Stanford | Institute for Economic Policy Research (SIEPR)

MEASURING AND MITIGATING RACIAL DISPARITIES IN TAX AUDITS

Hadi Elzayn
Stanford University

Arun Ramesh
University of Chicago

Robin Fisher
Stanford University

Evelyn Smith
University of Michigan

Jacob Goldin
University of Chicago and U.S. Treasury Department.

Thomas Hertz
U.S. Treasury Department, Office of Tax Analysis

Daniel E. Ho
Stanford University

Measuring and Mitigating Racial Disparities in Tax Audits *

Hadi Elzayn[†] Evelyn Smith[‡] Thomas Hertz[§] Arun Ramesh[¶]
Robin Fisher[§] Daniel E. Ho[‖] Jacob Goldin[**]

January 30, 2023

Abstract

Government agencies around the world use data-driven algorithms to allocate enforcement resources. Even when such algorithms are formally neutral with respect to protected characteristics like race, there is widespread concern that they can disproportionately burden vulnerable groups. We study differences in Internal Revenue Service (IRS) audit rates between Black and non-Black taxpayers. Because neither we nor the IRS observe taxpayer race, we propose and employ a novel partial identification strategy to estimate these differences. Despite race-blind audit selection, we find that Black taxpayers are audited at 2.9 to 4.7 times the rate of non-Black taxpayers. The main source of the disparity is differing audit rates by race among taxpayers claiming the Earned Income Tax Credit (EITC). Using counterfactual audit selection models for EITC claimants, we find that maximizing the detection of underreported taxes would not lead to Black taxpayers being audited at higher rates. In contrast, in these models, certain policies tend to increase the audit rate of Black taxpayers: (1) designing audit selection algorithms to minimize the "no-change rate"; (2) targeting erroneously claimed refundable credits rather than total under-reporting; and (3) limiting the share of more complex EITC returns that can be selected for audit. Our results highlight how seemingly technocratic choices about algorithmic design can embed important policy values and trade-offs.

*The views presented here are those of the authors and do not necessarily represent the position of the Treasury Department. We thank Emily Black, Edith Brashares, Geoffrey Gee, Robert Gillette, Alissa Graff, John Guyton, Anne Herlache, Jim Hines, Tatiana Homonoff, Barry Johnson, Julian Nyarko, Kit Rodolfa, Joel Slemrod, Megan Stevenson, and Alex Turk for helpful comments. We are grateful for financial support from the Hoffman-Yee Research Grant for Stanford's Institute for Human-Centered Artificial Intelligence and from Arnold Ventures.

[†]Equal first author, Stanford University

[‡]Equal first author, University of Michigan

[§]U.S. Treasury Department, Office of Tax Analysis

[¶]University of Chicago

[‖]Equal co-supervisor, Stanford University. Email: dho@law.stanford.edu

[**]Equal co-supervisor, University of Chicago and U.S. Treasury Department. Corresponding author: jsgoldin@uchicago.edu

1 Introduction

The U.S. federal government annually collects $4 trillion in tax revenue to fund public programs ranging from economic regulation to military defense. Increasingly, policymakers have also come to rely on the tax system for implementing a host of social programs; for example, the Earned Income Tax Credit (EITC) has replaced welfare as the largest cash-based safety net program in the United States. The Internal Revenue Service (IRS) administers these programs and is also charged with ensuring that individuals meet their taxpaying obligations. Like tax authorities around the world, it relies on audits to detect underreporting of tax liabilities and to verify that taxpayers qualify for the benefits they claim.

The task of selecting which taxpayers to audit is in large part an exercise in predicting which taxpayers have underreported tax obligations that an audit would uncover. Modern machine learning methods offer the potential to enhance the efficiency of audits by improving predictive accuracy, but a growing literature in algorithmic fairness warns that policies based upon such predictions may inadvertently reinforce disadvantages against historically marginalized groups (Angwin et al., 2016; Buolamwini and Gebru, 2018; Rambachan et al., 2020). Such concerns are particularly acute for tax audits, which can exacerbate financial strain for the lowest income taxpayers – whose tax refunds are typically frozen while an audit is in place – and can dissuade individuals from participating in safety net programs for which they qualify (Guyton et al., 2018; National Taxpayer Advocate, 2019a).

In this paper, we investigate racial disparities in the selection of taxpayers for audit, focusing on differences in the selection of Black and non-Black taxpayers. Because the IRS does not collect information about taxpayers' race, identifying differences in audit rates by race is itself a significant challenge. Researchers have developed a range of tools for imputing race from other observed characteristics but such approaches can lead to biased estimates for the parameters of interest unless restrictive assumptions are satisfied (Chen et al., 2019; Knox

et al., 2022). A second methodological challenge is that even if we could observe taxpayer race, the fact that underreporting is observed only for those returns that were selected for audit (the so-called selective labels problem) (Kleinberg et al., 2018) makes it difficult to understand why disparities emerge or which policies would lessen them.

Through a unique partnership with the Treasury Department, we investigate these questions using comprehensive microdata on approximately 148 million tax returns and 780,000 audits.[1] To circumvent the selective labels problem, we also leverage nearly 72,000 audits of randomly selected taxpayers to investigate the effects of counterfactual audit selection policies. To address the problem of missing race, we use Bayesian Improved First Name and Surname Geocoding (BIFSG), imputing race based on full name and census block groups (Imai and Khanna, 2016; Voicu, 2018). We then propose and implement a novel approach for bounding the true audit disparity by race from the (imperfectly measured) BIFSG proxy. By individually matching a subset of the tax data to self-identified race data from other administrative sources, we provide evidence that the assumptions underlying our bounding approach are satisfied in practice.

We report a number of new results about racial disparities in tax audits. First, we estimate that the audit rate for returns filed by Black taxpayers is between 0.81 and 1.34 percentage points higher than the audit rate for non-Black taxpayers. This (unconditional) disparity is substantial when compared to the base audit rate of 0.54% for the overall U.S. population. In relative terms, our estimates imply that Black taxpayers are audited at between 2.9 to 4.7 times the rate of non-Black taxpayers.

Second, we find that the difference in audit rates for Black and non-Black taxpayers is primarily driven by the difference in audit rates among taxpayers who claim the EITC. Others have speculated that Black taxpayers may be audited at higher rates because they are more likely than non-Black taxpayers to claim the EITC, and EITC claimants (of any race) are audited at higher rates than most other taxpayers (Bloomquist, 2019; Kiel and

[1] We primarily focus on tax year 2014 – the most recent year for which audit outcome data was complete at the time of our analysis. We find similar results for other years spanning the 2010–2018 time period.

Fresques, 2019). However, we find that this channel explains only a small portion of the disparity (14%). Instead, the larger source of the disparity (78%) stems from the selection of taxpayers for audit *within* the population of EITC claimants: Black taxpayers claiming the EITC are between 2.9 and 4.4 times as likely to be audited as non-Black EITC claimants. In contrast, we observe a much smaller, though still statistically significant, difference in audit rates between Black and non-Black taxpayers who do not claim the EITC.

Third, we explore the contribution of group-level differences between Black and non-Black taxpayers to the observed disparity in audit rates. We find that the disparity cannot be fully explained by racial differences in income, family size, or household structure, and that the observed audit disparity remains large after conditioning on these characteristics. For example, among unmarried men with children, Black EITC claimants are audited at more than twice the rate of their non-Black counterparts. More strikingly, the disparity cannot be fully explained even by accounting for group-level differences in tax under-reporting: we estimate that Black EITC claimants are audited at higher rates than non-Black EITC claimants within each decile of under-reported taxes.

Because EITC audit selection is largely automated, and because the IRS does not observe race, the status quo racial audit disparity is unlikely to be driven by disparate treatment of Black and non-Black taxpayers. Nonetheless, there may be opportunities for the agency to adjust its enforcement policies in ways that alleviate the disparity without significantly undermining other policy goals. Our final set of analyses explores this possibility and the associated trade-offs.

To do so, we use data from randomly selected audits of returns claiming the EITC to explore counterfactual audit selection policies. We first show that if it were possible to select audits in descending order of true under-reported tax liability, Black taxpayers would be audited at a *lower* rate than non-Black taxpayers. The same is true if audits were selected according to predicted underreporting based on the pre-audit tax return characteristics that

IRS observes and machine-learned models.[2]

However, we show that changing the prediction model's objective can dramatically increase the share of Black taxpayers selected for audit. In particular, if the model is trained to predict *whether* the taxpayer under-reports, instead of the expected magnitude of underreported taxes, Black taxpayers are audited at higher rates. Intuitively, the choice of label — whether to solve a classification or regression problem — matters for shaping disparity because the taxpayers with the highest under-reported taxes tend to be non-Black, but the available data allow the classifier model to assign the highest probabilities of underreporting to more Black than non-Black taxpayers. We find a similar effect on racial disparity from training the algorithm to detect over-claiming of refundable tax credits rather than total under-reporting due to any error on the return.

Finally, we consider the effect of policies that constrain the allocation of audits across distinct categories of EITC-claiming tax returns. In particular, tax returns with substantial business income tend to be more complicated to audit than those without such income. The IRS may be limited in its ability to reallocate audits from non-business to business returns, at least in the short-term, due to a lack of auditors with the required expertise. We find that operational constraints like these may shape racial disparities: among EITC returns, imposing a constraint that maintains the status quo fraction of audited returns that report business income increases the share of audits focused on Black taxpayers, relative to an unconstrained baseline.

Our results contribute to an empirical literature that studies the distributional effects of tax policy by race, with seminal contributions by Moran and Whitford (1996) and Brown (2021), among others.[3] However, the unavailability of race-linked administrative microdata has limited the questions that prior studies could address and the alternative policies they

[2]We obtain similar results when focusing on detected under-reporting *net* of IRS auditing costs.

[3]This literature has primarily focused on substantive tax provisions such as the mortgage interest deduction (Brown, 2009, 2018), the EITC (Brown, 2005; Hardy et al., 2021), and the Child Tax Credit (Collyer et al., 2019; Goldin and Michelmore, 2022).

could evaluate (Bearer-Friend, 2019; Brown, 2021; Dean, 2021).[4] We build on this literature by linking imputed race estimates with anonymized administrative data on tax returns and audits. Doing so allows us to produce the most direct evidence to date on longstanding questions about racial disparities in the administration of the U.S. income tax. In addition, by drawing on detailed data from random audits, our results shed light on how alternative policies may affect audit rate disparities while promoting or hindering other IRS objectives.

A second contribution of our paper is to introduce a novel partial identification approach for conducting algorithmic disparity assessments along the lines of a protected class, when that class is unobserved to the researcher. This challenge arises in a wide range of settings, including voting rights (Imai and Khanna, 2016), regulatory policy (CFPB, 2014; Anson-Dwamena et al., 2021; Haas et al., 2019), and a wide variety of industry settings (Alao et al., 2021; Andrus et al., 2021). For example, although many U.S. agencies are required by federal law to conduct disparity assessments of the algorithmic decision tools they employ, protected characteristics like race are often missing from administrative records (G.A.O., 2020; Exec. Order. 13985, 2021). Our approach relies on weaker conditions than those required to point-identify difference in outcomes by unobserved protected class (e.g., Chen et al., 2019; Fong and Tyler, 2021), while still yielding bounds that may be informative for policy.[5]

Finally, our results highlight how seemingly technocratic algorithmic design decisions can embed important policy values and trade-offs. Governments are increasingly employing modern prediction methods to allocate public benefits and enforcement resources. While prior research in the algorithmic fairness literature has demonstrated how ostensibly

[4]Prior analyses have yielded some suggestive evidence, however. For example, Bloomquist (2019) studies regional bias in IRS audits using estimated county-level data, and notes that the ten most heavily audited counties were predominantly comprised of Black taxpayers, whereas the ten least heavily audited counties were predominantly non-Black. In addition, prior research has linked tax data with Census records on self-reported race to study primarily non-tax outcomes (e.g., Chetty et al., 2020); however, various legal and institutional constraints currently limit our ability to apply this approach to our setting.

[5]Kallus et al. (2021) also consider a partial identification approach to estimating disparity when the protected characteristic must be imputed. Unlike our approach, their bounds cover all joint distributions consistent with the observed marginals. In our setting, the Kallus et al. bounds are largely uninformative as to the magnitude or even direction of the audit rate disparity; they cannot rule out Black taxpayers facing either higher or lower audit rates than non-Black taxpayers. Our approach requires additional structure, but the payoff to that structure is a significantly more informative estimate when our assumptions hold.

neutral algorithms can replicate or exacerbate biases present in the data, our results highlight the importance of a different aspect of algorithmic design: namely the definition of the outcome being predicted (i.e., the choice of "label"). Moreover, while much of the algorithmic fairness literature is concerned with an accuracy-fairness trade-off (Corbett-Davies et al., 2017; Kearns et al., 2018; Menon and Williamson, 2018; Zhao and Gordon, 2019; Chen et al., 2018; Elzayn et al., 2019), our results show how the extent and even existence of such trade-offs can depend on seemingly minor differences in how the label is defined.[6]

The paper proceeds as follows. Section 2 provides background on the U.S. tax system and taxpayer audits. Section 3 describes our empirical strategy. Section 4 describes our data. Section 5 provides results relating to estimated race probabilities and statistical bias of our proposed estimators. Section 6 estimates differences in audit rates between Black and non-Black taxpayers. Section 7 investigates the potential role of algorithmic design and other enforcement policies in generating and mitigating disparities. Section 8 concludes.

2 Institutional Background

This section provides background regarding the U.S. individual income tax, taxpayer audits, and the EITC.

2.1 The U.S. Income Tax

The individual income tax is the primary source of revenue for the United States federal government. Most U.S. citizens, as well as some non-citizens, are required to file an income tax return each year, on which they calculate and report their tax liability for the year

[6]The algorithmic fairness literature has proposed a large range of fairness metrics (Mitchell et al., 2021). In our setting, there are multiple normative baselines one might adopt, such as the magnitude of the disparity or the difference in audit rates relative to ground-truth differences in underreporting. We do not take a stance on the proper normative baseline but rather show how different audit selection policies would shape differences in the audit rate between Black and non-Black taxpayers.

based on their income as well as any deductions or credits for which they qualify. Unmarried taxpayers file individual returns, whereas most taxpayers who are married file a joint return with their spouse. Taxpayers with children or other dependents may claim them on their own return to qualify for various tax benefits. The vast majority (over 95%) of taxpayers prepare and file their returns with the help of a professional tax preparer or using guided tax preparation software.

