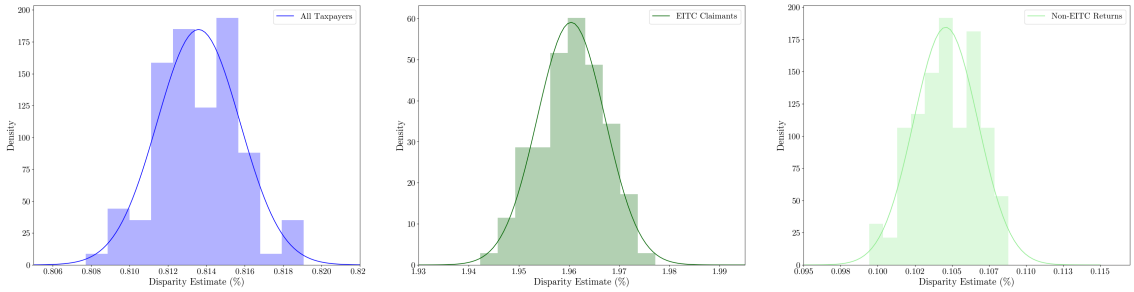
Online Appendix to Measuring and Mitigating Racial Disparities in Tax Audits

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

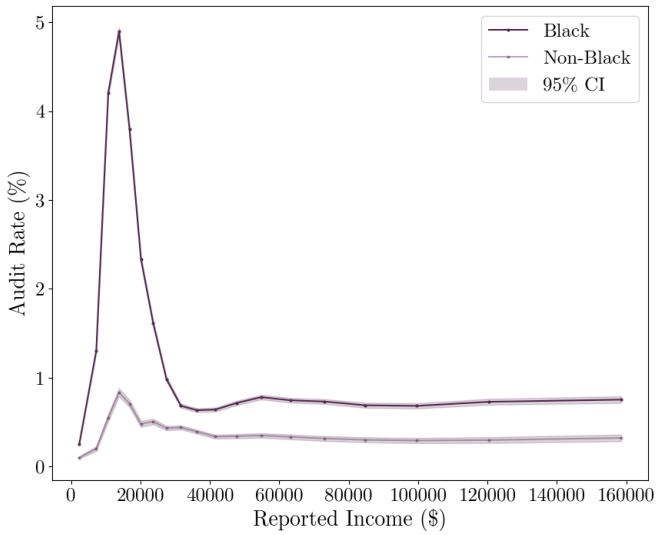
A Additional Tables and Figures

Figure A.1: Statistical Variation of Imputation-Based Disparity Estimates



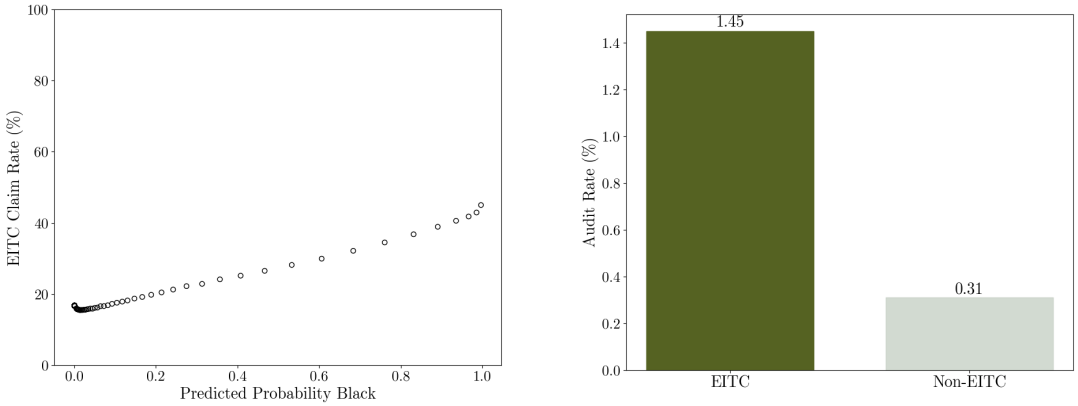
Notes: The figure shows statistical variation in the estimated audit rate disparity as estimated by the method of composition (Tanner, 1996); it captures uncertainty both due to measurement uncertainty (uncertainty around true race) as well sampling variability (i.e. the probabilistic nature of an audit). To obtain these estimates, we begin with the operational audit data and construct a counterfactual data set. This data set is constructed by realizing, for every individual, a Bernoulli-drawn indicator for Black self-report status with probability given by each individual's BIFSG-predicted probability of Black self-report status (but leaving the audit status as is). Given this counterfactual data set, we re-estimate the Black/non-Black disparity using linear regression with a dummy; we can interpret the coefficient and standard deviation as parameters for a posterior distribution on disparity *given* this counterfactual data set. We then draw an observation from a normal distribution with mean and variance parameterized accordingly, and save this observation as a single realized disparity estimate. We then repeat this entire process 100 times, obtaining a histogram and fitting a normal distribution to these drawn observations to obtain overall uncertainty. The panels show the disparity estimate distributions generated as the outcome of this process estimated for all 2014 taxpayers, EITC claimants, and non-EITC claimants, respectively. The y-axis reports the frequency of the disparity estimates displayed in percentage points on the x-axis.

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



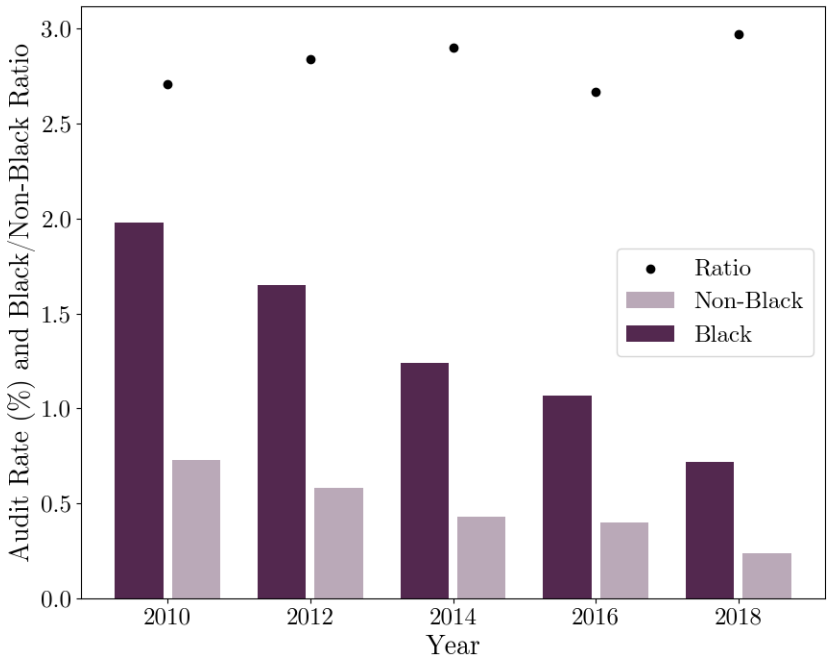
Notes: The figure shows the estimated disparity in audit rates, using the linear disparity estimator, between Black and non-Black taxpayers across bins of reported Adjusted Gross Income (AGI) for returns filed in tax year 2014. Taxpayers are grouped into 20 equal-sized bins. Disparity is calculated within each bin from the linear disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. The x-axis is limited to returns that report AGI under $200,000. The shaded area around the line shows the 95% confidence interval, derived from the asymptotic distributions described in Appendix B.3.

