

These equations are similar to those of Proposition 1, but there are two potential challenges to overcome before applying the recalibrated proxies in the same manner as our initial probability estimates. The first potential challenge is that it might be more difficult to reason about the sign of the covariances between outcome and re-calibrated proxy than the original proxy, since this recalibrated proxy could differ in its conditional relationship to Y given B and vice versa.

The following Lemma addresses this issue; it shows that the main covariance terms $\mathbb{E}[\mathrm{Cov}(Y, B|b^*)]$ and $\mathbb{E}[\mathrm{Cov}(Y, b^*|B)]$ will have the same signs as their non-recalibrated counterparts, under the minor condition that $\mathrm{Cov}(B, b) > 0$.

Lemma 3 (). Suppose that b is a (possibly mis-calibrated) estimate of the probability that a taxpayer is Black based on some observable characteristics Z and b^* is the re-calibrated proxy which can be written as an orthogonal projection:

$$B = \mu + \rho b + \varphi,$$

i.e.

$$b^*(b) = \mu + \rho b.$$

Suppose further that $\mathrm{Cov}(B, b) > 0$. Then

$$\mathrm{sign}(\mathbb{E}[\mathrm{Cov}(Y, B|b)]) = \mathrm{sign}(\mathbb{E}[\mathrm{Cov}(Y, B|b^*)])$$
$$\mathrm{sign}(\mathbb{E}[\mathrm{Cov}(Y, b|B)]) = \mathrm{sign}(\mathbb{E}[\mathrm{Cov}(Y, b^*|B)])$$

Proof. We note that

$$\mathrm{Cov}(Y, b^*|B) = \mathrm{Cov}(Y, \mu + \rho b|B) = \rho \mathrm{Cov}(Y, b|B)$$

and

$$\mathrm{Cov}(Y, B|b^*) = \mathrm{Cov}(Y, B|b).$$

Then the signs of $\mathbb{E}[\mathrm{Cov}(Y, B|b)]$ and $\mathbb{E}[\mathrm{Cov}(Y, B|b^*)]$ are identical, while the signs of $\mathbb{E}[\mathrm{Cov}(Y, b|B)]$ and $\mathbb{E}[\mathrm{Cov}(Y, b^*|B)]$ will agree if $\rho \geq 0$. Since ρ is the coefficient on b in said regression, it is given by $\mathrm{Cov}(B, b)/\mathrm{Var}(b)$, which is positive if and only if $\mathrm{Cov}(B, b) > 0$.

□

The second potential difficulty in applying the recalibrated proxy is that the additional covariance term that appears in the expression for D_p's asymptotic bias, $\mathrm{Cov}(\mathbb{E}[\varepsilon^*|b^*], \mathbb{E}[\eta^*|b^*])$, may also be difficult to reason about on the basis of theory. To interpret this term, note that when $\mathbb{E}[B|b]$ is a linear function of b, then (asymptotically) recalibrating via linear regression will recover the true conditional expectation function. By construction in that case, $\mathbb{E}[\varepsilon^*|b^*]$ will be 0, since b^* *is* the CEF. Thus, under linearity, the nuisance covariance term would be 0. Of course, we do not expect *exact* linearity to hold, but we will see in the recalibration exercise below that the CEF is close to linear, so that our estimate of $\mathrm{Cov}(E[\eta^*|b^*], E[\varepsilon^*|b^*])$ is close to 0 and certainly negligible compared to the other terms.

The upshot of the results in this subsection is that *if* we expect the covariance conditions to hold with our original proxy, we can also expect them to hold for our recalibrated proxy. This in turn allows us to treat disparity estimates arrived at using the recalibrated proxies in the same manner as estimates obtained using the original proxies. We now turn to our empirical approach.

Empirical Approach We now describe we apply the aforementioned strategy to re-calibrate the BIFSG-predicted probability Black in the North Carolina dataset. We consider North Carolina as a whole as well as EITC and non-EITC specific approaches.

First, we calculate $\widehat{\rho}$ as the coefficient from regressing an indicator for whether a taxpayer self reports as Black on the BIFSG-predicted probability that a taxpayer is Black. That is, we run the regression:

$$B = \alpha_0 + \rho b + \varphi,$$

with $\widehat{\rho}$ estimated once via ordinary least squares and separately via weighted least squares using the North Carolina weights. We also repeat both estimations separately for EITC taxpayers and non-EITC taxpayers. (The additional weighted/non-weighted and EITC/non-EITC calculations will be repeated throughout; where required, weighted estimates will be computed using the estimators described in B.4 above.) These estimates are reported in the first line of Table B.1.

Next, we assign each individual

$$b_i^* := \widehat{\alpha}_0 + \widehat{\rho} b_i,$$

and

$$\varepsilon_i^* = B_i - b_i^*;$$

we then estimate $\widehat{\text{Cov}}(b, \varepsilon^*)$ in the straightforward manner of using sample unweighted and weighted averages and product of b and ε^*, again separating out by EITC status. These estimates are reported in the second line of Table B.1.

The next four lines of Table B.1 are computed in a similar manner. That is, we compute the covariance within a given realization of the conditioning variable (e.g. for the set of non-Black taxpayers, $B = 0$) and then weight these estimates by the estimated share of taxpayers they represent. Importantly, we discretize both b_i^* and b_i by rounding to the nearest percentage point in order to create realizations to average over; this approach may introduce some arbitrariness to the analysis, but avoids making parametric assumptions.

Next, we run the regression:

$$Y = \alpha^* + \beta^* b^* + \eta^*$$

and interpret the estimated $\widehat{\beta}^*$; that is, $\widehat{\beta}^*$ is the linear estimator of disparity as applied to the re-calibrated b^*. We then obtain

$$\widehat{\eta}_i^* = Y_i - \widehat{\alpha}^* - \widehat{\beta}^* b_i^*.$$

We use this to compute the next line of Table B.1 in the following manner: first, we estimate $\mathbb{E}[\eta^*|b^*]$ and $\mathbb{E}[\varepsilon^*|b^*]$ by computing the sample averages of η^* and ε^* within each discretized b^* category. We then assign each individual their respective sample averages based on their value of b_i^*, and then compute the overall covariance estimate over the population using these features.

The next three lines of Table B.1 are computed straightforwardly - i.e. $\widehat{D}$ is based on the ground truth,while $\widehat{D}_l^*$ and $\widehat{D}_p^*$ are computed according to the formulas in equations 2 and 3 above and the appropriate values from previously computed rows of Table B.1.

Table B.1: Estimates from the Re-calibration Exercise

Value	Overall		EITC		Non-EITC	
	NC	Reweighted NC	NC	Reweighted NC	NC	Reweighted NC
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{\rho}$	0.828	0.872	0.923	0.964	0.767	0.802
$\widehat{\text{Cov}}(b, \varepsilon)$	-0.000	-0.000	-0.000	-0.000	0.000	-0.000
$\widehat{E}[\text{Cov}(Y, b\|B)]$	0.000	0.000	0.001	0.001	0.000	-0.000
$\widehat{E}[\text{Cov}(Y, B\|b)]$	0.001	0.001	0.001	0.001	0.000	0.000
$\widehat{E}[\text{Cov}(Y, b^*\|B)]$	0.000	0.000	0.001	0.001	0.000	-0.000
$\widehat{E}[\text{Cov}(Y, B\|b^*)]$	0.001	0.001	0.001	0.001	0.000	0.000
$\widehat{\text{Cov}}(\widehat{\mathbb{E}}[\eta^*\|b^*], \widehat{\mathbb{E}}[\varepsilon^*\|b^*])$	0.000	0.000	-0.000	-0.000	0.000	0.000
$\widehat{D}_l^*$	0.011	0.016	0.026	0.031	0.002	0.002
$\widehat{D}$	0.008	0.012	0.018	0.024	0.002	0.002
$\widehat{D}_p^*$	0.005	0.007	0.012	0.017	0.001	0.001

Notes: The table details the estimates for the disparities and covariance terms obtained from re-calibration, for the North Carolina dataset (both unweighted (odd columns) and re-weighted (even columns). The estimates are calculated for the overall population (columns 1 and 2), the EITC population (columns 3 and 4), and the non-EITC population (columns 5 and 6).

B.6 Gibbs Sampling

In addition to taxpayers' first name, surname, and geographic location, the IRS has access to additional information that may correlate to race. In principle, leveraging such additional information could lead to better estimates of race probabilities and thus of disparity. Additionally, it is possible that a finer breakdown of self-identified race and ethnicity could contain additional information that may affect our disparity estimates. Hence, as an additional robustness check, we leverage income (bucketed into 14 categories) and marital status, abbreviated "MARS" (Single, Married Filling Jointly, or Other) to obtain more accurate race/ethnicity estimates (at the more granular level of Hispanic, non-Hispanic White, non-Hispanic Black, and Other).

To our knowledge, there are no readily available marginal distributions of race/Hispanic probabilities conditional on income or marital status (and said distributions may differ among taxpayers than the general population); hence, we use Gibbs sampling to obtain approximate probabilities from the IRS' data and BIFSG. Gibbs sampling is a Bayesian algorithm that reduces the problem of sampling from complicated joint distributions to sampling from simpler marginal ones; in this section, we describe in detail this procedure and how we apply it to our setting.

As a starting point, we take the conditional distribution of race and Hispanic origin (RH) given first name, surname, and geography (F, S, and G, respectively, and, collectively, FSG), implied by BIFSG to be correct. We model the joint distribution of (RH, FSG, X), where X represents $(income, MARS)$, as a decomposable model with generating components $\{[RH, F][RH, S][RH, G][RH, X]\}$. (In other words, we make a similar naive Bayes assumption as in BIFSG, but treating X as a unit and allowing a more general relationship between income and MARS.) Given this model, we can write the conditional distribution of RH given (X, FSG) as

$$\Pr(RH|X, FSG) = \text{Multi}\left(n, C\boldsymbol{\theta}_{(i,j)}\frac{\Pr(RH|G)}{\Pr(RH)}\frac{\Pr(RH|F)}{\Pr(RH)}\frac{\Pr(RH|S)}{\Pr(RH)}\right)$$

where $\boldsymbol{\theta}_{i,j}$ is a vector of probabilities for the RH categories, given $(X_1 = i, X_2 = j)$; that is, $\boldsymbol{\theta}_{i,j} = \Pr(RH|X_1 = i, X_2 = j)$. Note also that $\text{Multi}(n, \mathbf{p})$ represents the multinomial distribution with n draws and class probabilities $\mathbf{p}$, and C is a normalizing constant.

We estimate the parameters in the model with a Bayesian procedure, so we need a prior on the unknown parameter $\boldsymbol{\theta}_{(i,j)}$. We set that to the Dirichlet prior with vector parameter $\boldsymbol{\alpha}_0 = (1, \ldots, 1)$, denoted $\boldsymbol{\theta}_{(i,j)} \sim \text{Dir}(\boldsymbol{\alpha}_0)$. This value for $\boldsymbol{\alpha}_0$ was chosen to contribute a small amount of information to the model while ensuring that the posterior is well-behaved. Denote the unobserved vector of counts in the RH categories as $\mathbf{n}_{i,j}$ for $(X_1 = i, X_2 = j)$. Given the form of the model, $\mathbf{n}_{i,j}|X \sim \text{Multi}(n, \boldsymbol{\theta}_{i,j})$, the Dirichlet distribution was chosen because it is the conjugate prior for the multinomial distribution and $\boldsymbol{\theta}_{(i,j)}|\mathbf{n}_{i,j} \sim \text{Dir}(\boldsymbol{\alpha}_0 + \mathbf{n}_{i,j})$. Now we have the full conditional distributions for the unobserved variables in the model.

$$\Pr\left(RH|X, F = f, S = s, G = g\right) = \text{Multi}\left(n, C\boldsymbol{\theta}_{(i,j)}\frac{\Pr(RH|g)}{\Pr(RH)}\frac{\Pr(RH|f)}{\Pr(RH)}\frac{\Pr(RH|s)}{\Pr(RH)}\right) \quad (4)$$

and

$$\Pr\left(\boldsymbol{\theta}_{(i,j)}|\mathbf{n}_{i,j}\right) = \text{Dir}(\boldsymbol{\alpha}_0 + \mathbf{n}_{i,j}),$$

which make a Gibbs sampling algorithm available for estimation.

An outline of the Gibbs Sampling algorithm used here is provided below. Note the superscript (b) indexes the iteration number; it is not an exponent.

- Initialization
  - For each record, indexed by m, generate $RH_m^{(0)}$ from

$$\Pr(RH_m|f_m, s_m, g_m) \sim \text{multi}\left(1, C\frac{\Pr(RH|g_m)}{\Pr(RH)}\frac{\Pr(RH|f_m)}{\Pr(RH)}\frac{\Pr(RH|s_m)}{\Pr(RH)}\right)$$

where again Multi$(n, \mathbf{p})$ represents the multinomial distribution with size n and probability $\mathbf{p}$ and C is a normalizing constant.

