Testing for Racial Bias in Police Traffic Searches

Joshua Shea*

July 5, 2023

Abstract

I develop a framework to detect and measure bias amid sample selection and statistical discrimination, and apply this framework to study racial bias in police traffic searches. I model the search decision with a threshold model and allow the threshold to be random. This allows the direction and intensity of bias to depend on the officer's belief of the probability that a driver carries contraband. This framework also allows me to evaluate each officer separately, thereby allowing for unrestricted heterogeneity in officer search preferences and beliefs. Sharp bounds on various measures of bias can be derived using bilinear programs. I use this framework to evaluate 50 officers in the Metropolitan Nashville Police Department and find 6 officers to be biased. For each of these officers, I construct sharp bounds on how search rates for minority drivers would change if they were treated as white drivers, and vice versa. The estimates suggest the intensity of bias depends on the officer's belief of the probability that a driver carries contraband.

*Department of Economics, University of Illinois Urbana Champaign. I am deeply grateful to Alexander Torgovitsky, Stéphane Bonhomme, and Peter Hull, who have provided invaluable guidance and support. I would also like to thank Jeffrey Grogger, Derek Neal, Jack Mountjoy, Evan Rose, Guillaume Pouliot, Azeem Shaikh, Max Tabord-Meehan, Jiaying Gu, Alexandre Poirier, Francesca Molinari, Lee Lockwood, Elias Bouacida, participants at the Becker Applied Economics Workshop at the University of Chicago, and participants at the European Winter Meeting of the Econometric Society 2022 for generously providing helpful feedback. A special thank you to Laura Sale, Francisco del Villar, Dan Kashner, Nadav Kunievsky, and the Econometrics Student Group at the University of Chicago for the regular and thoughtful discussions of my research. Any and all errors are my own.

1 Introduction

Disparities across race, sex, and other protected classes arise in many settings, including the labor market (Card et al., 2016; Agan and Starr, 2018; Kline et al., 2022), the criminal justice system (Arnold et al., 2018; Feigenberg and Miller, 2022), healthcare (Obermeyer et al., 2019; Wasserman, 2023), credit attribution (Sarsons et al., 2021; Ductor et al., 2021; Onuchic and Ray, 2023), and lending markets (Bhutta and Hizmo, 2021; Bartlett et al., 2022). However, due to possible unobserved confounders, it is often difficult to determine whether the disparities are a result of bias against particular groups of individuals, or measure the extent to which bias contributes to the disparities.

In this paper, I develop a framework to test for and measure bias, and I apply this framework to study racial bias in police traffic searches. Under this framework and setting, officers have preferences for searching white and minority drivers who are stopped. These preferences govern an economic choice model for searching drivers that depends on the officer's belief of the probability that a driver carries contraband (e.g., drugs or weapons). Although I do not observe these beliefs, I am able to make sharp inferences on how the officer's search decisions depend on his beliefs. This is achieved using a partial identification approach. This approach also allows restrictions on the officer's preferences and the probability that drivers carry contraband to be layered in a flexible and transparent manner. The econometric methods do not require officers to be randomly assigned to drivers and may be used to evaluate each officer separately. The methods can also be applied to study discrimination in other settings, such as the labor market and healthcare.

The test for bias checks whether the sharp identified set for the officer's search preferences (i.e., the smallest set of preferences consistent with the model and data) includes an equivalent pair of preferences for white and minority drivers. If not, then the officer's preferences must differ by race, implying he is biased. The intensity of bias may then be inferred from how dissimilar white and minority search preferences are. The partial identification approach permits the test to be valid even when officers have different beliefs about the probability that white and minority drivers carry contraband, which can occur for several reasons, including sample selection and statistical discrimination. Implementing this approach entails solving bilinear programs, a type of non-convex problem that can be solved to provable global optimality. Bilinear programs are not only novel in the context of discrimination, but also in the context of partial identification and econometrics in general.

A distinguishing feature of the test is how I model an officer's search decision. Similar to earlier papers, the officer is modeled to search drivers only if their probability of carrying contraband ("risk") exceeds a threshold, where the threshold represents the officer's search

preference. However, whereas recent papers have required or assumed fixed thresholds (see Canay et al. (2020a,b, 2022) and Hull (2021) for a discussion on this restriction), I use a random threshold. Consequently, there is no longer a single driver at the margin of search, but a "marginal driver" at every level of risk. This means a biased officer is not restricted to searching all drivers of one race with a given level of risk, while searching none of the equally risky drivers of the other race, as implied by a fixed threshold. Instead, the officer can search both groups of drivers at different intensities, e.g., whites with 10% risk are searched 20% of the time, whereas equally risky minorities are searched 40% of the time. Officers can even change direction of bias depending on the level of risk, e.g., whites with 10% risk are half as likely to be searched compared to equally risky minorities, but whites with 20% risk are twice as likely to be searched compared to equally risky minorities. The random threshold therefore permits a richer analysis of racial bias, where the direction and intensity of bias may depend on unobserved (to the researcher) characteristics of the driver. Moreover, the methods proposed allow me to learn about this dependence.

Identification is aided by an instrumental variable (IV) that shifts the distribution of risk among drivers stopped without shifting the officer's preferences. For each race of drivers, this generates a sequence of data points that must be consistent with a single preference, thereby constraining what the officer's preferences can be. This is similar to how an IV is used in demand estimation, where the instrument shifts supply without shifting demand, generating a sequence of equilibria tracing out the demand curve. Since it is possible to vary the risk of drivers stopped for each officer separately, it is possible to apply the proposed methods on each officer separately, thus permitting unrestricted heterogeneity across officers.

I apply the methods on a panel data set tracking officers in the Metropolitan Nashville Police Department (MNPD) between 2010 and 2019. I restrict my attention to the 50 officers with the most number of searches, who have made over 2,100 stops and 250 searches on average for each group of drivers. Across two sets of estimates, there are six officers who fail the test at the 5% significance level. For each of these officers, I estimate the average intensity of bias, as well as how the intensity of bias varies with the risk of the driver.

The paper proceeds as follows. Section 2 reviews the literature on testing for racial bias; Section 3 presents the model of an individual officer's search decision; Section 4 formalizes how bias may be detected and measured; Section 5 discusses the application; and Section 6 concludes.

2 Literature review

It is well documented that Black civilians are more likely to be stopped (Gelman et al., 2007), searched (Pierson et al., 2020), and killed by police officers than white civilians.[1] It is challenging to determine whether these disparities stem from racial bias because officers practice discretion when making their decisions, and researchers do not know what officers are thinking when those decisions are made. In this section, I summarize earlier approaches to overcoming these difficulties in detecting racial bias.

Knowles et al. (2001) lay the foundation for detecting racial bias in traffic searches using the outcome test proposed by Becker (1957, 1993). Officers are modeled as being homogeneous and only search drivers whose perceived risk of carrying contraband exceeds a fixed threshold. If the thresholds differ for white and minority drivers, then officers are racially biased. The researcher's objective is thus to recover these thresholds. If risk is observed by the researcher and continuously distributed over the unit interval, then the thresholds are identified from the risk of the white and minority drivers at the margin of search.

However, because risk is unobserved, the researcher must use a different strategy. Knowles et al. (2001) form a game-theoretic argument that drivers of the same race have the same risk in equilibrium, placing every driver at the margin of search.[2] The authors show that if officers are unbiased, then all white and minority drivers carry contraband with equal probability. This results in a straightforward test for bias: if officers have different success ("hit") rates when searching white and minority drivers, then officers are biased. However, the model's implication of homogeneous drivers within race, as well as its assumption of homogeneous officers, may both be rejected using officer-level data. For instance, the variation across MNPD officers in their hit rates reveals that drivers are not homogeneous in risk. In addition, Ba et al. (2021) find that the rate at which officers stop, arrest, and use force against civilians varies with the race and sex of officers.[3]

Anwar and Fang (2006) propose an alternative test that allows for heterogeneity in officer decisions and driver risk. By extending the model of Knowles et al. (2001) to allow different officers to have different thresholds, Anwar and Fang (2006) test for bias using pairwise comparisons of search decisions across groups of officers (e.g., white officers versus Black

[1] Source: Fatal Force, Washington Post.

[2] The argument is that drivers who are more likely to carry contraband will be searched more frequently. These drivers are therefore discouraged from carrying contraband. In equilibrium, all drivers of the same race carry contraband with equal probability and officers search each race at random.

[3] From surveys conducted on officers, Morin et al. (2017) find that men are three times more likely than women to have discharged their service weapon while on duty (30% versus 11%). White officers are also 80% more likely than Black officers to have been in an altercation with a civilian within a month prior to the interview (36% versus 20%).

officers). If both groups of officers are unbiased, then the ranking of their search rates should be the same regardless of the race of the driver. While this approach can detect bias, it cannot determine which group of officers is biased, nor which group of drivers is being discriminated against.