2.2 Taxpayer Audits

The IRS is the federal agency responsible for promoting and enforcing compliance with the income tax law. One channel through which it does so is taxpayer audits. Taxpayers selected for audit are required to provide additional information to the IRS or otherwise verify the accuracy of the tax liability or refund reported on their tax return.[7] Audits may occur by mail ("correspondence examinations") or through in-person (or virtual) meetings with IRS employees ("field" or "office" examinations). In recent years, approximately 70% of audits have been correspondence audits. Audits of this form tend to focus on a small number of issues and require a response, with substantiation. If the IRS does not receive a response by the due date, it will generally disallow the claimed item and issue the taxpayer with a notice of deficiency. Correspondence audits are substantially cheaper than other forms of audits for the IRS to conduct, since much of the process is automated and does not require the involvement of human auditors (IRS, 2011). At the same time, correspondence audits can be particularly burdensome for lower-income households, who may face additional barriers understanding the audit notice, acquiring the required documents, or obtaining expert assistance (G.A.O., 2016; National Taxpayer Advocate, 2021).

If an audit results in an adjustment of the tax liability reported on a taxpayer's return, the taxpayer is responsible for remitting the difference to the IRS and may face additional

[7]Distinct from formal audits and hence not counted as audits in our data, the IRS operates several programs through which it screens submitted returns for potential identity theft or makes adjustments to taxpayers' submitted returns based on information reported to it by third parties, the detection of math errors, or other factors.

penalties, as well as, in rare cases, criminal sanctions. Audits may occur pre- or post-refund; in the former case, applicable refunds are not issued until after the audit is resolved. Hence, taxpayers who fail to respond to a pre-refund audit typically forego the tax benefits they claimed on their return. Taxpayers who disagree with the results of an audit may appeal the determination with the IRS office of appeals and/or in federal court.

At a high level, audits can be categorized into two groups: research and operational. Research audits are conducted through the National Research Program (NRP), which consists of a stratified random sample of the tax filing population.[8] NRP audits seek to estimate the correctness of the whole return via a close to line-by-line examination. In part because they are so intensive to conduct, research audits constitute a small minority of the audits that the IRS performs each year. For example, about 2% of audited tax year 2014 returns were selected through NRP.

We refer to audits that are not research audits as "operational audits." Operational audits constitute the vast majority of audits that the IRS performs. Tax returns are selected for an operational audit through a wide variety of processes, the details of which are kept confidential. These processes can range from simple decision rules to manual examination to prioritization based on model-estimated risk scores. To facilitate our research, we were given access to information on some but not all of these processes.[9] Thus, we are able to observe whether a return was selected for an operational audit, but typically we cannot observe the reason why the audit occurred.

2.3 The Earned Income Tax Credit

One tax credit that will be important for our analysis is the Earned Income Tax Credit (EITC). The EITC provides support to low- and middle-income taxpayers with earnings

[8]Tax returns are selected for inclusion in the NRP program prior to entering the pipeline for operational audit selection.

[9]To the best of our knowledge, we observe all training data and the full set of taxpayer features for returns claiming the EITC – our focus below – with the exception of audit referrals from whistleblowers or law enforcement (which are present in a very small share of EITC audits).

from work. The EITC is designed as a refundable credit, meaning that taxpayers receive a payment or "refund" from the government if the credit brings their tax liability for the year below zero. Today, the EITC constitutes the largest cash-based safety net program in the United States, lifting an estimated 5.6 million households out of poverty at an annual budgetary cost of approximately $70 billion (Center on Budget and Policy Priorities, 2019; J.C.T., 2020).

The amount of EITC for which a taxpayer qualifies is based on the taxpayer's income and family size. The credit amount initially increases with income for lower-income taxpayers, plateaus, and then decreases with income once a taxpayer's income surpasses a specified threshold. The maximum EITC amount varies based on the number of children a taxpayer claims; in 2014 (the year that most of our analysis focuses upon), the maximum EITC ranged from $496 for taxpayers without children to $6,143 for taxpayers with three or more children, and was available to taxpayers with combined incomes up to $52,247 if married, or $46,997 if single. Taxpayers without income from market work do not qualify for any EITC amount. Approximately 19% of taxpayers claimed the EITC in 2014, with an average credit of $2122 among claimants (IRS, 2016).

To claim a child for the EITC, the child must satisfy several eligibility tests with respect to the taxpayer. In particular, the taxpayer must be related to the child through one of a specified set of relationships (e.g., the child's parent, stepparent, grandparent, aunt, uncle, or sibling) and must reside with the child for over half of the tax year. In addition, the child must generally be below the age of 19, or below the age of 24 if the child is a full-time student. In cases in which two or more taxpayers qualify to claim a single child, a series of "tie-breaker" rules govern which claim takes priority.

Non-compliance with the EITC rules has been a persistent subject of policy concern. Estimates of the EITC improper payment rate hover around 25% (U.S. Treasury Department, 2022), which many observers attribute to the complicated rules governing eligibility for the credit (Holtzblatt and McCubbin, 2004; National Taxpayer Advocate,

2019b). Federal law requiring agencies to measure improper payments effectively imposes a minimal number of research audits of EITC returns that IRS must conduct, but such rules do not directly constrain IRS's ability to adjust the number of EITC returns selected for operational audit (OMB, 2021). In addition to the EITC, the federal government designates two other refundable credits—the American Opportunity Tax Credit and Additional Child Tax Credit—as high-priority programs that are susceptible to significant improper payments.

In recent years, approximately one-third or more of individual taxpayer audits have been devoted to returns claiming the EITC (TIGTA, 2021). EITC returns are selected for audit through a variety of enforcement programs. Many of the audited EITC returns are selected through the Dependent Database (DDb) program, which flags returns based on a set of rules and heuristics as well as various proprietary risk scores. The features that are used to calculate these proprietary risk scores are drawn from taxpayers' filed returns as well as other administrative data about taxpayers or their children available to IRS.[10] Although its details are confidential, the model underlying the DDb program is calibrated (at least in part) to minimize the no-change rate – i.e., the probability that an audit ends with no change to the tax assessment that the taxpayer reported (G.A.O., 2015).

The vast majority (94% in 2014) of audits of EITC claimants are correspondence examinations and approximately two-thirds occur pre-refund (see Appendix Table A.1). Taxpayers who fail to respond to an EITC audit letter (either intentionally, or unintentionally – e.g., because they changed address) are treated as ineligible for the credit and forego their claimed refund.

[10]These administrative data sources include child custody data from the Department of Health and Human Services, child birth data from the Social Security Administration, and prisoner data from the National Prisoner File.

3 Empirical Framework

In this section, we provide results relating to the identification of audit rate disparities on the basis of protected characteristics (like race) when the characteristic in question cannot be directly observed by the researcher. Specifically, we show that under certain conditions, one may use individually estimated race probabilities to obtain informative bounds on the racial audit rate disparity.

3.1 Basic Notation

Tax returns are indexed by i, and have observable characteristics X_i. We use $B_i \in \{0, 1\}$ to indicate whether the primary filer on tax return i is Black, and $Y_i \in \{0, 1\}$ to indicate whether i is audited. The audit rate for Black taxpayers is $Y^B = \mathbb{E}[Y|B = 1]$, and the audit rate for non-Black taxpayers is $Y^{NB} = \mathbb{E}[Y|B = 0]$.

Our goal is to estimate the audit disparity faced by Black taxpayers, D, which we define as the (unconditional) difference in audit rates between Black and non-Black taxpayers:

$$D = Y^B - Y^{NB} = \mathbb{E}[Y|B = 1] - \mathbb{E}[Y|B = 0] \qquad \text{(Audit Disparity)}$$

We focus on this quantity, as law and policy mandate evidence on such "disparate impact" or racial disparity (U.S. Executive Order 13985, 2021). Later, we also examine evidence for disparities conditional on taxpayer characteristics.

An important barrier to studying differences in audit rates by race is that neither we nor the IRS observe taxpayer race.[11] To overcome this challenge, we first estimate the probability that a taxpayer is Black using a subset of characteristics we do observe, and second, use the resulting race probabilities, along with data on actual audits, to estimate differences in audit

[11] At various stages of the audit process following selection, individual examiners may be able to observe a taxpayer's race or proxies for it; we lack access to sufficient institutional details to assess whether such knowledge could translate into differential treatment by race of taxpayers during audit processes following audit selection.

rates by race. That is, our approach is to:

1. Estimate $b_i = Pr[B_i = 1|Z_i]$, where $Z_i \subseteq X_i$ is a subset of i's observable characteristics
2. Use estimated b_i and observed Y_i to estimate D

In the remainder of this section, we describe these two steps – imputing race and using imputed race to estimate disparity – in additional detail.

3.2 Imputing Race

To impute race, we draw on Bayesian Improved Surname Geocoding, which uses name and geolocation to probabilistically infer race (Imai and Khanna, 2016). This method has been widely applied in academic studies and is recommended when race is missing by the National Academy of Medicine (Nerenz et al., 2009). Recent work has shown that first names are more informative than surnames for identifying Black individuals (Voicu, 2018), so we incorporate first name information as well, applying Bayesian Improved First Name Surname Geocoding (BIFSG). The method is "naive" in the sense that it assumes that first name, surname, and geography are independent after conditioning on race:

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

where F is first name, S is surname, and G is geography. Using Bayes' rule, this assumption implies

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

and similarly for $\Pr[B = 0|F, S, G]$. See Appendix B for a formal derivation. Estimating these terms by name and geography allows us to impute individual-level race probabilities. The data used to construct these estimates are described in Section 4. Because audits occur

at the level of the tax return, we estimate a single race probability per return, focusing on the primary filer in cases of joint returns by married spouses.

3.3 Estimating Disparity using Imputed Race

After estimating the probability that each taxpayer is Black, we next consider how to use those estimated probabilities to identify the difference in average audit rates by race. We consider two estimators: the *probabilistic disparity estimator* and the *linear disparity estimator*. Below, we characterize the bias of each estimator, provide conditions under which the two estimators bound the true audit disparity, and provide empirical support for those conditions obtaining in our setting. We derive the asymptotic distributions of the probabilistic and linear disparity estimators in Appendix B.

The probabilistic estimator calculates average audit rates by race by weighting each observation's contribution by the probability that the taxpayer is Black. Formally, given estimated race probability b_i and audit status Y_i, we define the probabilistic audit rate estimators as

$$\widehat{Y}_p^B = \frac{\sum_i b_i Y_i}{\sum_i b_i} \qquad \widehat{Y}_p^{NB} = \frac{\sum_i (1-b_i) Y_i}{\sum_i (1-b_i)}$$

where B and NB refer to the estimated audit rate among Black and non-Black taxpayers, respectively. The probabilistic estimator of the disparity in audit rates between Black and non-Black taxpayers, $\widehat{D}_p$, is then given by the difference in the probabilistic audit rates by race:

$$\widehat{D}_p = \widehat{Y}_p^B - \widehat{Y}_p^{NB} = \frac{\sum_i b_i Y_i}{\sum_i b_i} - \frac{\sum_i (1-b_i) Y_i}{\sum_i (1-b_i)}$$

The second racial audit disparity estimator we consider is the linear disparity estimator,

$\widehat{D}_l$. Consider the regression of Y on b:

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

The linear disparity estimator is derived from the estimated coefficient on b in this regression:

$$\widehat{D}_l = \widehat{\beta} = \frac{\sum_i (Y_i - \overline{Y})(b_i - \overline{b})}{\sum_i (b_i - \overline{b})^2}$$

where $\overline{Y}$ and $\overline{b}$ denote the sample averages of Y and b, and where the corresponding linear estimators of Y^B and Y^{NB} are given by $\widehat{Y}_l^B = \widehat{\alpha} + \widehat{\beta}$ and $\widehat{Y}_l^{NB} = \widehat{\alpha}$.

In Appendix B, we show that the two pairs of audit rate estimators ($\widehat{Y}_l^B, \widehat{Y}_l^{NB}$ and $\widehat{Y}_p^B, \widehat{Y}_p^{NB}$) converge to the true audit rates (Y^B and Y^{NB}) under differing conditions; in particular, the linear audit rate estimators are unbiased when $\mathbb{E}[\mathrm{Cov}(Y, b|B)] = 0$, whereas the probabilistic audit rate estimators are unbiased when $\mathbb{E}[\mathrm{Cov}(Y, B|b)] = 0$. But in general, they, and consequently the disparity estimators derived from them ($\widehat{D}_l$ and $\widehat{D}_p$), will be biased. We characterize the biases of the disparity estimators in the following proposition:

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

$$D_l \leq D \leq D_p \tag{1.4}$$

We provide the proof of this proposition, and derive similar results for the audit rate estimators, in Appendix B.2.

Proposition 1 characterizes the bias of the probabilistic disparity estimator and the linear disparity estimator.[12] For the probabilistic disparity estimator to be asymptotically unbiased (Proposition 1.1), it must be that any association between race and audits is mediated through predicted race. In our setting, this condition would be violated if Black taxpayers were selected for audit at different rates than non-Black taxpayers with identical names and living in identical neighborhoods. Proposition 1.1 is related to a result in Chen et al. (2019), which substitutes $\mathbb{E}[\text{Cov}(Y, B|Z)]$ for $\mathbb{E}[\text{Cov}(Y, B|b)]$ in our expression, similar to conditioning on combinations of covariates in lieu of the propensity score. The difference between the two expressions is significant in practice, since it may not be practical to calculate $\text{Cov}(Y, B|Z)$ when Z takes on many values, such as in the common circumstance in which race is imputed from name and geography.