  - Tabulate $\mathbf{n}_{i,j}^{(0)} = \sum_{X_m=(i,j)} RH_m^{(0)}$

- Main Loop
  - for $b = 1, ..., B + b_0$:
    - $*$ generate $\boldsymbol{\theta}_{i,j}^{(b)} \sim \text{Dir}\left(1, \boldsymbol{\alpha}_0 + \mathbf{n}_{i,j}^{(b-1)}\right)$ for each i, j
    - $*$ generate $RH_m^{(b)}$ as

$$RH_m^{(b)} \sim \text{Multi}\left(1, C\boldsymbol{\theta}_{(i,j)_m}^{(b)} \frac{\Pr(RH|g_m)}{\Pr(RH)} \frac{\Pr(RH|f_m)}{\Pr(RH)} \frac{\Pr(RH|s_m)}{\Pr(RH)}\right)$$

    - $*$ tabulate $\mathbf{n}_{i,j}^{(b)} = \sum_{X_m=(i,j)} RH_m^{(b)}$

This generates a sequence of values $(\boldsymbol{\theta}_{i,j}^{(b)}, b = 1, ..., B + b_0)$. Here, b_0 is called the *burn-in time.* If the initial values, where $b = 0$, are far from the center of the posterior distribution, it may take several iterations for the sequence to move toward the mode of the posterior. It can be shown that, after a long enough burn-in time b_0, the set $\{\boldsymbol{\theta}_{i,j}^{(b)}, b = 1, ...B + b_0\}$ will be a sample from the target distribution, that is, the posterior distribution of $\boldsymbol{\theta}_{i,j}$, conditioned on the data. (Technical conditions for this are given in e.g. Geman & Geman (1984).) Then if B is large,

$$E(\boldsymbol{\theta}_{i,j}|\text{Data}) \approx \frac{1}{B} \sum_{b_0}^{B+b_0} \boldsymbol{\theta}_{i,j}^{(b)}$$

The new probabilities for RH are calculated for each record using Equation 4 above. For more details on the Gibbs sampling technique, see e.g. Casella & George (1992).

Using these probabilities, we re-compute the linear and probabilistic estimators in the same manner as described before; the results are given in Column (2) of Table A.12. These estimates are similar to those calculated with vanilla and re-calibrated BIFSG. Note that in practice, the algorithm described can be computationally expensive; we thus perform the entire procedure on a 1% sample of the population.

B.7 Estimating Audit Disparity Conditional on True Underreporting

In Section 7.1.1 we describe at a high level how we can combine operational audit data, NRP data, and baseline taxpayer data to estimate the audit rate of taxpayers conditional on a given (binned) amount of underreporting. Here, we provide additional detail. Note that the audit rate of taxpayers of group g and underreporting k is simply $\Pr[Y = 1|B = g, K = k]$.

By Bayes' rule:

$$\Pr[Y = 1|B = g, K = k] = \frac{\Pr[K = k|Y = 1, B = g]\Pr[Y = 1|B = g]}{\Pr[K = k|B = g]}.$$

Each of the quantities on the right-hand side of the equation includes self-reported race as a conditioning variable. Since we do not have access to self-reported race, we must use our predicted race probability, like in our estimates, to measure these quantities.

We consider each quantity in turn. Consider first $\Pr[K = k|B = 1]$, i.e. the probability that a taxpayer has non-compliance K given that they are Black. In practice, we bin taxpayers' underreporting amounts rather than viewing them as exact figures (as exact repeated amounts of underreporting are rare). Viewing $K = k$ as membership in the set of taxpayers whose underreporting is in bin k, we can thus apply either the probabilistic or linear estimator to obtain this quantity:

$$\begin{aligned}\widehat{\Pr}^p[K = k \big| B = 1] &:= \frac{\sum_{i \in \text{NRP}} b_i \cdot \mathbf{1}[K_i = k]}{\sum_{i \in \text{NRP}} b_i} \\ \widehat{\Pr}^p[K = k \big| B = 0] &:= \frac{\sum_{i \in \text{NRP}} (1 - b_i) \cdot \mathbf{1}[K_i = k]}{\sum_{i \in \text{NRP}} (1 - b_i)},\end{aligned}$$

where we limit our summation to NRP to ensure that our estimates are representative of the overall population, or the linear estimator, by regressing:

$$\mathbf{1}[K_i = k] = \alpha_k + \beta_k \cdot b_i + \xi_i$$

and taking:

$$\begin{aligned}\widehat{\Pr}^\ell[K = k \big| B = 1] &:= \widehat{\alpha}_k + \widehat{\beta}_k \\ \widehat{\Pr}^\ell[K = k \big| B = 0] &:= \widehat{\alpha}_k,\end{aligned}$$

where again the regression is run over EITC claimants in NRP. (As before, we can modify our estimators to take into account sample weights accordingly.)

Next, consider $\Pr[K = k|Y = 1, B = g]$. This quantity is just as $\Pr[K = k|B = g]$, but limited to taxpayers who were audited; thus, we can again apply the probabilistic and linear estimators, but run them over the operational audit data rather than NRP.

Finally, $\Pr[Y = 1|B = g]$ is simply the overall audit probability conditional on race, which is the main focus of the paper.

Combining the weighted estimators together, we can write:

$$\widehat{\Pr}^p[Y = 1|K = k, B = 1] := \frac{\frac{\sum_{i \in \text{NRP}} b_i \cdot \mathbf{1}[K_i = k]}{\sum_{i \in \text{NRP}} b_i} \cdot \frac{\sum_i b_i Y_i}{\sum_i b_i}}{\frac{\sum_{i \in \text{OP}} b_i \cdot \mathbf{1}[K_i = k]}{\sum_{i \in \text{OP}} b_i}}$$

and similarly for conditioning on $B = 0$. We can similarly combine the linear estimates.

Because we have combined the estimators together, Proposition 1 does not directly apply.

Additionally, estimates are not independent across different underreporting amounts. These factors make the behavior of this estimator more difficult to analyze. Thus, in order to obtain confidence intervals we do not attempt to characterize the standard errors analytically, but instead use the bootstrap. That is, we draw 100 re-samplings (of each dataset) without replacement, and re-compute the estimates for each subsample. We then take the mean of these estimates for each bin k to be our estimated audit rates, and add/subtract 1.96 times standard error to obtain the confidence intervals. We note that, while we do not have a formal statement about the direction of bias these combined estimators may have and whether the ground truth need lie between the combined probabilistic and linear estimators, the probabilistic estimator tends to produce smaller estimates of within-bin audit rate disparities than the linear estimator does, at least for the bulk of the distribution.

B.8 Conditional Disparity Estimators and Decomposition

We can generalize the disparity estimators straightforwardly to answer the following question: What is the racial disparity in audits *conditional* on some covariates? Formally, let X represent a (possibly vector-valued) feature, taking on values $x \in \mathcal{X}$. We define the *audit disparity at* x as:

$$D_x := \mathbb{E}[Y|B = 1, X = x] - \mathbb{E}[Y|B = 0, X = x].$$

To obtain estimates for these quantities, we again apply our linear and probabilistic estimators. That is, we define the probabilistic estimator for audit disparity at x as:

$$\widehat{D}_x^P := \frac{\sum_{i:X_i=x} Y_i b_i}{\sum_{i:X_i=x} b_i} - \frac{\sum_{i:X_i=x} Y_i(1-b_i)}{\sum_{i:X_i=x}(1-b_i)},$$

and the linear estimator for audit disparity at x as the estimated coefficient $\widehat{\beta}_x$ in the regression of Y on b over the set of observations i such that $X_i = x$. Formally, that is,

$$\widehat{D}_x^l = \frac{\sum_{i:X_i=x}(b_i - \bar{b}_x)(Y_i - \bar{Y}_x)}{\sum_{i:X_i=x}(b_i - \bar{b}_x)^2},$$

where $\bar{\cdot}_x$ indicates the average taken with respect to observations i having $X_i = x$.

It is easy to see that, assuming that $\Pr[X_i = x] > 0$, we will have that:

$$\widehat{D}_x^p \overset{n\to\infty}{\longrightarrow} \frac{\operatorname{Cov}(Y, b|X = x)}{\mathbb{E}[b|X = x](1 - \mathbb{E}[b|X = x])} \equiv D_x^p$$

by applying the Law of Large Numbers to observations with $X_i = x$. Similarly, we can observe that

$$\widehat{D}_x^l \overset{n\to\infty}{\longrightarrow} \frac{\operatorname{Cov}(Y, b|X = x)}{\operatorname{Var}(b|X = x} \equiv D_x^l,$$

and that the multiplicative relationship between D_x^l and D_x^p holds with x-specific constants:

$$D_x^p = D_x^l \cdot \frac{\mathrm{Var}(b|X = x)}{\mathbb{E}[b|X = x](1 - \mathbb{E}[b|X = x])}.$$

Accordingly, we can obtain x-specific covariance conditions for these estimators to write the following theorem:

Proposition 7 (Conditional Disparity Bounds)**.** Suppose that b is a taxpayer's probability of being Black given some observable characteristics Z, so that $b = \Pr[B = 1|Z, X]$. Define D_x^p as the asymptotic limit of the probabilistic disparity estimator at $X = x$, $\widehat{D}_p$, and D_x^l as the asymptotic limit of the linear disparity estimator at $X = x$, $\widehat{D}_x^l$. Then:

1.

$$D_x^p = D_x - \frac{\mathbb{E}[\mathrm{Cov}(Y, B|b, X)|X = x)]}{\mathrm{Var}(b|X = x)} \qquad \mathbf{(7.1)}$$

2.

$$D_x^l = D_x + \frac{\mathbb{E}[\mathrm{Cov}(Y, b|B, X)|X = x)]}{\mathrm{Var}(B|X = x)} \qquad \mathbf{(7.2)}$$

3. Suppose $\mathbb{E}[\mathrm{Cov}(Y, B|b, X)|X = x)] \geq 0$ and $\mathbb{E}[\mathrm{Cov}(Y, b|B, X)|X = x] \geq 0$. Then

$$D_x^p \leq D \leq D_x^l \qquad \mathbf{(7.3)}$$

4. Suppose $\mathbb{E}[\mathrm{Cov}(Y, B|b, X)|X = x] \leq 0$ and $\mathbb{E}[\mathrm{Cov}(Y, b|B, X)|X = x] \leq 0$. Then

$$D_x^l \leq D \leq D_x^p \qquad \mathbf{(7.4)}$$

Importantly, Proposition 7 requires that b be $\Pr[B = 1|Z, X]$. In other words, if we wish to apply it, we must not only have an accurate measurement of the probability a taxpayer is Black given their name and geography, but also including additionally the feature of interest. Moreover, there is again the possibility of bias and noise, and thus the "at-x" analogue of Proposition 2; a more general formulation and associated proof follow similarly. Thus, in applying this result, recalibration within the values taken on by X may be important.

Now, Proposition 7 provides an analogue of our bounds for each x. But we might wish to summarize D_x, which is a function, into a single number, e.g. by estimating $\mathbb{E}[D_x]$. Again, it is easy easy to see that

$$\frac{1}{n}\sum_x n_x \widehat{D}_x^p \overset{n\to\infty}{\longrightarrow} \mathbb{E}[D_x^p] \qquad \frac{1}{n}\sum_x n_x \widehat{D}_x^l \overset{n\to\infty}{\longrightarrow} \mathbb{E}[D_x^l],$$

but can we similarly say something about $\mathbb{E}[D_x^l]$ vs $\mathbb{E}[D_x]$ vs $\mathbb{E}[D_x^p]$? The following theorem provides both necessary and sufficient conditions; it boils down to asking that the covariance conditions hold *on average*.

Theorem 1. Let D_x^p, D_x^l, D_x be as above. Let $\mathcal{P}$ be an arbitrary distribution over X. Then:

$$\mathbb{E}_{\mathcal{P}}[D_x^p] \leq \mathbb{E}_{\mathcal{P}}[D_x] \leq \mathbb{E}_{\mathcal{P}}[D_x^l]$$

whenever:

1. (Sufficient, but not necessary:)

$$\mathbb{E}[\text{Cov}(Y, b|B, X)] \geq 0, \mathbb{E}[\text{Cov}(Y, B|b, X)] \geq 0 \ \forall x$$

2. (Necessary and sufficient):

$$\mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\text{Cov}(Y, b|B, x)|X = x]}{\text{Var}(B|X = x)}\right] \geq 0, \mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\text{Cov}(Y, B|b, x)|X = x]}{\text{Var}(b|X = x)}\right] \geq 0$$

The sufficient, but not necessary, condition is easy to reason about because it only requires an assumption about the sign of covariance terms, but of course less likely to hold given its strength. By contrast, the necessary condition is more reasonable in that it allows for the covariance conditions to be negative over some values of X as long as it is positive over enough of the others; however, ascertaining this requires not merely knowledge of signs, but quantitative variation.