Recently, Arnold et al. (2018) made an important contribution to the literature by using random assignment of defendants to judges as an instrument to detect racial bias in bail decisions. The authors extend the model of Anwar and Fang (2006) by allowing thresholds to be distributed continuously across decision makers. Under conditions formalized by Canay et al. (2022), the thresholds of all decision makers can be point identified using the marginal treatment effect framework of Heckman and Vytlacil (2005). These conditions include decision makers facing identical distributions of risk (hence the importance of random assignment) and modeling decision makers using the Extended Roy Model (i.e., fixed thresholds). This method is referred to as the marginal outcome test.

To see whether the marginal outcome test extends to the context of police traffic searches, Gelbach (2021) tests three implications of the marginal outcome test framework on police traffic data from Florida and Texas.[4] The implications are not satisfied and the author points to different distributions of risk across officers as the potential reason. Such differences can arise if officers are not randomly assigned to drivers or vary in their ability to assess the risk of drivers. Papers using the marginal outcome test to study bias in policing therefore require restrictions on the distributions of risk. For example, Marx (2022) requires the distributions of risk to be common across officers. Feigenberg and Miller (2022) allow the distributions of risk to vary across officers, but rule out sample selection on unobservables.[5] Arnold et al. (2022) also allow decision makers to face different distributions of risk, but require parametric assumptions on the joint distribution of thresholds and risk.[6]

Other papers have used statistical approaches to test whether civilian race has an effect on police decisions, including stop-and-frisk and use of force (Ridgeway, 2006; Grogger and Ridgeway, 2006; Gelman et al., 2007; Ridgeway and MacDonald, 2009; Goel et al., 2016a,b; Fryer Jr, 2019; MacDonald and Fagan, 2019; Knox et al., 2020a). These papers either assume that the distribution of risk may be balanced across races, or cannot attribute the effect of race to racial bias. Knox et al. (2020a) is noteworthy for emphasizing the difficulty of identifying the effect of race on post-stop decisions alone (e.g., use of force, traffic searches)

[4] Frandsen et al. (2023) propose a test for the exclusion and monotonicity assumptions of the marginal outcome test in the setting where legal cases are randomly assigned to judges.

[5] The difference-in-differences strategy used by Goncalves and Mello (2021) to study racial bias among officers writing speeding tickets also rules out sample selection on unobservables.

[6] See also Simoiu et al. (2017), Pierson et al. (2018), Pierson et al. (2020), and Chan et al. (2022), who impose similar parametric restrictions to identify thresholds of decision makers.

because of sample selection. The authors show that, under a principal strata framework, this is only possible in the knife-edge scenario where the biases from sample selection and omitted variables cancel each other out (Knox et al., 2020b; Gaebler et al., 2020).

3 Model

In this section, I model the search decision of a single officer (he) for drivers who are stopped (she). Since the analysis allows for unrestricted heterogeneity across officers, I suppress the officer indexes for brevity. Similar to Knowles et al. (2001) and Anwar and Fang (2006), I also suppress the notation indicating the analysis is conditional on drivers who are stopped.

3.1 Setup and notation

For each stop i, the officer observes the driver's race $R_i \in \{w, m\}$ (white or minority), and a set of non-race characteristics $V_i \in \mathcal{V}$ that may include the driver's demeanor, the direction of travel, and any other details the officer notices. Components of V_i may be observed by the officer prior to the stop. Some components of V_i may also be observable to one officer but not another. This allows officers to vary in their perceptiveness and form different beliefs about the driver's risk. The researcher only observes R_i but not V_i; any other characteristics of the driver and the stop observed by the researcher are implicitly conditioned on throughout. In Section 5, I discuss the variables being conditioned on in the application.

The driver may carry contraband (e.g., drugs, weapons), denoted by $Guilty_i \in \{0, 1\}$. The officer does not know whether the driver is guilty unless he performs a traffic search, denoted by $Search_i \in \{0, 1\}$. At the end of each traffic stop, the officer reports in the data whether a search was conducted and whether contraband was found. Finding contraband is referred to as a "hit,"

$$Hit_i \equiv Search_i \times Guilty_i.$$

I assume that the officer finds contraband if and only if he searches a guilty driver, as in Knowles et al. (2001) and Anwar and Fang (2006).

I assume drivers are drawn from a distribution that depends on the setting of the stop, $Z_i \in \mathcal{Z}$. For example, Z_i may be the hour and day of the stop, and the interpretation of this assumption would be that different types of drivers are stopped at different times. This may be because the composition of drivers on the road changes with time, or because the officer's stop decision changes with time.[7] The setting is observed by both the officer and

[7] If there are variables that inform the officer's stop decision and are visible for one value of Z_i but not

researcher, and will play the role of an instrument.

When deciding whether to search, the officer considers four possible outcomes of his decision: (i) searching an innocent driver; (ii) searching a guilty driver; (iii) not searching an innocent driver; and (iv) not searching a guilty driver. Associated with each outcome is an *ex post* utility that the officer learns after interacting with the driver and observing all of her characteristics, but prior to making his search decision. Let $\mathcal{U}_i^s(g;r)$ denote this utility when $Search_i = s$ and $Guilty_i = g$ for a driver with race $R_i = r$. These utilities are random and can vary across drivers who are observationally equivalent to the officer. The distributions of these utilities represent the officer's search preferences, and the objective of the test is to detect whether race has a direct effect on these distributions. To do this, I make the following assumption about the utilities $\{\mathcal{U}_i^0(g;r), \mathcal{U}_i^1(g;r)\}_{(g,r)\in\{0,1\}\times\{w,m\}}$, which I denote by $\{\mathcal{U}_i\}$ for brevity.

Assumption 1.

(i) $\mathcal{U}_i^1(1;R_i) - \mathcal{U}_i^1(0;R_i) > 0$ *and* $\mathcal{U}_i^0(1;R_i) - \mathcal{U}_i^0(0;R_i) < 0$ *for all* i.

(ii) $\{\mathcal{U}_i\}$ *are identically distributed across stops* i.

(iii) $\{\mathcal{U}_i\} \perp\!\!\!\perp (Z_i, Guilty_i, V_i) \mid R_i$.

Assumption 1(i) states that, for all drivers, the officer prefers to make the correct decision by searching guilty drivers and not searching innocent drivers. This implies that officers are more likely to search drivers who have greater probability of carrying contraband.

Assumption 1(ii) states that the utilities across stops are drawn from a common distribution. This allows me to pool the drivers of the same race together to infer the officer's preferences. Conditioning the analysis on observed variables that affect the distribution of utilities (e.g., age and sex of the driver) helps to satisfy this assumption. If instead $\{\mathcal{U}_i\}$ and $\{\mathcal{U}_{i'}\}$ were drawn from different distributions for every $i \neq i'$, there would be no way to use multiple stops to infer preferences.

Assumption 1(iii) states that the utilities $\{\mathcal{U}_i\}$ are independent of the joint distribution of the setting Z_i, the guilt of the driver $Guilty_i$, and the unobserved driver characteristics V_i after conditioning on the driver's race R_i. This is the key assumption of the model and determines how racial bias is defined and how it can be detected. I discuss Assumption 1(iii) in detail in Section 3.3.

another (e.g., race is visible during the day before stopping a driver, but is not visible at night), then the distribution of drivers stopped will vary with Z_i even if the composition of drivers on the road do not. This type of variation is used in the Veil of Darkness test by Grogger and Ridgeway (2006) to test whether race affects the stop decision.

Under Assumption 1, any dependence between the officer's preferences and the driver's race can only be through race, leading to the following definition of racial bias.

Definition 1. *The officer is racially biased in traffic searches if* $\{\mathcal{U}_i^s(g;w)\}_{(g,s)\in\{0,1\}^2}$ *and* $\{\mathcal{U}_i^s(g;m)\}_{(g,s)\in\{0,1\}^2}$ *do not have the same distribution.*

The objective of the test is thus to determine whether the distribution of utilities depends on race. In Section 3.3, I provide a more nuanced definition of bias that depends on the probability that a driver carries contraband.

3.2 Search decision

I assume that an officer seeks to maximize his utility when faced with a traffic search decision. Since the driver's guilt is not known to the officer, he chooses the decision that maximizes his expected utility. For search decision s, his expected utility is

$$\begin{aligned}&\mathbb{E}[\mathcal{U}_i^s(\mathit{Guilty}_i;R_i)\mid R_i=r,Z_i=z,V_i=v]\\&=G(r,z,v)\,\mathcal{U}_i^s(1;R_i)\;+(1-G(r,z,v))\,\mathcal{U}_i^s(0;R_i),\end{aligned}$$

where

$$G(r,z,v)\equiv\mathbb{P}\{\mathit{Guilty}_i=1\mid R_i=r,Z_i=z,V_i=v\}$$

is the officer's belief of the probability that the driver caries contraband, which I refer to as the "risk" of the driver. The officer's search decision may then be written as

$$\begin{aligned}\mathit{Search}_i&\equiv\arg\max_{s\in\{0,1\}}\mathbb{E}[\mathcal{U}_i^s(\mathit{Guilty}_i;R_i)\mid R_i,Z_i,V_i]\\&=\mathbb{1}\left\{G(R_i,Z_i,V_i)\geq T_i\right\},\end{aligned}\tag{1}$$

where

$$T_i\equiv\frac{\mathcal{U}_i^0(0;R_i)-\mathcal{U}_i^1(0;R_i)}{[\mathcal{U}_i^1(1;R_i)-\mathcal{U}_i^1(0;R_i)]-[\mathcal{U}_i^0(1;R_i)-\mathcal{U}_i^0(0;R_i)]}$$

is a random utility threshold representing the officer's preferences. See Appendix A for the full derivation. The officer thus searches a driver if and only if her risk is sufficiently large, and how large that risk must be can vary across stops. The researcher observes neither $G(R_i,Z_i,V_i)$ nor T_i.