Figure A.3: Estimated Audit Rates by EITC Claim Status



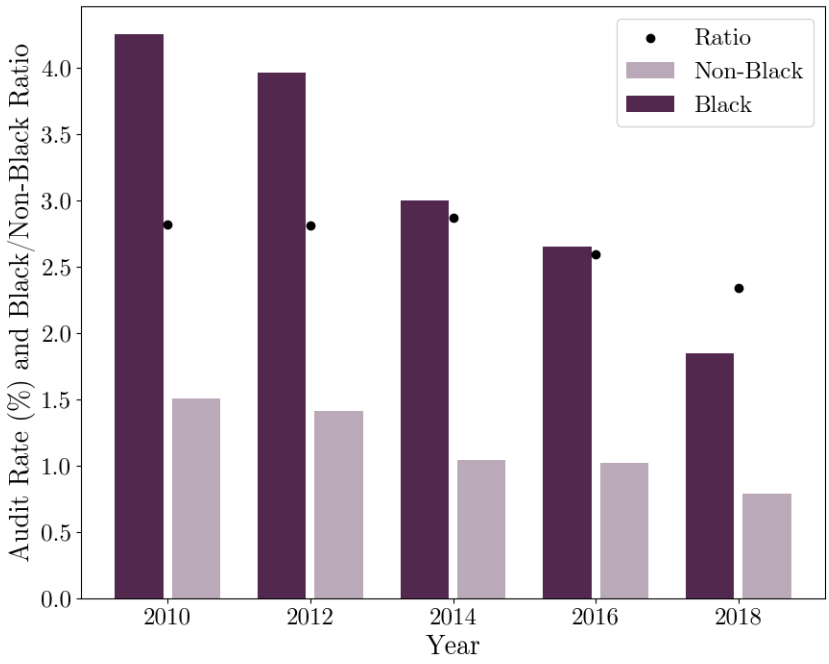
Notes: The figure shows the relationship between audits and EITC claim status among taxpayers filing returns for tax year 2014. Left: Binned scatterplot of EITC claim rate by BIFSG-predicted probability that a taxpayer is Black. Taxpayers have been grouped into 100 equal-sized bins. Right: Audit rates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

Figure A.4: Estimated Audit Rate Disparity by Year



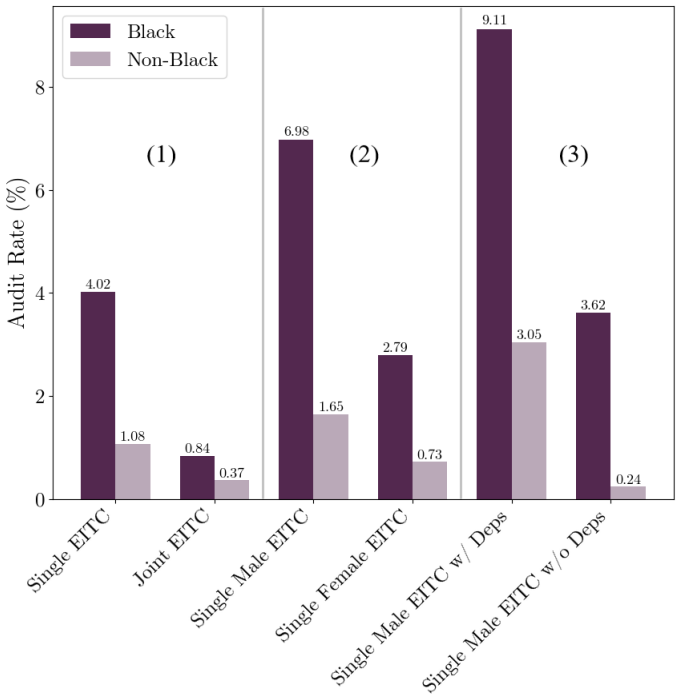
Notes: The figure reports the estimated audit rates among Black and non-Black taxpayers for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). "Ratio"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

Figure A.5: Estimated Audit Rate Disparity Among EITC Claimants by Year



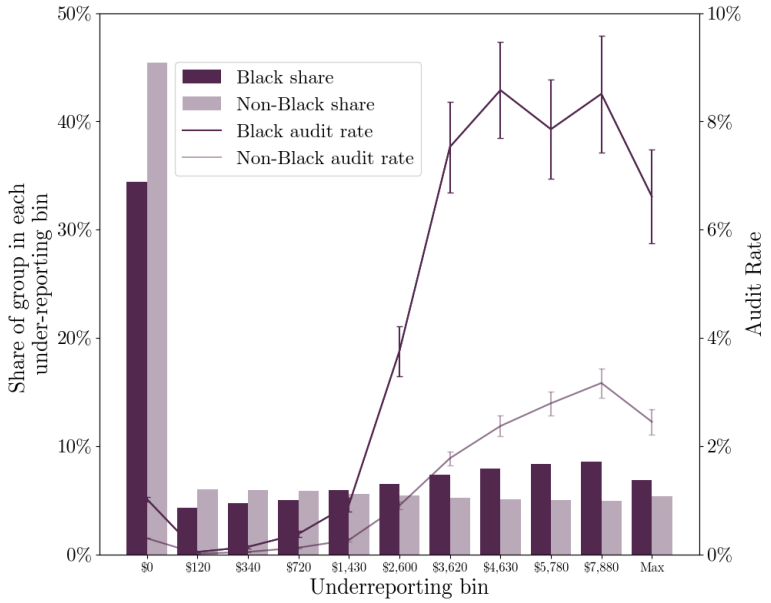
Notes: The figure reports the estimated audit rates among Black and non-Black EITC claimants for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). "Ratio"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