Proof. As noted, that

$$D_x = D_x^l - \frac{\mathbb{E}[\text{Cov}(Y, b|B, x)|X = x]}{\text{Var}[b|X = x]}$$

and

$$D_x = D_x^p + \frac{\mathbb{E}[\text{Cov}(Y, B|b, x)|X = x]}{\text{Var}[B|X = x]}$$

follows from the same logic, conditioned on the event $X = x$, as used in the proof of Proposition 2. Then:

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}}[D_x^l] = \sum_x \mathcal{P}(x) D_x^p = \sum_x \left[\mathcal{P}(x)\left[D_x + \frac{\mathbb{E}[\text{Cov}(Y, b|B, X)|X = x]}{\text{Var}[b|X = x]}\right]\right] \\
= \mathbb{E}_{\mathcal{P}}[D_x] + \sum_x \mathcal{P}(x) \frac{\mathbb{E}[\text{Cov}(Y, b|B, X)|X = x]}{\text{Var}[b|X = x]} \\
= \mathbb{E}_{\mathcal{P}}[D_x] + \mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\text{Cov}(Y, b|B, X)|X = x]}{\text{Var}[b|X = x]}\right].
\end{aligned}$$

Thus:

$$\mathbb{E}[\text{Cov}(Y, b|B, X)|X = x] \geq 0 \ \forall X \implies \mathbb{E}_{\mathcal{P}}[D_x^l] \geq \mathbb{E}_{\mathcal{P}}[D_x],$$

But more weakly, it also holds that:

$$\mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\mathrm{Cov}(Y, b|B, x)|X = x]}{\mathrm{Var}[b|X = x]}\right] \geq 0 \implies \mathbb{E}_{\mathcal{P}}[D^l_x] \geq \mathbb{E}_{\mathcal{P}}[D_x].$$

Similarly, we can write that:

$$\mathbb{E}_{\mathcal{P}}[D^p_x] = \mathbb{E}_{\mathcal{P}}[D_x] - \mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\mathrm{Cov}(Y, B|b, x)|X = x]}{\mathrm{Var}[B|X = x]}\right]$$

and so again,

$$\mathbb{E}[\mathrm{Cov}(Y, B|b, X)|X = x] \geq 0 \; \forall X \implies \geq \mathbb{E}_{\mathcal{P}}[D^p_x] \leq \mathbb{E}_{\mathcal{P}}[D_x],$$

and the weaker version

$$\mathbb{E}_{\mathcal{P}}\left[\frac{\mathbb{E}[\mathrm{Cov}(Y, B|b, x)|X = x]}{\mathrm{Var}[B|X = x]}\right] \geq 0 \implies \mathbb{E}_{\mathcal{P}}[D^p_x] \leq \mathbb{E}_{\mathcal{P}}[D_x]$$

also holds.

□

B.8.1 Using D_x to decompose D into constituent parts

If the distribution over X is the same for both groups, then $D = \mathbb{E}[D_x]$, where the expectation is taken over the population distribution of X, $\mathcal{P}_X$. Otherwise, there is a *compositional effect*; that is, unless audit rates are constant for all values in $\mathcal{X}$, groups will have different total audit rates because individuals tend to have values of X associated with higher audit rates in one group more than another, even if audit rates are the same for both groups conditional on every x. Arguably, one might take the point of view that the conditional disparity (either averaged over $\mathcal{P}_X$ or some other reference distribution, e.g. $\mathcal{P}_{X|B=1}$) is of primary importance, as it captures the portion of disparity stemming from differences in the outcome of interest given the same individual features. We do not take the view that the overall disparity is unimportant for many reasons; for example, even compositional effects can be related to systemic inequality and may thus have relevant policy implications. But it still useful to understand how much of disparity is driven by differing audit rates with the same features rather than compositional effects.

In Proposition 8, we thus decompose disparity into a portion coming from $\mathbb{E}[D_x]$ and a portion coming from the compositional effects. To understand this decomposition, we define $\mathcal{P}_g(X)$ to be the conditional distribution $\Pr[X = x|B = g]$ and, as mentioned, $\mathcal{P}_X$ to be the unconditional distribution $\Pr[X = x]$. We will use $\mathcal{P}$ for an arbitrary distribution over X. We also define an operator Δ representing the difference of two probability distributions weighted by a function of a random variable. In the discrete case, that is:

$$\Delta_{\mathcal{P},\mathcal{P}'}(f(x)) := \sum_x f(x)\left[\mathcal{P}(x) - \mathcal{P}'(x)\right].$$

With this notation, we can state and prove the following proposition capturing the decomposition with respect to an arbitrary "reference" distribution over X:

Proposition 8. Suppose $\mathcal{P}$ is an arbitrary distribution over X. Then overall disparity $\mathcal{D}$ can be decomposed with respect to $\mathcal{P}$ as:

$$D = \mathbb{E}_{\mathcal{P}}[D_x] + \Delta_{\mathcal{P}_1,\mathcal{P}}(\mathbb{E}[Y|B=1, X=x]) - \Delta_{\mathcal{P}_0,\mathcal{P}}(\mathbb{E}[Y|B=0, X=x]).$$

Proof. The law of iterated expectations says that:

$$\mathbb{E}[Y|B=g] = \mathbb{E}_{\mathcal{P}_g}\mathbb{E}[Y|B=g, X=x],$$

so

$$\begin{aligned} D &:= \mathbb{E}[Y|B=1] - \mathbb{E}[Y|B=0] \\ &= \mathbb{E}_{\mathcal{P}_1}\left[\mathbb{E}[Y|B=1, X=x]\right] - \mathbb{E}_{\mathcal{P}_0}\left[\mathbb{E}[Y|B=0, X=x]\right]. \end{aligned}$$

Now, adding and subtracting both $\mathbb{E}_{\mathcal{P}}\left[\mathbb{E}[Y|B=1, X=x]\right]$ and $\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}[Y|B=0, X=x]\right]$ we get:

$$\begin{aligned} D = {} & \mathbb{E}_{\mathcal{P}}[\mathbb{E}[Y|B=1, X=x]] - \mathbb{E}_{\mathcal{P}}[\mathbb{E}[Y|B=0, X=x]] \\ & + \mathbb{E}_{\mathcal{P}_1}[\mathbb{E}[Y|B=1, X=x]] - \mathbb{E}_{\mathcal{P}}[\mathbb{E}[Y|B=1, X=x]] \\ & + \mathbb{E}_{\mathcal{P}}[\mathbb{E}[Y|B=0, X=x]] - \mathbb{E}_{\mathcal{P}_0}[\mathbb{E}[Y|B=0, X=x]] \end{aligned}$$

The terms in the first line is simply $\mathbb{E}_{\mathcal{P}}[D_x]$. For the second term, note that

$$\begin{aligned} & \mathbb{E}_{\mathcal{P}_1}[\mathbb{E}[Y|B=1, X=x]] - \mathbb{E}_{\mathcal{P}}[\mathbb{E}[Y|B=1, X=x]] \\ & \quad = \sum_x \mathcal{P}_1(x)\mathbb{E}[Y|B=1, X=x] - \sum_x \mathcal{P}(x)\mathbb{E}[Y|B=1, X=x] \\ & \quad = \sum_x \mathbb{E}[Y|B=1, X=x]\left[\mathcal{P}_1(x) - \mathcal{P}(x)\right] \\ & \quad = \Delta_{\mathcal{P}_1,\mathcal{P}}(\mathbb{E}[Y|B=1, X=x]) \end{aligned}$$

and similarly the third simplifies to $\Delta_{\mathcal{P},\mathcal{P}_0}(\mathbb{E}[Y|B=0, X=x])$. □

The most natural choices to use as reference distributions in Proposition 8 are either $\mathcal{P}_X$, $\mathcal{P}_1$, or $\mathcal{P}_0$. For the latter two, the form becomes slightly simpler:

Corollary 1. By noting that $\Delta_{\mathcal{P},\mathcal{P}} = 0$, we can also write that:

$$D = \mathbb{E}_{\mathcal{P}_1}[D_x] + \Delta_{\mathcal{P}_1,\mathcal{P}_0}(Y|B=0, X=x)$$

and

$$D = \mathbb{E}_{\mathcal{P}_0}[D_x] + \Delta_{\mathcal{P}_0,\mathcal{P}}(Y|B=1, X=x)$$

Regardless of the reference distribution, Proposition 7 and Corollary 1 show that overall disparity can be decomposed into an average conditional disparity, under some distribution, and compositional effects that capture the difference of one or both groups' distribution over X with respect to that reference distribution.

To apply this decomposition, we must know or estimate the constituent components. Non-group-specific distributions, e.g. $\mathcal{P}(X)$ and $\mathbb{E}[Y|X]$, can be simply estimated from the data in the usual manner, but we must also estimate group-specific distributions over X and group-specific audit rates given X. We turn to estimation concerns in the following section.

B.8.2 Estimating decomposition-relevant quantities

We focus on the case where $\mathcal{X}$ is discrete and has a relatively small number of unique values. In this case, we need to estimate D_x, $\mathbb{E}[Y_x|B = 1, X = x]$,$\mathbb{E}[Y_x|B = 0, X = x]$, $\mathcal{P}_1(x)$, $\mathcal{P}_0(x)$. Once we obtain these estimates, we combine them in a plug-in manner to estimate the expression in Proposition 8:

$$\begin{aligned}\widehat{D} = \sum_x \widehat{\mathcal{P}}(x)\widehat{D}_x + \sum_x \widehat{\mathbb{E}}[Y|B = 1, X = x] \cdot \left[\widehat{\mathcal{P}}_1(x) - \widehat{\mathcal{P}}(x)\right] \\ + \sum_x \widehat{\mathbb{E}}[Y|B = 0, X = x] \cdot \left[\widehat{\mathcal{P}}(x) - \widehat{\mathcal{P}}_0(x)\right]\end{aligned} \tag{5}$$

As with the overall disparity estimation and related problems, we can obtain these estimates via either the probabilistic disparity estimator or the linear disparity estimator. For the probabilistic, we estimate $\widehat{D}^p_x$ by computing the probabilistic estimator for individuals with $X_i = x$; we estimate the audit rate conditional at each x using the levels ; and we estimate the distribution using the probabilistic estimator over the full dataset with an indicator for having $X_i = x$ as the outcome of interest. That is:

$$\begin{aligned}\widehat{D}^p_x &:= \frac{\sum_{i:X_i=x} b_i Y_i}{\sum_{i:X_i=x}} - \frac{\sum_{i:X_i=x} b_i Y_i}{\sum_{i:X_i=x}} \\ \widehat{E}[Y|B = 1, X = x] &:= \frac{\sum_{i:X_i=x} b_i Y_i}{\sum_{i:X_i=x} b_i} \\ \widehat{E}[Y|B = 0, X = x] &:= \frac{\sum_{i:X_i=x}(1 - b_i) Y_i}{\sum_{i:X_i=x}(1 - b_i)} \\ \widehat{\mathcal{P}}_1(x) &:= \frac{\sum_i b_i \mathbf{1}[X_i = x]}{\sum_i b_i} \\ \widehat{\mathcal{P}}_0(x) &:= \frac{\sum_i (1 - b_i)\mathbf{1}[X_i = x]}{\sum_i (1 - b_i)}.\end{aligned}$$

For the linear, we estimate, two regressions for each value of $x \in \mathcal{X}$:

$$Y_i = \alpha_x + \beta_x b_i$$

over all $i : X_i = x$, and

$$\mathbf{1}[X_i = x] = \zeta_x + \xi_x b_i$$

over all i. We then construct have:

$$\begin{aligned}\widehat{D}^p_x &=: \widehat{\beta}_x \\ \widehat{E}[Y|B=1,X=x] &:= \widehat{\alpha}_x + \widehat{\beta}_x \\ \widehat{E}[Y|B=0,X=x] &:= \widehat{\alpha}_x \\ \widehat{\mathcal{P}}_1(x) &:= \widehat{\zeta}_x + \widehat{\xi}_x \\ \widehat{\mathcal{P}}_0(x) &:= \widehat{\zeta}_x\end{aligned}$$

B.9 Tightening Bounds with a Linear Program

In this subsection we use a linear program to calculate the maximum and minimum values of disparity that are consistent with the Proposition 1.3 assumptions and the observed joint distribution of b and Y. To implement this approach, we discretize the b distribution into n mutually exclusive and equal width bins, labeling them with the sample average $\bar{b}_i$ of taxpayers in the bin. The share of taxpayers in each bin is given by $p_1, ...p_n$. As in Proposition 1, we assume race probabilities for each bin are perfectly calibrated, so that $P(B=1|b=b_i)=b_i$ for all i from 1 to n. Along with the marginal distribution of b, we observe the distribution of audits conditional on b. Denote the audit rate in each bin as $Y_1, ..Y_n$, and the overall audit rate as Y. The audit rate among Black taxpayers in bin i is denoted by Y_i^B and among non-Black taxpayers as Y_i^{NB}. Finally, as in Proposition 1.3, we assume the sign of the covariance conditions, i.e. that $\mathbb{E}[\mathrm{Cov}(Y,b|B)] \geq 0$ and $\mathbb{E}[\mathrm{Cov}(Y,B|b)] \geq 0$.