From its definition, T_i inherits the properties of $\{\mathcal{U}_i\}$ stated in Assumption 1 and may be used to define and detect racial bias.

Corollary 1.

(i) $T_i \mid R_i = r$ *is identically distributed across stops* i *for* $r \in \{w, m\}$.

(ii) $T_i \perp\!\!\!\perp (Z_i, Guilty_i, V_i) \mid R_i = r$ *for* $r \in \{w, m\}$.

(iii) *The officer is racially biased in traffic searches if* $T_i \not\perp\!\!\!\perp R_i$.

Proof. The random threshold T_i is a deterministic function of the utilities $\{\mathcal{U}_i\}$. Properties (i)–(ii) of the corollary follow immediately from Assumptions 1(ii)–1(iii). Property (iii) of the corollary follows immediately from Definition 1. ■

Instead of comparing the distribution of $\{\mathcal{U}_i\}$ across races to detect bias, it suffices to compare the distribution of T_i across races.

3.3 Discussion

Whereas earlier papers assume the threshold T_i is a deterministic function of race (Knowles et al., 2001; Anwar and Fang, 2006; Arnold et al., 2018),[8] I allow $T_i \mid R_i = r$ to be random. I show below that this stochastic threshold permits a new form of heterogeneity in bias where the severity and direction of bias can change with the risk of the driver. Nevertheless, as discussed by Canay et al. (2022), restrictions on the distribution of $T_i \mid R_i$ are required for there to be testable implications of racial bias. In the remainder of this section, I elaborate on the restrictions I impose on $T_i \mid R_i$ before deriving the test for racial bias in Section 4.

There are two key properties of $T_i \mid R_i$ that allow me to test for racial bias. The first is the independence property $T_i \perp\!\!\!\perp V_i \mid R_i$, which follows from the independence between $\{\mathcal{U}_i\}$ and V_i stated in Assumption 1(iii). This property is akin to the restriction separating the Extended Roy Model from the Generalized Roy Model highlighted by Canay et al. (2022). It is imposed in existing tests and is required to make a direct link between an officer's preferences and a driver's race. The property rules out cases where, for example, the officer is biased against facial tattoos (a driver characteristic unobserved by the researcher), which may be more common in one race than the other. In this example, bias against facial tattoos can be conflated with racial bias, since differences between $T_i \mid R_i = w$ and $T_i \mid R_i = m$ may stem from racial disparities in the prevalence of facial tattoos rather than race itself. However, if $T_i \perp\!\!\!\perp V_i \mid R_i$, then the unobserved characteristics V_i affect the search decision exclusively through the risk of the driver, $G(R_i, Z_i, V_i)$. This in turn implies that omitted variables, sample selection, and statistical discrimination—usual confounders of bias that

[8] A threshold that is a deterministic function of race can be obtained by assuming $\{\mathcal{U}_i\}$ are degenerate random variables.

operate through V_i—only affect the search decision through $G(R_i, Z_i, V_i)$. The econometric challenge of detecting bias is thus to separately infer the distributions of T_i and $G(R_i, Z_i, V_i)$.

To elaborate on how the three confounders affect the distribution of risk, consider an example where V_i is the condition of the vehicle, and aging vehicles are more likely to contain contraband.

Omitted variable bias pertains to differences in the distribution of V_i across races in population, e.g., whites are twice as likely to drive aging vehicles than minorities in population. So even if the officer stops drivers at random, the distribution of risk may differ across race since the underlying determinant V_i differs across race.

Sample selection pertains to differences in the distribution of V_i across race for drivers who are stopped, e.g., the officer may prefer to stop minority drivers in aging vehicles, so conditional on being stopped, whites are only half as likely to be in aging vehicles than minorities, despite how whites are twice as likely to drive aging vehicles in population.

Finally, statistical discrimination (in the sense of Aigner and Cain (1977)) pertains to how V_i maps to risk differently for white and minority drivers, e.g., aging vehicles are correlated with possessing contraband for whites but not for minorities. This notion of statistical discrimination also extends to other officers, where different officers observe different components of V_i (Hull, 2021; Arnold et al., 2022). For example, an experienced officer may know to consider the direction of travel along a highway when assessing the driver's risk (Barnes, 2004), whereas an inexperienced officer may not. Since the test may be applied to each officer separately, I place no restrictions on how different officers infer the risk of drivers.

The second key property of $T_i \mid R_i$ that allows me to test for racial bias is that $T_i \perp\!\!\!\perp Z_i \mid R_i$, which follows from the independence between $\{\mathcal{U}_i\}$ and Z_i stated in Assumption 1(iii). This property implies that Z_i affects the search decision exclusively through $G(R_i, Z_i, V_i)$, and is akin to an exogeneity condition that allows Z_i to shift the distribution of risk (by changing the types of drivers stopped) without shifting the distribution of thresholds. Such variation is helpful in partially identifying the distribution of T_i for each race. The intuition for this is similar to that of using an IV to identify a demand curve, where the instrument exclusively shifts the supply curve, generating a sequence of equilibria tracing out the demand curve. In my setting, Z_i exclusively shifts the distribution of risk, generating a sequence of search and hit rates that constrain what the distribution of T_i can be for each race. In Section 4, I show an example where bias can be detected even without variation in Z_i. However, such a test relies only on the variation in search decisions generated by R_i and can be weak.

Conditional on race, Z_i may shift $G(R_i, Z_i, V_i)$ in two ways. The first is through shifting the distribution of V_i, e.g., $G(R_i, Z_i, V_i)$ does not depend on Z_i but $Z_i \not\perp\!\!\!\perp V_i \mid R_i$. An example of this is if Z_i is the time of the traffic stop, and the time of the stop contains no information

on whether the driver is guilty, but criminals tend to drive at night. The second way Z_i may shift risk is to have a direct effect on $G(R_i, Z_i, V_i)$, i.e., $G(R_i, z_1, V_i) \neq G(R_i, z_2, V_i)$ for $z_1 \neq z_2$. This reflects how the same signals can be interpreted differently depending on the setting of the stop (Engel and Johnson, 2006; Novak and Chamlin, 2012). For example, stopping a white driver in a predominantly white suburb may not be suspicious, whereas stopping the same driver in a predominantly Black neighborhood may be more suspicious. Similarly, stopping a high school student in the afternoon shortly after school has ended is less suspicious than stopping the same student late into the night.

This instrument separates my approach from those using random assignment of decision makers as the instrument (Arnold et al., 2018, 2022). In my setting, the alternative instrument of random assignment would require officers to be randomly assigned to drivers and would imply that all officers share a common distribution of risk. This provides a way for the researcher to vary the officer's preferences without affecting the distribution of risk. However, this alternative instrument is difficult to justify in my setting since police traffic data are conditional on drivers who are stopped. Stopping a driver is not a random decision, and different officers may choose to stop different types of drivers, resulting in different distributions of risk.

Another distinguishing feature of the instrument I propose is that it allows me to test each officer separately for bias. This is because my identification strategy exploits variation in search and hit rates, and my instrument is able to generate such variation within officer by shifting the distribution of risk. I therefore allow unrestricted heterogeneity in both officer preferences and how officers infer the risk of drivers.

Under Assumption 1 and decision rule (1), the probability that a driver is searched may be written as

$$\begin{aligned}
&\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z, V_i = v\} \\
&= \mathbb{P}\{G(R_i, Z_i, V_i) \geq T_i \mid R_i = r, Z_i = z, V_i = v\} \\
&= \mathbb{P}\{G(r, z, v) \geq T_i \mid R_i = r, Z_i = z, V_i = v\} \\
&= \mathbb{P}\{G(r, z, v) \geq T_i \mid R_i = r\} \\
&= F_{T|R}(G(r, z, v) \mid r),
\end{aligned}$$

where the third equality follows from Assumption 1(iii),[9] and $F_{T|R}$ denotes the CDF of random variable T_i conditional on R_i. The probability a driver is searched is therefore equal

[9]The independence between $\{\mathcal{U}_i\}$ and $Guilty_i$ stated in Assumption 1(iii) requires the officer to infer the probability a driver carries contraband using only details from the traffic stop and not his utilities. This rules out clairvoyance, where the officer infers the driver's guilt using information beyond what is provided by the traffic stop.

to the probability the officer's threshold falls below the driver's risk. This permits a more nuanced definition of bias.