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



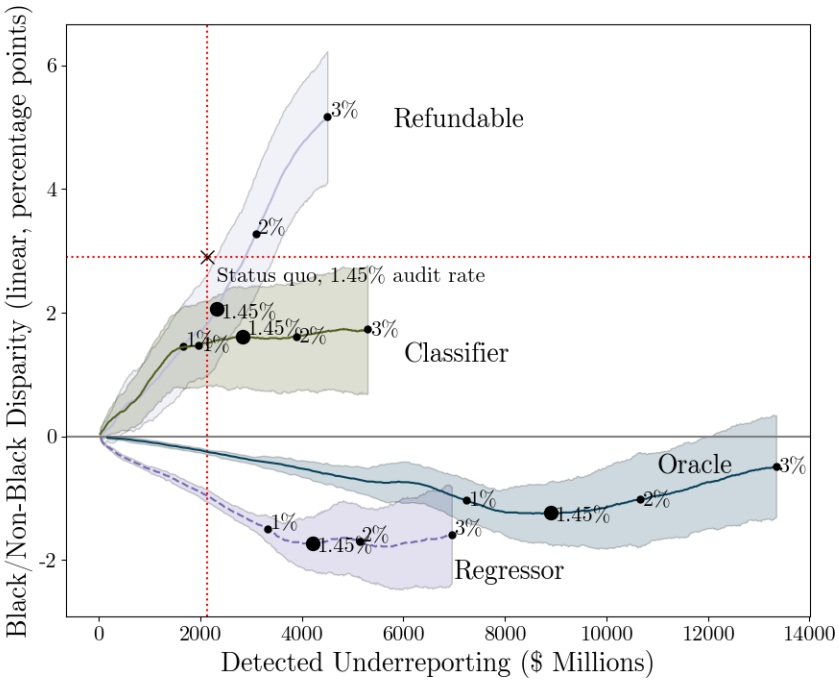
Notes: The figure shows the estimated audit rate among the specified subgroups of Black and non-Black taxpayers. Conditional audit rates by race are calculated using the linear audit rate estimator applied to BIFSG-predicted probabilities that a taxpayer is Black. Panel (1) splits EITC claimants by single 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: Racial Audit Disparity Among EITC Claimants by Underreported Taxes (Linear Estimator)



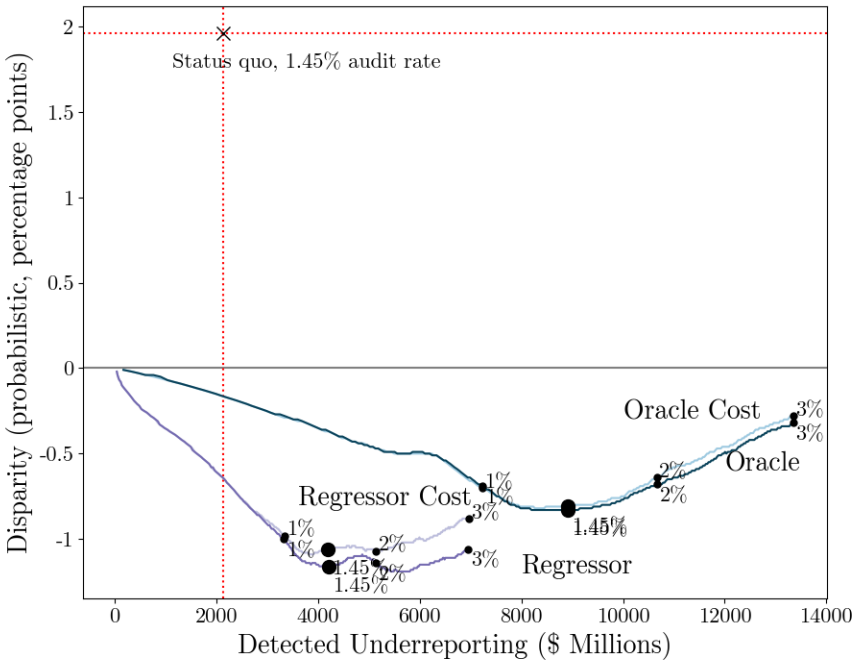
Notes: The figure shows the estimated audit rates for Black and non-Black EITC claimants, respectively, by under-reported taxes. Taxpayers are binned into 11 categories: those with less than $1 of under-reporting, and 10 equal deciles of taxpayers with positive under-reporting. Under-reporting deciles are defined based on the NRP. Estimated audit rates by race are calculated using the linear disparity 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.8: Detected Underreporting and Disparity by Algorithm (Linear Estimator)



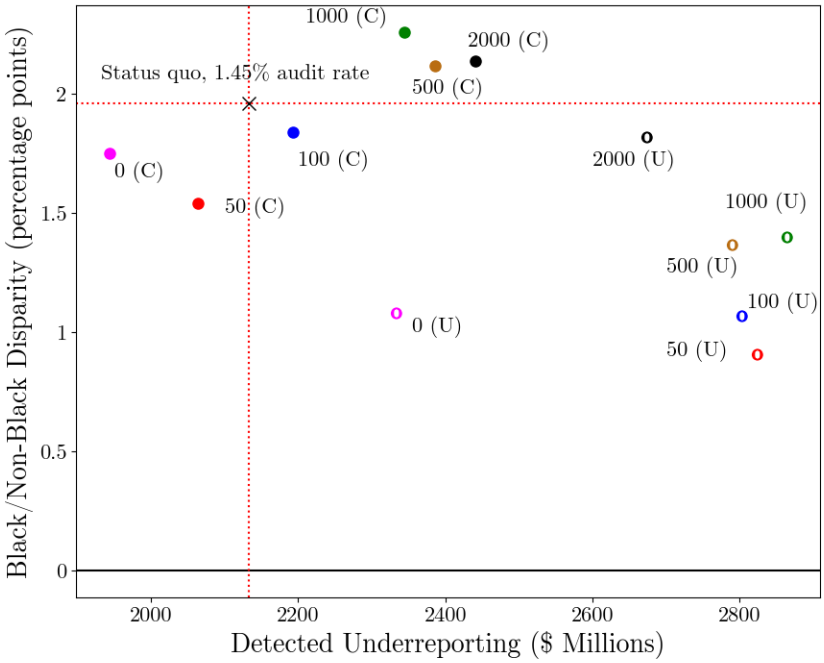
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative models for selecting EITC audits and under alternative audit rates. The trajectories correspond to the oracle (blue), random forest regressor (purple), and random forest classifier (green) models. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified model at the audit rate specified in the label. For each model, 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 regression model is trained to predict total adjustments. The classification model is trained to predict whether or not total adjustments exceed $100. The oracle selects returns according to their true underreporting. Disparity is calculated from the linear 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 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 F. 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 using the standard deviation of estimated disparity across bootstrap samples from the full set of NRP EITC claimants; see Appendix F for details. The p-value for the difference in disparity induced by the classifier and regressor models at the status quo EITC audit rate is less than 0.001; it is obtained from the distribution of the difference in audit rates for Black and non-Black taxpayers from each bootstrapped sample.