We ask: if we can allocate the Y_i into Y_i^B and Y_i^{NB} for each bin i, what is the maximum and minimum disparity we could obtain consistent with the results? This is a constrained optimization problem. Disparity is given by:

$$Y^B - Y^{NB} = \sum_i \Pr[b_i|B=1] \cdot Y_i^B - \sum_i \Pr[b_i|B=0] \cdot Y_i^{NB}$$

Noting that

$$b_i Y_i^B + (1-b_i) Y_i^{NB} = Y_i \implies Y_i^{NB} = \frac{Y_i - b_i Y_i^B}{1-b_i}$$

we can replace the Y_i^{NB} into the objective to get:

$$\sum_i \left[\Pr[b_i|B=1] Y_i^B - \Pr[b_i|B=0] \left(\frac{Y_i - Y_i^B b_i}{1-b_i}\right)\right]$$

and we may as well drop Y_i from the objective (since it will not change based on our choice variables) to get:

$$f(\vec{Y}^B) = \sum_i \left[\Pr[b_i|B=1] Y_i^B + \frac{\Pr[b_i|B=0] b_i}{1-b_i} Y_i^B\right]$$

Now by Bayes' rule, we can write that:

$$\Pr[b_i|B=1] = \frac{\Pr[B=1|b_i]p_i}{\Pr[B=1]} = \frac{b_i p_i}{\Pr[B=1]} \qquad \Pr[b_i|B=0] = \frac{\Pr[B=0|b_i]p_i}{\Pr[B=0]} = \frac{(1-b_i)p_i}{\Pr[B=0]}.$$

If we substitute these into the objective function, we have:

$$\sum_i \left[\frac{b_i p_i}{\Pr[B=1]} Y_i^B + \frac{(1-b_i)p_i}{\Pr[B=0]} \frac{b_i Y_i^B}{1-b_i}\right] = \sum_i Y_i^B b_i p_i \left[\frac{1}{\Pr[B=1]} + \frac{1}{\Pr[B=0]}\right]$$

The constant multiplier will also not be affected by our decisions, so we can drop it, and our objectives will be:

$$\max_{<Y_i^B>} \sum_i p_i Y_i^B b_i \qquad \text{and} \qquad \min_{<Y_i^B>} \sum_i p_i Y_i^B b_i$$

These are intuitive - to maximize (minimize) disparity, we allocate as much of the audit rate as possible to bins of b_i that are more likely to contain Black (non-Black) taxpayers, either because b_i or $\Pr[b_i]$ is higher (lower) than others.

As for constraints: First, we need Y_i^B to be in the right range. $[0,1]$ is a first pass, but note that this does not guarantee consistency with the data. For instance, if Y_i^B is 0, even $Y_i^{NB} = 1$ might not be enough to satisfy that the overall audit rate in the bin matches Y_i. So in particular, the lowest that Y_i^B could be is $\max\left\{0, \frac{Y_i-(1-b_i)}{b_i}\right\}$, while the most that it could be is $\min\left\{1, \frac{Y_i}{b_i}\right\}$. These numbers can be calculated from the data, so we will simply refer to them as upper bound u_i and lower bound l_i.

Second, the first covariance constraint, $\mathbb{E}[\mathrm{Cov}(Y,B|b)] \geq 0$.

We note that:

$$\mathrm{Cov}(Y,B|b) = \mathbb{E}[YB|b] - \mathbb{E}[Y|b]\mathbb{E}[B|b].$$

Now, $\mathbb{E}[Y|b=b_i]$ is simply Y_i, and $E[B|b=b_i]$ is b_i by assumption of perfect calibration. $E[YB|b=b_i] = b_i Y_i^B$, since $YB = Y_i^B$ whenever $B=1$ and 0 otherwise, and again $\Pr[B|b=b_i] = b_i$ by perfect calibration.

Then we can write that:

$$\mathrm{Cov}(Y,B|b) = b_i Y_i^B - Y_i b_i \implies \mathbb{E}[\mathrm{Cov}(Y,B|b)] = \sum_i p_i b_i (Y_i^B - Y_i)$$

Finally, the third condition: we need that $\mathbb{E}[\mathrm{Cov}(Y,b|B)] \geq 0$. Let's consider just $B=1$ first.

$$\mathbb{E}[Yb|B=1] - \mathbb{E}[Y|B=1]\mathbb{E}[b|B=1] =$$

$$\sum_i Y_i^B b_i \Pr[b_i|B=1] - \left(\sum_i \Pr[b_i|B=1] Y_i^B\right)\left(\sum_i \Pr[b_i|B=1] b_i\right)$$

Define $\bar{b}_B = \sum_i \Pr[b_i|B=1]b_i$, which has the interpretation of the average probability of being Black given ground truth race being Black.

Then we can write that the above is:

$$\begin{aligned}
\sum_i Y_i^B b_i \Pr[b_i|B=1] - &\left(\sum_i Y_i^B \Pr[b_i|B=1]\right)\left(\sum_i b_i \Pr[b_i|B=1]\right) \\
&= \sum_i Y_i^B b_i \Pr[b_i|B=1] - \sum_i Y_i^B \Pr[b_i|B=1]\bar{b}_B \\
&= \sum_i Y_i^B (b_i - \bar{b}_B) \Pr[b_i|B=1]
\end{aligned}$$

Using our Bayes' rule calculation as above, we can write:

$$\sum_i Y_i^B (b_i - \bar{b}_B) \Pr[b_i|B=1] = \frac{1}{\Pr[B]} \sum_i Y_i^B (b_i - \bar{b}_B) b_i p_i$$

For given $B = 0$, we have that:

$$\begin{aligned}
\sum_i Y_i^{NB} b_i \Pr[b_i|B=0] - &\left(\sum_i Y_i^{NB} \Pr[b_i|B=0]\right)\left(\sum_i b_i \Pr[b_i|B=0]\right) \\
&= \sum_i Y_i^{NB} b_i \Pr[b_i|B=0] - \sum_i Y_i^{NB} \Pr[b_i|B=0]\bar{b}_{NB} \\
&= \sum_i Y_i^{NB} (b_i - \bar{b}_{NB}) \Pr[b_i|B=0]
\end{aligned}$$

and again applying Bayes' rule we can write the above as:

$$\frac{1}{1-\Pr[B]} \sum_i Y_i^{NB} (b_i - \bar{b}_{NB})(1-b_i) p_i.$$

Then the overall constraint is:

$$\begin{aligned}
\Pr[B] \cdot \frac{1}{\Pr[B]} &\sum_i Y_i^B \cdot b_i \cdot (b_i - \bar{b}_B) \cdot p_i \\
&+ (1-\Pr[B]) \frac{1}{1-\Pr[B]} \sum_i Y_i^{NB} (1-b_i)(b_i - \bar{b}_{NB}) p_i \\
&= \sum_i p_i \left(Y_i^B b_i (b_i - \bar{b}_B) + Y_i^{NB}(1-b_i)(b_i - \bar{b}_{NB})\right) \geq 0
\end{aligned}$$

Noting that $Y_i^{NB} = (Y_i - Y_i^B b_i)/(1-b_i)$ and factoring out, we can rewrite the last inequality as:

$$\sum_i p_i \left(Y_i^B b_i (\bar{b}_{NB} - \bar{b}_B) + Y_i(b_i - \bar{b}_{NB}\right) \geq 0$$

Putting these together, our problem is:

Program 1 (Maximum Consistent Disparity)**.**

$$\begin{aligned}
\max_{<Y_i^B>} \sum_i p_i Y_i^B b_i \text{ s.t. } Y_i^B &\leq u_i \\
l_i &\leq Y_i^B \\
0 &\leq \sum_i p_i b_i (Y_i^B - Y_i) \\
0 &\leq \sum_i p_i \big(Y_i^B b_i (\bar{b}_{NB} - \bar{b}_B) + Y_i (b_i - \bar{b}_{NB} \big)
\end{aligned}$$

to obtain the maximum disparity consistent with the information, and minimizing the same (or maximizing the negative) to get the minimum disparity.

B.9.1 Linear Program Results

In Table B.2, we report the information that we observe, including the average probability Black within each bin, the audit rate within each bin, and the share of taxpayers that fall into each bin. Using these along with the assumption of calibration, we can compute the conditional probability of a taxpayer falling into each bin given that they are Black and the upper and lower bounds on Y_i^B. These together constitute the requisite input to Program 1 and its counterpart minimization. To solve these programs, we use the SciPy (Virtanen et al., 2020) library's linprog function. The results are given in the following table:

The solutions to the linear programs closely match the disparity estimates obtained by the probabilistic and linear estimators. Compare Column 1 of Table B.3 to Column 1 of Panel A of Table A.6 and Column 2 of Table B.3 to Column (1) of Panel A of Table A.5.

Table B.2: Inputs to Linear Program for Bounding Disparity

Bin	Probability Black	Fraction in Bin	Audit Rate
1	0.0057	0.7130	0.0039
2	0.0722	0.0574	0.0039
3	0.1232	0.0353	0.0047
4	0.1737	0.0233	0.0049
5	0.2241	0.0179	0.0055
6	0.2740	0.0143	0.0060
7	0.3245	0.0119	0.0068
8	0.3743	0.0101	0.0072
9	0.4246	0.0087	0.0077
10	0.4746	0.0079	0.0082
11	0.5248	0.0073	0.0090
12	0.5748	0.0068	0.0095
13	0.6249	0.0065	0.0101
14	0.6750	0.0064	0.0111
15	0.7251	0.0065	0.0118
16	0.7753	0.0068	0.0127
17	0.8255	0.0075	0.0136
18	0.8759	0.0089	0.0149
19	0.9268	0.0123	0.0165
20	0.9820	0.0312	0.0199

Notes: The Table reports statistics used to bound disparity through the linear program described in this subsection. Taxpayers have been binned into 20 groups based on their estimated probabilities of being Black, with each group corresponding to the values of b in a 5 percentage point range. Probability Black denotes the average estimated probability of being Black within the bin. Fraction in Bin denotes the share of the overall taxpayer population in the specified bin.

Table B.3: Maximum and Minimum EITC Audit Disparity Estimates

	Maximum Disparity	Minimum Disparity
	(1)	(2)
Black Audit Rate	1.710	1.241
Non-Black Audit Rate	0.362	0.427
Disparity	1.347	0.813
Audit Rate Ratio	4.7	2.9

Notes: The table reports the results of the Maximium Disparity and Minimum Disparity constrained optimization problems described in Section B.9. Units are percentage points (0-100).

C North Carolina Match and Bias Correction

C.1 North Carolina Match

In our North Carolina voter registration data (reflecting the state's voter registration file as of the close of 2020), we observe individuals' first names, last names, zip codes, residential street addresses, and mailing addresses at time of registration or filing. We match these data to IRS data using these common features according to the following procedure:

1. First, look for exact match on zip code, first name, last name, and full text of residential street address. Remove matched records from both datasets and append matched records to output file.

2. Among unmatched records, look for match on zip code, first four characters of first and last name, and full text of residential street address (after minor data cleaning).

3. Among unmatched records, look for match on zip code, first character of first name, first four characters of last name, and full text of residential street address.

4. Among unmatched records, look for match on zip code, first four characters of first and last name, residential street number, and city.

5. Among unmatched records, look for match on zip code, first character of first name, first four characters of last name, and full text of mailing address.

Using this procedure, we are able to match 2.5 million taxpayer and voter records, or approximately 47% of the population of North Carolina taxpayers for tax year 2014. We use the same procedure to match taxpayers to the 2023 voter registration data from other states reported in Appendix Figure A.13.

C.2 North Carolina Reweighting

When specified, we use inverse-probability weighting to align the composition of the North Carolina matched sample with that of the full population of tax returns for 2014. The weights are generated from a linear probability model whose binary outcome equals one for records appearing in the IRS-matched North Carolina sample, and whose features are chosen to reflect observable taxpayer characteristics that we would like to align with their US means. These are entered as categorical variables and are fully interacted with one another, resulting in a flexible nonparametric model of the conditional probability of appearing in the North Carolina data. Features include quintiles (as calculated on the full population) of the BIFSG-predicted probability that a taxpayer self reports as black; four activity code groupings[41]; gender; the presence of dependents; joint/non-joint filing status; and whether a taxpayer was audited. The weights are then given by the inverse of these conditional probabilities. The weights were successful in aligning the weighted sample proportions along all included dimensions to within 0.02% of their U.S. population means.

[41] Activity codes are grouped as: 270-271 (EITC claimants), 272 (1040 filers without additional schedules or very high income), 273-278 (filers with Schedule C etc. but not very high income), and 279-281 (filers with very high ($1M) or more income or high ($ > 250K) with additional schedules).

D EITC Disparity Decomposition

This appendix provides additional detail regarding the decomposition of the total racial audit disparity into components associated with the disparity within and between EITC claimants presented in Section 6.

As in the main text, Y_C^B and Y_C^{NB} refer to the average audit rates for Black and non-Black EITC claimants, respectively; Y_{NC}^B and Y_{NC}^{NB} refer to the average audit rates for Black and non-Black EITC non-claimants, respectively; and C_B and C_{NB} refer to the respective probabilities that Black and non-Black taxpayers claim the EITC.

