# Measuring and Mitigating Racial Disparities in Tax Audits *

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

July 19, 2024

**Abstract**

Tax authorities around the world rely on audits to detect underreported tax liabilities and to verify that taxpayers qualify for the benefits they claim. We study differences in Internal Revenue Service 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. An important driver of the disparity is differing audit rates by race among taxpayers claiming the Earned Income Tax Credit (EITC). Using counterfactual audit selection models to explore why the disparity arises, we find that maximizing the detection of underreported taxes would not lead to Black EITC claimants being audited at higher rates. Rather, the audit disparity among EITC claimants stems in large part from a policy decision to prioritize detecting overclaims of refundable credits over other forms of noncompliance. Modifying the audit selection algorithm to target total underreported taxes while holding fixed the number of audited EITC claimants would reduce the share of audited taxpayers who are Black, and would lead to more audits focused on accurate reporting of business income and deductions; fewer audits focused on the eligibility of claimed dependents; higher per-audit costs; and more detected noncompliance.

*The views presented here are those of the authors and do not necessarily represent the position of the Treasury Department. The Treasury Department reviewed these results to ensure compliance with statutory prohibitions on the disclosure of taxpayer information and with policy prohibitions on the disclosure of confidential information related to IRS audit processes. For helpful comments, we thank Emily Black, Edith Brashares, Dorothy Brown, Jonathan Choi, Geoffrey Gee, Robert Gillette, Alissa Graff, John Guyton, Anne Herlache, Jim Hines, Janet Holtzblatt, Hilary Hoynes, Tatiana Homonoff, Barry Johnson, Larry Katz, Elaine Maag, Michael Morse, Julian Nyarko, Nina Olson, Derek Ouyang, Claire Lazar Reich, Kit Rodolfa, Dan Rosenbaum, Joel Slemrod, David Splinter, Megan Stevenson, Alex Turk, Caroline Weber, the staff members of the Joint Committee on Taxation who contributed to a consolidated set of comments, and seminar participants. 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.

Elzayn: First author; Stanford University. Smith: University of Michigan. Hertz: Internal Revenue Service. Guage: Stanford. Ramesh: UCLA. Fisher: U.S. Treasury Department. Ho: Equal co-supervising author; Stanford University. Goldin: Equal co-supervising author; University of Chicago, American Bar Foundation, and NBER; Corresponding author: jsgoldin@uchicago.edu.

# 1 Introduction

In recent decades, U.S. policymakers have increasingly relied on the income tax system to implement 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 partly a prediction exercise—which taxpayers have underreported tax obligations that an audit would uncover?—and partly a policy determination—what type of underreporting should be predicted and pursued? Evidence from many domains in which data-driven algorithms are used to allocate enforcement resources suggests that both aspects of this process may inadvertently reinforce disadvantages against historically marginalized groups (Angwin et al., 2016; Buolamwini and Gebru, 2018; Obermeyer et al., 2019). 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., 2018a) makes it difficult to understand why disparities emerge or which policies would lessen them.

We investigate this topic using comprehensive administrative 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) to impute race based on taxpayers' names 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 an important contributor to the racial audit disparity is a difference in audit rates among Black and non-Black 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 Fresques, 2019). However, we find that this channel does not fully explain the disparity we observe. Rather, we estimate a substantial disparity in the selection of taxpayers for audit *within* the population of EITC claimants: Black taxpayers claiming the EITC are between 2.9 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.

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 factors contributing to the observed audit disparity among taxpayers claiming the EITC. Because EITC audit selection is largely automated, and because the IRS does not observe race, the disparity is unlikely to be driven by disparate treatment of Black and non-Black taxpayers.[2] At the same time, we find that the disparity cannot be explained by group-level differences in total dollars of tax underreporting: we estimate that Black EITC claimants are audited at higher rates than non-Black EITC claimants within each decile of under-reported taxes.

To better understand the source of the audit disparity in light of this finding, we simulate counterfactual audit selection algorithms for EITC claimants, using data from audits of randomly selected tax returns. We explore potential explanations related to both aspects of the audit selection process described above: predicting which taxpayers have underreported taxes that an audit would uncover and the policy decision concerning which types of underreporting to predict and pursue (the algorithmic objective).

Beginning with the algorithmic objective, recall that taxpayers may underreport taxes by misreporting their taxable income and/or by overstating their eligibility for tax-administered benefits, often in the form of refundable credits. Although details of the IRS audit selection algorithm are confidential, the agency publicly reports that the model underlying its primary program for auditing EITC claimants is designed to detect noncompliance from this latter source of underreporting — i.e., overclaimed refundable credits. Using our simulated audit algorithms, we find that when audits of EITC claimants are allocated based on this objective, Black taxpayers are audited at higher rates than non-Black taxpayers. In contrast, if audits were allocated based on the objective of maximizing total dollars of detected underreporting — from any source of noncompliance — we find that Black EITC claimants would be audited

[2]By disparate treatment, we mean different audit rates for taxpayers who file identical tax returns and differ only with respect to their race.

at *lower* rates than non-Black EITC claimants.

Why does the algorithmic objective shape the distribution of audits by race? Among taxpayers claiming the EITC, the highest underreporting is associated with taxpayers who have substantial independent contractor or other business income, but who underreport this income to such a degree that they appear to qualify for the EITC. Among this group, we find that those with the most underreporting are disproportionately non-Black. Despite their high levels of noncompliance, audit algorithms focused exclusively on refundable credits under-prioritize these taxpayers because only a portion of their underreporting takes the form of overclaimed credits – the rest stems from underreported taxable income. At the same time, we find that EITC returns filed by Black taxpayers appear more likely to violate the legal requirements for claiming children, which translates into larger average dollar adjustments to the refundable credits claimed on their returns. As a result, audit algorithms that are exclusively focused on this form of noncompliance select Black taxpayers at higher rates.

With any potential change to the audit selection algorithm, it is important to consider the consequences for the operational process through which audits are conducted. Along these lines, we show that the objective of the audit selection algorithm shapes the composition of audited issues, with substantial downstream implications for audit costs and detected underreporting. In this way, our results highlight an important connection between racial audit disparities and ongoing policy debates about the proper level of IRS funding (e.g., Boning et al., 2023).

Distinct from the objective of the audit selection algorithm, we also explore whether the architecture of the prediction model employed as part of the audit selection process contributes to the observed disparity. We find evidence that it does: although Black EITC claimants tend to overclaim refundable credits at higher rates, the observed audit disparity is larger than what these differences in overclaiming would imply. The confidential nature of the IRS audit selection process limits our ability in this paper to explore the details of the actual predictive model used by IRS to allocate EITC audits; however, we provide

suggestive evidence that differences by race in the predictive power of the available features—such as proxies for child custody or parental status—may contribute to higher audit rates for Black taxpayers. Subject to the same caveat, we provide suggestive evidence that it may be possible to modify the predictive model to lessen the racial audit disparity without substantially reducing the amount of overclaimed refundable credits detected on audit.

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 administrative microdata with information about race has limited the questions that prior studies could address and the alternative policies they could evaluate (Bearer-Friend, 2019; Brown, 2021; Dean, 2021).[4] We build on this literature by linking imputed race estimates with 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.

A second contribution of our paper is to introduce a novel partial identification approach for conducting algorithmic disparity assessments with respect to 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 industry (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).

[3]Much of this literature focuses on racial differences in the benefits from substantive tax provisions like 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).

[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.

Our approach relies on weaker conditions than those required to point-identify differences 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 relate to a growing literature studying how the choice of outcome to be predicted by an algorithm shapes the distributional properties of procedures based upon that algorithm (Barocas and Selbst, 2016; Kleinberg et al., 2018b; Passi and Barocas, 2019). For example, Obermeyer et al. (2019) link racial disparities in a health care setting to an algorithm that is trained to predict health care expenditures rather than direct health outcomes. With respect to tax audits, Black et al. (2022) study how the choice between regression and classification prediction tasks shapes the distribution of audits by taxpayer income. Related to this literature, our results highlight how policy decisions to prioritize certain forms of legal noncompliance over others can shift the distribution of enforcement burdens.

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 source of the observed audit disparity. An Online Appendix contains proofs and additional results.

# 2 Institutional Background

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

[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.

## 2.1 The U.S. Income Tax

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 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 IRS Audits

The IRS is the federal agency responsible for promoting and enforcing compliance with the tax law. One channel through which it does so is by employing 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. 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 conducted through correspondence. 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. At the same time, correspondence audits can be particularly burdensome for lower-income households, who may face additional barriers to 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 to the (net) tax reported on a taxpayer's return, the taxpayer is responsible for remitting the difference to the IRS and may face additional penalties, as well as, in rare cases, criminal sanctions. Audits may occur pre- or post-

refund; in the former case, 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. 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, research audits constitute a small minority of the audits that the IRS performs each year. For example, about 2% of audited returns for tax year 2014 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.[6] Different audit programs use different processes to select which returns to audit. For some programs, tax returns are ranked in a manner that focuses on total noncompliance, aggregated across issues on the return; selected returns are then "classified" by IRS examiners to identify the most promising potential noncompliance issues on which to focus. Other audit programs are focused more narrowly on a particular issues or set of issues; returns audited through such programs are selected based on criteria that relate to the specific noncompliance issues of focus.[7]

To facilitate our research, the IRS shared information on operational audit selection processes with members of our research team; however, IRS policy limits our ability to

[6]We observe all training data and the full set of taxpayer features for EITC audits – 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).

[7]The set of issues upon which an audit focuses may constrain the process through which the audit is conducted; for example, audits of claimed business deductions may be more amenable to correspondence audit than audits of unreported business income.

disclose information about these processes that would allow taxpayers to manipulate their risk of selection, such as tax return characteristics that may drive audit selection. The federal government itself has publicly disclosed some inputs into audit selection, such as the use of child custody and child birth records to assess whether a taxpayer is eligible to claim a child for a particular credit (G.A.O., 2015), and has also made clear that certain taxpayer characteristics are not taken into account, such as race or where the taxpayer lives (Kiel and Fresques, 2019).

Distinct from formal audits, 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 (including returns claiming benefits for which the taxpayer does not qualify based on observable characteristics such as the taxpayer's age or reported income), or other factors. In other cases, the IRS will flag a potential compliance issue on a submitted return through a "soft notice" or other process for the taxpayer to resolve the issue without undergoing a formal audit, such as certain situations in which multiple taxpayers claim the same child for the same year. Thus, many of the returns with "smoking gun" evidence of non-compliance are addressed outside of the formal audit process.