Figure A.9: Allocating Audits Based on Underreporting Net of Audit Costs



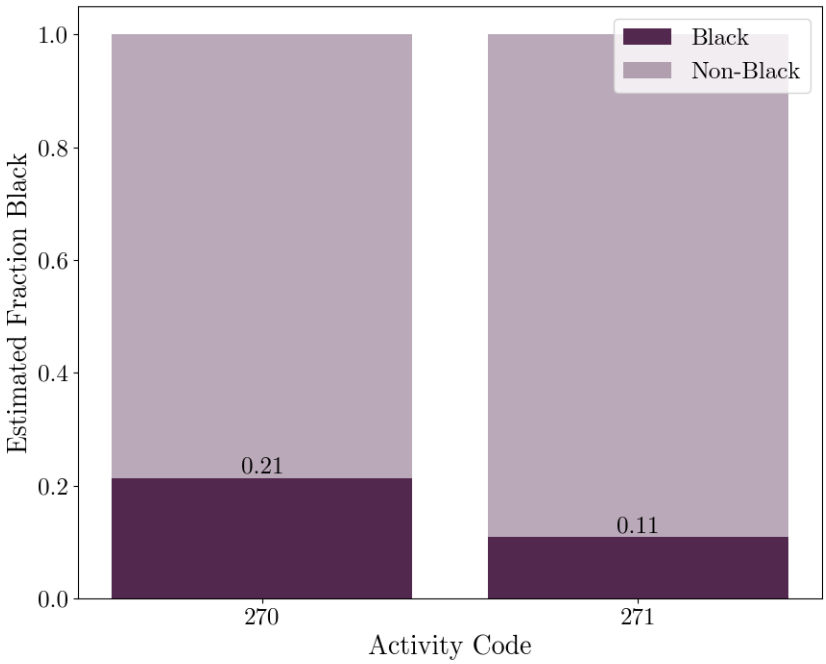
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative assumptions about whether returns are selected for audit based on detected underreporting (gross of audit costs, as in our other analyses) or based on detected underreporting minus expected audit costs. Underreporting is based on either the oracle or the random forest regressor model, 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 business return (activity code 270) is $385, whereas the average cost of auditing an EITC non-business return (activity code 271) is $29. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified model 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 model, 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 model. 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 F. 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.10: Detected Underreporting and Disparity by Classifier Threshold



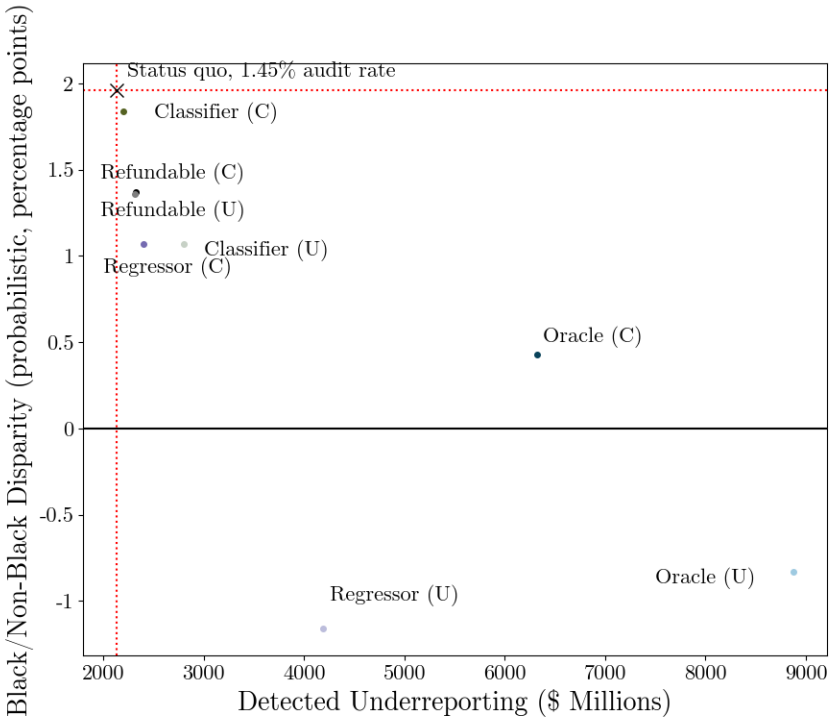
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) for random forest classification models trained on alternative dollar thresholds, under the assumption that 1.45% of the EITC population is selected for audit. Each point corresponds to a different classification model, trained to predict whether or not total adjustments exceed the specified dollar threshold. The solid dots correspond to models that are constrained to match the status quo allocation of audits between EITC business and non-business activity codes, as described in Section 3.4. The hollow dotes indicate unconstrained models. 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 F. 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.11: Racial Composition of EITC Activity Codes



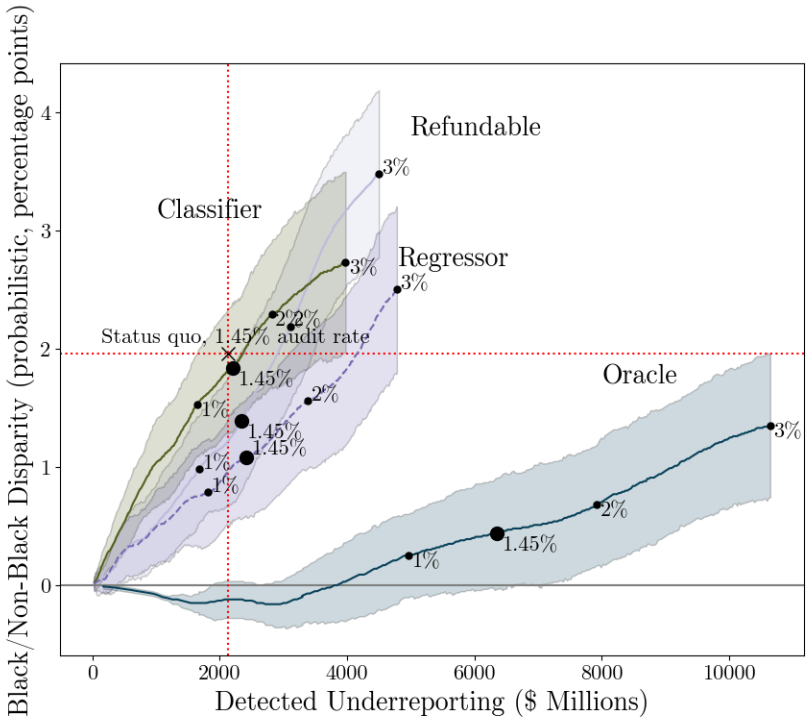
Notes: The figure shows the share of taxpayers who are Black and non-Black across the two activity codes into which EITC tax returns are categorized: returns with substantial business income (activity code 271) and returns without substantial business income (activity code 270). Shares are estimated using the probabilistic estimator described in Section 3.3.