Our preferred decomposition, presented in the main text, is given by:

$$
\begin{aligned}
Y^B - Y^{NB} = &\underbrace{\left(Y_C^B - Y_C\right) C_B - \left(Y_C^{NB} - Y_C\right) C_{NB}}_{(1)} \\
&+ \underbrace{\left(Y_{NC}^B - Y_{NC}\right) (1 - C_B) - \left(Y_{NC}^{NB} - Y_{NC}\right) (1 - C_{NB})}_{(2)} + \underbrace{(C_B - C_{NB}) (Y_C - Y_{NC})}_{(3)}
\end{aligned}
$$

The first component reflects racial differences in the audit rate among EITC claimants: if Black and non-Black EITC claimants were selected at the same rate, it would imply $Y_C^B = Y_C^{NB} = Y_C$, so that both terms in (1) would be zero. Similarly, the second component reflects racial differences in the audit rate among non-EITC claimants. The third component reflects compositional differences in the rate at which Black and non-Black taxpayers claim the EITC as well as differences in the audit rate of EITC versus non-EITC returns. This component would be zero if Black and non-Black taxpayers claimed the EITC at equal rates, or if EITC and non-EITC claimants (of any race) were audited at the same rate.

The following table presents estimates of the elements of the decomposition using both the linear and probabilistic audit rate estimators. These estimates are reported in Section 6 of the main text.[42]

An appealing feature of the above decomposition is that it weights the contribution of differences in the EITC and non-EITC audit rate for each racial group based on the share of that racial group claiming the EITC. On the other hand, to the extent that Black and non-Black taxpayers claim the EITC at different rates, there is a sense in which the first two components of the decomposition can be interpreted to reflect compositional differences in EITC claiming as well as differences in the audit rate. We consider two alternative decompositions below, which differ from the above in that they weight the within EITC and within non-EITC components of the disparity based on the EITC claim rates of either Black or non-Black taxpayers.

Using Black taxpayers as the reference group, we can decompose the total disparity as:

$$
Y^B - Y^{NB} = \underbrace{(Y_C^B - Y_C^{NB})\, C_B}_{(1)} + \underbrace{(Y_{NC}^B - Y_{NC}^{NB})\,(1 - C_B)}_{(2)} + \underbrace{(C_B - C_{NB})\ \left(Y_C^{NB} - Y_{NC}^{NB}\right)}_{(3)}
$$

[42] In finite samples, the individual components of the decomposition may not exactly sum to the estimated total disparity. The percentage contributions we report are calculated by dividing each component estimate by the sum of the three component estimates.

Here, the first component represents the contribution of the disparity within EITC claimants, the second represents the contribution of the disparity within EITC non-claimants, and the third component is the same as with the first decomposition presented, but with the difference in EITC versus non-EITC audit rates evaluated for Black taxpayers.

Using this decomposition with the estimates reported in Appendix Table D.1 implies a larger contribution to the overall disparity from the disparity among EITC claimants, between 78% and 83% depending on whether the probabilistic or linear estimator is used (see Appendix Table D.2).

Alternatively, using non-Black taxpayers as the reference group, we can decompose the total disparity as:

$$Y^B - Y^{NB} = \underbrace{\left(Y_C^B - Y_C^{NB}\right) C_{NB}}_{(1)} + \underbrace{\left(Y_{NC}^B - Y_{NC}^{NB}\right) \left(1 - C_{NB}\right)}_{(2)} + \underbrace{\left(C_B - C_{NB}\right) \left(Y_C^B - Y_{NC}^B\right)}_{(3)}$$

With this decomposition, the estimates reported in Appendix Table D.1 imply a smaller contribution to the overall disparity from the disparity within EITC claimants, and a larger contribution from racial differences in EITC claiming (see Appendix Table D.2). The explanation for this difference is that because non-Black taxpayers claim the EITC at lower rates, using that group as the reference leads to attaching less weight to the disparity among EITC claimants.

Table D.1: Decomposition of Disparity with Resepct to EITC Claiming

	Probabilistic Estimate	Linear Estimate
EITC Claim Rate Among...		
Black Taxpayers	32.43	41.14
Non-Black Taxpayers	17.26	16.06
Audit Rate Among...		
EITC Claimants	1.44	1.44
Black EITC Claimants	2.99	3.73
Non-Black EITC Claimants	1.04	0.85
EITC Non-Claimants	0.31	0.31
Black EITC Non-Claimants	0.40	0.47
Non-Black EITC Non-Claimants	0.30	0.29
Disaprity Contribution From...		
Within EITC Claimants	70%	72%
Within EITC Non-Claimants	9%	8%
Racial Differences in EITC Claiming	21%	20%

Notes: The table reports linear and probabilistic estimates of the terms used to calculate the decompositions of disparity described in this appendix section. The final three rows of the table report the contribution of terms (1), (2), and (3) of the preferred disparity decomposition described at the beginning of this Appendix section. Units are percentage points (0-100).

Table D.2: Alternative Disparity Decompositions

Reference Group	Estimator	Within-EITC Contribution	Outside-EITC Contribution	Differences in EITC Claiming Contribution
Black	Probabilistic	78%	8%	14%
Black	Linear	83%	7%	10%
Non-Black	Probabilistic	41%	10%	48%
Non-Black	Linear	32%	10%	57%

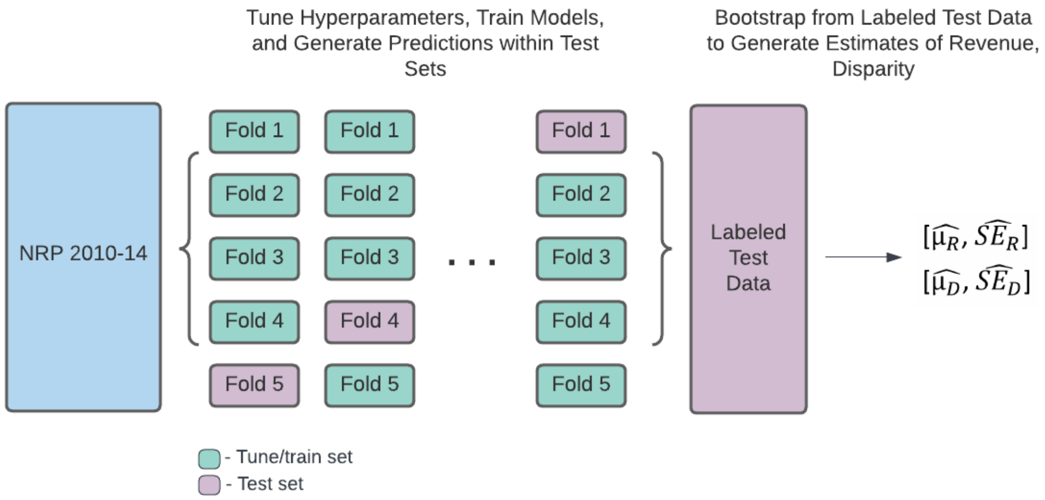
Notes: The table reports the decomposition of disparity under the alternative decomposition methods described in this appendix section and under alternative estimation approaches (linear versus probabilistic estimators). Estimates of the terms used to calculate each component of disparity are reported in Appendix Table D.1.

E Taxpayer Noncompliance Prediction Model

The outcomes of interest Y of the taxpayer non-compliance random forest models are the dollar amount of adjustment in either total tax liability or refundable credit amount following an audit (for the regression models) and a $\{0, 1\}$ indicator of whether non-compliance exceeds \$100 (for the classifier models). The inputs to the model are characteristics of the tax return, denoted X, which include wages and other sources of income, claimed deductions, and flags for whether dependents claimed on the return may violate IRS dependent rules. These features do not include race, gender, age, location, or other demographic variables.

To train and evaluate the models, we first subset the NRP data from tax years 2010-14 to taxpayers claiming the EITC. We then randomly divide the data into 5 folds. We designate 4 of these folds as the training set, and the remaining fold as the test set. We tune the hyperparameters of each model, including the total number of decision trees in each forest, the maximum depth of each decision tree, and the maximum number of features available to each decision tree, using 5-fold cross validation within the training set and random grid search over the space of hyperparameters we consider. We then fit each tuned model on the full set of training data to generate a function $\hat{Y} = m(X)$ which maps features into predictions, and we apply this model to the test set. We repeat this process 5 times, until each observation in the NRP data has a predicted label. The train-test splits are the same for both the regression and classification models. Figure E.1 provides an exposition of the data flow and model training process.

Figure E.1: Data Flow for Random Forest Models


Tune Hyperparameters, Train Models, and Generate Predictions within Test Sets
Bootstrap from Labeled Test Data to Generate Estimates of Revenue, Disparity
NRP 2010-14
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
. . .
Labeled Test Data
$[\widehat{\mu}_R, \widehat{SE}_R]$
$[\widehat{\mu}_D, \widehat{SE}_D]$
- Tune/train set
- Test set

To obtain estimates of disparity and annualized adjustments at each audit rate, we first bootstrap from the population of labeled test data, and then sort observations within the bootstrapped sample by either the magnitude of their predicted noncompliance (for the regression model) or the predicted likelihood of noncompliance above a \$100 threshold (for the classification model). Within each sample, annualized adjustments are given by:

$$R_s = \frac{1}{5} \sum_{t=2010}^{2014} \frac{W_t}{\sum_{i=1}^{n_{st}} w_{ist}} \sum_{i=1}^{n_{st}} (a_{ist} w_{ist} r_{ist}) \tag{6}$$

The rightmost sum computes the total weighted audit adjustments across observations in sample s from tax year t, where a_{ist} indicates whether individual i in sample s and tax year t was audited and $(w_{ist} r_{ist})$ is the weighted adjustment from the audit in 2014 dollars. The term to the left of this sum takes the total sample weights from NRP observations in year t (denoted W_t) over the total sample weight from this year included in sample s, to account for the fact that each fold only contains a portion of the total population available in each study year. We then sum across each of the 5 study years and divide by 5 to approximate one year of annual adjustments in 2014 dollars. Disparity measures are computed within each sample using both the linear and probabilistic estimators described in Section 3, adjusted to account for NRP sample weights. We take the mean and standard error of these measures across the bootstrapped samples to construct our trajectories and 95% confidence intervals.

Oracle adjustments and disparity calculations are analogous to the random forest calculations, with the exception that the data are sorted using true underlying noncompliance, rather than predicted amount or likelihood of noncompliance.

F Additional Factors Potentially Contributing to Audit Disparity

Section 7 of the main text provided evidence that the observed audit disparity among EITC claimants stems in part from audit selection algorithms designed to pursue refundable credit overclaims rather than total underreporting, in conjunction with differences by race in the prevalence of different types of noncompliance. This appendix section explores additional factors that may also contribute to the observed racial disparity.

F.1 Prediction Model Errors

Whereas the refundable credit oracle induces a smaller disparity than what we observe in the operational data, Figure 8 showed that focusing on *predicted* refundable overclaims – where the predictions are obtained from a random forest prediction model trained on the same data that is available to IRS – would lead to a disparity similar to that observed in the operational data. In this subsection, we attempt to better understand why focusing on predicted overclaims can exacerbate the disparity that would be implied by actual differences in refundable credit overclaiming by race.

As a starting point, note that we can express the disparity from a prediction-based algorithm relative to the disparity that would be induced by an oracle in terms of differential false positive and false negative rates by race, where a true positive is defined as a taxpayer's actual refundable credit overclaiming being in the top share of the refundable credit overclaiming distribution (see Appendix F.5 for details). Appendix Table F.1 reports the frequency of these prediction errors by race for our simulated refundable credit prediction algorithm. Relative to non-Black taxpayers, Black taxpayers experience higher false positive rates and lower false negative rates. Both of these errors push in the direction of higher audit rates for Black relative to non-Black taxpayers.

Of course, the distribution of prediction errors in our simulated algorithm might not translate to the actual audit selection algorithm employed by IRS. Motivated by the important role of child eligibility errors documented in the prior subsection, we next consider errors in the actual measure used by IRS to predict which refundable credit claims are based on ineligible children. This variable is used as part of the DDb audit selection process and is constructed to reflect the likelihood that a child claimed on a return violates the required residency or relationship test with respect to the taxpayer. Appendix Table F.3 links IRS-predicted ineligibility with ground-truth data on child eligibility, as determined through NRP audits. As with the simulated refundable overclaim algorithm explored above, this table also shows that the distribution of both false positive and false negative prediction errors push in the direction of higher audit rates for Black taxpayers.