## 2.3 The Earned Income Tax Credit

The Earned Income Tax Credit (EITC) is a tax credit designed to support low- and middle-income taxpayers with earnings from work. The credit is refundable, meaning that taxpayers receive a payment or "refund" from the government if it brings their tax liability for the year below zero. Today, the EITC constitutes the largest cash-based safety net program in the United States, benefiting approximately 31 million households at an annual budgetary cost of $64 billion (IRS, 2023a).

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 of up to $52,247 if married, or $46,997 if single. Taxpayers without income from work do not qualify for any EITC amount. Approximately 19% of taxpayers claimed the EITC in 2014, with an average credit of $2,122 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 40% or more of individual taxpayer audits have been

focused on returns claiming the EITC (Congressional Research Service, 2022).[8] EITC returns are selected for audit through a variety of enforcement programs; as such, EITC claimants may be audited in a manner that is narrowly focused on their eligibility for the EITC (e.g., their eligibility to claim a particular dependent) or in a manner that investigates other potential forms of noncompliance (e.g., the accuracy of their reported income or their eligibility to claim a business deduction). 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).

Approximately three-quarters of audited EITC returns in 2014 were 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.[9] For DDb audits, the selection step of the process is automated without manual review by human examiners (IRS, 2011).

# 3 Empirical Framework

In this section, we provide results relating to the identification of disparities in an outcome (like audits) with respect to a characteristic (like race) that cannot be directly observed by the researcher.

[8]The high EITC audit rate originates from a 1997 deal between House Republicans and the Clinton administration to preserve EITC funding in exchange for heightened enforcement (Johnston, 2000).

[9]These administrative data sources include state-provided data on child custody determinations that is compiled by the Department of Health and Human Services, child birth records from the Social Security Administration, and prisoner data from the Bureau of Justice Statistics.

## 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 the return 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 with respect to Black taxpayers, $D$, which we define as the difference in audit rates between Black and non-Black taxpayers:

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

An important barrier to studying differences in audit rates by race is that neither we nor the IRS observe taxpayer race. 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 administrative data on audits, to estimate differences in audit 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 in additional detail.

## 3.2 Imputing Race

To impute race, we apply 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.1 for a formal derivation. Estimating these terms by name and geography yields individual-level race probabilities. 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.

The independence assumption underlying BIFSG is strong, and is likely violated in practice (e.g., Greengard and Gelman, 2023). Still, prior research has found that the method performs well across a range of settings. Below, we validate the performance of BIFSG race probabilities as an input to our disparity estimators using a subset of tax records matched to non-IRS administrative data containing taxpayer race. We also verify the robustness of our results to alternative imputation methods, including an approach that sidesteps the independence assumption by predicting race solely on the basis of geographic information.

## 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 audit rates by race. We consider two estimators: the *probabilistic disparity estimator* and the *linear disparity estimator*. We

characterize the bias of each and provide conditions under which the two estimators bound the true audit disparity.

The probabilistic estimator calculates average audit rates by race by weighting each taxpayer's contribution to the average audit rate by the probability that the taxpayer is or is not 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 rates among Black and non-Black taxpayers, respectively. The probabilistic disparity estimator, $\widehat{D}_p$, is the difference in the probabilistic audit rate estimates for these groups:

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

The second 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 corresponds to 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 $b$. The corresponding linear estimators of $Y^B$ and $Y^{NB}$ are given by $\widehat{Y}_l^B = \widehat{\alpha} + \widehat{\beta}$ and $\widehat{Y}_l^{NB} = \widehat{\alpha}$.

The following proposition characterizes the asymptotic bias of the disparity estimators.

**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

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

Appendix B provides a proof of this proposition, derives similar results for the audit rate (level) estimators, and characterizes the estimators' asymptotic distribution.[10]

For the probabilistic disparity estimator to be consistent (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

[10]In some applications, we consider estimates of the *conditional* audit disparity with respect to a subset of taxpayer characteristics $x \in X$: $E[Y|B = 1, x \in X] - E[Y|B = 0, x \in X]$. Proposition 1 extends naturally to such applications, with the key covariance terms replaced by $\mathbb{E}[\mathrm{Cov}(Y, b|B, x \in X) \,|\, x \in X]$ and $\mathbb{E}[\mathrm{Cov}(Y, B|b, x \in X) \,|\, x \in X]$, respectively. In other applications, we consider weighted versions of these estimators. We discuss both of these cases in Appendix B.

rates than non-Black taxpayers with identical names and living in identical neighborhoods.[11]

The linear disparity estimator requires a different assumption for consistency, namely that there be no residual association between predicted race and audits after conditioning on the taxpayer's actual race (Proposition 1.2). This exclusion restriction would be violated 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.

Propositions 1.3 and 1.4 highlight a useful implication of Proposition 1.1 and 1.2: 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.[12] In our setting, using geography and name to predict race, both of these covariance terms are likely to be positive. 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). In addition, 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 Black Americans 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 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, suggesting $E[\text{Cov}(Y, b|B)] > 0$. Below, we

[11] Proposition 1.1 is related to a result in Chen et al. (2019) and Kallus et al. (2021), 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 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.

[12] Proposition 3 in Appendix B establishes that the Black and non-Black audit rate levels are similarly bounded by the linear and probabilistic estimators under the same conditions.

provide empirical support for these conditions using an auxiliary data set in which race is observed for a subset of taxpayers.

Finally, Proposition 1 requires that the individual race probability estimates are perfectly calibrated, $b = Pr[B = 1|Z]$ for each $Z$; in Appendix B, we derive the bias of the disparity estimators when this assumption fails (see Proposition 2). One method for researchers to assess the calibration of the individual race probabilities is by calculating those probabilities for a subset of the data for which individual-level race information is available. To the extent the race probabilities are found to be miscalibrated, Proposition 2 shows that a simple linear correction using the available race labels can remove much of the resulting bias (see Appendix B.5 for details). Below, we implement this approach to assess and re-calibrate our BIFSG estimates.

# 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 approximately 148 million individual income tax returns with valid social security numbers for 2014. We primarily 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 pool 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.[13] 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 collect data on taxpayer race, either directly via tax returns or indirectly via merging tax data with administrative data on race from other agencies. Therefore, we rely on a BIFSG approach to estimate the probability that a taxpayer self-reports as Black based on the first name, last name, and location of residence reported on the taxpayer's 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. 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).[14] Information

[13]More precisely, we observe whether an audit for the return was completed prior to 2023. Nearly all audits are completed within six years of a return being filed.

[14]For purposes of training and validating BIFSG, we follow the racial and ethnic classifications provided in the first name and surname files; this entails treating individuals who report their ethnicity to be Hispanic as non-Black, and treating individuals who report themselves to be multi-racial as non-Black. Note that because of how the Census-derived inputs to BIFSG are measured, the method is technically trained to predict the householder's report of the taxpayer's race rather than the taxpayer's own self-report. The steps

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 the available subset of these attributes to impute race.

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.[15] 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 statistical bias of the audit disparity estimators on which we rely.

## 5.1 Taxpayer Race Estimates

As described above, we use BIFSG to estimate the 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

of this analysis requiring non-anonymized taxpayer information were conducted by Treasury economists, with the (anonymized) results provided to the other members of our research team.

[15] 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.

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 2014 (ACS, 2014).[16]

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 and Appendix Figure A.1 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 the parameters shaping the bias of the disparity estimates from Proposition 1, Figure 2 bins the taxpayers from the North Carolina data set based on their estimated probability of being Black, and plots the fraction of Black and non-Black taxpayers audited

[16]The relatively small share of individuals estimated to be Black with very high probability may reflect a limitation of the BIFSG methodology, which outputs a probability near one only for a taxpayer living in a very high Black-share neighborhood *and* with very distinctively Black first and last names. We investigate the robustness of our results to alternative race imputation methods below.

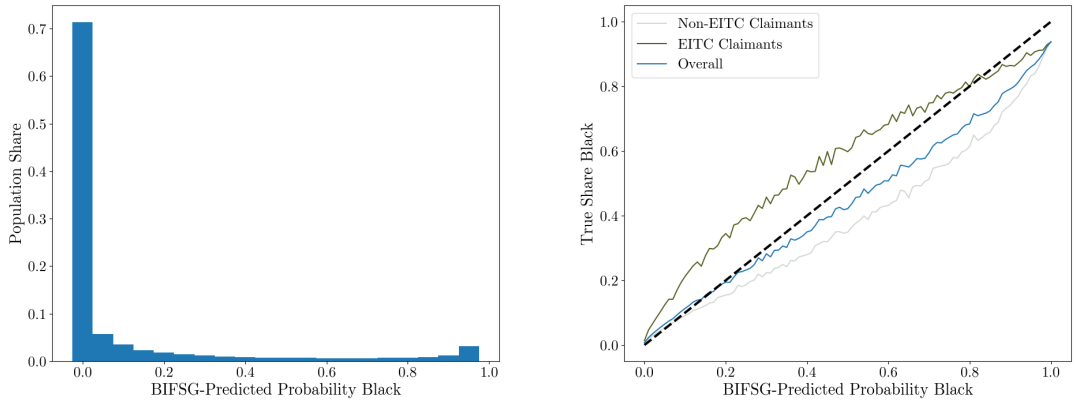
within each bin. The figure shows a positive residual correlation between audit probability and being Black after conditioning on the estimated race probability. From Proposition 1.1, this implies that the probabilistic disparity estimate is downward-biased, i.e., $E[\text{Cov}(Y, B|b)] > 0$. At the 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 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. 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 share of taxpayers in each race probability percentile. For both $E[\text{Cov}(Y, b|B)]$ and $E[\text{Cov}(Y, B|b)]$, we reject the null hypothesis that the parameter is less than or equal to 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 the second of 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.

Figure 1: Distribution and Calibration of Estimated Race Probabilities




*Notes:* Left: Nationwide histogram of BIFSG-predicted probability that a taxpayer is Black (non-Hispanic). The mean prediction is 12.2%. 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 self-reported race observed in the North Carolina matched sample, re-weighted to be representative of the overall United States (see Appendix C for details). A perfectly calibrated predictor would fall 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).