The linear disparity estimator requires different, but similarly strong, assumptions in order for it to yield an unbiased estimate of the true audit rate disparity. In particular, Proposition 1.2 shows that the linear disparity estimator is unbiased if and only if there is no residual association between predicted race and audits after conditioning on the taxpayer's actual race. This condition can be thought of as the exclusion restriction that would apply

[12] In some applications, we consider weighted versions of these estimators; we discuss properties of weighted probabilistic and linear disparity estimators in Appendix B.4.

when using imputed race as an instrument to study the effect of self-reported race on audit selection. In our setting, the linear disparity estimator may yield a biased estimate for the true disparity if some of the information used to predict race – e.g., name – is associated with audits through channels other than race, such as socioeconomic status.

Finally, Propositions 1.3 and 1.4 establish that the linear and probabilistic disparity estimators (asymptotically) bound the true disparity when $E[\text{Cov}(Y, b|B)]$ and $E[\text{Cov}(Y, B|b)]$ share the same sign.[13] To understand the intuition, suppose that both conditional covariance terms are positive. First, $E[\text{Cov}(Y, b|B)] > 0$ implies that the race probabilities are (positively) correlated with the outcome through channels other than B, so that regressing Y on b (as the linear disparity estimator does) yields an over-estimate for the association between Y and B. Second, $E[\text{Cov}(Y, B|b)] > 0$ implies that the association between B and Y is not perfectly mediated through b; hence, focusing only on the mediated portion of the association (as the probabilistic disparity estimator does) under-estimates the true association between B and Y. Thus, one can bound the true disparity by combining the results of the two estimators.

In our setting, using geography and names to predict race, it is reasonable to expect Proposition 1.3 to apply. For example, there are well-documented differences in marital patterns by race, with rates of marriage among Black households below those of other groups (Aughinbaugh et al., 2013). At the same time, unmarried taxpayers may be associated with higher rates of EITC audit risk due to the residency and relationship tests that govern the EITC qualifying child rules (Leibel et al., 2020). Hence, to the extent that race predictions derived from name and geography do not fully capture racial differences in household structure, it is likely that some residual correlation between audits and race would remain, so that $E[\text{Cov}(Y, B|b)] > 0$. At the same time, some research shows that individuals with more distinctively Black names have lower socioeconomic outcomes (Fryer and Levitt, 2004; Cook et al., 2016). To the extent that audit rates are declining in income

[13] Appendix B shows that under the same conditions, the Black and non-Black audit rates are similarly bounded by the linear and probabilistic audit rate estimators.

for some portions of the income distribution (Black et al., 2022), this suggests that, even after conditioning on race, taxpayers with higher predicted probabilities of being Black may be audited at higher rates than taxpayers with lower predicted probabilities of being Black, or $E[\text{Cov}(Y, b|B)] > 0$. Below, we provide empirical support for these conditions using an auxiliary data set in which race is observed for a subset of taxpayers.

4 Data

This section describes the IRS data relating to audits and other tax variables as well as the data we use to impute taxpayer race.

4.1 Tax Data

We begin with comprehensive, administrative, and anonymized IRS data from ∼148M individual income tax returns with valid social security numbers for 2014. We focus on tax year 2014 because it is the most recent year for which the vast majority of audits were complete and available to us at the time of analysis. For each return, we observe the amount and sources of reported income, deductions, and credits claimed. We also observe information returns for each taxpayer, such as employer-reported wages on Form W-2, and other administrative records, such as Social Security Administration data on gender and year of birth.

Among the returns in our data, there were 780,627 operational audits, constituting 0.53% of returns filed for the year. We also use data on the research audits conducted under the NRP, which, as described in Section 2, are selected using a stratified random sample of taxpayers. Between 2010 and 2014, there were between approximately 13,500 and 15,500 NRP audits per year. To increase the precision of our analyses that use NRP data, we use the 71,878 returns selected for NRP audit between 2010 and 2014.

For each filed return, we observe whether the return was selected for audit.[14] In addition, among audited returns, we observe the amount, if any, of the IRS-imposed adjustment to the originally filed return. Throughout, we report quantities in 2014 dollars, inflation-adjusting the NRP returns from prior years.

4.2 Race Data

As described above, the IRS does not systematically collect data on taxpayer race, either directly via tax returns or indirectly via merging tax data with administrative data on race from other agencies (Bearer-Friend, 2019). Instead, we rely on a BIFSG approach to estimate the probability that a taxpayer is Black (and non-Hispanic) based on the first name, last name, and location of the taxpayer.[15] First and last names were obtained from the tax return. The taxpayer's location was measured at the level of the Census Block Group, the smallest geographic unit with racial composition reported by the Census, which typically contain 600-3,000 individuals.[16] Data on the joint distribution of first names and race were obtained from Loan Application Registers under the Home Mortgage Disclosure Act from 2007-2010, following Tzioumis (2018); data on the joint distribution of last names and race were obtained from the 2010 Decennial Census Surname File (U.S. Census Bureau, 2021); and data on the racial make-up of Census block groups from the American Community Survey 5-year estimates (2010-2014). Information regarding the availability of these characteristics in our sample is described in Appendix Table A.2. We are able to estimate race probabilities based on all three attributes (first name, last name, and geolocation) for 73% of tax year 2014 returns. For the remaining returns, we use a subset of these attributes to impute race. We exclude taxpayers for whom no race probabilities could be assigned because the address could not be geocoded to a specific Census Block Group and neither first name nor last name

[14] More precisely, we observe whether the return was selected for audit prior to 2021. Nearly all audits are initiated within six years of a return being filed.

[15] The steps of these analyses requiring non-anonymized taxpayer information were conducted by Treasury economists, with the (anonymized) results provided to the other members of our research team.

[16] Addresses were mapped to Census Block Groups using an algorithm developed by Geoffrey Gee for the Treasury Department Office of Tax Analysis.

was available in our race-by-name datasets (0.2% of 2014 returns).

To assess the validity of our estimated race probabilities and the statistical bias of our audit disparity estimators, we obtained data on self-reported race for a subset of our sample by matching taxpayers to publicly available voter registration records from North Carolina. North Carolina required all registered voters to report race until 1993, after which reporting became optional. Taxpayer and voter records were matched using name and address, which resulted in a 47% unique match rate and ∼2.5M matched records.[17] In some of the calibration exercises using this data, we re-weight North Carolina taxpayers to match the overall U.S. population on observable demographic characteristics. Appendix C provides additional details regarding the data match and the construction of these weights.

5 Race Estimate Calibration and Statistical Bias Assessment

This section presents results relating to the calibration of our taxpayer race estimates and potential biases of the audit rate disparity estimators on which we rely.

5.1 Taxpayer Race Estimates

As described above, we calculate the predicted probability that a taxpayer is Black based on the taxpayer's first name, last name, and geography. Figure 1 summarizes the results of this exercise. The left panel of the figure presents the distribution of estimated race probabilities. The distribution is bi-modal, with 4.4% of taxpayers having 90% or higher predicted probability of being Black and 77.0% of taxpayers having 10% or lower predicted probability of being Black. The mean prediction is 12.2%, which corresponds closely to the 12.2% of the overall U.S. population that was estimated to be Black by the U.S. Census in

[17]Matching took place in a fire-walled environment, separately from the main analysis, under the direction of the Treasury Office of Tax Analysis. Information relating to voting and political party were excluded from all analyses.

2014 (ACS, 2014).

The right panel of Figure 1 assesses the calibration of the estimated race probabilities. It uses the matched North Carolina data to compare the true probability that a taxpayer identifies as Black with the BIFSG-predicted probability that the taxpayer does so. The figure shows that the predicted race probabilities are generally monotonic in true self-reported race and generally track the 45-degree line. We observe a similar pattern for the sub-population of taxpayers claiming the EITC, although here we observe some evidence that BIFSG may under-estimate the probability a taxpayer is Black; we explore several re-calibration methods below to address this issue. Overall, to the extent our matched North Carolina sample is representative of the population, this analysis suggests that the BIFSG-derived race estimates constitute a reasonably accurate approach for estimating taxpayer race in this setting (see Appendix Table A.3 for additional detail). Below, we consider several robustness checks that employ alternative methods for calculating race probability estimates and obtain qualitatively similar results.

5.2 Assessing Bias of the Audit Disparity Estimators

In this section we use the North Carolina data to shed light on the statistical bias of the audit disparity estimators described above.

To visualize these quantities, Figure 2 bins the taxpayers from the North Carolina data set based on the estimated probability of being Black, and plots the fraction of Black and non-Black taxpayers audited within each bin. The figure shows a positive residual correlation between audit probability and being Black after conditioning on the estimated race probability. From Proposition 1.1, this implies that the probabilistic disparity estimate is downward-biased, i.e., $E[\text{Cov}(Y, B|b)] > 0$. At the same time, the figure suggests an upward-sloping audit rate in predicted race, for both Black and non-Black taxpayers, after conditioning on self-reported race, or $E[\text{Cov}(Y, b|B)] > 0$. From Proposition 1.2, this implies that the linear disparity estimator is upward-biased.

Appendix Table A.4 reports the results of a more formal test for the sign of the key covariance terms. To estimate $E[\text{Cov}(Y, b|B)]$, we use the North Carolina data to directly calculate the covariance between audits and predicted race probabilities, separately for Black and non-Black taxpayers. We aggregate these estimated covariances into an estimate of $E[\text{Cov}(Y, b|B)]$ by weighting each race-specific covariance by the estimated proportion of all taxpayers that are Black or non-Black, respectively. We reject the null hypothesis that $E[\text{Cov}(Y, b|B)] \leq 0$ with $p < 0.01$. In other words, our results suggest that the estimated race probabilities are positively correlated with audits, even after conditioning on the portion of the association due to race.

In similar fashion, we estimate $E[\text{Cov}(Y, B|b)]$ by calculating the sample covariance between audits and self-reported race separately for each estimated race probability percentile, and then aggregate based on the estimated share of taxpayers with each race probability percentile. We reject the null hypothesis that $E[\text{Cov}(Y, B|b)] \leq 0$ with $p < 0.01$.

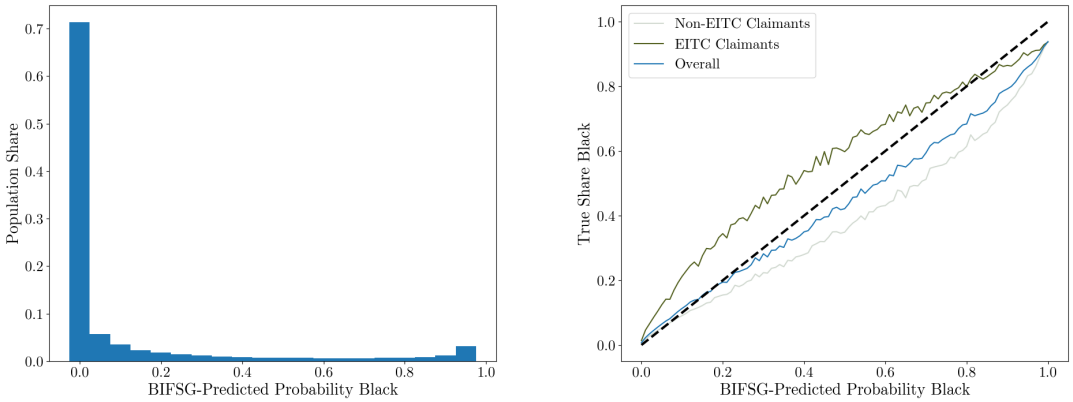
We obtain similar results when we re-weight the North Carolina data to match the U.S. population on a range of observable characteristics (Column 2 of Appendix Table A.4) and when we restrict the analysis to EITC claimants (Columns 3 and 4). In contrast, for the non-EITC population (Columns 5 and 6), we are unable to sign these covariance terms with statistical significance.

We interpret the results in this subsection to support the hypothesis that $E[\text{Cov}(Y, B|b)] > 0$ and $E[\text{Cov}(Y, b|B)] > 0$ for EITC taxpayers and the overall population, and therefore, that the probabilistic and linear disparity estimators bound the true audit rate disparity for these populations.

6 Audit Disparity Results

In this section, we report our estimates of the difference in audit rates between Black and non-Black taxpayers. We begin with the overall population of U.S. taxpayers before turning

Figure 1: Distribution and Calibration of Race Imputations



Notes: Left: Nationwide histogram of BIFSG-predicted probability that a taxpayer is Black (non-Hispanic). The mean prediction is 12.4%. Right: The figure shows the calibration of the BIFSG imputations for the taxpayers in the matched North Carolina data set. 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 ground-truth race observed in the North Carolina matched sample, re-weighted to be representative of the overall United States (see Appendix C.2 for details). A perfectly calibrated predictor would fall exactly 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).

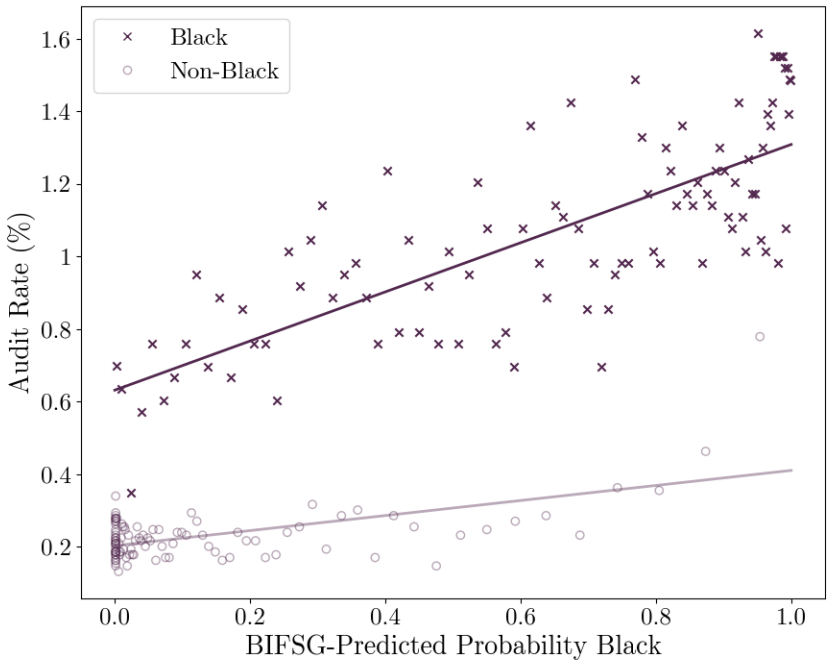
to EITC claimants.