Which features in the IRS algorithm are responsible for the prediction errors we observe? To maintain the confidentiality of the audit selection process, IRS policy prevents us from publicly disclosing which taxpayer characteristics drive audit selection or how its child eligibility prediction variable is constructed. However, the IRS publicly discloses that it draws on administrative social security records, including birth certificates, to help determine whether a taxpayer claiming a child satisfies the required relationship test

(G.A.O., 2015).[43] Appendix Table F.4 investigates missingness of parental information on the birth certificate data that SSA provides to IRS for EITC claimants; whereas mothers are missing at roughly equal rates, children claimed on the returns of Black taxpayers are substantially more likely to be missing information about the identity of their father on their birth certificate (53.6% vs 36.9%).[44] Because of this missingness, this data source is likely to be less reliable at identifying Black fathers who satisfy the relationship test as compared to non-Black fathers. Consistent with this hypothesis, Figure 6 showed that the racial audit disparity is especially large in percentage point terms among unmarried men claiming children for the EITC.

Finally, it may be that certain model features yield an outsized effect on disparity relative to their importance for accuracy; adding or subtracting features could differentially affect the ability of the model to detect overclaiming for Black and non-Black taxpayers, leading to different shares of each group being selected for audit. In addition, although dropping features generally leads to (weakly) lower model accuracy, some features may be important for minimizing overall MSE but less important for accurately ranking the very top of the overclaiming distribution. To explore this possibility, we identified the top 20 features of the refundable credit overclaiming model in terms of importance and plotted their unconditional correlations with race; six of these features exhibited markedly uneven distributions (Appendix Figure F.1). Re-training the refundable overclaiming prediction model without these six features yielded similar performance at the status quo audit rate but substantially lower disparity compared to the baseline refundable credit prediction algorithm (see Appendix Figure F.2).[45]

The finding that prediction errors are disproportionately concentrated among Black EITC claimants suggests there may be opportunities for policymakers to reduce disparities by modifying the predictive algorithm to improve accuracy, even conditional on the objective of prioritizing the detection of refundable credit overclaims. A full exploration of this possibility would require divulging more details of the EITC audit selection algorithm than is possible under IRS policy, however, so we defer additional exploration of this issue to internal IRS analyses.

F.2 High-Risk Signals

Section 7 explored whether differences by race in the dollar value of predicted noncompliance contribute to the observed racial audit disparity. A distinct factor that might contribute to the disparity are differences in informational signals that strongly indicate fraud or other errors. With respect to EITC eligibility, for example, this might take the form of multiple taxpayers claiming the same child in the same year (which is not allowed), or a claim by a

[43] As discussed above, the IRS also draws on administrative child custody data to predict credit eligibility; prior research has documented substantial inaccuracies in that data, but we lack direct evidence on the distribution of those inaccuracies by race (see National Taxpayer Advocate, 2018).

[44] Fathers are more likely to be listed on a child's birth certificate when the mother is married at the time of birth, and as discussed above, marriage rates tend to be lower among Black parents.

[45] We reiterate that because our refundable credit prediction algorithm represents an approximation to the algorithm underlying actual EITC audit selection, we do not conclude that the IRS could achieve this outcome by excluding the same six features we identified. Rather, this exercise illustrates how the IRS's current approach may not be at the accuracy-disparity frontier.

childless taxpayer whose age falls outside of the statutorily prescribed range for eligibility. In such cases, the IRS might choose to prioritize these likely-noncompliant returns for audit, even above other returns with larger expected (but less certain) adjustments.

For several reasons, it appears unlikely that differences by race in the distribution of these high-risk signals significantly contributes to the observed audit disparity. First, many of the apparent errors that fall into this category (such as claimed children with ages outside of the allowable range) are addressed by the IRS outside of the audit workstream, such as through automatic "math error" adjustments to the taxpayer's return, or "Automatic Underreporter" corrections that address discrepancies between what the taxpayer reports and information returns that the IRS receives.[46] As such, errors detected through these programs would not be counted as audits in our data. Second, with respect to multiple taxpayers claiming the same child, our disparity estimates are largely unchanged when we exclude from our analysis operational audits that are focused on this issue (Appendix Table F.5).

To further investigate the role of highly predictive (and therefore hard-to-ignore) signals in driving the observed disparity, we consider simple predictive models to identify features, which, on their own, are highly predictive of non-compliance ("smoking guns"). For each of the most important features in the underreporting prediction model, we train a minimalistic model to predict the presence of non-compliance above a $100 threshold using that feature alone.[47] Appendix Figure F.3 summarizes the performance and disparity of each feature, via these models, on the x-axis and the implied disparity on the y-axis. A "smoking gun" feature that drives disparity would appear as an outlier in terms of both the x and y axes — that is, a model that has much higher performance than others but also much higher disparity. We do not observe a feature that meets this criterion; the apparent outliers in terms of high disparity are near the center of mass of the performance distribution. We interpret this analysis as further evidence against the hypothesis that the disparity is primarily driven by differences in the distribution of high-information signals of noncompliance by race.

F.3 Regression vs Classification Prediction Task

A related possibility involves the IRS selecting audits to maximize the share of returns with a positive adjustment, rather than total dollars of detected overclaims or underreporting (Black et al., 2022). To investigate how this change to the audit objective shapes disparity, we trained random forest classifier models to predict whether an audited return would yield a positive adjustment of $100 or higher (see Appendix E for details).[48] Appendix Figure F.5 shows this aspect of the prediction model objective yields an ambiguous effect on disparity, depending on whether the underlying objective is to predict total underreporting

[46] Children of any age may be claimed for the EITC if they are "permanently and totally disabled", but taxpayers claiming these children must indicate on their return that this exception applies to avoid a math error correction.

[47] For continuous features (38 of the 40 we consider) we use a two-layer decision tree, and for discrete features we use a single layer decision tree. A one (two) layer decision tree can segment observations into two (four) categories; these models are thus extremely simple without imposing monotonicity, and are flexible in that they consider all potential splits for each layer.

[48] The classifier threshold of $100 does not necessarily correspond to the dollar amount used to train any classifier model that the IRS employs. We explore alternative thresholds in Appendix Figure F.4 and find no systematic relationship between disparity and threshold amount.

or refundable credit overclaims. That is, the underreporting classifier yields higher disparity than our baseline underreporting prediction models (which ranks return according to predicted dollars of underreporting), whereas the overclaims classifier yields lower disparity than our baseline overclaims prediction models. Thus, although the question of whether to select audits based on the expected dollars of noncompliance versus the probability of any non-compliance can shape disparity in some settings, it does not appear to explain the audit disparity we observe given IRS's focus on refundable credit overclaims.

F.4 Other Group-Level Differences in Taxpayer Characteristics

Without disparate treatment, the observed audit disparity must arise via group-level differences in observable characteristics between Black and non-Black taxpayers. Hence, with sufficient data and model flexibility (i.e., allowing non-linearities and interactions), it must be the case that controlling for all of the inputs available to IRS would eliminate any estimated independent effect of race in a model of audit selection. Rather than focus on which of the observable characteristics driving audit selection induce the disparity, our main approach has been to investigate how the disparity arises from IRS policy goals and differences in the distribution of types of noncompliance that the individual features are predicting.

Still, it may be helpful to understand the role played by several high-salience taxpayer characteristics in driving the disparity, such as the number of dependents a taxpayer claims or whether a taxpayer uses a tax preparer to file her return. Towards that end, in Appendix B.8 we propose a method for estimating the audit disparity after adjusting for certain observable characteristics between groups, and provide conditions under which we can interpret these adjusted disparity estimates as bounds (analogous to Proposition 1). Implementing this approach, we find that the observed disparity remains after adjusting for potential differences by race in the distribution of taxpayer characteristics related to income, household composition (i.e., marital status and number of dependents), and tax preparation method (Appendix Table F.6). These findings appear to undermine some commonly conjectured candidate sources of group-level differences that could drive the observed disparity.

F.5 Disparity Decomposition with Respect to Underreporting and Prediction Errors

This appendix section derives a simple decomposition (that was referred to in Appendix Section F.1) for an observed disparity in terms of the role of group-level differences in underreporting and errors in which taxpayers are selected. The decomposition also provides insight into the difference in disparity induced by an oracle versus the disparity induced by some other model, such as the refundable credit overclaim prediction model.

Let A^O indicate whether someone was selected by the oracle. Let A^R indicate whether someone selected by the other algorithm.

The rate that a taxpayer from group j is selected by the non-oracle model is given by:

$$
\begin{aligned}
p(A^R = 1 \,|\, B = j) &= P(A^R = 1 \,\&\, A^O = 1 \,|\, B = j) + P(A^R = 1 \,\&\, A^O = 0 \,|\, B = j) \\
&= P(A^O = 1 \,|\, B = j)\, p(A^R = 1 \,|\, A^O = 1, B = j) \\
&+p(A^O = 0 \,|\, B = j)\, p(A^R = 1 \,|\, A^O = 0, B = j) \\
&+ p(A^O = 1 \,|\, B = j) - p(A^O = 1 \,|\, B = j) \\
&= p(A^O = 1 \,|\, B = j) - p(A^O = 1 \,|\, B = j)\,(1 - p(A^R = 1 \,|\, A^O = 1, B = j)) \\
&+ P(A^O = 0 \,|\, B = j)\, p(A^R = 1 \,|\, A^O = 0\,,\, B = j) \\
&= p(A^O = 1 \,|\, B = j) - p(A^O = 1 \,|\, B = j)\, p(A^R = 0 \,|\, A^O = 1, B = j) \\
&+ P(A^O = 0 \,|\, B = j)\, p(A^R = 1 \,|\, A^O = 0, B = j) \\
&= c_j - c_j\, \nu_j + (1 - c_j)\, \pi_j
\end{aligned}
$$

where $c_j = p(A^O = 1 \,|\, B = j)$ is the rate that a taxpayer from group j would be selected by the oracle; $\nu_j = p(A^R = 0|A^O = 1, B = j)$ is the false negative rate (i.e., the probability that a non-compliant taxpayer does not get audited); and $\pi_j = p(A^R = 1|A^O = 0, B = j)$ is the false positive rate (i.e., the probability that a compliant taxpayer would be audited, where "compliant" here refers to whether the taxpayer would be selected by the oracle for audit.

Disparity induced by the oracle is:

$$D_O = p(A^O = 1|B = 1) - p(A^O = 1|B = 0) = c_B - c_N$$

Disparity induced by the other model is:

$$
\begin{aligned}
D_R &= p(A^R = 1|B = 1) - p(A^R = 1|B = 0) \\
&= c_B - c_B\nu_B + (1 - c_B)\pi_B - (c_N - c_N\nu_N + (1 - c_N)\pi_N) \\
&= D_O - (c_B\nu_B - c_N\nu_N) + ((1 - c_B)\pi_B - (1 - c_N)\pi_N)
\end{aligned}
$$

So the "excess disparity" (relative to oracle) of the other model is:

$$\Delta_R = D_R - D_O = ((1 - c_B)\pi_B - (1 - c_N)\pi_N) - (c_B\nu_B - c_N\nu_N)$$

We can add and subtract terms to obtain:

$$
\begin{aligned}
\Delta_R &= c_N\nu_N - c_B\nu_B + (1 - c_B)\pi_B - (1 - c_N)\pi_N \\
&= c_N\nu_N - c_N\nu_B + c_N\nu_B - c_B\nu_B \\
&+ (1 - c_B)\pi_B - (1 - c_N)\pi_B + (1 - c_N)\pi_B - (1 - c_N)\pi_N \\
&= c_N\,(\nu_N - \nu_B) + \nu_B\,(c_N - c_B) + \pi_B\,(c_N - c_B) + (1 - c_N)\,(\pi_B - \pi_N)
\end{aligned}
$$

$$= (1 - c_N)\,(\pi_B - \pi_N) + c_N\,(\nu_N - \nu_B) - D_O\,(\nu_B + \pi_B) \tag{7}$$

The three terms correspond to: (1) differences in the false positive rate by group; (2) differences in the false negative rate by group; and (3) to the extent the oracle would induce a disparity, that disparity will be attenuated by the non-oracle model to the extent that the latter induces accuracy mistakes (of either type).