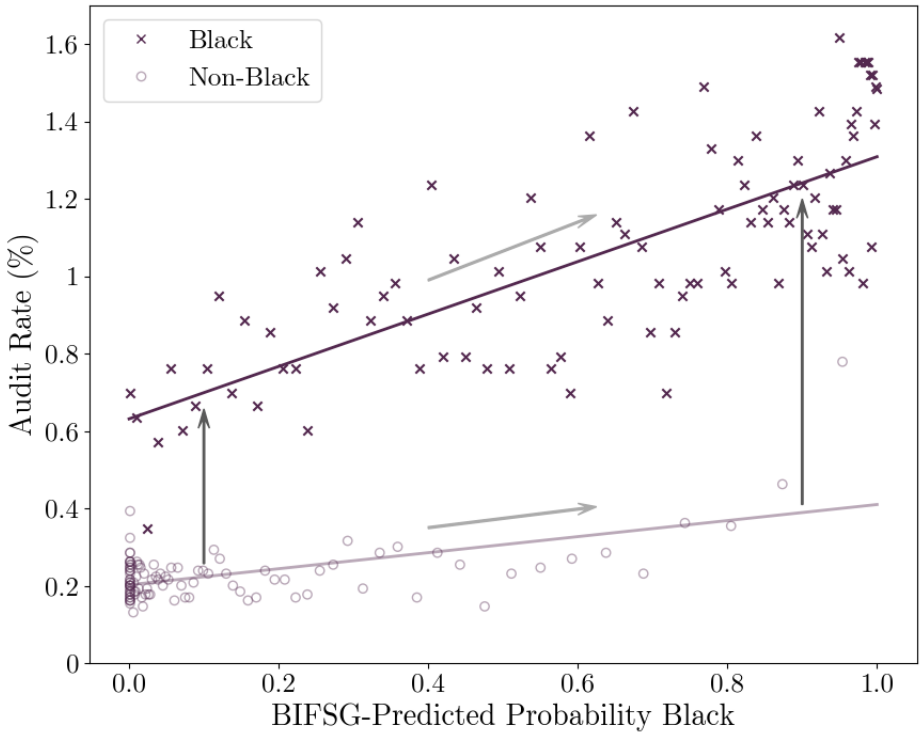
# 6 Audit Disparity Results

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 to EITC claimants.

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 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 and Self-Reported Race




*Notes:* The figure shows the relationship between audit incidence (y axis) and BIFSG-predicted probability that a taxpayer is Black (x axis) for taxpayers filing returns for tax year 2014. Audit rates are plotted separately for Black and non-Black taxpayers in the North Carolina matched sample. Black and non-Black taxpayers are each grouped into 100 equal-sized bins, with Black taxpayers indicated by dark purple x's and non-Black taxpayers indicated by light purple circles. The average vertical distance between the x's and circles (illustrated by the dark gray arrows) provides an estimate of the sign of $E[\text{Cov}(Y, B|b)]$. The average slope of the group-specific best-fit lines (illustrated by the light gray arrows) provides an estimate of the sign of $E[\text{Cov}(Y, b|B)]$.

probabilistic disparity estimate ranges from 0.81 to 0.82 percentage points.[17] As expected, the linear estimator implies an even larger racial audit disparity, of 1.34 percentage points, and is also precisely estimated. 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

[17]This confidence interval reflects sampling uncertainty in the outcome model, in the sense that even the universe of 2014 taxpayers may not perfectly reflect the underlying data generating process, but abstracts from uncertainty in the construction of the BIFSG-based probability estimates. We obtain slightly less precise disparity estimates after accounting for this source of uncertainty, following the dual-bootstrap method proposed by Lu et al. (2024) (Appendix Table A.10 and Appendix Figure A.12).

at between 2.9 and 4.7 times the rate of non-Black taxpayers.[18]

Figure 4 plots estimated audit rates by income and race. Black taxpayers appear more 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.5 shows that the audit rate among EITC claimants (of any race) is more than 4 times higher than among non-EITC claimants (1.45 vs. 0.31 percent). 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.

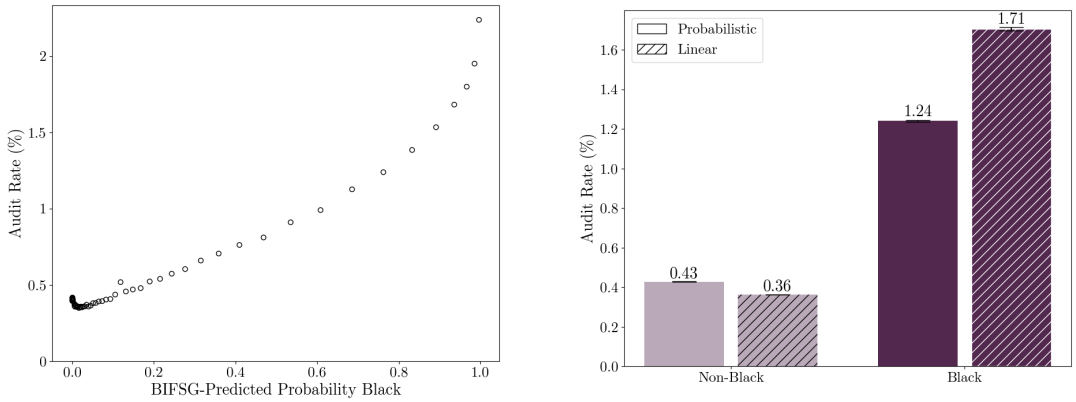
To assess this hypothesis, we estimate audit disparities by race separately for EITC claimants and non-claimants. If differences in EITC claiming rates by race account for the racial disparity, we would expect the difference in audit rates by race to be relatively small *within* the population of EITC claimants. However, Figure 5 shows this is not the case. Instead, the estimated disparity in audit rates between Black and non-Black EITC claimants is substantially larger in percentage point terms (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]

[18]We obtain comparable results from using a linear program to calculate maximum and minimum values of disparity that are consistent with the Proposition 1.3 assumptions and the observed joint distribution of $b$ and $Y$; see Appendix B.9.

[19]For ease of exposition, we focus on the probabilistic estimate of the audit rate; Appendix Figure A.2 shows a similar pattern using the linear estimator. As discussed in Appendix B.8, interpreting the estimators as bounds on a conditional audit disparity requires that the Proposition 1 conditions hold at the subgroup-level. Appendix Figure A.4 reports estimates of the relevant covariance terms by income bin from the NC data. For most bins, the Proposition 1.3 conditions appear satisfied, although the pattern is more consistent for the probabilistic estimator. For higher income taxpayers, the linear disparity estimator may not constitute an upper bound.

[20]These disparities are precisely estimated; refer to Table 1 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

Figure 3: 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 audit rate estimator with BIFSG-predicted probabilities. Error bars show the 95% confidence interval based on the distribution of estimates from 100 bootstrapped samples.

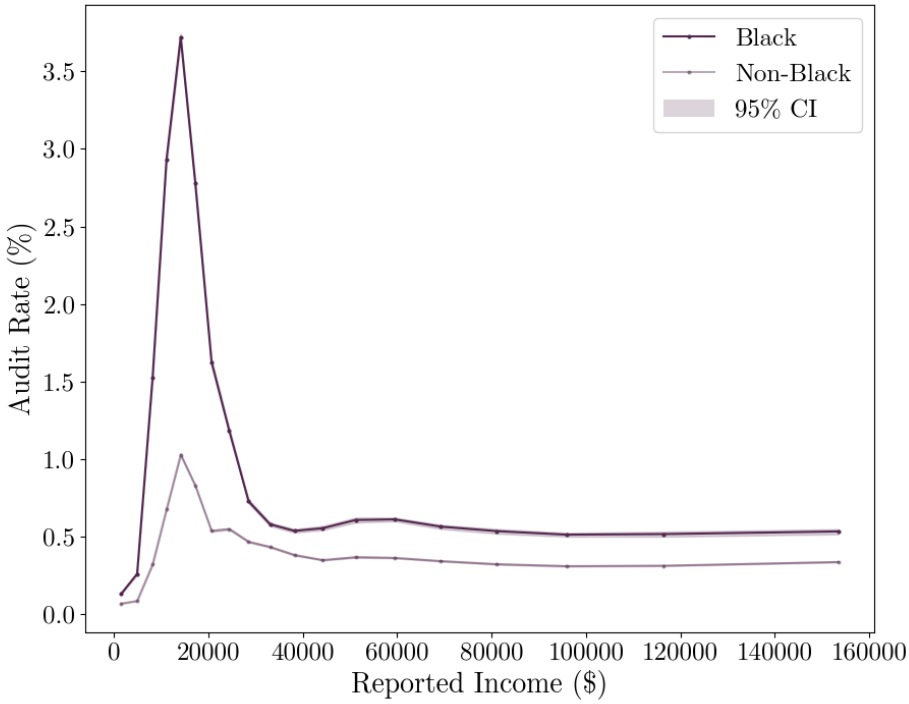
We can formally decompose the overall audit disparity into three components: (1) racial differences in the audit rate among EITC claimants; (2) racial differences in the audit rate among EITC non-claimants; and (3) racial differences in the rate at which taxpayers claim the EITC, scaled by differences in the audit rate for EITC versus non-EITC returns:

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

where $C_B$ and $C_{NB}$ denote the respective probabilities that Black and non-Black taxpayers claim the EITC; $Y_C^B$, $Y_C^{NB}$, and $Y_C$ denote the respective audit rates for Black, non-Black, and all EITC claimants; and similarly for $Y_{NC}^B$, $Y_{NC}^{NB}$, and $Y_{NC}$ with respect to EITC non-claimants. We estimate that racial differences in the audit rate within EITC claimants

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 (see Appendix Table A.7).