Definition 2. *The officer is racially biased at risk $g \in [0,1]$ if*

$$\beta(g) \equiv F_{T|R}(g \mid m) - F_{T|R}(g \mid w) \neq 0,$$

where $\beta(g) > 0$ ($\beta(g) < 0$) if the officer is biased against minority (white) drivers with risk g.

If $\beta(g) \neq 0$ for any $g \in [0,1]$, it immediately follows that the officer is biased, as defined in Definition 1. However, the converse does not hold. That is, an officer with different utilities for searching white and minority drivers can have identical distributions of T_i for both groups drivers.[10] I ignore these cases since the bias does not affect the search decision.

Since $\beta(g)$ can vary with g and even change sign, the intensity and direction of bias can vary with the unobserved (to the researcher) risk of the driver. This feature of the model arises from the random threshold and distinguishes my model from earlier models where, conditional on race, an officer searches all drivers with a given level of risk or none at all (Knowles et al., 2001; Anwar and Fang, 2006; Arnold et al., 2018; Hull, 2021). A random threshold thus extends the notion of the marginal driver to every level of risk and permits a more nuanced analysis of bias.[11] I show in Section 4 how sharp bounds on $\beta(\cdot)$ may be derived.

A concern with existing tests of racial bias is the accuracy of the decision maker's beliefs and whether it is possible to distinguish between inaccurate beliefs and racial bias (Bordalo et al., 2016; Bohren et al., 2022). To illustrate the problem, suppose an unbiased officer incorrectly believes minority drivers are twice as risky as they truly are. His search decision may then be written as

$$\begin{aligned} \mathit{Search}_i &= \mathbb{1}\left\{(1 + \mathbb{1}\{R_i = m\})\, G(R_i, Z_i, V_i) \geq T_i\right\} \\ &= \mathbb{1}\left\{G(R_i, Z_i, V_i) \geq \widetilde{T}_i\right\}, \end{aligned}$$

where $\widetilde{T}_i \equiv T_i/(1 + \mathbb{1}\{R_i = m\})$. In this example, the effect of inaccurate beliefs is observationally equivalent to the officer drawing thresholds that are half as large for minorities compared to whites. Earlier tests, as well as the one I propose, may incorrectly detect bias

[10] For example, suppose $\mathcal{U}_i^s(g; w) = k\mathcal{U}_i^s(g; m)$ for $(s, g) \in \{0,1\}^2$ and some constant $k \neq 0$. The thresholds for white and minority drivers will be identically distributed in this scenario.

[11] The model I propose nests the earlier models with fixed thresholds. In estimation, implementing a fixed threshold entails imposing integrality constraints on the officer's preferences.

in this example since $\widetilde{T}_i \not\perp\!\!\!\perp R_i$, although T_i is the true object of interest and may be independent of race. Nevertheless, these tests for bias are still valuable since the effects of inaccurate beliefs and bias are the same for drivers. These tests may serve as a preliminary check to determine which officers should be reviewed, and further investigation may reveal whether racial disparities in search behavior stem from bias or inaccurate beliefs. Given the difficulty of distinguishing between racial bias and inaccurate beliefs, researches have begun using experiments to elicit beliefs of decision makers when studying discrimination.[12]

4 Detecting and measuring racial bias

In line with Becker's (1957, 1993) outcome test, the test I propose checks whether an officer's search decisions are consistent with him being unbiased. If they are not, then the officer is deemed biased. To avoid conflating racial bias with omitted variable bias, sample selection, and statistical discrimination, I use a partial identification approach to infer the officer's preferences separately from the distribution of risk.

4.1 Defining the test

For each traffic stop, I observe the driver's race, R_i; the setting of the stop, Z_i; the search decision, $Search_i$; and whether contraband is found, Hit_i. From this, I am able to construct the officer's search and hit rates for race $r \in \{w, m\}$ and setting $z \in \mathcal{Z}$,

$$\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} = \int_{\mathcal{V}} F_{T|R}(G(r, z, v) \mid r)\ dF_{V|R,Z}(v \mid r, z), \tag{2}$$

$$\mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} = \int_{\mathcal{V}} G(r, z, v)\ F_{T|R}(G(r, z, v) \mid r)\ dF_{V|R,Z}(v \mid r, z). \tag{3}$$

These equations simply follow from the law of iterated expectations and Corollary 1.[13] The conditional hit rate is the probability that contraband is found conditional on a traffic search and is equal to the ratio of (2) and (3),

$$\mathbb{P}\{Hit_i = 1 \mid Search_i = 1, R_i = r, Z_i = z\} = \frac{\mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\}}{\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\}}.$$

The instrument Z_i varies the search and hit rates by varying the distributions of risk.

To define the identified set of the model, let $\mathcal{F}$ denote the space of distributions of

[12] See Bohren et al. (2019, 2022).

[13] See Appendix A for the full derivation.

$(V_i, T_i, Guilty_i) \mid R_i, Z_i$ satisfying Assumption 1. The sharp identified set is

$$\{F \in \mathcal{F} : (2) \text{ and } (3) \text{ are satisfied for all } (r, z) \in \{w, m\} \times \mathcal{Z}\}.$$

However, in testing for racial bias, the parameters of interest are only $F_{T|R}(\cdot \mid w)$ and $F_{T|R}(\cdot \mid m)$. So I consider a projection of the identified set when testing for bias.

To define this projection, let

$$G_i \equiv G(R_i, Z_i, V_i),$$
$$\sigma(\cdot; r) \equiv F_{T|R}(\cdot \mid r),$$

where G_i denotes the risk in stop i, and $\sigma(g; r)$ denotes the probability a driver with risk g and race r is searched. The function $\sigma(\cdot; r)$ represents the officer's search preference for race r and is the parameter of interest. Denote the distribution of risk conditional on race and setting by

$$F_{G|R,Z}(g \mid r, z) \equiv \int_{\mathcal{V}} \mathbb{1}\{G(r, z, v) \leq g\} \, dF_{V|R,Z}(v \mid r, z).$$

Equations (2)–(3) may then be written as

$$\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} = \int_0^1 \sigma(g; r) \, dF_{G|R,Z}(g \mid r, z), \tag{4}$$
$$\mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} = \int_0^1 g \, \sigma(g; r) \, dF_{G|R,Z}(g \mid r, z). \tag{5}$$

Let Σ denote the space of non-decreasing, right-continuous functions with domain and codomain equal to $[0, 1]$; and let $\mathcal{F}_G$ denote the space of distributions for scalar random variables with support $[0, 1]$. Then the sharp identified set for the officer's search preferences is

$$\Sigma^\dagger \equiv \left\{ (\sigma(\cdot; w), \sigma(\cdot; m)) \in \Sigma^2 : \begin{array}{c} \exists F_{G|R,Z}(\cdot \mid r, z) \in \mathcal{F}_G \text{ s.t. (4) and (5) are} \\ \text{satisfied for all } (r, z) \in \{w, m\} \times \mathcal{Z} \end{array} \right\}. \tag{6}$$

A testable implication for racial bias immediately follows from (6) (see Canay et al., 2013).

Corollary 2. *Define* $\Sigma^\star \equiv \{\sigma \in \Sigma : (\sigma, \sigma) \in \Sigma^\dagger\}$. *Under* (1) *and Assumption 1, if the officer is unbiased, then* $\Sigma^\star$ *is non-empty.*

Proof. Corollary 2 follows immediately from Definition 1 and property (iii) of Corollary 1. ■

Since $\Sigma^\dagger$ is sharp, Corollary 2 is the strongest testable implication of the model for unbiasedness.

4.2 Intuition

To build intuition for the test, consider a simple setting where risk is equal to 0, 0.5, or 1. The left panel of Figure 1 shows a search preference in this setting, with each square indicating the probability that the officer searches a driver with a given level of risk. The right panel shows the data that can be generated by this preference. The horizontal position of each square in the right panel is equal to the search probability $\sigma(g; r)$ for some risk g; and the vertical position is equal to the joint probability of searching the driver and finding contraband, $g\sigma(g; r)$.