6.1 Overall Audit Disparity

Figure 3 presents our main findings concerning racial audit disparities for the population of U.S. taxpayers. The left panel plots the mean audit rate against the binned estimated probability that a taxpayer is Black. The upward-sloping relationship shown in the figure suggests that Black taxpayers are audited at a higher rate than non-Black taxpayers.

The right panel of Figure 3 depicts the estimated audit rates among Black and non-Black taxpayers, respectively, obtained from the probabilistic and linear disparity estimators. Both estimators imply that Black taxpayers were audited at a higher rate than non-Black taxpayers. In particular, the probabilistic estimator implies a racial audit disparity of 0.81 percentage points: 1.24% of Black taxpayers were audited, compared to 0.43% of non-Black taxpayers. These audit rates are precisely estimated: the 95% confidence interval on the

Figure 2: Audit Rate by Predicted Race Conditional on Self-Reported Race



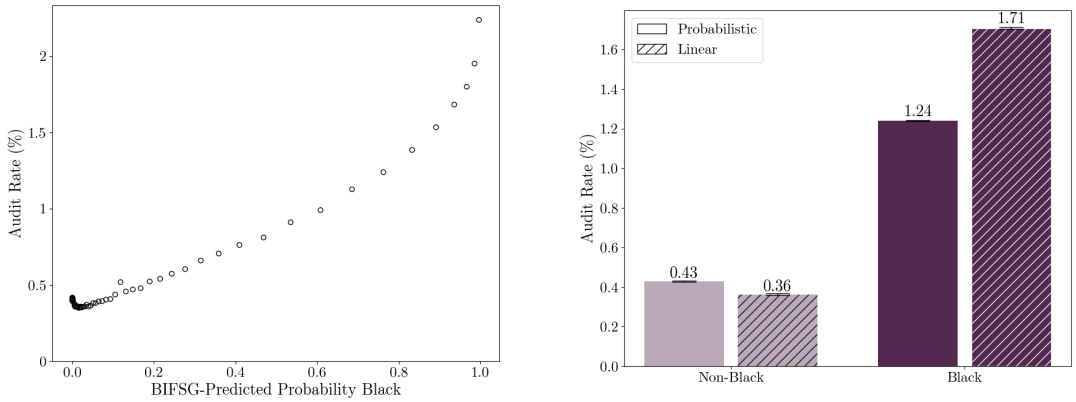
Notes: The figures show the relationship between audit incidence and BIFSG-predicted probability that a taxpayer is Black for taxpayers filing returns for tax year 2014. Audit incidence is plotted separately for Black and non-Black taxpayers in the North Carolina matched sample. Black and non-Black taxpayers 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.

probabilistic disparity estimate ranges from 0.81 to 0.82 percentage points.[18] As expected, the linear estimator implies an even larger racial audit disparity, of 1.34 percentage points, and is also precisely estimated. Appendix Table A.5 provides additional details. Because the conditions for Proposition 1.3 appear satisfied in our setting, we interpret the probabilistic and linear disparity estimates as bounds on the true racial audit disparity. Thus, our results suggest that Black taxpayers were audited at between 2.9 and 4.7 times the rate of non-Black taxpayers.

Figure 4 plots estimated audit rates by income and race. Black taxpayers appear more

[18]This confidence interval reflects sampling uncertainty, in the sense that even the universe of 2014 taxpayers may not perfectly reflect the underlying data generating process. We continue to obtain statistically precise results when incorporating uncertainty stemming from imperfect measurement of taxpayers' race; see Appendix Figure A.1.

Figure 3: Estimated Audit Rates by Race



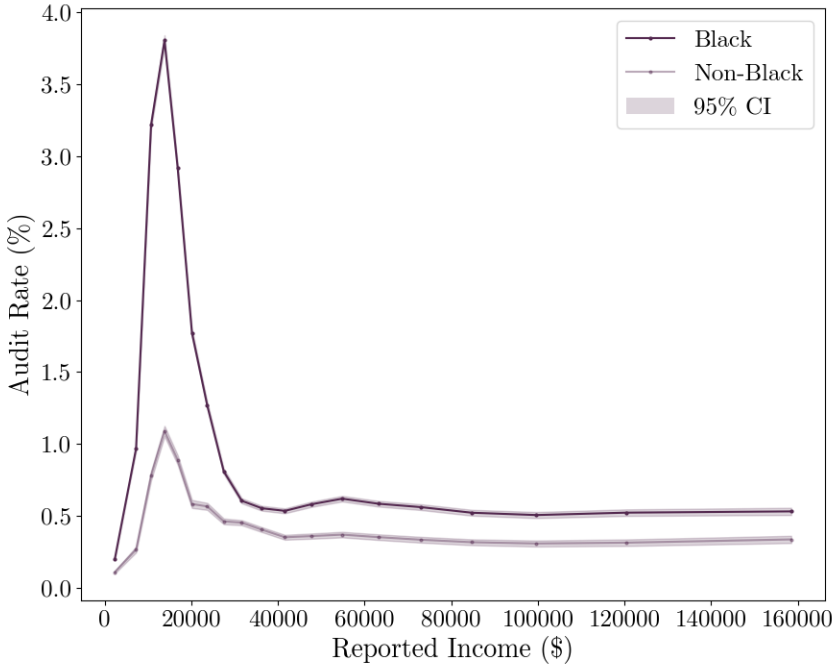
Notes: The figure shows the relationship between audits and race among taxpayers filing returns for tax year 2014. Left: Binned scatterplot of audit rate by BIFSG-predicted probability that a taxpayer is Black. Taxpayers have been grouped into 100 equal-sized bins. Right: Estimated audit rates among Black and non-Black taxpayers, calculated using the probabilistic audit rate estimator and the linear disparity estimator with BIFSG-predicted probabilities. Error bars show the 95% confidence interval, derived from the asymptotic distributions described in Appendix B.3.

heavily audited throughout the income distribution.[19] Notably, the difference in audit rates appears largest for taxpayers with incomes that potentially qualify for the EITC. To more directly explore the role of the EITC in the observed racial audit disparity, we investigate differences in audit rates by EITC claim status as well as differences in EITC claiming by race. The right panel of Appendix Figure A.3 shows that the audit rate among EITC claimants is more than 4 times higher than among non-EITC claimants (1.45 vs. 0.31 pp). In addition, the left panel of the figure shows that the EITC claim rate is increasing in the probability that a taxpayer is Black. Hence, one possibility is that the observed difference in audit rates could be due to EITC claimants being audited at higher rates and Black taxpayers being over-represented among that group.

To assess this hypothesis, we estimate audit disparities by race separately for EITC claimants and non-claimants. If differences in EITC claiming rates by race account for the racial disparity, we would expect the racial disparity to be relatively small *within* the population of EITC claimants. However, Figure 5 shows this is not the case. The estimated

[19]For ease of exposition, we focus on the more conservative probabilistic estimate of the disparity, which constitutes a lower bound in our setting. Appendix Figure A.2 shows a similar pattern using the linear disparity estimator.

Figure 4: Audit Rates by Income and Race



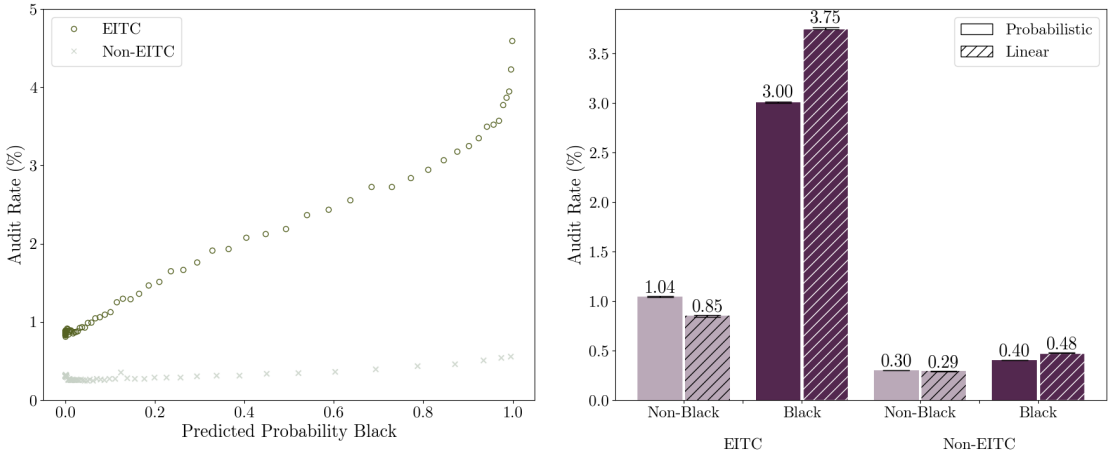
Notes: The figure shows the estimated audit rate among Black and non-Black taxpayers filing returns for tax year 2014. Income is measured according to the adjusted gross income (AGI) reported on the taxpayer's return (i.e., prior to audit adjustments). Binned audit rates by race are determined using the probabilistic estimator; Appendix Figure A.2 reports the corresponding analysis, but with audit rates estimated using the linear disparity estimator. The sample is limited to taxpayers reporting non-negative AGI. Taxpayers have been grouped into 20 equal-sized bins, based on their AGI. To facilitate presentation, the x-axis limited to taxpayers with reported AGI under $200,000. Error bars show the 95% confidence interval, derived from the asymptotic distributions described in Appendix B.3.

disparity in audit rates between Black and non-Black EITC claimants is substantially larger (between 1.96 and 2.90 p.p.) than the estimated disparity for the full population (between 0.81 pp and 1.34 p.p.). In contrast, we estimate significantly smaller racial audit disparities among taxpayers not claiming the EITC (between 0.10 and 0.18 p.p.).[20]

We can formally decompose the overall audit disparity into the following three components: (1) racial differences in the audit rate among EITC claimants; (2) racial differences in the audit rate among EITC non-claimants; and (3) racial differences in the

[20]These disparities are precisely estimated; refer to columns 2 and 3 of Appendix Table A.5 for standard errors. As discussed in Section 5, we lack empirical evidence that the linear disparity estimator yields an upper bound on the racial audit disparity for EITC non-claimants. In our linked North Carolina data set containing self-reported race, the true racial audit disparity for this group is slightly larger than, but similar in magnitude to, the disparity obtained from the linear disparity estimator.

Figure 5: Estimated Audit Rates by Race and EITC Claim Status



Notes: The figure shows the relationship between audits and race among taxpayers filing returns for tax year 2014, broken out by whether a taxpayer claims the EITC in that year. Left: Binned scatterplot of audit rate by BIFSG-predicted probability Black by EITC claim status, with EITC claimants and non-claimants each grouped into 100 equal-sized bins based on their estimated probability of being Black. EITC claimants are represented by dark green dots and non-claimants by light gray x's. Right: Estimated audit rate by race and EITC claim status, calculated using the probabilistic audit rate estimator and the linear audit rate estimator with BIFSG-predicted probabilities. Error bars show the 95% confidence interval, derived from the asymptotic distributions described in Appendix B.3.

rate at which taxpayers claim the EITC, scaled by differences in the audit rate for EITC versus non-EITC returns (see Appendix D for details). We estimate that the racial audit disparity within EITC returns contributes 78% of the overall disparity. The remainder is due to disproportionate auditing of EITC returns (14%), and to a lesser extent, disparities in audit selection among non-EITC returns (8%).[21]

We next explore the type of audits from which the disparity arises. Appendix Table A.6 shows that audit disparities appear to be largely driven by differences in the selection of correspondence audits, whereas Black and non-Black taxpayers appear to be selected for field and office audits at similar rates. We observe disparities in both pre-refund and post-refund audits, although the magnitude is larger in the former than in the latter even once we limit consideration to correspondence audits. Although audit disparities are largely concentrated among EITC claimants, correspondence audits appear to drive the smaller disparity we

[21]Note that the large contribution of the within-EITC disparity to the overall-population disparity is partly due to the relatively large share of audits that involve EITC-claiming returns; if this fraction was reduced, the percentage contribution of the within-EITC disparity would be reduced.

observe among non-claimants as well.

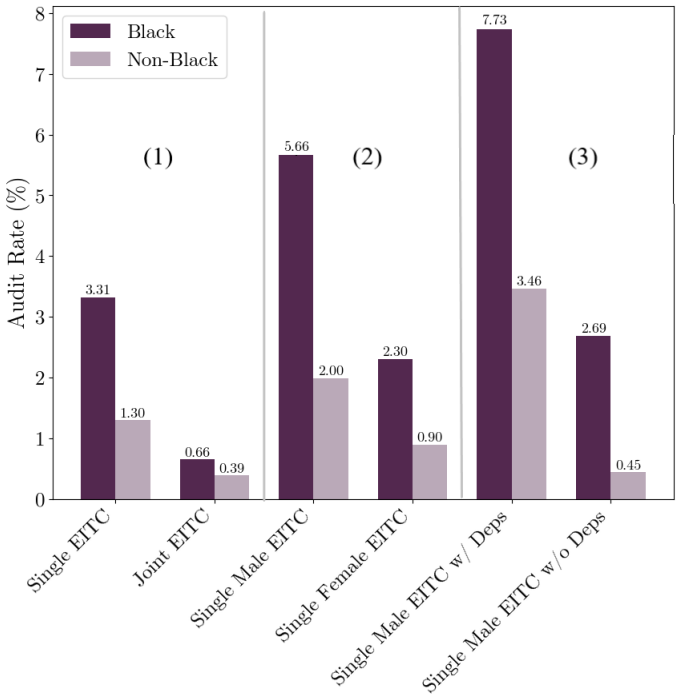
Finally, we conduct several analyses to assess the robustness of our disparity estimate. First, recall that 27% of the taxpayers in our sample are missing one or more of the variables used to impute race. Column 1 of Appendix Table A.7 re-estimates the racial audit disparity after excluding this group; the results are largely unchanged. Second, as an alternative to BIFSG, we re-estimate predicted race using additional individual characteristics and relaxing the naive Bayes independence assumption with Gibbs sampling (see Appendix B.6 for details). The resulting disparity estimate is similar to our main findings (Column 2). Third, our BIFSG race probability estimates are derived from samples that are designed to be representative of the overall U.S. population, not the subset of the population that files taxes or claims the EITC. If this causes our race probability estimates to be mis-calibrated, it could bias our disparity estimates. In Appendix B, we formally derive this bias and use the result to re-calibrate our race probability estimates, treating the North Carolina data as ground truth. Again, the results are largely unchanged from our baseline analysis (Columns 3 and 4 of Appendix Table A.7). Last, our results thus far have focused on tax returns filed for 2014. To confirm the patterns we observe are not limited to that year, we estimate disparity among all taxpayers (Appendix Figure A.4) and among EITC claimants (Appendix Figure A.5) for tax years 2010, 2012, 2016, and 2018. In each case, we obtain similar results as for 2014.