Figure A.12: Effect of Audit Allocation Constraints on Detected Underreporting and Disparity



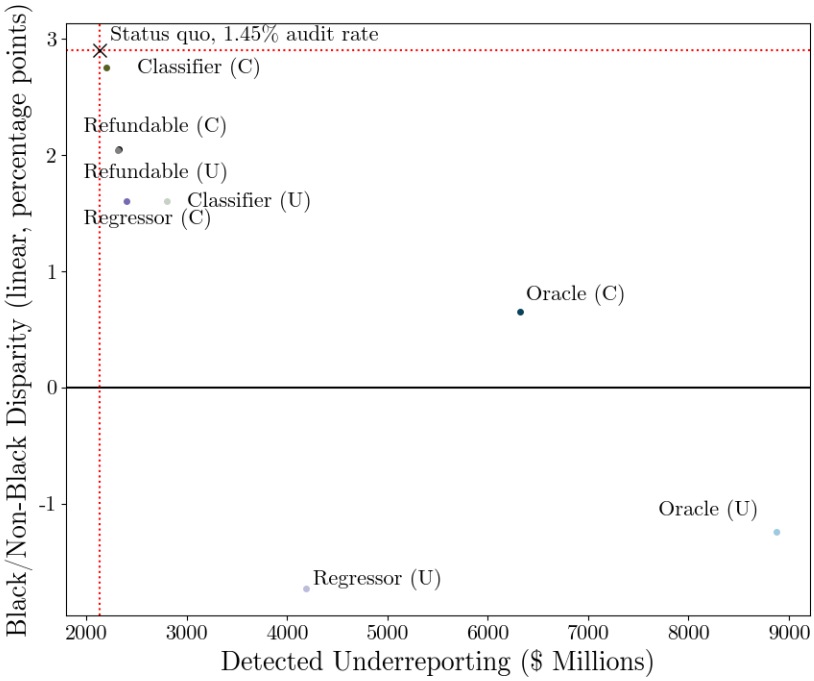
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) for alternative audit selection models, under the assumption that 1.45% of the EITC population is selected for audit. The points correspond to the unconstrained oracle (light blue), constrained oracle (dark blue), unconstrained random forest regressor (light purple), constrained random forest regressor (dark purple), unconstrained random forest classifier (light green), constrained random forest classifier (dark green), unconstrained refundable credit regressor (gray), and constrained refundable credit regressor (black) models at the status quo EITC audit rate of 1.45%. The random forest models are regression models (i.e., they are trained to predict total adjustments). "Constrained" indicates that the model's allocation of audits between EITC business and non-business activity codes is constrained to match the status quo allocation. Disparity is calculated from the weighted disparity estimator applied to BIFSG-derived probabilities that a taxpayer is Black. For a similar analysis using the linear disparity estimator, see Appendix Figure A.14. 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 F. 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.13: 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 models for selecting EITC audits and under alternative audit rates, where each model's allocation of audits between EITC business and non-business activity codes is constrained to match the status quo allocation. The trajectories correspond to the oracle (blue), random forest regressor (purple), and random forest classifier (green) models. The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified model at the audit rate specified in the label. For each model, 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 regression model is trained to predict total adjustments. The classification model is trained to predict whether or not total adjustments exceed $100. The oracle selects returns according to their true underreporting. 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 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 F. 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 using the standard deviation of estimated disparity across bootstrap samples from the full set of NRP EITC claimants; see Appendix F for details. The p-value for the difference in disparity induced by the classifier and regressor models at the status quo EITC audit rate is less than 0.001; it is obtained from the distribution of the difference in audit rates for Black and non-Black taxpayers from each bootstrapped sample.

Figure A.14: Effect of Audit Allocation Constraints on Detected Underreporting and Disparity (Linear Estimator)



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 alternative audit selection models, under the assumption that 1.45% of the EITC population is selected for audit. The points correspond to the unconstrained oracle (light blue), constrained oracle (dark blue), unconstrained random forest regressor (light purple), constrained random forest regressor (dark purple), unconstrained random forest classifier (light green), constrained random forest classifier (dark green), unconstrained refundable credit regressor (gray), and constrained refundable credit regressor (black) models at the status quo EITC audit rate of 1.45%. The random forest models are regression models (i.e., they are trained to predict total adjustments). "Constrained" indicates that the model's allocation of audits between EITC business and non-business activity codes is constrained to match the status quo allocation. Disparity is calculated from the linear 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 F. 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.

Table A.1: Audit Frequency by Timing and Type

Panel A: EITC Population			
	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%)
Panel B: Activity Code 270			
	Correspondence	Field/Office	All Audit Types
Pre-Refund	260,841	0	260,841
	(67.8%)	(0%)	(67.8%)
Post-Refund	109,549	14,355	123,904
	(28.5%)	(3.7%)	(32.2%)
All Audit Times	370,390	14,355	384,745
	(96.3%)	(3.7%)	(100%)
Panel C: Activity Code 271			
	Correspondence	Field/Office	All Audit Types
Pre-Refund	4,922	0	4,922
	(27.6%)	(0%)	(27.6%)
Post-Refund	2,943	9,959	12,902
	(16.5%)	(55.9%)	(72.4%)
All Audit Times	7,865	9,959	17,824
	(44.1%)	(55.9%)	(100%)

Notes: The table reports the frequency of audits of 2014 tax returns 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). Panel A reports audit frequencies for all taxpayers claiming the EITC; Panels B and C are limited to EITC claimants who fall into activity codes 270 (gross business income below $25,000) and 271 (gross business income above $25,000), respectively.

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. Our main sample is constituted by rows 1 through 7. 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.

Table A.3: Calibration Metrics for BIFSG Predictions.

Metric	Full Population		EITC Population	
	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	0.06317	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	0.05036	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	0.5170	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 as measured in the North Carolina data through various metrics which capture different aspects of performance. We use columns to demarcate different versions 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
	(1.95)	(1.90)	(8.00)	(7.92)	(1.32)	(1.12)
E[cov(Y,b)|B]	2.03***	2.28***	8.76***	9.62***	0.004	-0.3
	(0.20)	(0.33)	(0.91)	(1.35)	(0.13)	(0.24)
Weighted		x		x		x
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.2. Standard errors are displayed in parentheses. The displayed estimates and standard errors are multiplied by 10^4. Stars correspond to p-values derived from one-sided hypothesis tests. $^{*}: P < .10;^{**}: P < .05;^{***}: P < .01$.