Figure 4: Audit Rates by Income and Race



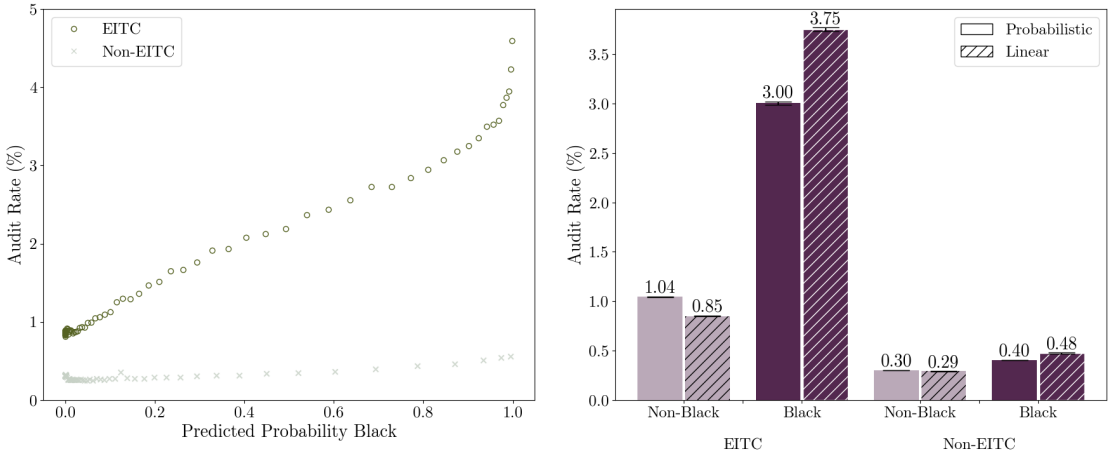

*Notes:* The figure shows the estimated audit rate by income among Black and non-Black taxpayers filing returns for tax year 2014. Income is measured according to the adjusted gross income (AGI) reported on the taxpayer's return (i.e., prior to audit adjustments). Binned audit rates by race are determined using the probabilistic estimator; Appendix Figure A.2 reports the corresponding analysis with the linear estimator. The sample is limited to taxpayers reporting non-negative AGI. Taxpayers have been grouped into 20 equal-sized bins, based on their AGI. The shaded area around each line shows the 95% confidence interval based on the distribution of estimates from 100 bootstrapped samples. To facilitate presentation, the x-axis is limited to bins with mean reported AGI under $200,000; a version of the figure with all income percentiles is presented in Appendix Figure A.3.

contribute between 70% and 73% of the total disparity, with the remainder primarily due to racial differences in the rate that taxpayers claim the EITC (20% to 21%) and a smaller portion due to racial differences in the audit rate among taxpayers not claiming the EITC (7% to 8%).[21]

We next explore the type of audits from which the disparity arises. Table 1 shows that audit disparities appear to be largely driven by differences in the selection of correspondence

[21]Appendix D provides additional detail, as well as other potential decompositions. The relative importance of these three components is sensitive to the decomposition considered and the choice of reference group, but a consistent finding is that differences in audit rates among EITC claimants are an important contributor to overall disparity—generating at least 32% and up to 83% of the total difference in audit rates.

Figure 5: Audit Rates by Race and EITC Claiming




*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 based on the distribution of estimates from 100 bootstrapped samples.

audits, whereas Black and non-Black taxpayers appear to be selected for field and office audits at roughly similar rates.[22] We observe disparities in both pre-refund and post-refund audits, although the magnitude is larger in the former category 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 observe among EITC non-claimants as well. Among EITC claimants, 78.5% of the observed audit disparity is attributable to audits conducted through the DDb program (Appendix Table A.11).

We next explore heterogeneity in audit disparities within distinct groups of EITC claimants. Figure 6 reveals 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 points

[22]The financial and time costs of being audited differ across these audit categories, as do the average amounts of back taxes, penalties, and interest that are imposed. Appendix Tables A.8 and A.9 provide back-of-the-envelope estimates for how audit burdens vary by race, accounting for these factors.

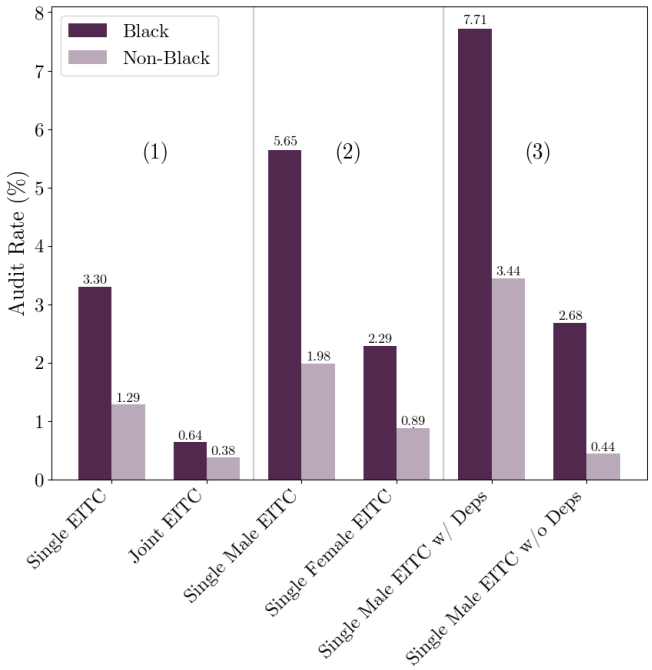
Table 1: Updated Estimated Audit Rate Disparity

| | | Audit Timing | | Audit Type | |
|---|---|---|---|---|---|
| | Any Audit | Pre Refund | Post Refund | Correspondence | Field/ Office |
| | (1) | (2) | (3) | (4) | (5) |
| Panel A: Full Population | | | | | |
| Probabilistic | 0.813 | 0.569 | 0.244 | 0.804 | 0.010 |
| | (0.003) | (0.002) | (0.002) | (0.003) | (0.001) |
| Linear | 1.345 | 0.941 | 0.403 | 1.328 | 0.016 |
| | (0.004) | (0.003) | (0.002) | (0.004) | (0.001) |
| Mean Audit Rate | 0.526 | 0.243 | 0.283 | 0.433 | 0.094 |
| N (Millions) | 148.3 | 148.3 | 148.3 | 148.3 | 148.3 |
| Panel B: EITC Population | | | | | |
| Probabilistic | 1.950 | 1.475 | 0.475 | 1.943 | 0.007 |
| | (0.008) | (0.007) | (0.004) | (0.008) | (0.001) |
| Linear | 2.885 | 2.182 | 0.703 | 2.875 | 0.010 |
| | (0.009) | (0.008) | (0.005) | (0.009) | (0.002) |
| Mean Audit Rate | 1.443 | 0.956 | 0.487 | 1.356 | 0.087 |
| N (Millions) | 28.3 | 28.3 | 28.3 | 28.3 | 28.3 |
| Panel C: Non-EITC Population | | | | | |
| Probabilistic | 0.102 | 0.005 | 0.097 | 0.088 | 0.013 |
| | (0.002) | (0.001) | (0.002) | (0.002) | (0.001) |
| Linear | 0.180 | 0.009 | 0.171 | 0.156 | 0.024 |
| | (0.003) | (0.001) | (0.002) | (0.002) | (0.001) |
| Mean Audit Rate | 0.310 | 0.075 | 0.235 | 0.215 | 0.096 |
| N (Millions) | 120.0 | 120.0 | 120.0 | 120.0 | 120.0 |

*Notes:* The table reports probabilistic and linear estimates of the difference in audit rates between Black and non-Black taxpayers filing income tax returns for tax year 2014. Units are percentage points (0-100). The category of audit considered varies across columns; for example, the results in column (4) show the estimated difference in the rate that Black versus non-Black taxpayers are selected for a correspondence audit. Panel A includes all taxpayers, whereas Panels B and C restrict the analysis to EITC claimants and non-claimants, respectively. Standard errors, reported in parentheses, are calculated from the asymptotic distributions described in Appendix B.3. See Appendix Tables A.5 and A.6 for estimates of audit rate levels by race.

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 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

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 EITC claimants. 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 estimator, is presented in Appendix Figure A.6

remains substantial among these groups.[23]

We conduct a number of analyses to assess the robustness of our results to alternative approaches for estimating taxpayer race, reported in Appendix Table A.12. First, recall that 27% of the taxpayers in our sample are missing one or more of the variables used to impute race; Column 1 re-estimates the racial audit disparity after excluding this group. Second, as an alternative to BIFSG, in Column 2 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). Third, our BIFSG race probability estimates are derived from samples that are designed to be representative of the overall U.S. population,

[23]Appendix Figure A.7 reports estimates of the conditional covariance terms from Proposition 1.3 for each subgroup in Figure 6. In each case, the terms are estimated to have the expected sign, with the exception of the term associated with joint filers claiming the EITC. Hence, the linear estimator may not identify an upper bound on disparity for that group.

not the subset of the population that files taxes or claims the EITC. As discussed in Section 3, such mis-calibration can shape the bias of our disparity estimators. 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 (Columns 3 and 4).[24] Fourth, we can avoid imposing BIFSG's conditional independence assumption if race is imputed based on geography alone rather than on geography and name. Column 5 replicates our main results using this simpler proxy.[25] Across columns, the results are largely unchanged from our baseline approach.

A different potential concern with our analysis is that the matched North Carolina data that we use to validate our identifying assumptions may not be representative of the national population, even after re-weighting. To assess this possibility, we matched our sample to voter registration data in six other states in which race is recorded and estimated the key covariance terms from Proposition 1. Across states, we obtain substantially similar results (Appendix Figure A.13).[26]

Finally, our results thus far have focused on returns for tax year 2014. To confirm the patterns we observe are not limited to that year, we estimate disparity among all taxpayers (Appendix Figure A.14) and among EITC claimants (Appendix Figure A.15) for tax years 2010, 2012, 2016, and 2018. In each case, we obtain comparable results to those from 2014.

[24] Appendix Figures A.8 and A.9 contain results from the analogous exercise applied to the income bins and demographic groups reported in Figures 4 and 6.

[25] For additional detail on the geography-based estimates, see Appendix Figures A.10 and A.11 and Appendix Table A.13

[26] A limitation of the matched data from states other than North Carolina, and the reason we do not rely on them for our main analysis, is that they are drawn from voting records for 2023; as expected, this leads to a much lower match rate. Separately, a recent working paper independently confirms the existence of the audit rate disparity in a matched national sample of residents in certain tax-subsidized apartments (Derby et al., 2024).

# 7 What Causes the Racial Audit Disparity Among EITC Claimants?

Because the racial audit disparity appears concentrated among EITC claimants, our remaining analyses focus on this population of taxpayers. As discussed above, we are confident that the disparity for this group is not due to disparate treatment in audit selection because the vast majority of EITC returns are selected for audit based on automated processes, and these processes do not include race as an input. In this section, we explore how group-level differences in taxpayer characteristics, as well as choices related to the design of the audit selection algorithm, might contribute to the observed disparity.To simplify exposition, in this section we rely primarily on the probabilistic estimator; we obtain qualitatively similar results using the linear disparity estimator (reported in Appendix A).

## 7.1 Differences in Underreporting

We first investigate whether the observed audit disparity can be explained by differences in the distribution of underreporting between Black and non-Black EITC claimants. By underreporting, we mean the difference between a taxpayer's correct income tax obligations for a tax year (which may be negative in the case of a taxpayer qualifying for refundable credits) and the 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 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.