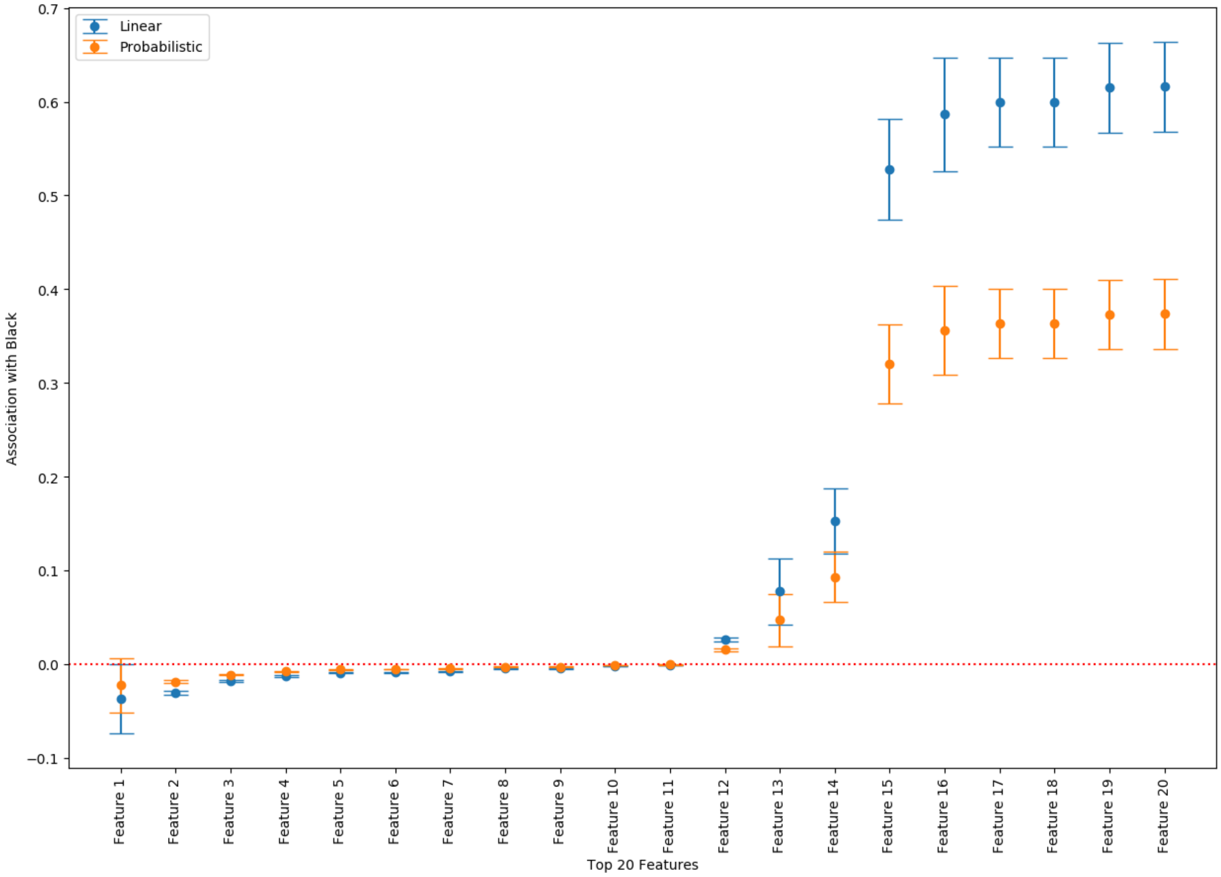
Finally, adding the oracle disparity to both sides of this equation yields an expression for the disparity induced by the non-oracle model:

$$D_R = (1 - c_N)(\pi_B - \pi_N) + c_N(\nu_N - \nu_B) + D_O(1 - \nu_B - \pi_B) \tag{8}$$

Intuitively, the disparity induced by a non-oracle model deviates from the disparity induced by the oracle based on average prediction errors (false positives and false negatives) as well as differences in the prediction errors by race.

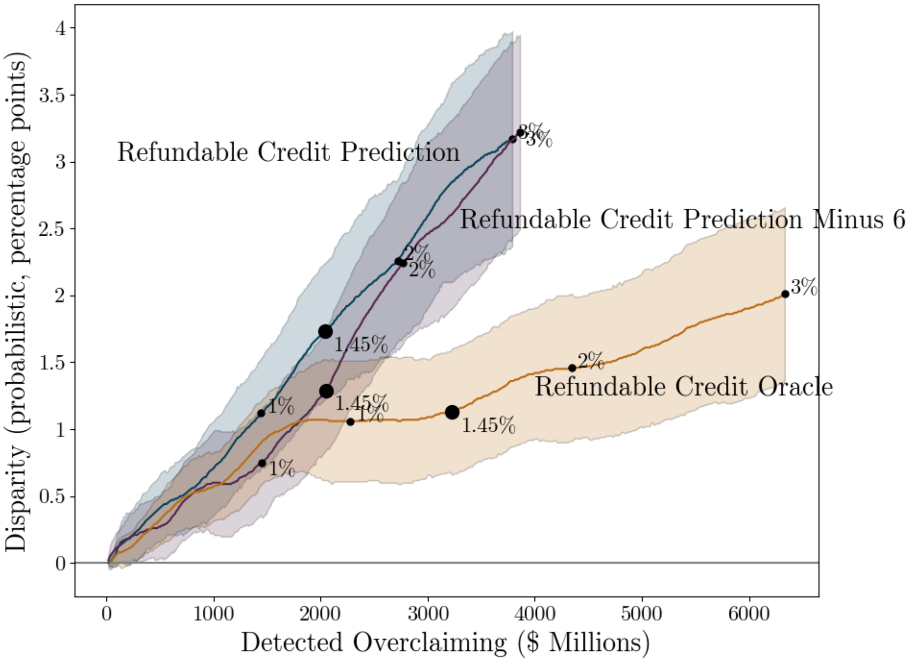
Figures and Tables for Appendix F

Figure F.1: Distributions by Race for the Important Features of the Refundable Credit Prediction Model



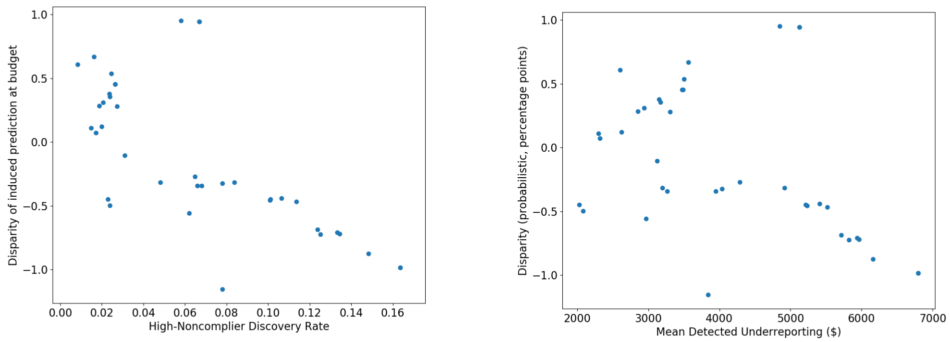
Notes: This figure shows the association between the top 20 most important features in the refundable credit model and the estimated probability a taxpayer is Black. Associations are calculated by first applying the linear and probabilistic estimators to each feature and then dividing the output by the standard deviation of the feature. Standard errors are calculated by dividing the standard error of the linear and probabilistic estimators (calculated from the asymptotic distributions described in Appendix B.3) by the standard deviation of the feature. Feature importance scores are computed as the mean decrease in impurity at each node of the decision tree, averaged over trees in the random forest model.

Figure F.2: Detected Underreporting and Disparity by Algorithm (Minus Identified Features)



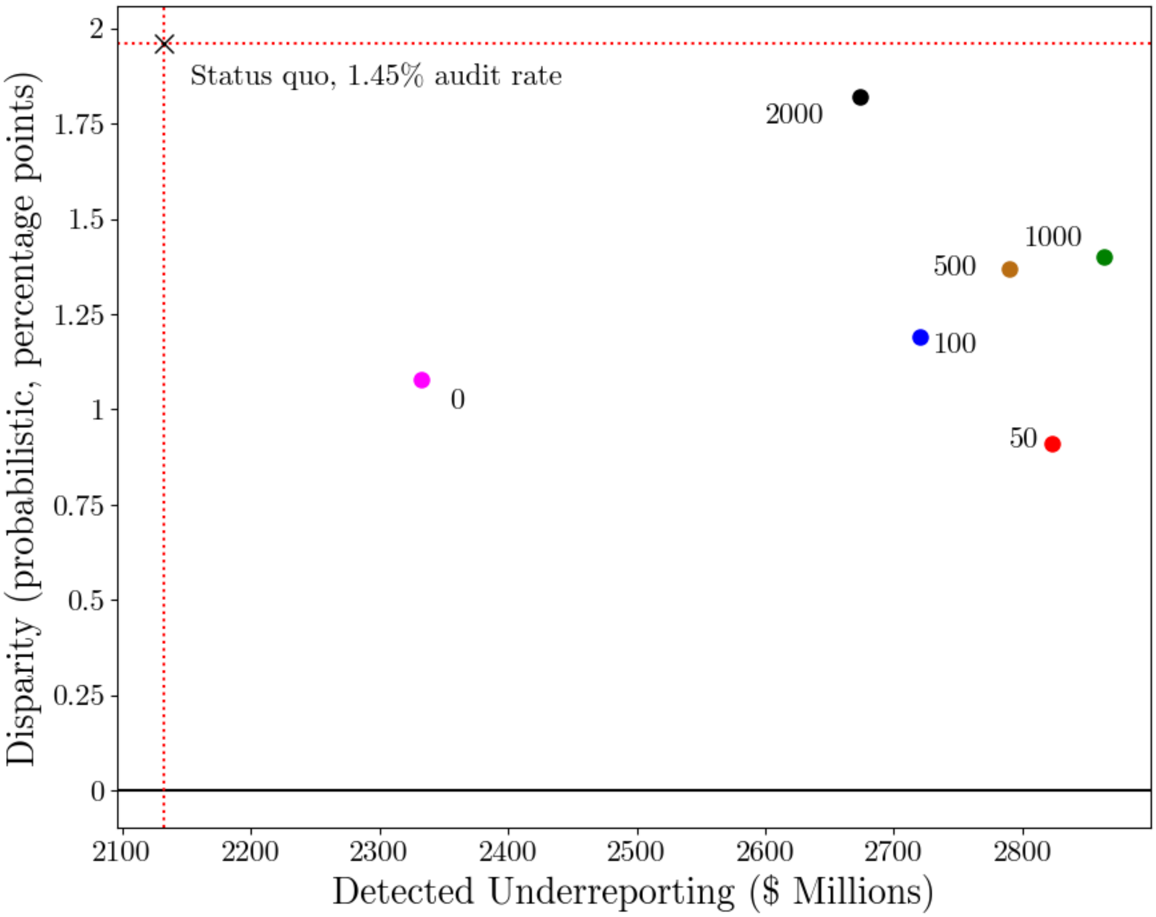
Notes: The figure shows the estimated difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates. Predictive models are trained and evaluated on the set of NRP EITC claimants from 2010-14; see Appendix E for details. The displayed trajectories correspond to the refundable credit prediction regressor (teal), total refundable credit oracle (orange), and refundable credit regressor trained without the 6 features most correlated with race among the twenty most important features of the refundable credit prediction regressor. These correspond to Features 15 through 20 in Appendix Figure F.1. The labeled points along each trajectory represent estimated detected overclaiming and disparity for the specified algorithm at the audit rate specified in the label. The audit rates considered range from 0.1% to 3%. The audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The refundable credit prediction algorithm is based on a random forest regressor trained to predict overclaiming of adjustments to EITC, CTC, and AOTC amounts. The refundable credit prediction minus 6 is the same, but is trained without the 6 features described above. The refundable credit oracle selects returns in descending order of true EITC, CTC, and AOTC overclaiming. Disparity is calculated using the probabilistic disparity estimator. Annualized detected underreporting is calculated as the total detected underreporting (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. The point labeled “Status quo” shows estimated disparity and total underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around disparity estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Figure F.3: Performance and Disparity of Single-Feature Models



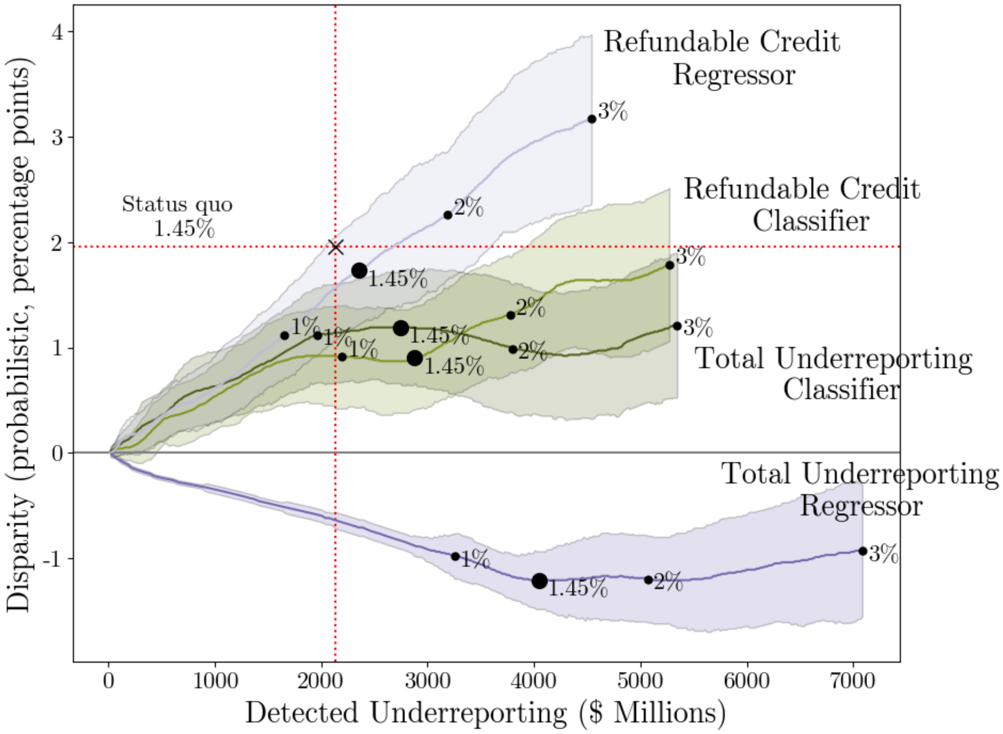
Notes: The figure shows performance and disparity obtained by the single-feature decision trees for predicting total underreporting described in Section 7. Split-points are selected based on the induced impurity of the resulting segmentation. Five instances of the model are trained, each using a different single held-out fold for evaluation and the remaining four folds for training. Model performance is evaluated by selecting the top 1.45% (weighted) EITC NRP returns as ordered by the model predictions. Within each fold, we repeat the selection process 1000 times and permute returns for tie-breaking purposes, and report the average of the specified metric. Each point represents a model trained with a given feature; the x-axis represents performance (right panel: mean non-compliance conditional on selection by the model; left panel: fraction of the top 1.45% of non-compliant taxpayers discovered) and the y-axis represents disparity (as measured by the probabilistic estimator).