Table A.5: Linear and Probabilistic Disparity Estimates

Estimator	Full Population (1)	EITC (2)	Non-EITC (3)
Linear	1.344	2.900	0.185
	(0.004)	(0.009)	(0.003)
Probabilistic	0.813	1.960	0.104
	(0.003)	(0.008)	(0.002)
N	148,305,318	28,145,049	118,936,909

Notes: The table shows estimated audit rate disparities using both the linear and the probabilistic estimators. 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 A.6: Audit Disparity by Audit Timing and Audit Type

Estimator	Audit Timing		Audit Type	
	Pre Refund	Post Refund	Correspondence	Field/ Office
	(1)	(2)	(3)	(4)
Panel A: Full Population				
Linear	0.941	0.403	1.328	0.016
	(0.003)	(0.002)	(0.004)	(0.001)
Probabilistic	0.569	0.244	0.804	0.010
	(0.002)	(0.002)	(0.003)	(0.001)
N	148,305,318	148,305,318	148,305,318	148,305,318
Panel B: EITC Population				
Linear	2.194	0.706	2.890	0.010
	(0.008)	(0.005)	(0.009)	(0.002)
Probabilistic	1.483	0.477	1.953	0.007
	(0.007)	(0.004)	(0.008)	(0.001)
N	28,145,049	28,145,049	28,145,049	28,145,049
Panel C: Non-EITC Population				
Linear	0.010	0.174	0.160	0.025
	(0.001)	(0.002)	(0.002)	(0.001)
Probabilistic	0.006	0.099	0.090	0.014
	(0.001)	(0.002)	(0.002)	(0.001)
N	118,936,909	118,936,909	118,936,909	118,936,909

Notes: The table shows estimated audit rate disparity by audit timing (pre-refund or post-refund audits, in columns 1 and 2) and by audit type (correspondence or field/office audits, in columns 3 and 4). In each column, the outcome is a binary indicator for being selected for the specified category of audit. Units are percentage points (0-100). Audit rate disparities are presented for both the linear and the probabilistic estimator. 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 A.7: Audit Disparity Robustness Checks

	BIFSG	Gibbs	Re-calibrated Unweighted	Re-calibrated Weighted
	(1)	(2)	(3)	(4)
Panel A: Full Population				
Linear	1.30	1.57	1.623	1.543
	(0.005)	(0.04)	(0.005)	(0.004)
Probabilistic	0.811	1.04	0.774	0.735
	(0.004)	(0.04)	(0.003)	(0.003)
N	107,624,714	1,379,130	148,305,318	148,305,318
Panel B: EITC population				
Linear	3.01	2.83	3.50	3.33
	(0.01)	(0.08)	(0.01)	(0.01)
Probabilistic	2.15	2.04	1.853	1.816
	(0.01)	(0.07)	(0.008)	(0.008)
N	19,234,523	281,297	28,145,049	28,145,049
Panel C: Non-EITC Population				
Linear	0.176	0.17	0.223	0.212
	(0.003)	(0.03)	(0.003)	(0.002)
Probabilistic	0.106	0.10	0.099	0.093
	(0.002)	(0.02)	(0.002)	(0.002)
N	87,632,261	1,097,833	118,936,909	118,936,909

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). Heteroskedasticity-robust 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 Section C.2. 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.8: Effect of Controls on Estimated Disparity in EITC Audit Rates.

	Baseline	Income Percentiles	Marital Status x EITC Dependents	Income Percentiles x Marital Status x EITC Dependents
	(1)	(2)	(3)	(4)
Disparity	2.900	2.635	2.381	2.072
	(0.009)	(0.009)	(0.009)	(0.009)
N	28,145,049	28,145,049	28,145,049	28,145,049

Notes: The table reports the coefficient on a taxpayer's estimated race from a linear regression of audit status on estimated race and the specified controls. Units are percentage points (0-100). The analysis is restricted to the EITC population. Heteroskedasticity-robust standard errors are reported in parentheses. Column 1 shows the baseline disparity (without additional controls). Column 2 includes fixed effects for income percentiles. Column 3 includes controls for family size (marital status interacted with the number of children claimed for the EITC). Column 4 includes each interaction of family size and income (marital status by children claimed by income percentile). Each displayed disparity estimate is statistically different from zero ($p < .01$).

B Additional Results Relating to Disparity Estimation

In this section, we provide additional results relating to disparity estimation. We first derive the BIFSG equation. We then provide a proof of Proposition 1 as a special case of a more general result, allowing for mis-calibration of the estimated taxpayer race probabilities (B.2). We next discuss statistical inference, and provide the asymptotic distributions of the linear and probabilistic disparity estimators (B.3). Third, we consider weighted versions of the linear and probabilistic disparity estimators (B.4). Finally, as a robustness check, we consider a linear re-calibration exercise for the estimated taxpayer race probabilities, using the North Carolina data as ground truth (B.5).

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.

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

2.

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

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

$$D_p \leq D \leq D_l \tag{1.3}$$

4. Suppose $\mathbb{E}[\mathrm{Cov}(Y, B|b)] \leq 0$ and $\mathbb{E}[\mathrm{Cov}(Y, b|B)] \leq 0$. 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 $D_p = D_l \cdot \frac{\mathrm{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}$ (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:

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

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 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:

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

Finally, we recall that $D_l = \frac{\mathrm{Cov}(Y,b)}{\mathrm{Var}(b)}$ by construction; substituting in $\mathrm{Cov}(Y,b) = D_l \mathrm{Var}(b)$ 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 $\mu = \mathrm{Cov}(\mathbb{E}[\eta|b], \mathbb{E}[\varepsilon|b])$, where η denotes the residual from the linear projection of Y on b.

Then:

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

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

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, $\mathrm{Cov}(b,\eta) = \mathrm{Cov}(B,\nu) = 0$. 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,

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

Using the law of total covariance, we can write:

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

The latter term can be expanded as:

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

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

Next, note that

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

Combining these results, we have:

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

From the definition of ε, we have:

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

Thus

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

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

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

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

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

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

Note that:

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

By the same logic:

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

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

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

Rearranging, and recalling that $D_p=\beta\frac{\operatorname{Var}(b)}{\mathbb{E}[b](1-\mathbb{E}[b])}$ 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

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

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

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

yields Proposition 1.2.