### 7.1.1 Differences in Actual Underreporting

A challenge in studying whether the observed disparity is due to racial differences in the distribution of underreporting is that we observe underreporting only among those taxpayers who were selected for audit. We can circumvent this obstacle by combining non-random operational audits with data from randomly selected NRP audits. That is, using Bayes rule, we can express the audit rate for taxpayers of race $j$ and underreporting amount $k$ as:

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

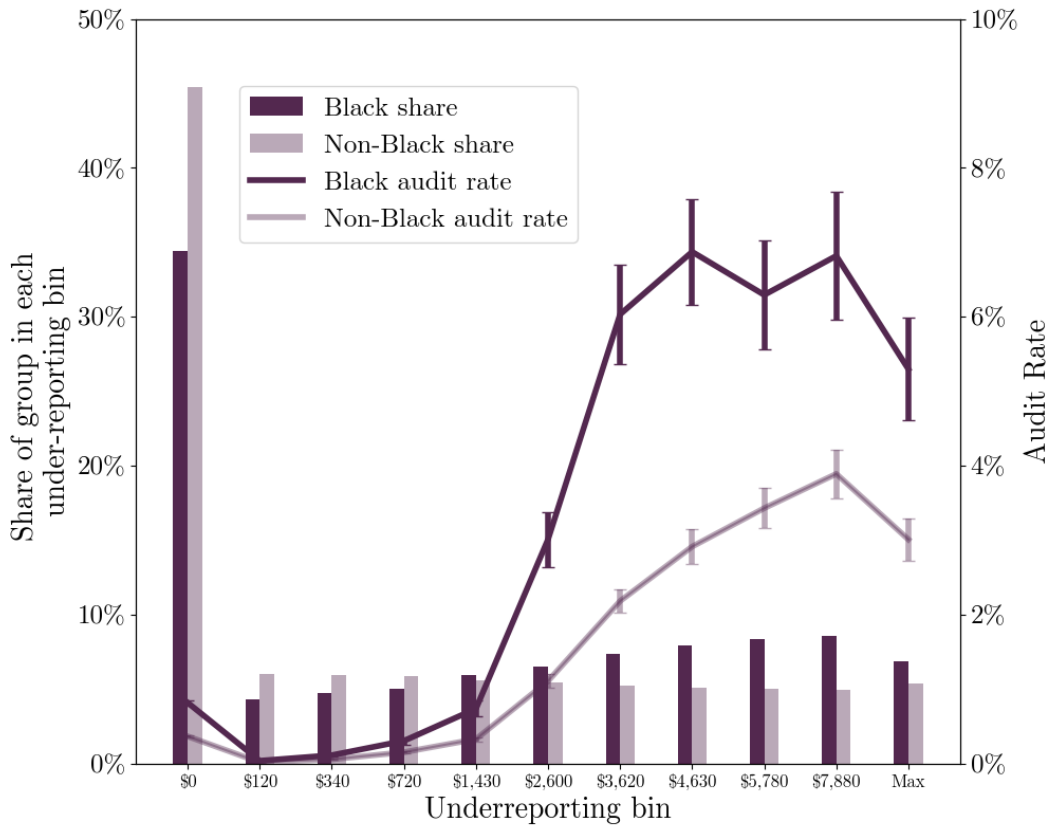
From this equation, we can estimate the racial audit disparity at a given level of underreporting using: the estimated audit rates by race for EITC claimants from Figure 5 for $\Pr(Y = 1|B = j)$; the NRP data to estimate $\Pr(K = k|B = j)$; and the detected underreporting from the operational audit results to estimate $\Pr(K = k|Y = 1, B = j)$.[27]

The results of this analysis are presented in Figure 7, with EITC claimants binned according to their underreported taxes. The figure shows that the distribution of underreporting for Black EITC claimants tends to be concentrated at higher values than for non-Black EITC claimants. However, within each underreporting bin, the estimated audit rate for Black taxpayers exceeds that of non-Black taxpayers, and for many bins, the difference is substantial. Figure 7 therefore provides evidence that differences in audit rates remain when comparing Black and non-Black taxpayers with similar levels of underreporting.[28]

[27] With respect to the last term, operational audits may not detect all of a taxpayer's underreporting because, unlike the NRP, operational 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. Conversely, although they are intensive, even NRP audits may miss some underreporting, and differential mis-measurement of underreporting by race could distort our results. For differences between NRP and operational audit scope to affect this analysis, those differences would have to vary by race. In addition, 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. See Appendix B.7 for further details.

[28] A limitation of this analysis is that the relatively small degree of overlap between the EITC NRP sample and the matched NC data set limits our ability to re-calibrate the disparity estimates by underreporting bin, as we did with the subgroup disparity estimates presented in Appendix Figures A.8 and A.9.

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 underreporting, and 10 equal deciles of taxpayers with positive underreporting. Underreporting deciles are defined based on the distribution of underreporting among EITC claimants, as measured by NRP audits. 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 estimator. All analyses account for NRP sampling weights. Brackets reflect the estimated 95% confidence interval, derived from bootstrapped standard errors (N=100). The bars show the estimated share of Black and non-Black taxpayers, respectively, that fall into each underreporting bin. A similar analysis, corresponding to the linear estimator, is presented in Appendix Figure A.16.

Figure 7 highlights the relationship between underreporting and audits throughout the underreporting distribution; in contrast, our next analysis zooms in on the extreme right tail of the underreporting distribution. The motivation for doing so is that the IRS selects less than 1.5% of EITC returns for audit; it could be that Black taxpayers are over-represented among EITC claimants with the highest underreporting. If so, selecting from this population could generate the higher observed audit rate for that group.

To explore this possibility, we simulate an "oracle" selection algorithm, which prioritizes returns for audit according to the actual dollar amount of underreporting that would be detected if the return were audited.[29] We treat the set of NRP-audited returns—for which we know true underreporting—to be the population, and pretend that we must select some subset of returns from this population to audit. To implement the oracle, we rank each return in this population based on its under-reported taxes. Then, in descending order of this ranking, we select returns for audit until some pre-specified audit rate has been reached. For each audit rate that we consider, we calculate (1) the sum of detected underreporting over the audited taxpayers, and (2) the racial audit disparity that would be induced by this selection process.[30]

The results of this analysis are shown by the line labeled "Total Underreporting Oracle" in Figure 8. As a benchmark, the Figure also reports total detected underreporting and disparity from audits of tax year 2014 returns claiming the EITC, indicated by the dashed red lines and labeled "Status Quo". At each audit rate considered, the underreporting oracle selects Black taxpayers at a lower rate than the status quo, and, more strikingly, at a lower rate than non-Black taxpayers. Thus, although Figure 7 shows that the distribution of underreporting for Black EITC claimants tends to be concentrated at higher values than

[29]Because the actual amount of underreporting on a return cannot be known before an audit is conducted, such an algorithm is not feasible for the IRS to implement; we consider it as a benchmark before turning to predicted underreporting below.

[30]Throughout, we use "detected underreporting" as shorthand for total recommended adjustments from audit, not accounting for appeals or uncollected tax debts. We report annualized detected underreporting, accounting for NRP sample weights, and adjusting for the fact that our NRP data pools multiple tax years; see Appendix E for details.

for non-Black EITC claimants, these results suggest that the EITC claimants with the very largest underreporting are disproportionately non-Black.

Figure 9 illustrates this contrast more directly. In Section 6, we estimated that the status quo audit rate for non-Black EITC claimants was 1.04% compared to 3.00% for Black EITC claimants. In contrast, if EITC claimants were selected for audit according to their true underreporting, the audit rate for non-Black EITC claimants would be 1.63% compared to 0.74% for Black EITC claimants. We interpret this result, in conjunction with the different audit rates by race within underreporting bins in Figure 7, as evidence that the observed audit disparity cannot be entirely explained by group-level differences in underreporting by race.

### 7.1.2 Differences in Predicted Underreporting

In practice, the actual amount of a taxpayer's underreporting is unknown at the time that auditing decisions are made. We now ask whether the observed audit disparity is due to the IRS relying on *predicted*, rather than actual, underreporting in selecting which returns to audit.

To study this question, we train a random forest model to predict taxpayer underreporting (in dollars) based on features that the tax authority can observe at the time of the audit decision. The model is trained on NRP returns claiming the EITC, and largely incorporates the same features available to the DIF and DDb programs to select audits of EITC claimants — information reported on tax returns supplemented with additional administrative data available to the IRS.[31] We then simulate selecting taxpayers in descending order of the model's predictions, until some specified audit rate has been reached. In other words, the approach is the same as the underreporting oracle, except that returns are selected for audit on the basis of predicted, rather than actual, underreporting.

The dark purple line in Figure 8 shows detected underreporting and disparity induced by

[31] Additional detail on the predictive model is provided in Appendix E.

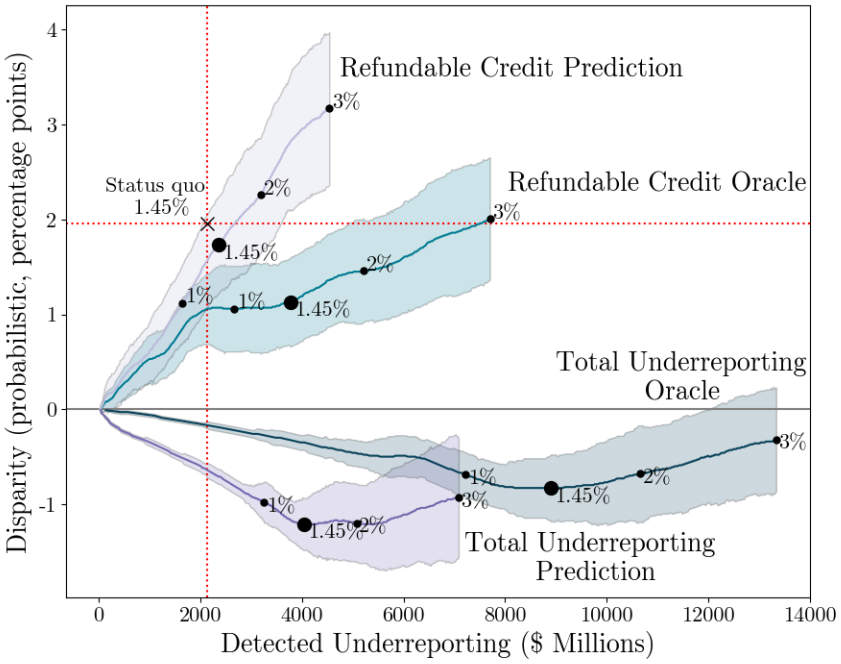
the predicted underreporting algorithm. Unsurprisingly, selecting audits based on predicted underreporting yields significantly less detected underreporting than the oracle at any given audit rate. However, like the underreporting oracle, the underreporting prediction algorithm selects Black taxpayers for audit at lower rates than non-Black taxpayers at each audit rate we consider. At the status quo audit rate, selecting EITC claimants for audit based on predicted underreporting would yield an audit rate of 0.45% for Black taxpayers and 1.71% for non-Black taxpayers. Hence, the fact that the IRS must select audits based on predicted rather than actual noncompliance does not, in itself, appear to explain the observed audit disparity.