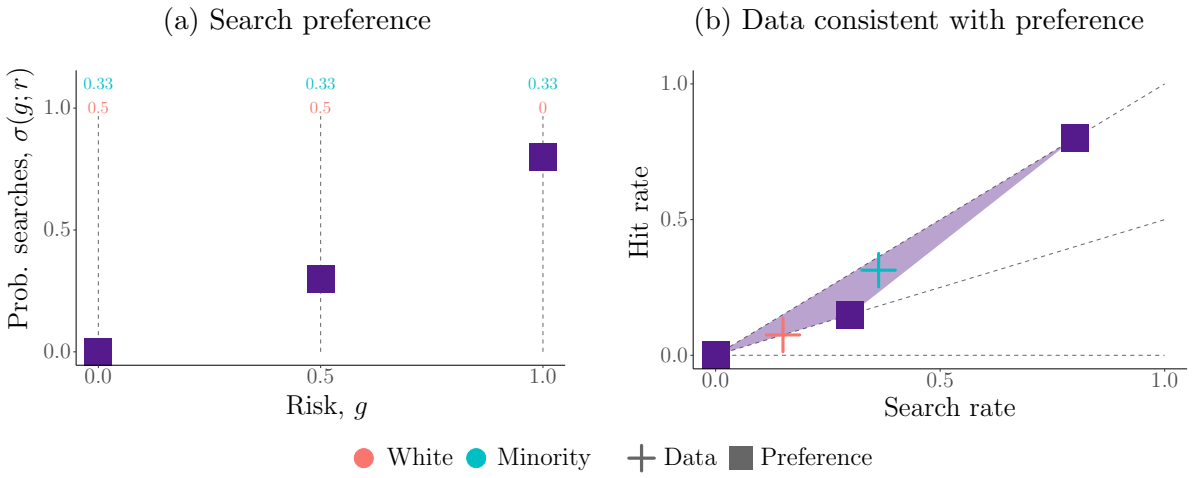
Equations (4)–(5) imply that the search and hit rates must lie in the convex hull of the three squares in the right panel, indicated by the purple region. Since the observed search and hit rates for both groups of drivers—represented by the crosses—indeed lie in the purple region, it is possible that both data points are generated by a common preference and it cannot be ruled out that the officer is unbiased. The colored numbers on the left panel indicate possible distributions of risk that generate the crosses of the same color.

Figure 2 presents the case where only the red cross lies in the convex hull generated by the preference in the left panel. This implies that the blue cross is generated by a different preference. Corollary 2 states that if the officer is unbiased, then there must exist a preference that generates a purple region in the right panel containing both data points, as in Figure 1. However, if no such preference exists, then the data for white and minority drivers must be generated by distinct preferences and the officer must be biased.

Figure 3 presents such a case, where no single preference is capable of generating the data for both groups of drivers. Racial bias is therefore detected. Note that this is possible even though I only have one data point for each race of drivers, which corresponds to the case where there is no variation in Z_i. If there is variation in Z_i, I would have multiple data points for each race of drivers. This strengthens the test as it is more difficult to find a single preference generating a larger number of data points.

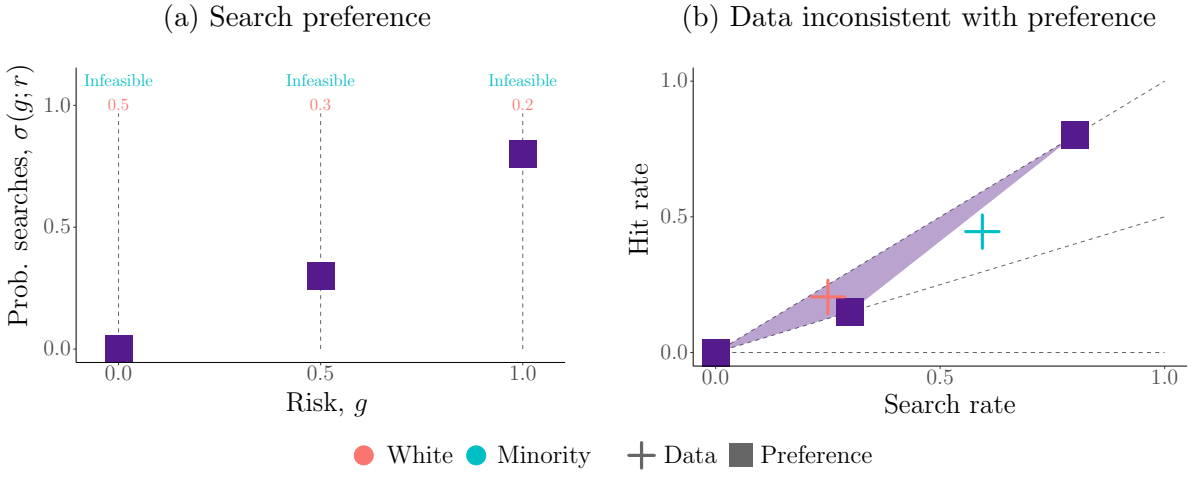
Beyond testing whether an officer is biased, my framework also allows me to obtain bounds on the intensity of bias. The left panel of Figure 3 shows two distinct preferences that could have generated the data in the right panel, as well as the implied intensity of bias at each level of risk. By considering different preferences and distributions of risk that are consistent with the data, I can derive bounds on various measures of bias.

Figure 1: How preferences generate search and hit rates



Note: The squares in each figure represent an officer's search preference. Data that are consistent with the officer's preference must lie inside the purple region in the right panel. The colored crosses in the right panel represent the observed search and hit rates. Since the data points lie inside the purple region, it is possible that they are generated by the preference shown in the left panel. The colored numbers in the left panel indicate possible distributions of risk generating the data points of the same color.

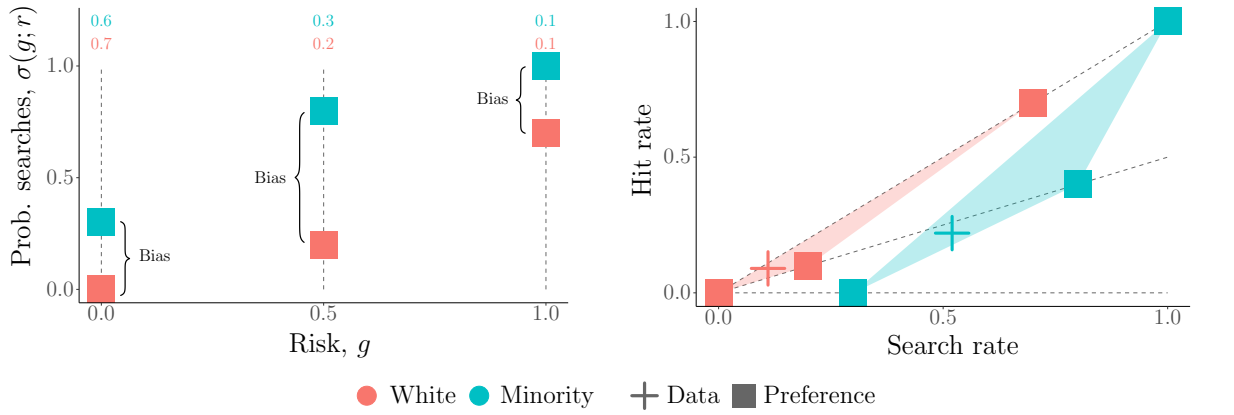
Figure 2: How search and hit rates are informative of preferences



Note: The red data point is consistent with the preference shown, whereas the blue data point is not. If the officer is unbiased, there must exist a different preference that generates a purple region in the right panel containing both data points. If no such preference exists, then the officer must have distinct preferences for white and minority drivers.

Figure 3: How search and hit rates are informative of bias

(a) Biased search preferences (b) Data only consistent with biased preferences



Note: Since it is impossible to find a single preference generating the data for both white and minority drivers, the officer must be biased. Any pair of preferences required to generate the data for both groups of drivers implies an intensity of bias at each level of risk. By exploring the space of preferences consistent with the data, I am able to derive bounds on various measures of bias (e.g., bias condition on risk, bias averaged over risk).

4.3 Implementation

Corollary 2 may be implemented as a bilinear programming (BP) problem. Despite being non-convex, bilinear programs can be solved to provable global optimality by commercial solvers.[14] For simplicity, suppose that G_i is discrete and $\text{supp}\,(G_i) = \{g_1, \ldots, g_K\}$ for finite K. Then (4)–(5) become

$$\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} = \sum_{k=1}^{K} \sigma(g_k; r)\ p_{r,z}(g_k), \tag{7}$$

$$\mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} = \sum_{k=1}^{K} g_k\ \sigma(g_k; r)\ p_{r,z}(g_k), \tag{8}$$

where

$$p_{r,z}(g) \equiv \mathbb{P}\{G_i = g \mid R_i = r, Z_i = z\}$$

denotes the distribution of risk conditional on the race of the driver and setting of the stop. Online Appendix B discusses how B-splines may be used to model preferences and risk if G_i

[14] Bilinear programs are solved to global optimality using a branch-and-bound algorithm. The domain space is partitioned, and convex McCormick relaxations of the original problem are solved across the partitions. See McCormick (1976), Mehlhorn et al. (2008), and Gurobi Optimization, Inc. (2021).

is continuously distributed over the unit interval.