Proposition 1.1 follows by again substituting in $\text{Cov}(b, \varepsilon) = 0$ 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 $\mathbb{E}[\text{Cov}(Y, b|B)]$ and $\mathbb{E}[\text{Cov}(Y, B|b)]$ are met. □

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

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

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}:

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

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

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

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

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

4. Suppose $\mathbb{E}[\text{Cov}(Y, B|b)] = 0$. 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 disparity 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 $D_l = \gamma + \frac{\mathbb{E}[\text{Cov}(Y,b|B)]}{\text{Var}(B)}$

Substituting this into the above, we have:

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

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 D_l and D_p 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:

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

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 and 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 $\widehat{Y}_{p,w}^B$ 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:

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

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

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

□

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 Re-calibrating a proxy for a robustness check

In general, we may not have access to b, but do have access to a re-calibrated b^*, i.e. the linear projection of B on to the space of b and a constant. We can use this re-calibrated proxy to obtain similar as those in Proposition 1 by again applying Proposition 2 and the particulars of re-calibration.

To see this, note that by construction, $\text{Cov}(b^*, \varepsilon^*) = 0$, so Proposition 2 applies. Moreover, $\mathbb{E}[b^*] = \mathbb{E}[B]$, so $\mathbb{E}[b^*](1 - \mathbb{E}[b^*]) = \mathbb{E}[B](1 - \mathbb{E}[B]) = \text{Var}(B)$. 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:

$$D^*_l = D + \frac{\mathbb{E}[\text{Cov}(Y, b^*|B)]}{\sigma^2_{b^*}} \tag{2}$$

and

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

These equations are as in the form of Proposition 1, but there are two apparent difficulties. The first 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 (which could be investigated from first principles or empirical evidence). The following lemma shows that as long as our initial proxy was positively correlated with B, the signs of these covariance terms will not change using the re-calibrated proxy.

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 $\text{Cov}(B, b) > 0$. Then

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

Proof. We note that

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

and

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

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

The second difficulty is the term $\text{Cov}(\mathbb{E}[\eta^*|b^*], \mathbb{E}[\varepsilon^*|b^*])$, which will not in general be 0 and may not be obvious even in sign. This term can, however, be estimated; we do so below, and observe that (at least in our context) it is exceedingly small.

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.

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

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).

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

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

    where again $\text{Multi}(n, \mathbf{p})$ 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 and 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 and 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.7. 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 6.2 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 disparity estimator to obtain this quantity:

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

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:

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

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 disparity 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.

C North Carolina Match and Bias Correction

C.1 North Carolina Match

In our North Carolina voter registration data, 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.

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[31]; 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.

[31] 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 Disparity Decomposition

This appendix shows how to decompose the overall disparity in audit rates with respect to the contribution from EITC audits.

D.1 Notation

As above, let $Y \in \{0, 1\}$ denote whether a taxpayer is audited and $b \in \{B, NB\}$ denote whether a taxpayer is Black. Y_B is the average audit rate among Black taxpayers and Y_{NB} is the average audit rate among non-Black taxpayers. $D = Y_B - Y_{NB}$ denotes the difference in average audit rates across Black and non-Black taxpayers. Let $D_B^C = Y_B^C - Y_{NB}^C$ denote the difference in Black versus non-Black audit rates among EITC claimants and $D_B^{NC} = Y_B^{NC} - Y_{NB}^{NC}$ denote the difference in Black versus non-Black audit rates among EITC non-claimants. Finally, let C_b denote the probability a taxpayer of race b claims the EITC, for $b \in \{B, NB\}$.

D.2 Decomposition

By the law of iterated expectations,

$$Y_B = Y_B^C \, C_B + Y_B^{NC} \, (1 - C_B) \tag{5}$$

and similarly,

$$Y_{NB} = Y_{NB}^C \, C_{NB} + Y_{NB}^{NC} \, (1 - C_{NB}) \tag{6}$$

Substituting (11) and (12) into the definition of D, we can write

$$D = Y_B^C \, C_B + Y_B^{NC} \, (1 - C_B) \; - Y_{NB}^C \, C_{NB} - Y_{NB}^{NC} \, (1 - C_{NB}) \tag{7}$$

Focusing on the first and third terms in (13), we can write:

$$\begin{aligned} Y_B^C \, C_B - Y_{NB}^C \, C_{NB} & \\ &= Y_B^C \, C_B - Y_{NB}^C \, C_B + Y_{NB}^C \, C_B - Y_{NB}^C \, C_{NB} \\ &= D^C \, C_B + Y_{NB}^C \, (C_B - C_{NB}) \end{aligned} \tag{8}$$

Similarly, focusing on the second and fourth terms in (13), we can write:

$$\begin{aligned} Y_B^{NC} \, (1 - C_B) - Y_{NB}^{NC} \, (1 - C_{NB}) & \\ = Y_B^{NC} \, (1 - C_B) - Y_{NB}^{NC} \, (1 - C_B) + Y_{NB}^{NC} \, (1 - C_B) - Y_{NB}^{NC} \, (1 - C_{NB}) & \\ = D^{NC} \, (1 - C_B) - Y_{NB}^{NC} \, (C_B - C_{NB}) \end{aligned} \tag{9}$$

Substituting (14) and (9) into (13) yields

$$D = \; D^C \, C_B + D^{NC} \, (1 - C_B) \; + (C_B - C_{NB}) \left(Y_{NB}^C - Y_{NB}^{NC} \right) \tag{10}$$

Equation 10 expresses the overall difference in the audit rate among Black and non-Black taxpayers in terms of three components. The first term reflects the difference in audit rates between Black and non-Black EITC claimants. The second term reflects the difference in audit rates between Black and non-Black taxpayers not claiming the EITC. The third term reflects differences in the rate at which Black and non-Black taxpayers claim the EITC and the extent to which EITC returns are audited at a different rate than non-EITC returns among taxpayers of the same race.

D.3 Empirical Implementation

Using the weighted estimator described in section 3, we estimate $D^C = 1.96\%$, $D^{NC} = 0.10\%$, $C_B = 32.23\%$, and $C_{NB} = 17.14\%$. We also estimate that EITC returns are audited at a higher rate than non-EITC returns among non-Black taxpayers, $Y^C_{NB} - Y^{NC}_{NB} = 1.05\% - 0.31\% = 0.74\%$. Following this approach, we estimate that the audit rate disparity within EITC returns (term 1) contributes 78% of the observed disparity, with the remainder due to the disproportionate auditing of EITC returns (term 3, 14% of the overall disparity) and, to a lesser extent, the disparity within non-EITC returns (term 2, 8% of the overall disparity).