## 7.2 Algorithmic Objective: Refundable Credit Overclaims

In this subsection, we explore the possibility that the observed audit disparity arises because the IRS selects among EITC returns for audit based on some objective other than maximizing the detection of underreported taxes. In particular, we consider the effect on disparity of allocating audits based on (1) the amount of underreporting attributable to overclaimed refundable credits, rather than (2) the total amount of underreported taxes (from whatever source).

To do so, we compare the audit selection algorithms described in the prior subsection that focus on total underreporting with algorithms that focus exclusively on refundable credit overclaims. Specifically, we consider two new algorithms (both of which are depicted in Figure 8). The first is a refundable credit oracle, which ranks returns by the sum of actual underreporting (in dollars) attributable to the three refundable credits designated by OMB as high-priority programs susceptible to significant improper payments: the EITC, the Additional Child Tax Credit, and the American Opportunity Tax Credit. As with the total underreporting oracle, the refundable credit oracle is not feasible to implement in practice because the actual amount of overclaimed refundable credits cannot be known with certainty before the audit occurs, but serves as a useful benchmark. The

Figure 8: Detected Underreporting and Disparity by Algorithm



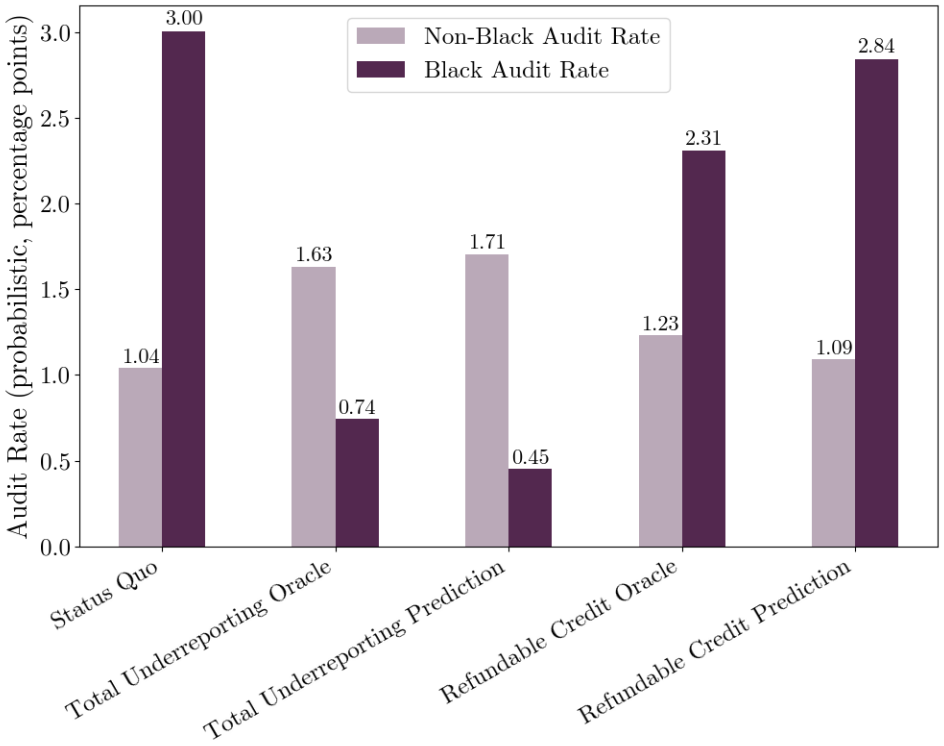

*Notes:* The figure shows estimated disparity ($y$-axis) and annualized detected underreporting ($x$-axis) under alternative algorithms for selecting audits of EITC claimants and under alternative audit rates (0.1% to 3%). For each algorithm, returns claiming EITC in the NRP sample are first ranked and then selected for audit in descending order of the ranking until the specified audit rate is reached. The ranking varies by algorithm and is based on: total underreporting (dark blue line), predicted total underreporting (dark purple line), underreporting due to overclaimed refundable credits (light blue line), and predicted underreporting due to overclaimed refundable credits (light purple line). 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. The labeled points along each line correspond to the audit rate specified in the label; the audit rate corresponding to the status quo EITC audit rate (1.45%) is denoted by a larger dot. Disparity is calculated using the probabilistic disparity estimator; see Appendix Figure A.17 for results using the linear disparity estimator. Shaded regions around each line correspond to 95% confidence intervals for disparity calculated based on the distribution of estimates from 100 bootstrapped samples. Annualized detected underreporting is the sum of detected underreporting (positive or negative) among returns selected for audit under the specified algorithm, scaled to reflect our use of five years of NRP data. All analyses incorporate NRP sampling weights. The two prediction algorithms are based on random forest regression models. See Appendix E for additional detail.

second new algorithm is based on predicted refundable credit overclaims; it ranks returns by the output of a random forest regressor trained to predict the sum of underreporting attributable to the same three refundable credits. After ranking, both algorithms select returns for audit in descending rank order until a specified audit rate is reached. In other words, the two refundable credit algorithms mirror the two total underreporting algorithms described above; the only difference is that the ranking used to selected audits is based on overclaimed refundable credits—actual overclaims in the case of the oracle and predicted overclaims in the case of the prediction algorithm—rather than total underreporting.

As shown in Figure 8, the refundable credit algorithms detect substantially less underreporting than the algorithms focused directly on that objective, indicating that refundable credit overclaims are not the only important source of underreporting among EITC claimants. With respect to disparity, both refundable credit algorithms yield opposite-signed results compared to the algorithms focused on total underreporting: they select Black taxpayers at higher rates than non-Black taxpayers. As highlighted in Figure 9, selecting EITC claimants for audit based on actual refundable credit overclaims while holding the overall audit rate of EITC claimants fixed at the status quo level would lead to an audit rate of 2.31% for Black EITC claimants, compared to 1.23% for non-Black EITC claimants. Similarly, selecting audits according to predicted refundable credit overclaims would lead to an audit rate of 2.84% for Black EITC claimants, compared to 1.09% for non-Black EITC claimants. We interpret these results as evidence that targeting the detection of overclaimed refundable credits leads to a larger share of Black taxpayers being selected for audit compared to targeting the detection of total underreporting.

Does the objective of the audit selection algorithm actually contribute to the audit disparity we document in Section 6? Several pieces of evidence suggest that it does. First, public governmental documents describe the goal of the DDb audit program — the IRS's primary EITC audit selection tool during our sample period — to be the identification of

Figure 9: Group-Specific Audit Rates by Algorithm




*Notes:* The figure reports estimated audit rates for Black and non-Black EITC claimants that would be induced by the algorithms considered in Figure 8 under an assumed audit rate of 1.45%. The population of tax returns available for audit is based on the NRP sample of taxpayers claiming the EITC; see notes to Figure 8 for additional detail. Status quo refers to the estimated audit rates by race for tax year 2014 returns claiming the EITC, reported in Figure 5. Audit rates by race are estimated using the probabilistic estimator; see Appendix Figure A.18 for results using the linear estimator.

taxpayers who do not meet refundable credit eligibility requirements (G.A.O., 2015).[32] Second, the results in Figure 8 suggest that the predicted refundable credit algorithm serves as a good proxy for the amalgamation of (confidential) IRS programs and algorithms through which EITC audit selection occurs. In particular, the Figure shows that when we select audits based on our predictions of refundable credit overclaiming, where our predictions are based on largely the same features and training data that are available to IRS, we obtain a similar disparity as we observe among the returns actually selected by IRS for audit. Finally, as shown in Appendix Figure A.19, operational audits of EITC returns are strongly associated with predicted refundable credit overclaims; in contrast we observe a much lower association between operational audits and predicted total underreporting. For these reasons, we interpret our results to support the conclusion that the IRS's choice of algorithmic objective is an important contributor to the observed disparity in EITC audit rates.

The next subsection unpacks this conclusion, exploring why the audit selection objective shapes the distribution of audits by race.

### 7.2.1 Why Algorithmic Objective Shapes Audit Disparity

To understand why the objective of the audit selection algorithm shapes the distribution of audits of EITC claimants by race, we explore racial differences in the distribution of various forms of tax non-compliance. To do so, we draw on the richness of the NRP data to identify specific types of errors present on EITC claimants' returns, and consider the prevalence of these errors among EITC claimants that would be selected under each of the audit selection algorithms considered in Figure 8. The results are presented in Table 2. As expected, returns selected by the refundable credit oracle tend to have higher refundable credit overclaims than those selected by the total underreporting oracle ($8,191 vs $3,909 on average), but smaller total adjustments to the taxpayer's claimed refund or balance due ($9,595 vs $22,578 on

[32] Our main results are not limited to DDb-selected audits but Appendix Table A.11 confirms the presence of a disparity for this subset of audited returns.

average).

Table 2: Audit-Selected Tax Returns by Algorithm

| | Total Underreporting Oracle | Refundable Credit Oracle | Total Underreporting Prediction | Refundable Credit Prediction |
|---|---|---|---|---|
| Any Underreporting (%) | 100.0 | 100.0 | 90.2 | 89.1 |
| Mean Underreporting ($) | 22,578 | 9,595 | 10,164 | 5,952 |
| Any Refundable Credit Overclaiming (%) | 91.4 | 100.0 | 81.8 | 85.5 |
| Mean Refundable Credit Overclaiming ($) | 3,909 | 8,191 | 2,174 | 5,144 |
| Dependent Error Rate (%) | 27.1 | 79.8 | 3.6 | 33.1 |
| Head of Household Error Rate (%) | 17.7 | 71.3 | 4.3 | 61.8 |
| Any Business Income Underreporting (%) | 85.7 | 27.0 | 83.9 | 31.4 |
| Probabilistic Disparity (p.p.) | -0.9 | 1.1 | -1.3 | 1.8 |
| Linear Disparity (p.p.) | -1.3 | 1.7 | -1.9 | 2.6 |

*Notes:* The table reports characteristics of the tax returns that would be selected for audit by each of the algorithms considered in Figure 8 under an assumed audit rate of 1.45%. The population of tax returns available for audit is based on the NRP sample of taxpayers claiming the EITC; see notes to Figure 8 for additional detail. Noncompliance is measured based on NRP audit adjustments. "Dependent Error Rate" refers to the share of tax returns that had one or more dependents living with the taxpayer reduced upon audit. "Head of Household Error Rate" refers to the share of tax returns filing as head of household but found ineligible to do so upon audit. "Any Business Income Underreporting" refers to the share of tax returns with positive underreporting of Schedule C Net Income. The final two rows report estimated disparity; units are percentage points (0-100).