To specify the BP problem, I introduce the following notation.

$$\begin{aligned}
\mathbf{m}^S_{r,z} &\equiv \mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} \\
\mathbf{m}^H_{r,z} &\equiv \mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} \\
\mathbf{g} &\equiv (g_1, \ldots, g_K)' \\
\boldsymbol{\sigma}_r &\equiv (\sigma(g_1; r), \ldots, \sigma(g_K; r))' \\
\mathbf{p}_{r,z} &\equiv (p_{r,z}(g_1), \ldots, p_{r,z}(g_K))'
\end{aligned}$$

The moments $\mathbf{m}^S_{r,z}$, $\mathbf{m}^H_{r,z}$ are the search and hit rates for each race r and setting z and are identified from the data. The vector $\mathbf{g}$ is the support of G_i, which I assume is known to the researcher. The unknown parameters of the BP problem are $\{\boldsymbol{\sigma}_r\}_{r\in\{w,m\}}$, which are the values of preferences $\sigma(\cdot; r) \in \Sigma$ evaluated at each point of $\mathbf{g}$; and $\{\mathbf{p}_{r,z}\}_{(r,z)\in\{w,m\}\times\mathcal{Z}}$, which are the distributions of risk conditional on race and setting. For brevity, I refer to the preferences by $\{\boldsymbol{\sigma}_r\}$ and the distributions of risk by $\{\mathbf{p}_{r,z}\}$.

To ensure these parameters are consistent with the model, I impose two baseline sets of constraints. Both sets of constraints are linear in the parameters of the model. The first set is

$$0 \leq \boldsymbol{\sigma}_{r,k} \leq \boldsymbol{\sigma}_{r,k+1} \leq 1 \text{ for } r \in \{w, m\} \text{ and } k = 1, \ldots, K-1, \tag{9}$$

where $\boldsymbol{\sigma}_{r,k}$ denotes the k^{th} component of $\boldsymbol{\sigma}_r$, i.e., $\boldsymbol{\sigma}_{r,k} = \sigma(g_k; r)$. This ensures $\sigma(\cdot; r) \in \Sigma$, as required by Corollary 2. The second set of constraints is

$$\mathbf{p}_{r,z,k} \in [0, 1] \text{ for } (r, z) \in \{w, m\} \times \mathcal{Z} \text{ and } k = 1, \ldots, K, \tag{10}$$

$$\sum_{k=1}^{K} \mathbf{p}_{r,z,k} = 1 \text{ for } (r, z) \in \{w, m\} \times \mathcal{Z}, \tag{11}$$

where $\mathbf{p}_{r,z,k}$ denotes the k^{th} component of $\mathbf{p}_{r,z}$. This ensures $\mathbf{p}_{r,z} \in \mathcal{F}_G$ for $(r, z) \in \{w, m\} \times \mathcal{Z}$, as required by the definition of $\Sigma^\star$. To simplify the discussion, I assume that $\text{supp}\,(Z_i \mid R_i = w) = \text{supp}\,(Z_i \mid R_i = m)$, but this assumption is not necessary.

Define the population criterion function as

$$Q(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) \equiv \sum_{r,z} \left|\boldsymbol{\sigma}_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z}\right| + \sum_{r,z} \left|(\mathbf{g} \odot \boldsymbol{\sigma}_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z}\right|,$$

where $\odot$ denotes the Hadamard (element-wise) product. The criterion function measures

how much (7)–(8) are violated. The following proposition describes how to test for bias in population using Corollary 2.

Proposition 1. *Define* $Q^\star$ *as*

$$Q^\star \equiv \min_{\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}} Q(\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}) \tag{12}$$
$$\text{s.t.} \quad \boldsymbol{\sigma}_w = \boldsymbol{\sigma}_m, \ (9), \ (10), \ (11).$$

The officer is biased if $Q^\star > 0$.

Proof. The constraint $\boldsymbol{\sigma}_w = \boldsymbol{\sigma}_m$ restricts the officer to be unbiased. If $Q^\star > 0$, then (7) or (8) is violated for some $(r, z) \in \{w, m\} \times \mathcal{Z}$. Then by Corollary 2, the officer is biased. ■

The criterion $Q^\star$ in Proposition 1 is the minimum ℓ^1-norm between the moments of the model and the moments of the data when the officer is restricted to be unbiased. Since the ℓ^1-norm can be reformulated as being linear, the criterion function in (12) is bilinear. Other norms may be used but may be more computationally demanding.

4.3.1 Adding restrictions

It is straightforward to strengthen the test by adding restrictions to Σ and $\mathcal{F}_G$. This can be done in a transparent, modular fashion.

For example, consider restricting the mass of drivers to be decreasing as risk increases,

$$\mathbf{p}_{r,z,k} \geq \mathbf{p}_{r,z,k+1} \text{ for } (r, z) \in \{w, m\} \times \mathcal{Z} \text{ and } k = 1, \ldots K - 1. \tag{13}$$

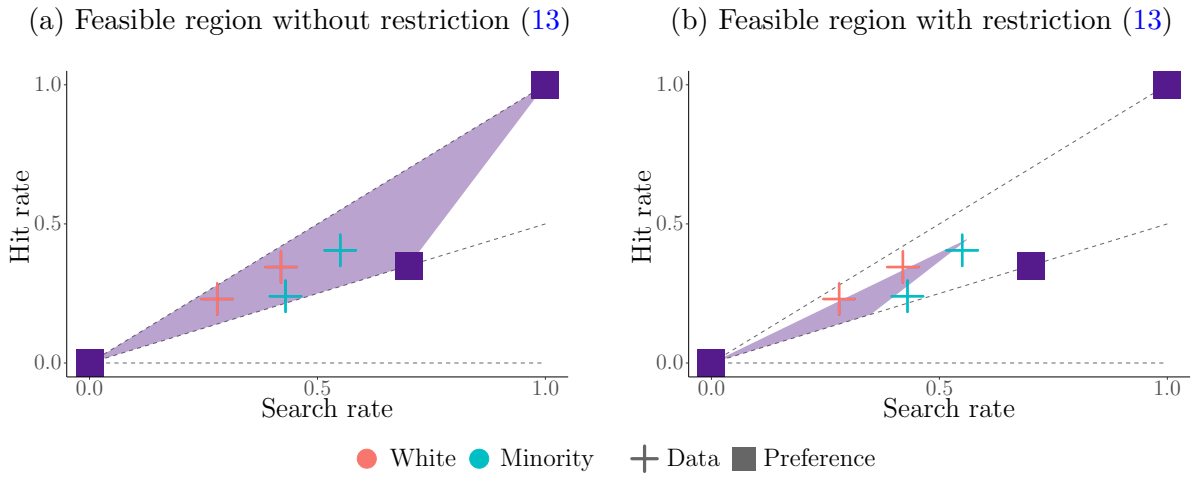
Such an assumption is suitable when the mass of low-risk drivers in population is large compared to the mass of high-risk drivers. Even if the officer is much more likely to stop high-risk drivers, the greater volume of low-risk drivers on the road may result in a distribution of risk (conditional on being stopped) where the mass of drivers decreases as risk increases.[15]

Figure 4 demonstrates how this restriction strengthens the test. The same preference is depicted in the left and right panel. However, the range of data that can be generated by the preference is reduced when (13) is imposed. In fact, while there exist preferences capable of generating the data for both races when there are no restrictions on the distributions of risk, it is no longer the case once (13) is imposed.

For more examples of imposing restrictions on the model, see Online Appendix A.

[15] See Online Appendix A.3 for a numerical example.

Figure 4: Strengthening the test by restricting $\{\mathbf{p}_{r,z}\}$



Note: The purple region in the left panel shows the possible data points generated by a particular preference when there are no restrictions on the distribution of risk. The purple region in the right panel shows the possible data points generated by the same preference, except the mass of drivers is restricted to be decreasing as risk increases. Reducing the size of the purple region strengthens the test for racial bias by making it easier to rule out preferences from $\Sigma^\star$.

4.4 Determining the direction and intensity of bias

If bias is detected, the next step is to determine how the officer is biased. This can be done in several ways. Below, I first introduce a general measure of bias and show how it can be bounded. I then show some restrictions that can be imposed to obtain specific measures of bias.

4.4.1 Bounding a general measure of bias

The general measure of bias takes the form

$$\theta \equiv \boldsymbol{\omega}'(\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w), \tag{14}$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K)'$ is a vector of weights with $\boldsymbol{\omega}_k \in [0,1]$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \boldsymbol{\omega}_k = 1$. θ is thus a weighted average of the bias at each level of risk and $\boldsymbol{\omega}$ is a counterfactual distribution of risk.[16] The choice of $\boldsymbol{\omega}$ determines the measure of bias, and the weights can be chosen beforehand or treated as variables in the BP problem. If $\theta > 0$, then the officer is biased against minorities given $\boldsymbol{\omega}$. If $\theta < 0$, then the officer is biased

[16] Oaxaca (1973), Blinder (1973), and DiNardo et al. (1996) decompose average outcomes into structural and composition effects. By reweighting the structural effects, the authors are able to construct counterfactuals. θ is constructed in a similar way, where $\boldsymbol{\omega}$ reweights the effect of race on search rates captured by $\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w$. See Fortin et al. (2011) for a summary of decomposition methods in economics.

against whites.