E Decomposition of Classifier-Induced Disparity

This appendix shows how to decompose the overall disparity in audit rates that is induced by the classifier, as discussed in the main text.

E.1 Notation

As above, let $Y \in \{0, 1\}$ denote whether a taxpayer is audited and $b \in \{B, NB\}$ denote whether a taxpayer is Black. Y_B is the audit rate among Black taxpayers and Y_{NB} is the audit rate among non-Black taxpayers. $D_B = Y_B - Y_{NB}$ denotes the difference in average audit rates across Black and non-Black taxpayers. Let f_b denote the share of compliant taxpayers in group b who are audited, which is a false-positive rate. Let s_b denote the sensitivity, or recall, of the audit selection process, or the share of non-compliant taxpayers from group b who are selected at a given audit budget. Finally, let c_b denote the probability a taxpayer of race b is compliant, defined as under-reported tax liabilities of less than \$100.

E.2 Decomposition

The audit rate for each group is the weighted sum of the rates at which its compliant and non-compliant taxpayers are audited:

$$Y_B = f_B\, c_B + s_B\, (1 - c_B) \tag{11}$$

$$Y_{NB} = f_{NB}\, c_{NB} + s_{NB}\, (1 - c_{NB}) \tag{12}$$

Subtracting the two gives:

$$D = Y_B - Y_{NB} = f_B\, c_B - f_{NB}\, c_{NB} + s_B\, (1 - c_B) - s_{NB}\, (1 - c_{NB}) \tag{13}$$

This expression can be rewritten by adding and subtracting the cross-terms $c_B f_{NB}$ and $c_B s_{NB}$ and rearranging, which gives:

$$D = c_B\, (f_B - f_{NB}) + (1 - c_B)\, (s_B - s_{NB}) + (c_{NB} - c_B)(s_{NB} - f_{NB}) \tag{14}$$

Equation 14 expresses the overall difference in the audit rate between Black and non-Black taxpayers in terms of three components. The first term reflects that part of exam-rate disparity that is proportional to the group difference in audit rates for compliant taxpayers. The second term reflects the part that is proportional to the group difference in audit rates for non-compliant taxpayers. The third term is proportional to the difference between compliance rates for Black and non-Black claimants. The latter is scaled by the difference between the sensitivity and the false positive rate for the reference group, which will be larger for more accurate models. Note that either group can serve as the reference group; results were qualitatively similar when the reference group was switched.

Table E.1: Unconstrained Classifier Disparity Decomposition

	Black	Non-Black	Contribution to Disparity (percentage points)
False-positive Rate	0.0015	0.0017	-0.0076
Sensitivity	0.0367	0.0231	0.8431
Share Compliant	0.3823	0.5060	0.2651
Observed Disparity			1.1006

Notes: The table decomposes the overall disparity induced by the unconstrained random forest classifier into three sub-components: The share accounted for by differences in underlying rates of compliance, the share accounted for by differences in the sensitivity of the model to noncompliance within each group, and the share accounted by differences in the rate of false positives across groups. "Compliant" taxpayers are those whose overall audit adjustments do not exceed $100. Sensitivity refers to the share of non-compliant taxpayers selected for audit. False positive rate refers to the share of compliant taxpayers selected for audit. Contribution to Disparity refers to the term from Equation 14 corresponding to the specified row. All values are computed at the status-quo EITC audit rate of 1.45%. Values in the last column of the table are represented in percentage points.

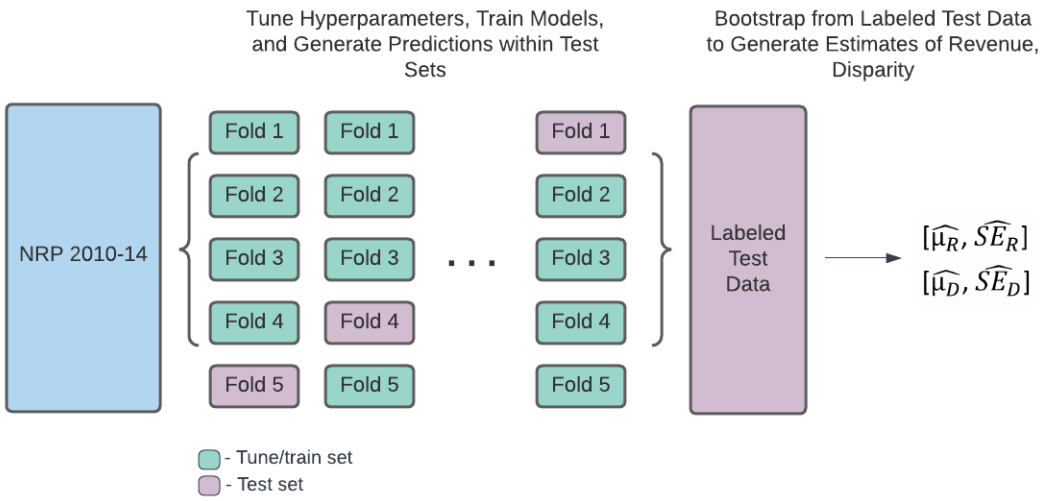
E.3 Empirical Implementation

Table E.2: Constrained Classifier Disparity Decomposition

	Black	Non-Black	Contribution to Disparity (percentage points)
False-positive Rate	0.0036	0.0011	0.0958
Sensitivity	0.0444	0.0208	1.4563
Share Compliant	0.3823	0.5060	0.2429
Observed Disparity			1.7950

Notes: The table decomposes the overall disparity induced by the constrained random forest classifier into three sub-components: The share accounted for by differences in underlying rates of compliance, the share accounted for by differences in the sensitivity of the model to noncompliance within each group, and the share accounted by differences in the rate of false positives across groups. "Compliant" taxpayers are those whose overall audit adjustments do not exceed $100. Sensitivity refers to the share of non-compliant taxpayers selected for audit. False positive rate refers to the share of compliant taxpayers selected for audit. Contribution to Disparity refers to the term from Equation 14 corresponding to the specified row. All values are computed at the status-quo EITC audit rate of 1.45%. Values in the last column of the table are represented in percentage points.

Figure E.1: Data Flow for Random Forest Models



F Taxpayer Non-compliance Model

The outcomes of interest Y of the taxpayer non-compliance random forest models are the dollar amount of adjustment following an audit (for the regression model) and a [0,1] indicator of whether non-compliance exceeds $100 (for the classifier model). 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.

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:

$$R_s = \frac{1}{5} \sum_{t=2010}^{2014} \frac{W_t}{\sum_{i=1}^{n_{st}} w_{ist}} \sum_{i=1}^{n_{ft}} (a_{ist} w_{ist} r_{ist}) \tag{15}$$

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.