Figure F.4: Detected Underreporting and Disparity by Classifier Threshold



Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) for random forest classification models trained on alternative dollar thresholds, under the assumption that 1.45% of the EITC population is selected for audit. Each point corresponds to a different classification model, trained to predict whether or not total adjustments exceed the specified dollar threshold. Disparity is calculated from the probabilistic disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. Annualized detected underreporting is calculated as the total amount of adjustments (positive or negative) imposed on returns selected for audit under the specified audit selection model. Detected underreporting and disparity estimates for all models are constructed using the full set of NRP EITC observations from 2010-14. Details relating to the predictive models and associated estimates are contained in Appendix E. The point labeled "Status quo" shows the estimated disparity in audit rates between Black and non-Black taxpayers and detected underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights.

Figure F.5: Detected Underreporting and Disparity by Algorithm (Regression vs. Classification)



Notes: The figure shows the estimated difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates. Predictive models are trained and evaluated on the set of NRP EITC claimants from 2010-14; see Appendix E for details. The displayed trajectories correspond to the total underreporting classifier (dark green), total underreporting regressor (dark purple), refundable credit classifier (light green), and refundable credit regressor (light purple) algorithms. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified algorithm at the audit rate specified in the label. The audit rates considered range from 0.1% to 3%. The audit rate corresponding to the status quo (1.45%) is denoted by a larger dot. The total underreporting classifier is based on a random forest model trained to predict whether or not underreporting exceeds $100. The total underreporting regressor is based on a random forest model trained to predict total underreporting. The refundable credit classifier is based on a random forest model trained to predict whether or not total adjustments to EITC, CTC, and AOTC amounts exceed $100. The total underreporting regressor is based on a random forest model trained to predict total adjustments to EITC, CTC, and AOTC amounts. Disparity is calculated from the probabilistic disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. Annualized detected underreporting is calculated as the total detected underreporting (positive or negative) imposed on returns selected for audit under the specified audit selection algorithm, scaled to reflect our use of five years of NRP data. The point labeled “Status quo” shows estimated disparity and total underreporting from the 1.45% of EITC returns selected for audit from the population of tax year 2014 returns. All analyses incorporate NRP sampling weights. Bars around each trajectory represent 95% confidence intervals around disparity estimates; they are calculated based on the distribution of estimates from 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix E for details.

Table F.1: Refundable Credit “Excess” Disparity Decomposition (Probabilistic)

	Black	Non-Black
Oracle Selection Rate	2.309	1.230
False Positive Rate	2.236	0.886
False Negative Rate	71.547	82.282
Excess Disparity Contribution from ...		
Scaled Difference in False Positive Rates	1.333	
Scaled Difference in False Negative Rates	0.132	
Attenuation of Oracle Disparity	-0.796	
Total Excess Disparity	0.669	
Oracle Disparity	1.079	
Prediction Model Disparity	1.748	

Notes: The table decomposes the excess disparity between the Refundable Credit Prediction Model and the Refundable Credit Oracle, as observed in Figure 8. All analyses uses NRP EITC claimants 2010-14, and incorporate NRP sampling weights. Quantities are percentage points (0-100). All estimates are based on the probabilistic estimator. The first row reports the audit rate by race when the refundable credit oracle selects 1.45% of EITC claimants. The second row reports the fraction of taxpayers who would be selected by the refundable credit prediction algorithm among those who would not be selected by the refundable credit oracle (“false positives”). The third row reports the fraction of taxpayers who would be not be selected by the refundable credit prediction algorithm among those who would be selected by the refundable credit oracle (“false negatives”). The next three rows correspond to the three terms in the excess disparity decomposition (see Equation (7) in Appendix F.5). The fifth row corresponds to the first term in the decomposition, $(1 - c_N)(\pi_B - \pi_N)$. The sixth row corresponds to the second term in the decomposition, $c_N(\nu_N - \nu_B)$. The seventh row corresponds to the third term in the decomposition, $D_O(\nu_B + \pi_B)$. The eighth row corresponds to the “excess” disparity and is the sum of the three previous rows. The ninth row corresponds to the disparity induced by the Refundable Credit Oracle. The final row corresponds to the disparity induced by the Refundable Credit Prediction Model and is the sum of the two previous rows. Table F.2 presents an analog to this analysis using the linear estimator.

Table F.2: Refundable Credit “Excess” Disparity Decomposition (Linear)

	Black	Non-Black
Oracle Selection Rate	2.732	1.121
False Positive Rate	2.772	0.750
False Negative Rate	68.927	83.543
Excess Disparity Contribution from ...		
Scaled Difference in False Positive Rates	1.999	
Scaled Difference in False Negative Rates	0.164	
Attenuation of Oracle Disparity	-1.155	
Total Excess Disparity	1.007	
Oracle Disparity	1.611	
Prediction Model Disparity	2.611	

Notes: The table replicates Appendix Table F.1 using the linear estimator. Specifically, the table decomposes the excess disparity between the Refundable Credit Prediction Model and the Refundable Credit Oracle, as observed in Figure 8. All analyses uses NRP EITC claimants 2010-14, and incorporate NRP sampling weights. Quantities are percentage points (0-100). All estimates are based on the linear estimator. The first row reports the audit rate by race when the refundable credit oracle selects 1.45% of EITC claimants. The second row reports the fraction of taxpayers who would be selected by the refundable credit prediction algorithm among those who would not be selected by the refundable credit oracle (“false positives”). The third row reports the fraction of taxpayers who would be not be selected by the refundable credit prediction algorithm among those who would be selected by the refundable credit oracle (“false negatives”). The next three rows correspond to the three terms in the excess disparity decomposition (see Equation (7) in Appendix F.5). The fifth row corresponds to the first term in the decomposition, $(1 - c_N)(\pi_B - \pi_N)$. The sixth row corresponds to the second term in the decomposition, $c_N(\nu_N - \nu_B)$. The seventh row corresponds to the third term in the decomposition, $D_O(\nu_B + \pi_B)$. The eighth row corresponds to the “excess” disparity and is the sum of the three previous rows. The ninth row corresponds to the disparity induced by the Refundable credit Oracle, and the final row corresponds to the disparity induced by the Refundable Credit Prediction Model.

Table F.3: High-Risk Classification in NRP and DDB

Panel A: Black			
	Not High-Risk (NRP)		High-Risk (NRP)
Not High-Risk (DDb)	0.68		0.22
High-Risk (DDb)	0.04		0.07
False Positive Rate		0.05	
False Negative Rate		0.76	
Panel B: Non-Black			
	Not High-Risk (NRP)		High-Risk (NRP)
Not High-Risk (DDb)	0.79		0.16
High-Risk (DDb)	0.03		0.03
False Positive Rate		0.03	
False Negative Rate		0.84	

Notes: The table displays the estimated distribution of Black (Panel A) and non-Black (Panel B) taxpayers for two tests of high-risk classification, one imputed in the Dependent Database, and one determined by line-by-line audits in NRP. Each cell shows the fraction of the group which was classified as either high risk or not high-risk according to the test result in DDb and high risk or not high-risk according to the test result as verified in the NRP. Estimates are computed using the probabilistic estimator and weighted using NRP weights to be representative of the full taxpayer population of EITC claimants with dependents. We report results for all taxpayers in our sample that have a non-missing high-risk indicator in the NRP, and in cases when such a taxpayer does not have a risk-indicator in DDb, we impute not high-risk. The false positive rate is calculated as the share of Black/non-Black taxpayers that are classified as high-risk by DDb but not high-risk by NRP, divided by the share of Black/non-Black taxpayers that are classified as not high-risk by NRP. The false negative rate is calculated as the share of Black/non-Black taxpayers that are classified as high-risk by NRP but not high-risk by DDb, divided by the share of Black/non-Black taxpayers that are classified as high-risk by NRP.

Table F.4: Missingness of Parental Social Security Numbers for EITC-Claimed Dependents

	Overall	Black	Non-Black
Missing Mother's SSN	15.40	15.02	15.51
Missing Father's SSN	40.56	53.55	36.85

Notes: The table reports the rate of missing parental information on birth certificates for EITC-claimed children on returns for tax year 2014. Units are percentage points (0-100). Rates are calculated at the return-level for the overall population (column 1), the population of Black taxpayers (column 2), and the population of non-Black taxpayers (column 3). The last two columns are calculated using BIFSG estimates and the probabilistic estimator. Note that taxpayer race in columns 2 and 3 refers to the estimated race of the taxpayer claiming the child, not the race of the child.

Table F.5: Disparity Estimates (Excluding Duplicate Child Claim Audits)

Estimator	Full Population	EITC	Non-EITC
	(1)	(2)	(3)
Linear	1.283	2.792	0.165
	(0.004)	(0.009)	(0.003)
Probabilistic	0.776	1.887	0.093
	(0.003)	(0.008)	(0.002)
N	148,305,318	28,338,472	119,966,846

Notes: The table shows estimated audit rate disparities using both the linear and the probabilistic estimators for audits initiated because the same child was claimed as a dependent on multiple returns. Units are percentage points (0-100). The Black/non-Black audit disparity is shown for the full population (column 1), the EITC population (column 2) as well as the non-EITC population (column 3). Standard errors, reported in parentheses, are calculated from the asymptotic distributions described in Appendix B.3. Each displayed disparity estimate is in terms of percentage points and is statistically different from zero ($p < .01$).

Table F.6: Subgroup Disparity Estimates

Estimator	Overall	Income Category	Family Type	Preparer	Combined
Linear	2.900	2.379	2.413	2.899	2.087
	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)
Probabilistic	1.960	1.627	1.636	1.963	1.420
	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)

Notes: The table reports the conditional disparity as described in Section B.8.2. For each feature considered, we compute the D_x^p and D_x^l by applying the probabilistic and linear estimators, respectively, to the set of taxpayers whose value of the feature is x. Then we estimate bounds on $\mathbb{E}[D_x]$ by averaging D_x^p or D_x^l over x. The standard error on each bound is the square root of the sum over all subgroups of the standard error of the disparity estimate within the subgroup squared times the share of taxpayers in that subgroup squared. We repeat this for: the overall population (as a reference); Income Category, constructed as a cross-product of total adjusted gross income quintiles and Schedule C Income Status (positive amount, non-positive amount, or none); Family Type, constructed as a cross-product of family status (married, single male, or single female) and number of dependents claimed (0, 1, 2, or 3+); Preparer, i.e. whether the taxpayer prepares their own taxes; and Combined, i.e. the cross-product of all of the above.

G Appendix References

References

Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference, 67*(2):209–226. (Cited on Appendix-49)

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1479–1503. (Cited on 7, Appendix-78)

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174. (Cited on Appendix-56)

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pages 339–348). (Cited on 2, 7, 17, Appendix-47)

Delevoye, A. & Sävje, F. (2020). Consistency of the Horvitz–Thompson estimator under general sampling and experimental designs. *Journal of Statistical Planning and Inference*, 207:190–197. Elsevier. (Cited on Appendix-49)

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741. https://doi.org/10.1109/TPAMI.1984.4767596 (Cited on Appendix-56)

Government Accountability Office. (2015). IRS return selection: Wage and investment division should define audit objectives and refine other internal controls. (Cited on 10, 41, Appendix-77)

Guyton, J. & Hodge, R. (2014). The compliance costs of IRS post-filing processes. *IRS Research Bulletin.* (Cited on Appendix-34)

Lu, B., Wan, J., Ouyang, D., Goldin, J., & Ho, D. E. (2024). Quantifying the uncertainty of imputed demographic disparity estimates: The dual-bootstrap. (Cited on 24, Appendix-13, Appendix-35)

National Taxpayer Advocate. (2018). Annual report to Congress 2018: Most serious problem 11. (Cited on Appendix-77)

Robinson, P. M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2):234–238 (Cited on Appendix-49)

Tzioumis, K. (2018). Demographic aspects of first names. *Scientific Data*, 5(1):1–9 (Cited on 19, Appendix-29)

Virtanen, P., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272. (Cited on Appendix-67)