6.2 Disparity Among EITC Claimants

Because the racial audit disparity appears concentrated among EITC claimants, our remaining analyses home in on this population of taxpayers.

Figure 6 reports estimates of the conditional audit disparity among various demographic sub-populations of EITC claimants. We observe significant absolute and relative disparities by race among unmarried EITC claimants, particularly unmarried men. Strikingly, among unmarried EITC claimants with dependents, the audit rate for Black men is over 4 percentage

Figure 6: Audit Rate Disparities by EITC Subgroup



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 probabilistic 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. A similar analysis, corresponding to the linear disparity estimator, is presented in Appendix Figure A.6

points larger than the audit rate for non-Black men, and both are an order of magnitude larger than the audit rate for the overall U.S. population. In contrast, we observe smaller (but still positive) racial audit disparities among joint filers, unmarried women, and unmarried men who do not claim dependents (although the ratio of audit rates among Black to non-Black taxpayers remains substantial among these groups).

The fact that the disparities persist after conditioning on these demographic characteristics suggests that racial differences in the distribution of these characteristics are not sufficient to explain the observed disparity in audits. Similarly, differences in audit rates by race among EITC claimants persist when flexibly controlling for income, family composition, or their interaction (Appendix Table A.8).

To better understand the source of the observed racial audit disparity, we next explore

whether it can be explained by group-level differences in tax underreporting between Black and non-Black taxpayers. By underreporting, we refer to the difference between a taxpayer's correct federal income tax obligations for a year (which may be negative in the case of a taxpayer qualifying for refundable credits) and the federal tax obligations reported on the taxpayer's return. For example, underreporting may arise from reporting too little income, too many deductions, or from claiming a tax credit for which the taxpayer does not qualify. Underreporting may be intentional or inadvertent, and may be due to decisions by either the taxpayer or a tax preparer.

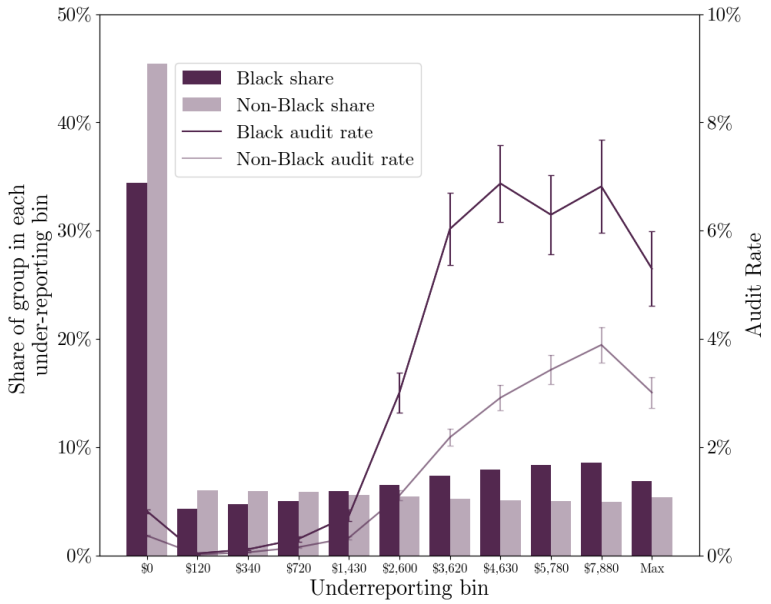
A challenge in exploring whether differences in underreporting are generating the observed disparity in audit rates is that we observe underreporting only among those taxpayers who were selected for audit. However, we can circumvent this obstacle by combining NRP and operational audit data. That is, using Bayes rule, we can express the audit rate for a given group (j) and amount of underreporting (k) as:

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

From this, we can estimate the racial audit disparity at a given level of underreporting using the disparity estimators along with the operational audit data results reported in Figure 3 for $\Pr(Y = 1|B = j)$; the NRP data to estimate $\Pr(K = k|B = j)$; and the detected underreporting from the operational audit results to estimate $\Pr(K = k|Y = 1, B = j)$.[22] See Appendix B.7 for further details. The results of this analysis are presented in Figure 7, with EITC claimants binned according to their underreported taxes. The distribution of underreporting among Black taxpayers appears shifted to the right relative to non-Black taxpayers. However, within each underreporting bin, the estimated audit rate for Black taxpayers exceeds that of non-Black taxpayers, and for many bins, the difference

[22]With respect to the last of these, operational audits may not detect all of a taxpayer's underreporting because unlike the NRP, such audits do not assess each issue on the return. In practice, this concern is lessened by the fact that the issues worked by operational audits tend to be those where underreporting is suspected.

Figure 7: Racial Audit Disparity Among EITC Claimants by Underreported Taxes



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. Bin labels on the x-axis reflect the upper dollar limit of each underreporting bin (rounded for confidentiality). Estimated audit rates by race are calculated using the probabilistic 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. A similar analysis, corresponding to the linear disparity estimator, is presented in Appendix Figure A.7.

is substantial. We interpret this result as evidence that the observed audit rate disparity cannot be entirely explained by group-level differences in underreporting by race.

7 Factors that May Influence Disparity

To better understand how alternative policies may shape racial disparities in audit rates, we next consider the role of several factors: (1) differences by race in the distribution of tax underreporting; (2) differences by race in the distribution of *predicted* tax underreporting; (3) the objective of the prediction model used to allocate audits; and (4) operational decisions that govern the allocation of audits across categories of returns. The models we will use to study these questions do not aim to replicate IRS's existing operational audit selection

processes, which are confidential. Rather, we build our own (counterfactual) risk prediction models to enable us to study various audit policy design choices. All of the algorithms considered in this section are trained and evaluated using returns that were randomly selected for inclusion into the NRP. We restrict our focus to audits of EITC returns, as our earlier results show their out-sized role in driving status quo audit disparities.

In measuring the racial audit disparity that an algorithm would produce, we focus on the results given by the probabilistic disparity estimator. As discussed above, this approach likely yields a conservative estimate of the racial audit disparity in our setting. We obtain qualitatively similar results using the linear disparity estimator (reported in Appendix A).

7.1 Does Maximizing Detected Underreporting Generate a Racial Disparity?

We first investigate the racial audit disparity that would result if taxpayers could be selected for audit in descending order of the true amount of their under-reported taxes. By under-reported taxes, we refer to any additional tax liability uncovered during the NRP audit process.[23] If this approach generated significant racial disparities, it would suggest a fundamental trade-off between detecting under-reported taxes and avoiding racial disparities in audit selection.

To study this question, we employ an "oracle" audit selection algorithm, which prioritizes returns for audit according to the total dollar amount of underreporting that would be found if the return was audited. Specifically, we rank taxpayers based on their under-reported taxes and then, in descending order, select taxpayers for audit until some pre-specified audit rate has been reached. For each audit rate that we consider, we calculate (1) total (annualized)

[23] Although they are intensive, even NRP audits may miss some underreporting, and differential mismeasurement of underreporting by race could distort our results (e.g., Obermeyer et al., 2019). However, the factors that have been found to induce inaccuracies in NRP audit results are concentrated at the top of the income distribution (Guyton et al., 2021); hence, we do not expect this concern to be important for the EITC population – our focus here. Note also that detected underreporting is distinct from revenue to the agency; for example, the former does not account for whether the detected underreporting is collected from the taxpayer or incorporate IRS administrative costs incurred by conducting the audit.

detected underreporting by summing the audit adjustments for the selected taxpayers, and (2) the racial audit disparity that would be induced by this selection process.[24] We focus on oracle-induced disparities rather than simple comparisons of underreporting by group in order to capture the relevant decision margin at operationally relevant audit rates (Simoiu et al., 2017).

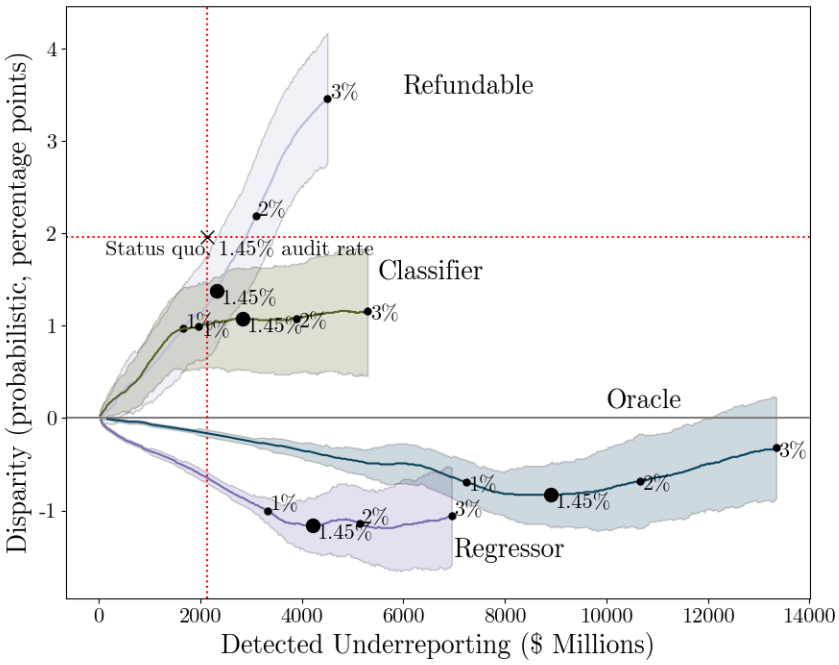
Figure 8 shows the results of this analysis. At each rate considered, the oracle selects Black taxpayers for audit at a *lower* rate than non-Black taxpayers. For comparison, the detected underreporting and disparity from audits of EITC returns for 2014 (the "status quo") are marked by the dashed red lines. At the status quo audit rate of 1.45%, the oracle obtains much more detected underreporting (as expected), while selecting Black taxpayers at a *lower* rate than non-Black taxpayers. Because the oracle represents an upper bound on the under-reported taxes that an audit algorithm can detect, these results suggest it may be possible to reduce racial audit disparities without sacrificing on the accuracy of audit selection.

7.2 Does Selection Based on Predicted Underreporting Generate a Racial Disparity?

The previous subsection provides evidence that selecting audits according to underreporting amount would not result in Black taxpayers being audited at higher rates than non-Black taxpayers. In practice, of course, the amount of underreporting that would be detected on audit is unavailable at the time the audit decision is made. We now ask whether the same pattern holds when audits are selected according to *predicted* underreporting. To study this question, we train a random forest model to predict taxpayer underreporting based on features that the tax authority can observe at the time of the audit decision, and simulate selecting taxpayers in descending order of this prediction, until some specified audit rate

[24]Throughout, we use "detected underreporting" as shorthand for total recommended adjustments from audit, not accounting for appeals or rates of collection. We provide additional details for how we calculate annualized detected underreporting using NRP sample weights in Appendix F.

Figure 8: Detected Underreporting and Disparity by Algorithm



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 models for selecting EITC audits and under alternative audit rates. Models are trained and evaluated on the set of NRP EITC claimants from 2010-14; see Appendix F for details. The displayed trajectories correspond to the oracle (blue), random forest regressor (purple), random forest classifier (green), and refundable credit models (light purple). The labeled points along each trajectory represent estimated detected underreporting and disparity for the specified model 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 regression model is trained to predict underreporting. The classification model is trained to predict whether or not underreporting exceeds $100. The oracle selects returns in descending order of true underreporting. The refundable credit model is trained to predict total adjustments to EITC, CTC, and AOTC amounts. Disparity is calculated using the probabilistic disparity estimator; Appendix Figure A.8 replicates this analysis using the linear 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 model, 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 using the standard deviation of estimated disparity across 100 bootstrapped samples from the full set of NRP EITC claimants; see Appendix F for details.

has been reached. The model is trained on NRP EITC returns to predict underreporting as closely as possible; hence, it is a "regression model," in the sense that its objective is to predict the expected dollar value of underreporting, not merely the presence or absence thereof.[25] The model incorporates the same type of features used by the DIF and DDb programs to select EITC audits – information reported on tax returns supplemented with additional administrative data available to the IRS. For additional details on the model, refer to Appendix F.

The detected underreporting and disparity induced by the random forest regressor model at various audit rates are shown by the dashed purple line in Figure 8. Unsurprisingly, the regressor achieves significantly less detected underreporting than the oracle at any given audit rate, but like the oracle, the regressor selects Black taxpayers for audit at lower rates than non-Black taxpayers for the range of audit rates we consider. The random forest regressor model detects roughly twice as much underreporting as the status quo at the same audit rate; in so doing it audits Black taxpayers at a lower rate than non-Black taxpayers. Hence, the fact that IRS does not observe true underreporting does not, in itself, appear to explain the status quo disparity in EITC audits.[26]

7.3 Does Prediction Model Objective Shape Racial Disparity?

We next explore how the objective of the prediction model that is used to select audits shapes the racial audit disparity. First, we investigate the effect of allocating audits based on algorithms trained to predict a binary versus continuous measure of underreporting (classification versus regression). Second, we investigate the effect of focusing audits on predicted underreporting due to over-claimed refundable credits.

[25] This use of the word regression (common in the machine learning literature) differs from how the term is typically employed in the economics and statistics literatures, and does not refer to a linear regression.