Proposition 2. *The sharp bounds on θ are obtained by solving the following BP problem,*

$$\begin{aligned} \theta_{\text{lb}}, \theta_{\text{ub}} \equiv \min_{\boldsymbol{\omega}, \{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}} / \max_{\boldsymbol{\omega}, \{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}} \; & \boldsymbol{\omega}' (\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w) \\ \text{s.t.} \quad & Q(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) = 0, \; (9), \; (10), \; (11). \end{aligned} \tag{15}$$

Proof. The objective in (15) defines the measure of bias, θ. Since the constraints characterize the sharp identified set $\Sigma^\dagger$, the bounds on θ are sharp by construction. ■

Let Θ denote the identified set for θ. The bounds in Proposition 2 are sharp in the sense that they are the smallest and largest values of Θ. However, because bilinear programs are non-convex, Θ need not be the full interval $[\theta_{\text{lb}}, \theta_{\text{ub}}]$. I focus the discussion on the bounds in Proposition 2, although Θ can be constructed by "inverting" (15), similar to how a confidence interval can be constructed by inverting a statistical test. See Appendix B for how to fully recover Θ.

When there are no restrictions on $\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w$, the officer can be biased against one group of drivers for a given level of risk and reverse their direction of bias at another level of risk. If the researcher has a strong prior on the direction of bias, then a sign restriction on the elements of $\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w$ can easily be imposed. For example, bias against white drivers can be ruled out if every element of $\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w$ is restricted to be non-negative.

4.4.2 Bounding bias conditional on risk

A parameter of interest may be the bias conditional on risk, $\beta(\cdot)$, as defined in Definition 2. The bounds on $\beta(g_k)$ are obtained by setting

$$\theta = \sigma(g_k; m) - \sigma(g_k; w) = \beta(g_k).$$

This corresponds to setting $\boldsymbol{\omega} = \mathbf{e}_k$, where $\mathbf{e}_k \in \mathbb{R}^K$ is the k^{th} standard basis vector. The researcher can therefore bound the bias at every level of risk. It is possible for $0 \in [\theta_{\text{lb}}, \theta_{\text{ub}}]$ for every level of risk even if the officer fails the test in Proposition 1. This corresponds to the case where bias is detected, but the direction of bias is undetermined.

4.4.3 Bounding average bias

Another parameter of interest is the average bias under a counterfactual distribution of risk. A specific distribution of risk can be imposed by setting the weights $\boldsymbol{\omega}$ equal to that

distribution. For example, the average bias under the counterfactual where risk is uniform for both groups of drivers corresponds to the constraint

$$\boldsymbol{\omega}_k = \frac{1}{K} \text{ for all } k = 1, \dots K.$$

A more interesting measure of bias is one that uses the actual unobserved distribution of risk for white or minority drivers. For example, the following constraint sets $\boldsymbol{\omega}$ equal to the distribution of risk for white drivers in the data,

$$\begin{aligned}
\boldsymbol{\omega}_k &= \mathbb{P}\{G_i = g_k \mid R_i = w\} \\
&= \sum_{z \in \mathcal{Z}} \mathbb{P}\{G_i = g_k \mid R_i = w, Z_i = z\} \, \mathbb{P}\{Z_i = z \mid R_i = w\} \\
&= \sum_{z \in \mathcal{Z}} \mathbf{p}_{w,z,k} \, \mathbb{P}\{Z_i = z \mid R_i = w\}, \qquad (16)
\end{aligned}$$

where $\mathbb{P}\{Z_i = z \mid R_i = w\}$ is identified from the data. This choice of $\boldsymbol{\omega}$ implies that $\theta = \mathbb{E}[\beta(G_i) \mid R_i = w]$, where θ measures how search rates would change for white drivers if they were treated as minorities.

4.5 Estimation and inference

In this section, I discuss how these methods can be performed on a sample. Statistical inference is based on the Re-Sampling (RS) test of Bugni et al. (2015), who propose a specification test for partially identified models defined by moment inequalities; as well as Bugni et al. (2017), who propose an inference method for subvectors of partially identified parameters defined by moment inequalities.

4.5.1 Testing for bias

To adapt the RS test to my setting, several terms must first be defined. Let $\mathbf{m} \equiv (\mathbf{m}^S_{r,z}, \mathbf{m}^H_{r,z})'_{(r,z) \in \{w,m\} \times \mathcal{Z}}$ denote the vector of search and hit rates for all races and settings. Similarly, let $\mathbf{D}$ denote a diagonal matrix containing $Var[Search_i \mid R_i = r, Z_i = z]$ and $Var[Hit_i \mid R_i = r, Z_i = z]$ for all races and settings. Let $\widehat{\mathbf{m}}$ and $\widehat{\mathbf{D}}$ denote consistent estimates of $\mathbf{m}$ and $\mathbf{D}$. Let $m(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\})$ denote the vector of search and hit rates implied by the model parameters, i.e., the right hand sides of (7)–(8). Finally, define the scaled sample criterion as

$$\widehat{Q}(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) \equiv \sqrt{n} \left\| \widehat{\mathbf{D}}^{-1/2} \left(m(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) - \widehat{\mathbf{m}} \right) \right\|_1,$$

where n is the total number of traffic stops and $\|\cdot\|_1$ denotes the ℓ^1-norm.[17]

To test the null hypothesis that the officer is unbiased, define

$$\widehat{Q}^{\star}_{\text{Unbiased}} \equiv \min_{\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}} \widehat{Q}(\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\})$$
$$\text{s.t.} \quad \boldsymbol{\sigma}_w = \boldsymbol{\sigma}_m, \ (9), \ (10), \ (11),$$

and

$$\widehat{Q}^{\star}_{\text{Biased}} \equiv \min_{\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}} \widehat{Q}(\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\})$$
$$\text{s.t.} \quad (9), \ (10), \ (11).$$

Then the test statistic

$$\widehat{\tau} \equiv \widehat{Q}^{\star}_{\text{Unbiased}} - \widehat{Q}^{\star}_{\text{Biased}} \tag{17}$$

compares the fit of the model when the officer is restricted to be unbiased against the fit without the restriction. A large test statistic suggests the fit of the model is affected by the restriction of unbiasedness and is evidence against the null hypothesis.

To estimate the distribution of $\widehat{\tau}$ under the null hypothesis, I resample the data B times. The data are resampled at the weekly level to account for possible dependencies over time. For each resampled dataset, indexed by $b = 1, \ldots, B$, I calculate (17) and denote its value by $\widehat{\tau}_b$. Define $\widehat{\tau}^{\text{Null}}_b \equiv \widehat{\tau}_b - \widehat{\tau}$. I reject the null hypothesis at the α significance level if $\widehat{\tau}$ exceeds the $1-\alpha$ quantile of $\{\widehat{\tau}^{\text{Null}}_b\}$.

4.5.2 Estimating the intensity of bias

I estimate the bounds on the bias by solving

$$\theta_{\text{lb}}, \theta_{\text{ub}} \equiv \min/\max_{\boldsymbol{\omega},\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}} \boldsymbol{\omega}'(\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w)$$
$$\text{s.t.} \quad \widehat{Q}(\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}) \leq \widehat{Q}^{\star}_{\text{Biased}}, \ (9), \ (10), \ (11).$$

[17] $\widehat{Q}$ is based on the scaled sample criterion proposed by Bugni et al. (2015), which requires a test function. I use a variant of the Modified Method of Moments test function from Andrews and Guggenberger (2009), with the ℓ^1-norm being used instead of the squared Euclidean norm. This test function satisfies the regularity conditions in Bugni et al. (2015) (see Andrews and Soares (2010)). In addition, the matrix $\mathbf{D}$ does not depend on the model parameters, although it can in general. $\mathbf{D}$ does not depend on the model parameters in my setting because the model may defined using moment equalities where the model parameters are separable from the data (see (7)–(8); see Example 6.1 of Bugni et al. (2015) for another example).

I construct the confidence interval for the intensity of bias by inverting the test for bias. That is, rather than test the specification that the officer is unbiased, I test the specification that the intensity of bias is equal to $t \in [-1, 1]$. If the test does not reject the specification at the α significance level, then t enters the $(1 - \alpha)$ confidence interval. See Appendix B for a full description of this procedure.

5 Application

I apply the test to police traffic data from the Metropolitan Nashville Police Department (MNPD). The data contain records of traffic stops for over 2,200 MNPD officers between 2010 and 2019 and is made available by the Stanford Open Policing Project (Pierson et al., 2020).

5.1 Data

Each observation in the data represents a traffic stop made by an officer. I observe the driver's race, age, sex, state of registration, and an anonymized officer identifier. I observe the logistical details of the traffic stop, including the time and geocoordinates of the stop; the reason for the traffic stop; whether a search occurred, and if so, why the search occurred and whether any contraband was found. I categorize the reason for stop into three groups: driving-related reasons, non-driving reasons, and investigative reasons.[18] Reasons for traffic searches include driver consent, probable cause, and plain view of contraband. Although the data categorize contraband into weapons and drugs, I treat all forms of contraband as being the same.

I supplement the traffic data with data provided by the MNPD on criminal incidents and calls for services,[19] as well as local measures of racial composition and median household income from the American Community Survey (ACS). Both the MNPD and ACS supplemental data are at the census tract level, and they allow me to control for environmental factors that potentially correlate with the setting of the stop and the officer's search preferences.