More striking is the different profile of taxpayer errors pursued by the different audit selection algorithms. The vast majority (80%) of returns selected by the refundable credit oracle contained a dependent error – that is, the audit determined that at least one of the dependents claimed on the return was not eligible to be claimed by the taxpayer – compared to only 27% of returns selected by the underreporting oracle. In contrast, 86% of returns selected by the underreporting oracle underreported income from a taxpayer's

business, compared to only 27% of returns selected by the refundable credit oracle.[33] This pattern arises because eligibility for many refundable credits is linked to children; hence, the detection of erroneously claimed dependents tends to be associated with large reductions in refundable credits. At the same time, the contribution of refundable credit overclaims to total underreporting is necessarily bounded by the maximum credit amount, whereas detecting underreported income can lead to arbitrarily large increases in tax liability. Hence, the largest adjustments to total tax – i.e., those prioritized by the underreporting oracle – primarily stem from returns with large amounts of underreported income.

We next explore whether the type of error on which an algorithm focuses shapes the racial distribution of taxpayers selected for audit. We find evidence that it does. In particular, we estimate that erroneously claimed dependents are more common among Black taxpayers than non-Black taxpayers (Table 3). As a result, audit selection processes that are (implicitly) trained to detect dependent mistakes tend to select Black taxpayers at higher rates.[34]

In contrast, the bottom rows of Table 3 suggest that Black taxpayers are less likely to be selected by audit algorithms focused on total underreporting because they are disproportionately under-represented among those taxpayers with the largest amounts of underreported business income. That is, among EITC claimants are a group of taxpayers who have high (actual) incomes but who underreport their incomes to such a degree that they appear eligible for the EITC. Auditing the business income and deductions claimed by taxpayers in this group yields large upward adjustments to tax liability, which is why they are prioritized by algorithms that focus on total underreporting. Because taxpayers in this group are disproportionately non-Black, audit algorithms focused on total underreporting tend to select Black taxpayers at lower rates.

[33]The two prediction algorithms differ from one another in a similar manner.

[34]It is beyond the scope of this paper to explore why child-claiming errors vary by race, but one factor that likely contributes is the lower marriage rate among Black Americans in conjunction with credit eligibility rules that prevent cohabitating individuals from claiming the child of an unmarried partner, even if they contribute to the child's support (Maag et al., 2016; Goldin and Jurow Kleiman, 2021). Along these lines, Michelmore and Pilkauskas (2022) find that Black children are more likely to reside in families with complex or ambiguous tax filing situations.

Table 3: Types of Tax Noncompliance by Race

| | Probabilistic Estimate | | Linear Estimate | |
|---|---|---|---|---|
| | Black Taxpayers | Non-Black Taxpayers | Black Taxpayers | Non-Black Taxpayers |
| Claims Dependent | 73.0 | 69.4 | 74.6 | 69.3 |
| Dependent Error Rate | 26.6 | 16.3 | 30.8 | 15.4 |
| Head of Household Error Rate | 33.5 | 19.3 | 39.1 | 17.8 |
| Any Business Income | 19.0 | 21.7 | 17.9 | 22.0 |
| Any Business Income Underreporting | 15.9 | 18.1 | 15.0 | 18.4 |
| Underreported Business Income is Among Top... | | | | |
| 10% | 1.03 | 1.97 | 0.66 | 2.06 |
| 5% | 0.40 | 1.01 | 0.17 | 1.07 |
| 1% | 0.05 | 0.21 | -0.01 | 0.23 |

*Notes:* The table reports estimates of the characteristics of tax returns filed by Black and non-Black taxpayers. Analyses are based on the NRP sample of taxpayers claiming the EITC. All analyses account for NRP sampling weights. "Dependent Error Rate" refers to the share of tax returns that had one or more dependents living with the taxpayer reduced upon audit. "Head of Household Error Rate" refers to the share of tax returns filing as head of household but found ineligible to do so upon audit. Business Income refers to net income reported on Schedule C of a tax return. The final three rows report the shares of taxpayers with business income underreporting within the top 90th, 95th, and 99th percentiles of the distribution of positive business income underreporting.

### 7.2.2 Algorithmic Objective and IRS Operational Considerations

Modifying the objective of the IRS audit selection algorithm could have a range of downstream implications for the agency's enforcement operations beyond the change in the identity of the audited taxpayers. One channel through which this could occur is a change in the composition of the issues upon which audits focus. As discussed in the prior subsection, switching from an algorithm focused on overclaimed refundable credits to one

focused on total underreporting is likely to shift the focus of examinations away from dependent eligibility issues and toward issues related to the proper reporting of business income and deductions.

What would this change in focus mean from an operational perspective? A likely consequence would be an increase in per-return auditing costs.[35] EITC returns differ from one another in terms of the number of hours required to conduct an audit as well as the training and experience required of the auditor. An important driver of this variation is the presence of business income: on average, EITC returns with substantial business income cost the agency $369.70 per return to audit, compared with $23.09 for other EITC returns.[36] By increasing the share of audits of EITC claimants that are focused on issues relating to business income, the change in algorithmic objective would increase average auditing costs by increasing the examiner resources required to conduct each exam.

To quantify the magnitude of this effect, Appendix Figure A.20 plots audit costs by algorithm based on the share of returns with substantial business income that each algorithm selects. Holding the current EITC audit rate fixed, switching from an algorithm focused on refundable credit overclaims to one focused on total underreporting would increase the share of audited returns with substantial business income from 3% to 93%, and would raise EITC examination costs by nearly an order of magnitude (Appendix Table A.15). This result suggests that switching algorithmic objectives may, in practice, require that the IRS reduce the total number of EITC returns it audits, at least in the short-term where the amount of examiner resources available for EITC audits may be relatively fixed.[37]

[35]The merits of accounting for auditing costs in selection raises difficult normative considerations that are outside the scope of this project; for example, it may be unfair (or generate perverse incentives) for the IRS to avoid auditing those taxpayers most likely to vigorously contest their assessment. Similarly, the cheapest correspondence audits for the IRS to conduct are those for which a refundable credit is disallowed because the taxpayer does not respond. However, non-response to an audit does not necessarily signal ineligibility for a credit, and is more common among Black taxpayers.

[36]These cost estimates are calculated from operational audits based on the average time logged by IRS employees dealing directly with the case multiplied by the applicable General Schedule payscale given the employee level. The classification of returns into those with and without substantial business income follows the IRS's internal classification of EITC returns into *activity codes* 270 and 271 based on whether the return reports gross business receipts in excess of $25,000 (see Appendix Table A.14).

[37]To illustrate the potential effects of this type of constraint, Appendix Figure A.21 replicates Figure 8,

At the same time, Appendix Figure A.20 also shows that the increase in detected underreporting from the change in algorithmic objective would substantially exceed the increase in audit costs. Intuitively, EITC returns with business income are more difficult to audit, but yield much higher adjustments on average when an audit is undertaken. As such, the qualitative pattern in Figure 8 is largely unchanged when audits are allocated in a manner that accounts for the higher cost of auditing EITC returns claiming business income (Appendix Figure A.22). We therefore conclude that although prioritizing total underreporting would increase per-return audit costs, reforms to reduce disparity need not conflict with the policy goal of deploying audits to efficiently detect tax noncompliance.[38]

## 7.3 Algorithmic Bias in Refundable Credit Prediction

Thus far we have focused on the difference between the disparity induced by the total underreporting algorithms compared to the disparity induced by the overclaimed refundable credit algorithms. However, Figure 8 also highlights that the observed status quo disparity (1.96 p.p.) is substantially larger than the disparity induced by the refundable credit oracle (1.08 p.p.). This suggests that some of the observed disparity is attributable to factors beyond the choice of algorithmic objective. In other words, the status quo disparity is larger than what we would expect to observe even conditioning on the audit selection objective being the detection of refundable credit overclaims. Additional evidence for this hypothesis is provided in Appendix Figure A.25, which shows that racial disparities in audit rates remain when comparing EITC claimants with similar levels of refundable credit overclaims.

Beyond algorithmic objective, one factor that could contribute to the observed audit

holding fixed the share of audits allocated to returns with substantial business income. Doing so raises the share of Black taxpayers selected by the total underreporting algorithms, but not to the level of the refundable credit algorithms or to the status quo disparity.

[38]The analyses in this subsection assume the operational audits from which the cost estimates were derived were focused on the same issues as would be selected under the various algorithms considered. We obtain qualitatively similar results when relying on a more conservative (potentially inflated) measure of auditing costs, derived from the hours that auditors report spending on NRP audits (which cover nearly all potential issues on a return); see Appendix Figures A.23 and A.24.

disparity is if the prediction errors underlying the status quo audit selection algorithm were unevenly distributed by race. A large literature in algorithmic fairness documents how data-driven predictive models may exacerbate ground-truth group-level differences in the outcome being predicted (e.g., Leino et al., 2018; Reich, 2021). Figure 8 provides some initial evidence that this may be occurring in our setting: the disparity induced by the refundable credit prediction model (1.75 p.p.) is similar in magnitude to the status quo disparity and approximately 60% larger than the disparity induced by the refundable credit oracle. Since the refundable credit model shares the same objective and data as the IRS, this raises the possibility that the predictive models used by IRS for pursuing refundable credit overclaims may be amplifying the audit disparity that would emerge due to actual differences by race in the distribution of refundable credit overclaiming and selection along those lines.

In Online Appendix F, we investigate the role of differential prediction errors by race in generating the observed disparity. Beginning with our simulated algorithms, we find that errors in predicting refundable credit overclaims are unevenly distributed by race in a manner that may contribute to higher audit rates for Black taxpayers. Turning to IRS operational models, we document uneven errors in the risk measure used by the IRS's DDb model to predict the likelihood that a child has been incorrectly claimed. We also investigate missingness of parental information on the birth certificate data that IRS uses to help determine whether the children claimed on tax returns meet the required eligibility criteria. We find that whereas birth certificates are missing maternal information at roughly equal rates by race, the birth certificates of children claimed on the returns of Black taxpayers are substantially more likely to be missing information about the identity of the father. Finally, we provide suggestive evidence that there may be opportunities to reduce audit disparities without substantially degrading accuracy by adjusting the features used to form predictions about overclaimed refundable credits.[39]

[39]The Online Appendix explores a number of additional factors apart from prediction errors that could potentially also contribute to the observed audit disparity. As detailed in that appendix, we find no evidence

# 8 Conclusion

In this paper, we found that Black taxpayers are audited at a higher rate than non-Black taxpayers, and that this is primarily due to differences in the audit rate between Black and non-Black EITC claimants. In addition, we found that the audit selection objective for EITC returns contributes to this observed disparity, in conjunction with differences by race in the types of errors that EITC claimants tend to make. In particular, we found that designing audit selection algorithms to maximize the detection of overclaimed refundable credits leads to Black EITC claimants being selected at higher rates, whereas designing such algorithms to maximize the detection of total underreporting (from any source) yields the opposite pattern. We interpret our results to suggest that policymakers seeking to reduce the observed audit disparity should consider reorienting audit selection (at least in part) around the goal of maximizing the detection of total underreporting rather than prioritizing noncompliance from refundable credit overclaims. More generally, while the proper focus of tax enforcement has been debated for decades, our findings shed new light on how competing enforcement priorities shape the distribution of audits by race.