[26] We also investigated whether disparity would be affected by allocating audits based on actual or predicted underreporting *net* of IRS audit costs. Fully accounting for audit costs is beyond the scope of this project; for example, there may be fairness concerns with auditing those taxpayers less likely to vigorously contest their assessment. In any event, Appendix Figure A.9 shows that accounting for heterogeneous auditing costs at the level of the activity code does not appear to greatly affect the disparity induced by either the oracle or regressor model at any of the audit budgets we consider.

7.3.1 Classification versus Regression

In this section, we compare risk models trained to predict the expected dollar amount of underreporting versus models trained to predict a binary indicator for the presence of any underreporting. Models trained for the latter task (i.e., "classification models") may yield fewer false positives but also tend to detect less total underreporting.[27]

To study these questions, we train a random forest classifier on the NRP data to predict whether a taxpayer under-reports by at least \$100, using the same features that were used to build the regression model.[28] In Figure 8, the classifier corresponds to the solid green line. Across audit rates, the classifier yields lower detected underreporting than the regressor, as expected. With respect to disparity, the difference between the models is stark; whereas the regression model selects Black taxpayers at lower rates than non-Black taxpayers (and at similar rates as the oracle), the classifier selects Black taxpayers at higher rates than non-Black taxpayers. The difference between the disparity induced by the classifier and regressor models at the status quo audit rate is statistically significant ($p < 0.001$).

To better understand the audit rate disparity induced by the classifier, we derive the following decomposition (see Appendix E for details):

$$D = c_B\left(f_B - f_{NB}\right) + \left(1 - c_B\right)\left(s_B - s_{NB}\right) + \left(c_{NB} - c_B\right)\left(s_{NB} - f_{NB}\right)$$

where c_j denotes the share of group j that is compliant (defined as having underreporting less than \$100); f_j denotes the false positive rate for group j (i.e., the share of compliant taxpayers in group j that are selected by the algorithm for audit); and s_j denotes the sensitivity for group j (the share of non-compliant taxpayers in group j selected by the algorithm for audit). In words, the audit rate disparity can be expressed as the sum of three terms: the first is

[27] Black et al. (2022) compares the distributional effects by income of audit selection algorithms informed by classification versus regression models. Apart from the distribution of audited taxpayers, the two approaches may also differ in their deterrence effects, with regression models providing more deterrence against large evaders and classification models providing more deterrence to those who would evade by small amounts.

[28] For additional details on the random forest classification model, refer to Appendix F. We obtain qualitatively similar results with other non-zero classification thresholds (Appendix Figure A.10).

proportional to group differences in the share of compliant taxpayers selected for audit; the second is proportional to group differences in the share of non-compliant taxpayers selected for audit; and the final term is proportional to the difference in the mean compliance rates between groups. Appendix Table E.1 applies this decomposition to the audit rate disparity induced by the classifier model described in this section. The vast majority of exam-rate disparity (73%) is attributable to differences in the model's sensitivity to non-compliance by Black versus non-Black taxpayers – i.e., to the fact that non-compliant Black taxpayers are more likely to be audited than non-compliant non-Black taxpayers. A smaller portion (25%) of the disparity is attributable to group differences in the rate of non-compliance, with the remaining 2% attributable to group differences in the algorithm's false positive rate.

7.3.2 Targeting Refundable Credits

This section explores a different component of the prediction model objective: whether audits are allocated based on total predicted underreporting or based on the portion of predicted underreporting that is due to over-claiming of refundable tax credits. Although a dollar of detected underreporting is worth the same whether it arises from underreported income or over-claimed credits, the agency may prioritize enforcement activity around the latter due to various policy or political considerations.

To explore this issue, we train a random forest regression model as above, except that we now train the model to predict the sum of NRP adjustments to the three refundable credits designated by OMB as high-priority program susceptible to significant improper payments: the EITC, the Additional Child Tax Credit, and the American Opportunity Tax Credit. We then allocate audits in descending order of the predicted refundable credit over-claim, up to the specified audit budget. The results are shown as the light purple line in Figure 8. Compared to the baseline regression model trained on total underreporting, the model trained on refundable credit over-claims achieves substantially less total detected underreporting, indicating that over-claims of refundable credits are not the only important

source of underreporting even among EITC claimants. Even more striking is the difference between the models with respect to disparity: the model trained on refundable credits selects Black taxpayers at a higher rate for all audit budgets we consider.

Based on the results in Sections 7.3.1 and 7.3.2, we conclude that the objective of the prediction model used for audit selection may be one factor shaping the observed racial audit disparity.

7.4 Do Audit Allocation Constraints Generate a Racial Disparity?

As discussed in Section 2, some elements of the audit selection process are based on considerations distinct from the output of algorithmic prediction models, such as allocation decisions based on agency resource constraints like examiner expertise, audit costs, or other policy goals. To investigate the role of one such allocation policy in producing audit rate disparities, we focus on the allocation of audits 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).[29] Audits of non-business EITC returns tend to focus on issues relating to credit eligibility rules, such as whether EITC qualifying children were properly claimed, whereas audits of EITC returns in the business income category are more likely to focus on verifying reported income and claimed deductions.

While non-business audits constitute 93% of EITC audits, they would comprise only 21% of EITC audits if our random forest regression model was used as the basis for selection. Operational constraints (e.g., the grade, experience, subject-matter focus, and training of available examiners) could push toward selecting a larger share of non-business (activity code 270) than business (activity code 271) returns, at least in the short term. In addition, we estimate Black taxpayers constitute 21% of non-business EITC returns versus 11% of business EITC returns (Appendix Figure A.11). Hence, by shifting the allocation of EITC

[29]EITC claimants are classified in Activity Code 271 if their gross receipts on Schedule C or F exceed $25,000.

audits, operational constraints could affect the distribution of audits by race.

To explore this possibility, we consider versions of the counterfactual algorithms explored thus far that are constrained to select the status quo ratio of business to non-business EITC returns for audit. Appendix Figure A.12 plots the detected underreporting and disparity associated with these constrained models, and with their unconstrained analogs, at the status quo EITC audit rate.[30] For each model, imposing the status quo allocation constraint reduces detected underreporting (as expected) and results in a larger share of Black taxpayers being selected for audit. These results therefore suggest that operational considerations of the form described above can contribute to higher audit rates for Black taxpayers. At the same time, the disparity induced by the constrained classifier and constrained refundable credit models remain higher than that induced by the constrained regressor ($p < 0.001$), suggesting that algorithmic design may continue to play an important role once operational considerations are taken into account.

8 Discussion

In this paper, we have presented evidence that Black taxpayers are audited at higher rates than non-Black taxpayers, and that this disparity is primarily due to differences in the audit rate between Black and non-Black EITC claimants. Using counterfactual audit selection algorithms, we explored some of the potential factors that might contribute to or alleviate this disparity. Our results suggest that audit disparities are not driven exclusively by racial differences in the distribution of underreporting or in the distribution of underreporting that can be predicted based on the information that IRS observes. Instead, we find that the objective of the predictive model underlying audit selection, as well as operational considerations relating to the complexity of audited tax returns, can be critical drivers of disparity.

[30] Appendix Figure A.13 shows the detected underreporting and disparity induced by the constrained models at varying audit rates.

Concretely, these results suggest potential avenues through which the IRS may – depending on its current confidential practices – be able to alleviate racial audit disparities: shifting from more classification- to regression-based prediction algorithms, equally prioritizing underreporting from refundable credits and other sources, and expending agency resources to accommodate auditing more complex EITC returns. At the same time, some of the factors we identify are shaped by forces outside the IRS's control. For example, Congress determines the information reported to the IRS by third parties – which shapes the distribution of non-compliant taxpayers that classification models identify – the rules governing credit eligibility – which may contribute to more mistakes among Black taxpayers due to racial differences in family structure – and IRS funding levels – which shapes the ability of the agency to allocate resources to complex cases.

We note several limitations to our work. First, we have focused our investigation on audit disparities for Black taxpayers, which has been the subject of significant scholarly and policy interest. Yet there is comparable interest in disparities for other racial and ethnic groups, with potentially differing causes and opportunities for mitigation. Investigating those disparities is an important avenue for future work, to which our methods can readily be extended. Second, we have focused on the trade-off between racial disparities and detected underreporting, but policymakers may prioritize other objectives as well, such as deterring non-compliant behavior, avoiding audits on compliant taxpayers, transparency, and promoting a fair distribution of audited returns by income. Relatedly, our analysis of counterfactual audit algorithms does not account for the full set of constraints facing tax authorities like the IRS, such as the types of compliance issues that can be explored through correspondence audit, or differences in audit response rates depending on whether the audit is pre- versus post-refund. A more complete optimal policy analysis would require accounting for these additional objectives and constraints. Finally, audit selection constitutes only one dimension in which tax administration may differently affect taxpayers by race. Disparities may also exist with respect to such

processes as collections, appeals, settlements, and guidance (Bearer-Friend, 2021; Book, 2021). The approach described in this paper can serve as a foundation to explore disparities in these areas as well.

References

Alao, R., Bogen, M., Miao, J., Mironov, I., and Tannen, J. (2021). How meta is working to assess fairness in relation to race in the u.s. across its products and systems.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23(2016):139–159.

Anson-Dwamena, R., Pattath, P., and Crow, J. (2021). Imputing missing race and ethnicity data in covid-19 cases.

Aughinbaugh, A., Robles, O., and Sun, H. (2013). Marriage and divorce: Patterns by gender, race, and educational attainment. *Monthly Lab. Rev.*, 136:1.

Bearer-Friend, J. (2019). Should the IRS Know Your Race? The Challenge of Colorblind Tax Data. *Tax L. Rev.*, 73:1.

Bearer-Friend, J. (2021). Colorblind tax enforcement. *NYU Law Review*.

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.

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., and 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.

Bloomquist, K. M. (2019). Regional bias in irs audit selection. *Tax Notes*.

Book, L. (2021). Tax administration and racial justice: The illegal denial of tax based pandemic relief to the nation's incarcerated. *South Carolina Law Review*, 72.

Brown, D. A. (2005). The tax treatment of children: Separate but unequal. *Emory LJ*, 54:755.

Brown, D. A. (2009). Shades of the american dream. *Wash. UL Rev.*, 87:329.

Brown, D. A. (2018). Homeownership in black and white: The role of tax policy in increasing housing inequity. *U. Mem. L. Rev.*, 49:205.

Brown, D. A. (2021). *The Whiteness of Wealth: How the Tax System Impoverishes Black Americans–And How We Can Fix It*. Crown Publishing Group (NY).

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Center on Budget and Policy Priorities (2019). Policy basics: The earned income tax credit.

CFPB (2014). *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A methodology and assessment.* Consumer Financial Protection Bureau.

Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*.

Chen, J., Kallus, N., Mao, X., Svacha, G., and 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.

Chetty, R., Hendren, N., Jones, M. R., and Porter, S. R. (2020). Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783.

Collyer, S., Harris, D., and Wimer, C. (2019). Left behind: The one-third of children in families who earn too little to get the full child tax credit.

Cook, L. D., Logan, T. D., and Parman, J. M. (2016). The mortality consequences of distinctively black names. *Explorations in Economic History*, 59:114–125.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Dean, S. A. (2021). *Testimony to House Committee on Ways and Means*.

Delevoye, A. and Sävje, F. (2020). Consistency of the horvitz–thompson estimator under general sampling and experimental designs. *Journal of Statistical Planning and Inference*, 207:190–197.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. (2019). Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 170–179.

Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.

Fryer, R. G. and Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, 119(3):767–805.

Geman, S. and 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.

Goldin, J. and Michelmore, K. (2022). Who benefits from the child tax credit? *National Tax Journal*.

Government Accountability Office (GAO) (2015). IRS Return Selection.

Government Accountability Office (GAO) (2017). Refundable Tax Credits: Comprehensive Compliance Strategy and Expanded Use of Data Could Strengthen IRS's Efforts to Address Noncompliance.

Guyton, J., Langetieg, P., Reck, D., Risch, M., and Zucman, G. (2021). Tax evasion at the top of the income distribution: theory and evidence.

Guyton, J., Leibel, K., Manoli, D. S., Patel, A., Payne, M., and Schafer, B. (2018). The effects of eitc correspondence audits on low-income earners.

Haas, A., Elliott, M. N., Dembosky, J. W., Adams, J. L., Wilson-Frederick, S. M., Mallett, J. S., Gaillot, S., Haffer, S. C., and Haviland, A. M. (2019). Imputation of race/ethnicity to enable measurement of hedis performance by race/ethnicity. *Health Services Research*, 54(1):13–23.

Hardy, B., Hokayem, C., and Ziliak, J. P. (2021). Income inequality, race, and the eitc. *Working paper*.

Holtzblatt, J. and McCubbin, J. (2004). Issues affecting low-income filers. *The crisis in tax administration*, 148:148–49.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, pages 263–272.

Internal Revenue Service (IRS) (2011). Internal revenue manual 4.19.20.

Internal Revenue Service (IRS) (2016). Internal revenue service data book, 2015.

Joint Committee on Taxation (JCT) (2020). Estimates of federal tax expenditures for tax years 2020-2024.

Kallus, N., Mao, X., and Zhou, A. (2021). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

Kiel, P. and Fresques, H. (2019). Where in The U.S. Are You Most Likely to Be Audited by the IRS?

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1):null.

Leibel, K., Lin, E., and McCubbin, J. (2020). Social welfare considerations of eitc qualifying child noncompliance. *Treasury Office of Tax Analysis Working Paper*.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163.

Moran, B. I. and Whitford, W. (1996). A Black Critique of the Internal Revenue Code. *Wis. L. REv.*, page 751.

National Taxpayer Advocate (2019a). Annual report to congress 2019.

National Taxpayer Advocate (2019b). Report: Making the eitc work for taxpayers and the government.

National Taxpayer Advocate (2021). Annual report to congress 2021.

Nerenz, D. R., McFadden, B., Ulmer, C., et al. (2009). Race, ethnicity, and language data: standardization for health care quality improvement.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Office, G. A. (2021). Artificial intelligence: An accountability framework for federal agencies and other entities.

Office of Management and Budget (OMB) (2021). Circular a-123, appendix c: Requirements for payment integrity improvement.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. (2020). An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*, volume 110, pages 91–95.

Robinson, P. (1982). On the convergence of the horvitz-thompson estimator. *Australian Journal of Statistics*, 24(2):234–238.

Simoiu, C., Corbett-Davies, S., and Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer Series in Statistics. Springer, 3rd edition.

TIGTA (2021). The earned income tax credit examination compliance strategy can be improved.

Tzioumis, K. (2018). Demographic aspects of first names. *Scientific data*, 5(1):1–9.

U.S. Census Bureau (2021). Decennial census surname files (2010, 2000).". `https://www.census.gov/data/developers/data-sets/surnames.html`, Last accessed on 2023-01-12.

U.S. Executive Order 13985 (2021). Exec. order no. 13985 86 fed. reg. 7009, advancing racial equity and support for underserved communities through the federal government.

U.S. Treasury Department (2022). Agency financial report: Fiscal year 2021.

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13.

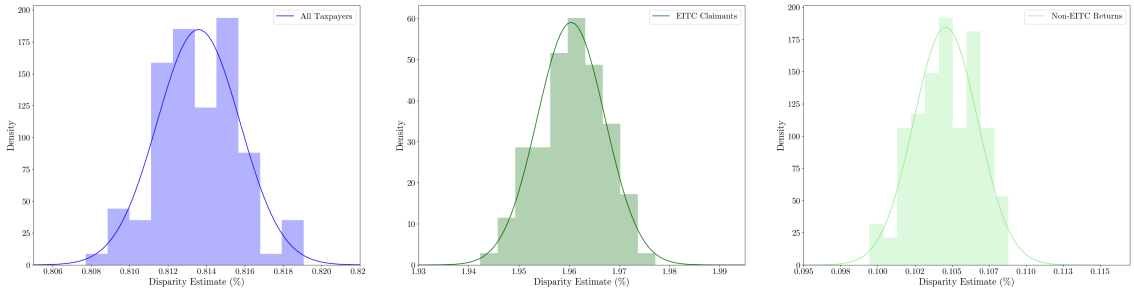
Zhao, H. and Gordon, G. (2019). Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685.

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

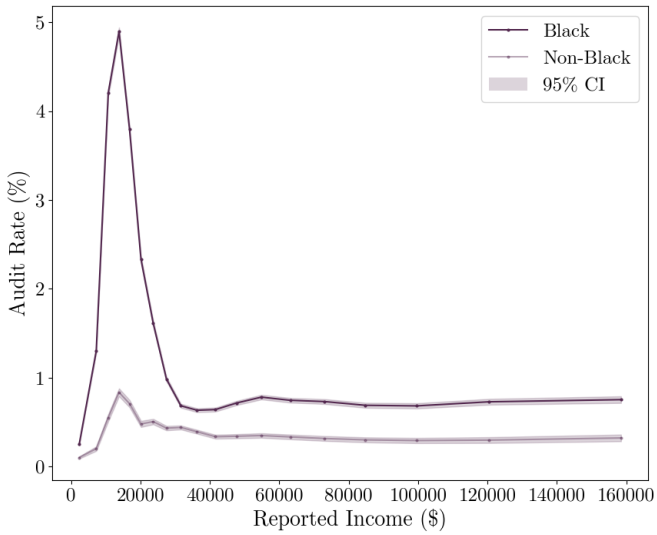
A Additional Tables and Figures

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



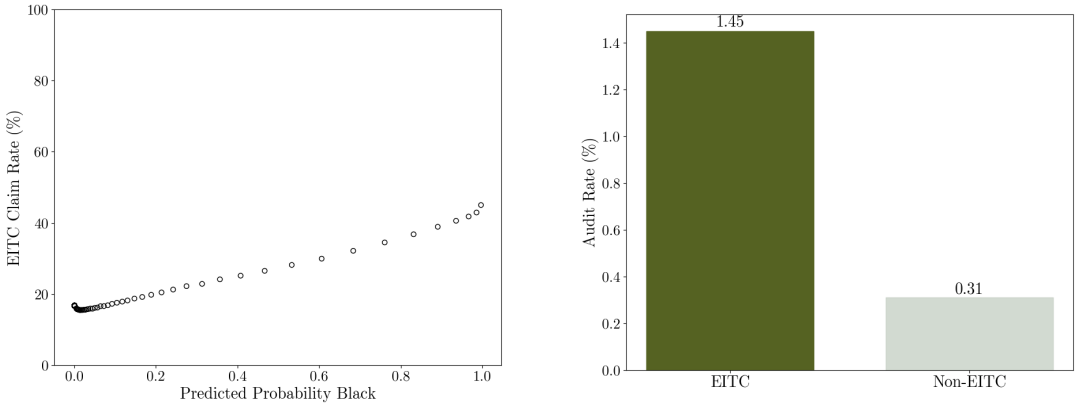
Notes: The figure shows statistical variation in the estimated audit rate disparity as estimated by the method of composition (Tanner, 1996); it captures uncertainty both due to measurement uncertainty (uncertainty around true race) as well sampling variability (i.e. the probabilistic nature of an audit). 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)



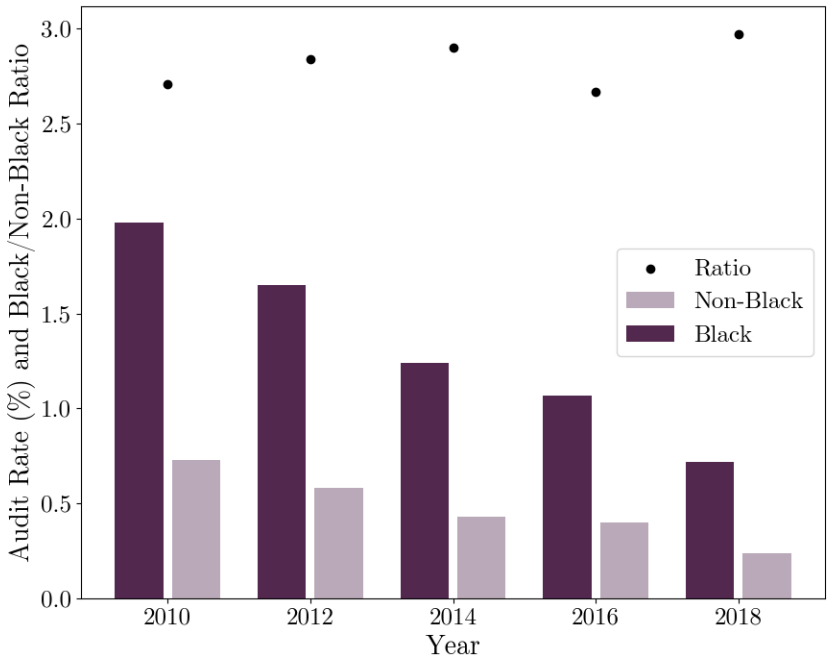
Notes: The figure shows the estimated disparity in audit rates, using the linear disparity estimator, between Black and non-Black taxpayers across bins of reported Adjusted Gross Income (AGI) for returns filed in tax year 2014. Taxpayers are grouped into 20 equal-sized bins. Disparity is calculated within 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



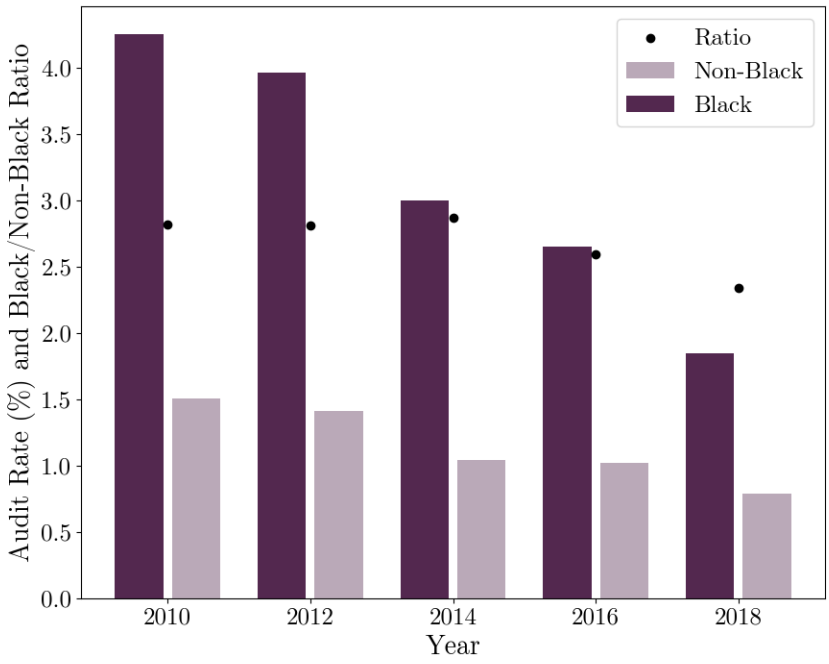
Notes: The figure shows the relationship between audits and EITC claim status among taxpayers filing returns for tax year 2014. Left: Binned scatterplot of EITC claim rate by BIFSG-predicted probability that a taxpayer is Black. Taxpayers have been grouped into 100 equal-sized bins. Right: Audit rates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

Figure A.4: Estimated Audit Rate Disparity by Year



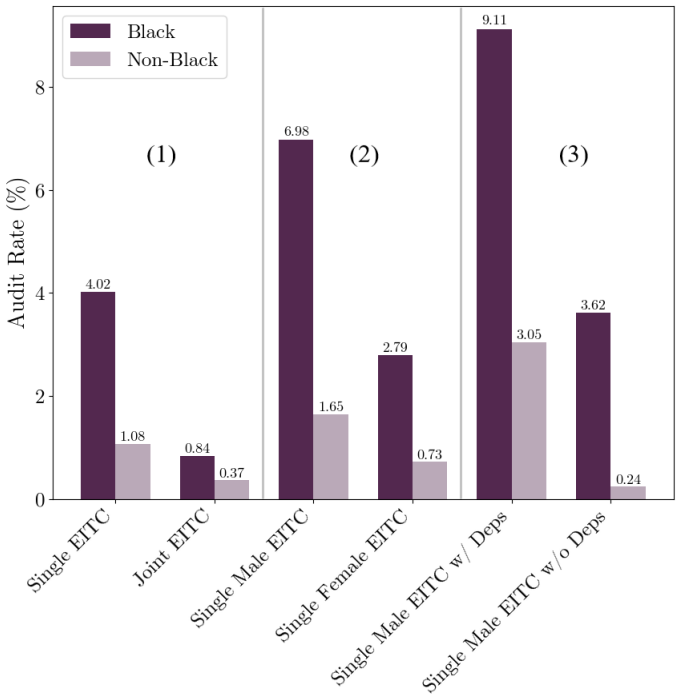
Notes: The figure reports the estimated audit rates among Black and non-Black taxpayers for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). "Ratio"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



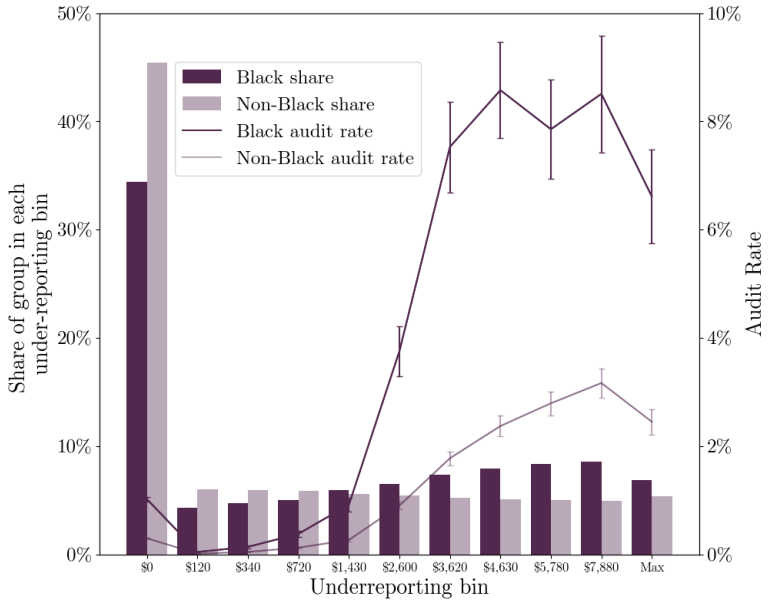
Notes: The figure reports the estimated audit rates among Black and non-Black EITC claimants for tax years 2010, 2012, 2014, 2016, and 2018, calculated using the probabilistic audit rate estimator applied to BIFSG-predicted probabilities (calculated using the data sources described in Section 4.2). "Ratio"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



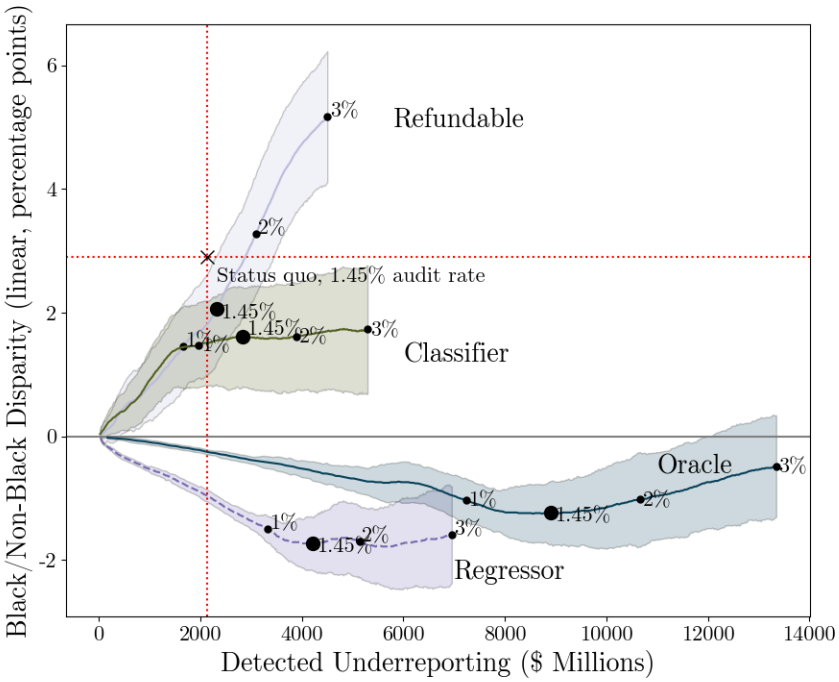
Notes: The figure shows the estimated audit rate among the specified subgroups of Black and non-Black taxpayers. Conditional audit rates by race are calculated using the linear audit rate estimator applied to BIFSG-predicted probabilities that a taxpayer is Black. Panel (1) splits EITC claimants by single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