We emphasize that implementing a change in audit selection objectives is not as simple as swapping one predictive model for another. In particular, we found that modifying the algorithmic objective to focus on total underreporting would have important downstream effects on the composition of audited EITC returns – shifting audits away from issues of dependent eligibility and toward issues relating to the accuracy of taxable business income. Audits in the latter category require more resources — both auditor time and expertise — compared to most audits of EITC returns today. Although we found that the increase in detected underreporting would exceed the increase in examiner costs, implementing this

that the observed disparity is substantially driven by racial differences in the distribution of EITC claimants' reported income, household composition, or use of a tax preparer. We also provide evidence that the observed disparity is not driven by racial differences in the distribution of high-information ("smoking gun") signals of non-compliance, or by the use of "regression" models trained to maximize dollars of detected refundable credit overclaims as opposed to "classification" models trained to maximize the probability that audited returns overclaim a refundable credit by at least some (specified) amount.

type of change in the short-term would likely require that the IRS conduct fewer audits of EITC returns while devoting more resources per audit. Longer term, the IRS may be able to reduce the cost of auditing the issues identified by algorithms focused on total underreporting by conducting a greater share of such audits by correspondence (e.g., mailing requests for documentation of claimed business deductions). The IRS recently announced that it plans to study this possibility (IRS, 2023b).

Although the IRS is responsible for audit selection, some of the factors we have identified as contributing to the observed audit disparity are shaped by forces outside the IRS's control. For example, Congress sets the rules governing credit eligibility — which may contribute to more mistakes among Black taxpayers due to racial differences in family structure — and IRS funding — which shapes the ability of the agency to allocate resources to complex cases. In contrast, the IRS is not compelled by Congress to prioritize refundable credit overclaims over other forms of tax non-compliance.

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. Investigating audit disparities for other racial and ethnic groups is an important avenue for future work, with potentially differing causes and opportunities for mitigation.[40] Similarly, although we have focused on audit disparities among EITC claimants, our analysis also suggests (smaller) disparities among higher income taxpayers, for whom we cannot rule out the possibility of disparate treatment (since audit selection of higher income taxpayers may not be fully automated).

Second, we have focused on audit selection algorithms designed to detect underreporting, but policymakers may prioritize other objectives as well, such as deterring non-compliant behavior, avoiding audits of 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

[40]See Derby et al. (2024) for a discussion of these issues and some evidence concerning audit disparities with respect to other racial and ethnic groups.

authorities like the IRS, such as the types of compliance issues that can be explored through correspondence audit, or differences in audit response rates or dollars collected depending on whether the audit is pre- versus post-refund. A more complete optimal policy analysis would require accounting for these additional considerations.

Finally, audit selection constitutes only one dimension in which tax administration may differently affect taxpayers by race. Disparities may also exist in how the audit is conducted, and 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. (Cited on 6)

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. (Cited on 6)

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23(2016):139–159. (Cited on 2)

Anson-Dwamena, R., Pattath, P., and Crow, J. (2021). Imputing missing race and ethnicity data in covid-19 cases. (Cited on 6)

Aughinbaugh, A., Robles, O., and Sun, H. (2013). Marriage and divorce: Patterns by gender, race, and educational attainment. *Monthly Lab. Rev.*, 136:1. (Cited on 17)

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California law review*, pages 671–732. (Cited on 7)

Bearer-Friend, J. (2019). Should the IRS Know Your Race? The Challenge of Colorblind Tax Data. *Tax L. Rev.*, 73:1. (Cited on 6)

Bearer-Friend, J. (2021). Colorblind tax enforcement. *NYU Law Review*. (Cited on 50)

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. (Cited on 7, Appendix-78)

Bloomquist, K. M. (2019). Regional bias in irs audit selection. *Tax Notes*. (Cited on 3, 6)

Boning, W. C., Hendren, N., Sprung-Keyser, B., and Stuart, E. (2023). A welfare analysis of tax audits across the income distribution. Technical report, National Bureau of Economic Research. (Cited on 5)

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. (Cited on 50)

Brown, D. A. (2005). The tax treatment of children: Separate but unequal. *Emory LJ*, 54:755. (Cited on 6)

Brown, D. A. (2009). Shades of the american dream. *Wash. UL Rev.*, 87:329. (Cited on 6)

Brown, D. A. (2018). Homeownership in black and white: The role of tax policy in increasing housing inequity. *U. Mem. L. Rev.*, 49:205. (Cited on 6)

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). (Cited on 6)

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. (Cited on 2)

CFPB (2014). *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A methodology and assessment.* Consumer Financial Protection Bureau. (Cited on 6)

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. (Cited on 2, 7, 17, Appendix-47)

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. (Cited on 6)

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. (Cited on 6)

Congressional Research Service (2022). Audits of eitc returns: By the numbers. (Cited on 12)

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. (Cited on 17)

Dean, S. A. (2021). *Testimony to House Committee on Ways and Means.* (Cited on 6)

Derby, E., Dowd, C., and Mortenson, J. (2024). Assessing statistical bias in racial and ethnic disparity estimates using bifsg. *Working Paper.* (Cited on 31, 49)

Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484. (Cited on 7)

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. (Cited on 17)

Goldin, J. and Jurow Kleiman, A. (2021). Whose child is this? improving child-claiming rules in safety-net programs. *Yale Law Journal*, 131:1719. (Cited on 43)

Goldin, J. and Michelmore, K. (2022). Who benefits from the child tax credit? *National Tax Journal.* (Cited on 6)

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

Government Accountability Office (2021). Artificial intelligence: An accountability framework for federal agencies and other entities. (Cited on 6)

Government Accountability Office (GAO) (2017). Refundable Tax Credits: Comprehensive Compliance Strategy and Expanded Use of Data Could Strengthen IRS's Efforts to Address Noncompliance. (Cited on 8)

Greengard, P. and Gelman, A. (2023). Bisg: When inferring race or ethnicity, does it matter that people often live near their relatives? *arXiv preprint arXiv:2304.09126*. (Cited on 14)

Guyton, J., Langetieg, P., Reck, D., Risch, M., and Zucman, G. (2021). Tax evasion at the top of the income distribution: theory and evidence. (Cited on 33)

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. (Cited on 2)

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. (Cited on 6)

Hardy, B., Hokayem, C., and Ziliak, J. P. (2021). Income inequality, race, and the eitc. *Working paper*. (Cited on 6)

Holtzblatt, J. and McCubbin, J. (2004). Issues affecting low-income filers. *The crisis in tax administration*, 148:148–49. (Cited on 11)

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, pages 263–272. (Cited on 3, 6, 13)

Internal Revenue Service (IRS) (2011). Internal revenue manual 4.19.20. (Cited on 12)

Internal Revenue Service (IRS) (2016). Internal revenue service data book, 2015. (Cited on 11)

Internal Revenue Service (IRS) (2023a). Commissioner Letter to Chairman Wyden. `https://home.treasury.gov/system/files/136/091823-Wyden-Letter-from-IRS-Commissioner-on-Audit-Disparites.pdf` (Accessed: Feb 2024). (Cited on 49)

Internal Revenue Service (IRS) (2023b). EITC Fast Facts. `https://www.eitc.irs.gov/partner-toolkit/basic-marketing-communication-materials/eitc-fast-facts/eitc-fast-facts` (Accessed: June 2023). (Cited on 10)

Johnston, D. C. (April 16, 2000). I.r.s. more likely to audit the poor and not the rich. *New York Times*. (Cited on 12)

Kallus, N., Mao, X., and Zhou, A. (2021). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*. (Cited on 7, 17)

Kiel, P. and Fresques, H. (2019). Where in The U.S. Are You Most Likely to Be Audited by the IRS? (Cited on 3, 10)

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018a). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293. (Cited on 3)

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174. (Cited on 7)

Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1):null. (Cited on 2)

Leibel, K., Lin, E., and McCubbin, J. (2020). Social welfare considerations of eitc qualifying child noncompliance. *Treasury Office of Tax Analysis Working Paper*. (Cited on 17)

Leino, K., Black, E., Fredrikson, M., Sen, S., and Datta, A. (2018). Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*. (Cited on 47)

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

Maag, E., Peters, H. E., and Edelstein, S. (2016). Increasing family complexity and volatility: The difficulty in determining child tax benefits. *Tax Policy Center*. (Cited on 43)

Michelmore, K. M. and Pilkauskas, N. V. (2022). The earned income tax credit, family complexity, and children's living arrangements. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 8(5):143–165. (Cited on 43)

Moran, B. I. and Whitford, W. (1996). A Black Critique of the Internal Revenue Code. *Wis. L. REv.*, page 751. (Cited on 6)

National Taxpayer Advocate (2019a). Annual report to congress 2019. (Cited on 2)

National Taxpayer Advocate (2019b). Report: Making the eitc work for taxpayers and the government. (Cited on 11)

National Taxpayer Advocate (2021). Annual report to congress 2021. (Cited on 8)

Nerenz, D. R., McFadden, B., Ulmer, C., et al. (2009). Race, ethnicity, and language data: standardization for health care quality improvement. (Cited on 13)

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. (Cited on 2, 7)

Office of Management and Budget (OMB) (2021). Circular a-123, appendix c: Requirements for payment integrity improvement. (Cited on 11)

Passi, S. and Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 39–48. (Cited on 7)

Reich, C. L. (2021). Resolving the disparate impact of uncertainty: Affirmative action vs. affirmative information. *arXiv preprint arXiv:2102.10019*. (Cited on 47)

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

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. (Cited on 19)

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. (Cited on 6)

U.S. Treasury Department (2022). Agency financial report: Fiscal year 2021. (Cited on 11)

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13. (Cited on 3, 13)