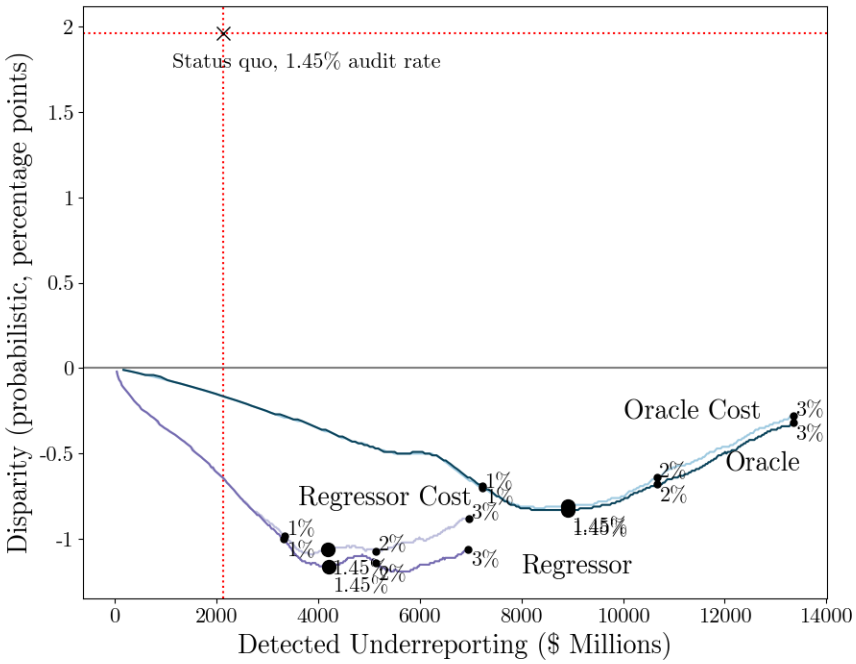
Notes: The figure shows the estimated audit rates for Black and non-Black EITC claimants, respectively, by under-reported taxes. Taxpayers are binned into 11 categories: those with less than $1 of under-reporting, and 10 equal deciles of taxpayers with positive under-reporting. Under-reporting deciles 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)



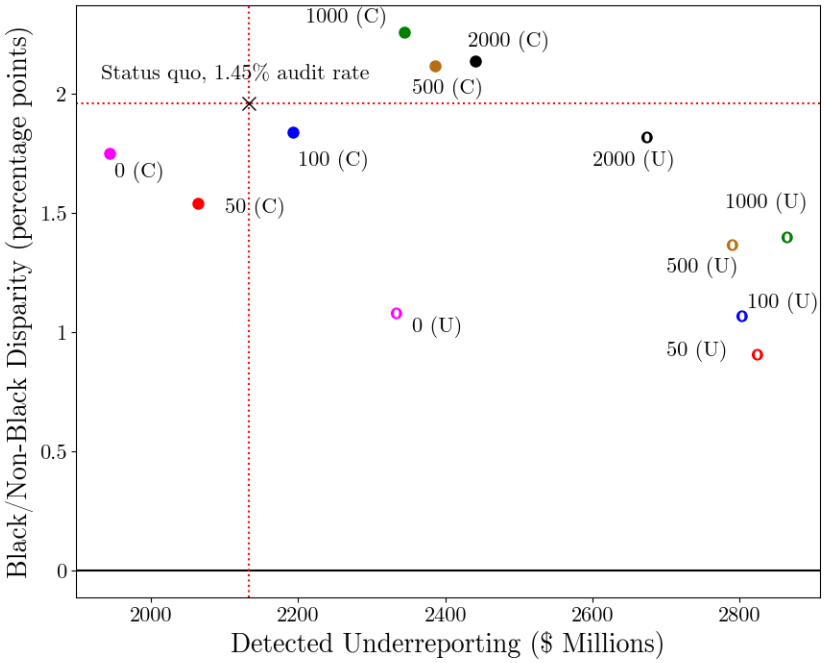
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative models for selecting EITC audits and under alternative audit rates. The trajectories correspond to the oracle (blue), random forest 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



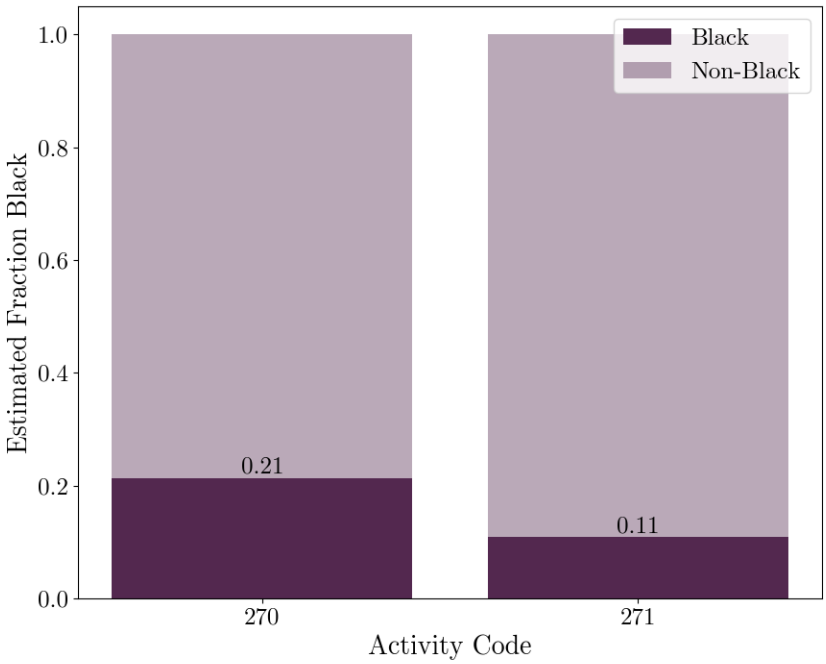
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) under alternative assumptions about whether returns are selected for audit based on detected underreporting (gross of audit costs, as in our other 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



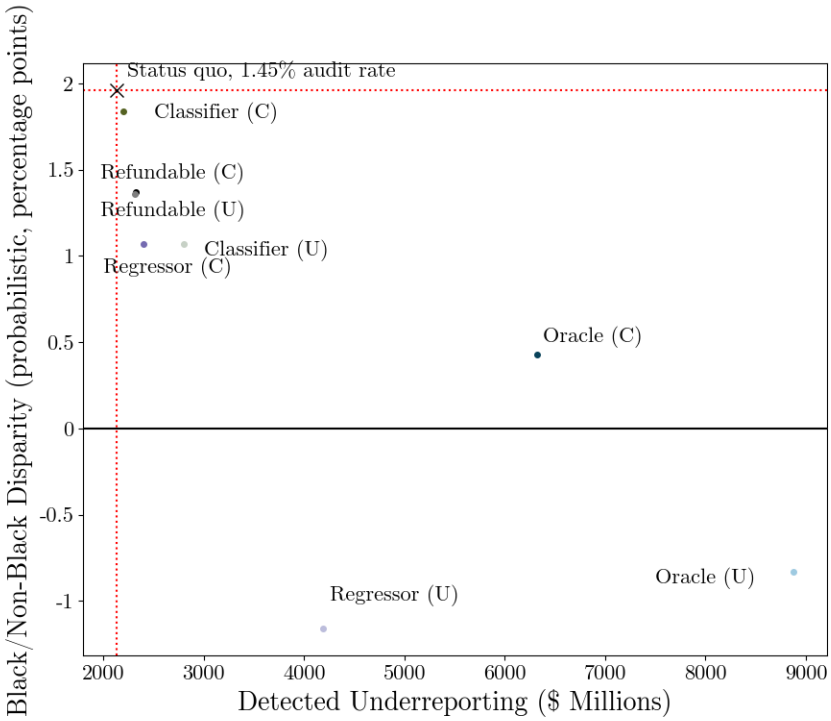
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) for random forest classification models trained on alternative dollar thresholds, under the assumption that 1.45% of the EITC population is selected 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



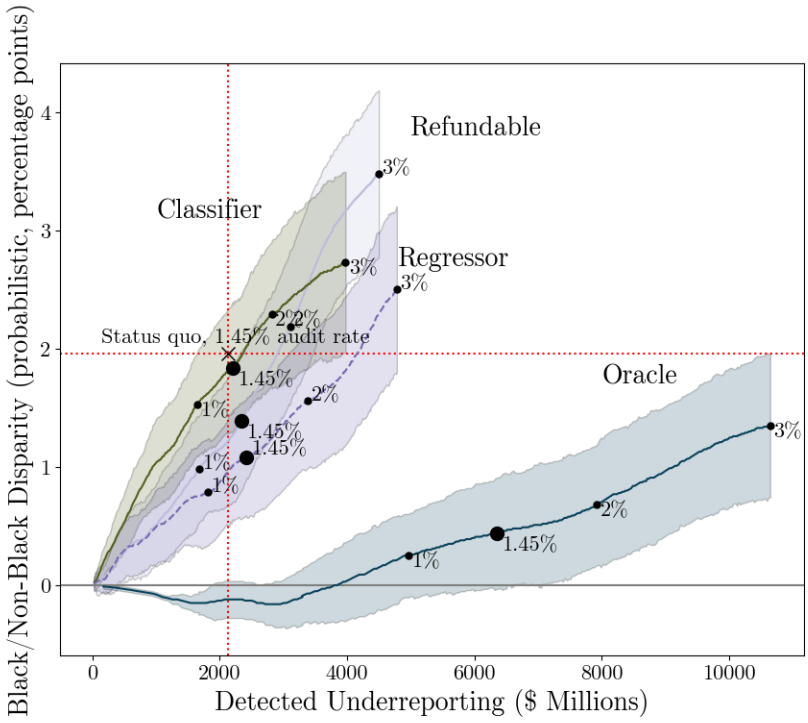
Notes: The figure shows the share of taxpayers who are Black and non-Black across the two activity codes into which EITC tax returns are categorized: returns with substantial business income (activity code 271) and returns without substantial business income (activity code 270). Shares are estimated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

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



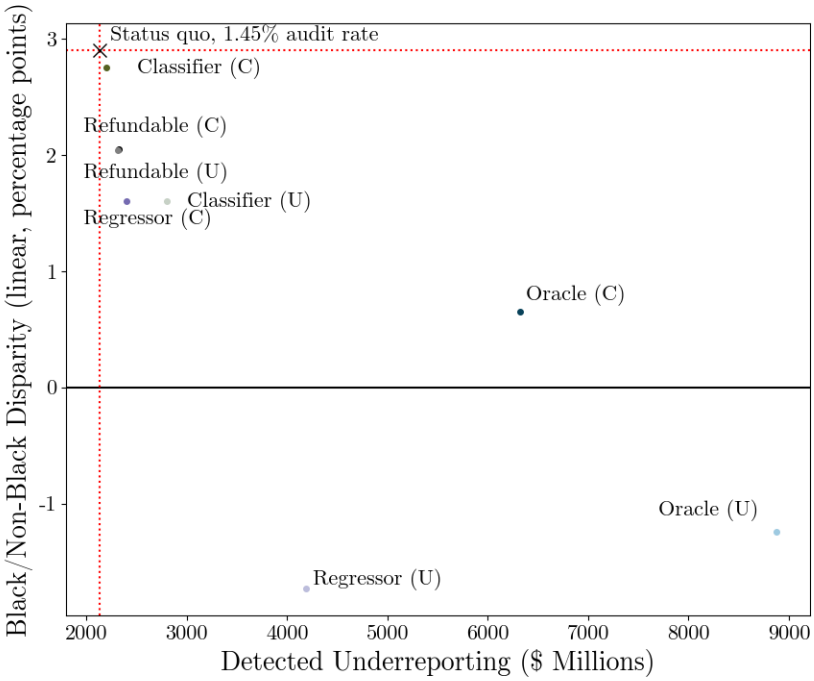
Notes: The figure shows the implied difference in audit rates between Black and non-Black taxpayers (y-axis) and annualized detected underreporting (x-axis) for alternative audit selection models, under the assumption that 1.45% of the EITC population is selected for audit. The points correspond to the 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.

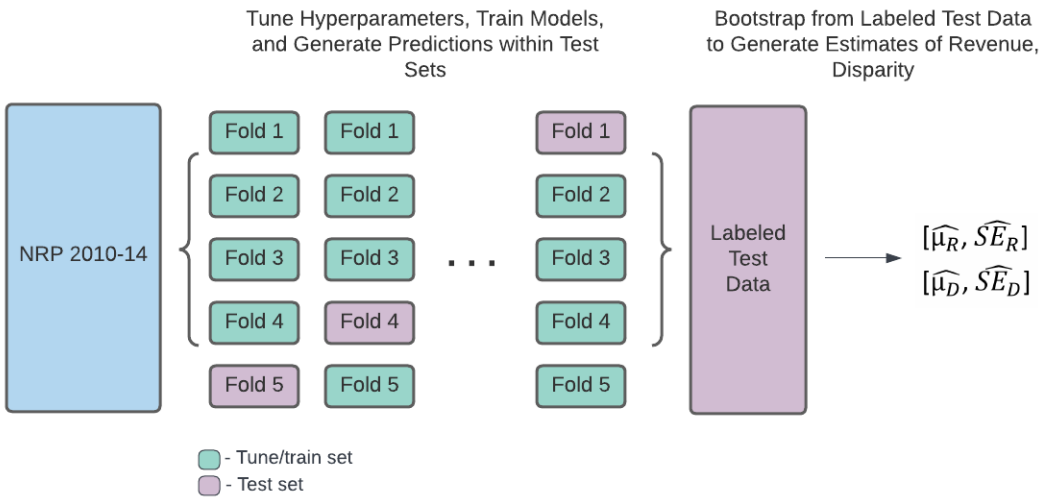
E.3 Empirical Implementation

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.