[18] Driving-related reasons correspond to how the driver maneuvers her vehicle and how she interacts with other drivers on the road. They include moving traffic violations, safety violations, and vehicle equipment violations. Non-driving reasons correspond to reasons unrelated to how the vehicle is driven, and include seat belt violations, parking violations, registration violations, and issues with child restraints. Investigative stops are its own category and not an aggregate of other reasons.

[19] I restrict criminal incidents and calls for services to those related to violent crimes, theft, or drugs, as these may affect an officer's decision to search for contraband.

Table 1: Summary of stops, searches, and hits for select 50 officers

	Full sample		Avg. by officer	
	White	Minority	White	Minority
Stops	109,023	113,405	2,180	2,268
Searches	12,622	15,732	252	315
Hits	1,831	2,741	37	55
Search rate	0.1158	0.1387	0.1546	0.1884
Uncon. hit rate	0.0168	0.0242	0.0277	0.0297
Con. hit rate	0.1451	0.1742	0.2431	0.2135

Notes: For each officer, the conditional hit rate can be calculated from the ratio of the unconditional hit rate and search rate.

5.2 Restricting the sample

To study bias in traffic searches, the searches used in the analysis must be discretionary. Traffic searches motivated by rules or mandates are therefore excluded from the study. This includes searches that are incidental to an arrest, inventory searches, and searches based on warrants.[20] In total, 72% of the traffic searches in the data are retained.

I restrict my attention to the 50 officers with the highest number of traffic searches. This is because the methods discussed in Section 4.3 are performed on each officer separately, and in order to reasonably estimate their search and hit rates, I require each of them to have made a large number of traffic stops and searches. On average, these officers have made 2,180 stops and 252 searches for white drivers, and 2,268 stops and 315 searches for minority drivers. Remarkably, this small fraction of officers make up one third of all the searches in the data.

Finally, I focus on comparing the officer's preferences for searching white drivers against that of Black and Hispanic drivers. "Minority" therefore exclusively refers to Black and Hispanic drivers.

Table 1 summarizes the number of traffic stops, searches, and hits in the restricted sample.

[20] Searches incidental to an arrest occur after a driver has been arrested. Inventory searches are required whenever a vehicle is impounded by the police. Warrants to search a driver are typically obtained before the traffic stop, suggesting that warrant-based searches are predetermined and non-discretionary. Hernández-Murillo and Knowles (2004) propose a methodology to incorporate non-discretionary searches into the analysis.

5.3 Context variable and controls

I choose Z_i to be combinations of the day of the week and the patrol shift. I divide the days into weekdays and weekends, and patrol shifts are either in the morning (7 a.m. to 3 p.m.), evening (3 p.m. to 11 p.m.), or night (11 p.m. to 7 a.m.). This generates up to six values of Z_i for each officer. Variation in the search and hit rates across settings strengthens the test by making it more difficult to find a single preference capable of generating the data across all settings for both groups of drivers. To support the independence condition in Assumption 1, I control for variables that may be correlated with both Z_i and officer preferences. These control variables are summarized in Table 2.

The first set of controls consists of observable characteristics of the driver besides race, which includes age, sex, and state of registration. This set of controls accounts for how officers may feel differently towards searching certain demographics who may drive during different times of the day and days of the week. For example, elderly female drivers may drive more often during the morning on weekdays, and officers may be reluctant to search elderly female drivers.

The second set of controls include the details of the traffic encounter, namely the reason for the stop and, if a search took place, the reason for the search.[21] These variables control for how certain aspects of the traffic stop (e.g., being stopped for driving-related reasons) or driver behavior (e.g., having contraband in plain view) might affect an officer's preferences and be correlated with the setting. For example, Makofske (2020) finds that officers in Louisville, Kentucky arrest 40% of drivers stopped for failing to signal, compared to 1% of drivers stopped for any other reason. This suggests that certain stops in Louisville are pretextual and the reason for stopping a driver can indicate an officer's search preference. Although the MNPD data do not show signs of pretextual stops, they show a 10% increase in the proportion of stops being attributed to driving-related reasons across the evening and night shifts,[22] as well as a 50% increase in searches attributed to contraband being in plain view across the same pair of shifts. Controlling for these features of the traffic stop reduces the concern that the test is detecting differences along these dimensions rather than detecting racial bias.

The final set of controls relates to the environment where the stop takes place. This includes whether the stop was made on a street or a highway; which police precinct the stop was made in; the racial composition, household income, and crime rate of the census

[21] Durlauf and Heckman (2020) raise concerns about the credibility of self-reported police data. While the concern is valid, there is currently not a good solution.

[22] In a study on endogenous driving behavior, Kalinowski et al. (2021) find that minority drivers adjust their driving behavior during the day, when their race is more visible to the officer.

Table 2: Summary of control variables

	Drivers stopped		Drivers searched	
	White	Minority	White	Minority
Driver characteristics				
Male	0.6032	0.6007	0.6613	0.7722
Age	37.28	34.64	32.31	30.49
Out of state	0.0638	0.0330	0.0490	0.0340
Reason for stop				
Driving	0.8803	0.8776	0.8668	0.8687
Non-driving	0.1070	0.1065	0.1072	0.1031
Investigation	0.0127	0.0159	0.0260	0.0282
Reason for search				
Plain view			0.4978	0.2606
Consent			0.4336	0.5938
Probable Cause			0.0686	0.1456
Location				
Highway	0.1228	0.0644	0.0759	0.0495
Precinct 1	0.0763	0.0509	0.0640	0.0521
Precinct 2	0.1190	0.1760	0.0882	0.1920
Precinct 3	0.1042	0.1446	0.0913	0.1377
Precinct 4	0.0395	0.0249	0.0789	0.0381
Precinct 5	0.3618	0.2567	0.2573	0.2227
Precinct 6	0.0400	0.1100	0.0257	0.0774
Precinct 7	0.1366	0.1528	0.1469	0.1540
Precinct 8	0.1225	0.0842	0.2477	0.1260
Census tract demographics				
Percent white	0.5901	0.4523	0.6028	0.4580
Median household income	49038	41170	48642	40029
Crime incident rate	0.0256	0.0369	0.0305	0.0400
Calls for MNPD services	0.0207	0.0216	0.0212	0.0227

Notes: Crime and call rates are per capita and are restricted to those pertaining to violent crimes, theft, or drugs.

tract; and the frequency of calls for MNPD services originating from the census tract. This accounts for the possible correlation between an officer's search preferences and Z_i induced by his surroundings. Such a correlation may arise because an officer is more hesitant to conduct a search when in a low-income or high-crime neighborhood, where he may find himself more often during the night.[23]

Some potential concerns may be that officers are not randomly assigned to shifts, or there are ticket quotas, or officers are instructed to search more aggressively during certain times. Each issue may be seen as a threat to Assumption 1. Regarding endogenous selection of shifts, Assumption 1 will hold as long as the officer is equally willing to search drivers during each shift. Regarding the ticket quotas, Tennessee has explicit laws banning quotas on traffic citations, although this has not stopped departments from implementing such quotas.[24] Nevertheless, ticket quotas target the stop decision of officers. As long as ticket quotas do not affect search preferences, then the quotas only impact the search decisions through changing the distribution of risk via sample selection. Finally, regarding the concern that officers are instructed to search more aggressively during different shifts, there were no such policies during the time frame of the data I analyze. To the best of my knowledge, such policies were only implemented beginning in July of 2019.[25]

5.4 Setting up the BP problem

I discretize the support of risk to be

$$\mathbf{g} = \{\underbrace{0, 0.025, 0.05, 0.075}_{\text{Increments of 0.025}}, \underbrace{0.1, 0.15, 0.20, 0.25}_{\text{Increments of 0.05}}, \underbrace{0.3, 0.4, 0.5, 0.6}_{\text{Increments of 0.1}}, 0.75, 1\}.$$

[23] Roh and Robinson (2009) find there to be spatial correlation in traffic search decisions even after controlling for driver characteristics. The authors attribute the correlation to similarities in environmental variables, such as the racial composition of the neighborhood and the volume of police allocated nearby. Novak and Chamlin (2012) also find that the police workload (measured via calls for services) and degree of 'social disorganization' (e.g., percentage of single parent households, percentage of residents in poverty) are predictive of officer behavior.

[24] For example, the mayor of Ridgetop, TN tried to have the city's police department enforce a ticket quota to raise money for the city, only to be turned in by the city's police chief (Ferrier, 2019). See Tennessee Code §39-16-516 (2014) for the law banning ticket quotas.

[25] In July of 2019, the MNPD introduced the Entertainment District Initiative, which assigned 17 additional officers to the Entertainment District on Fridays and Saturdays between 6 p.m. and 4 a.m. to improve public safety. These officers performed high-visibility patrols on foot, bike, and utility task vehicles, and would make unannounced visits to local establishments. In February of 2021, the MNPD introduced the Office of Alternative Policing Strategies to address an increase in violent crime in Nashville. A new shift of 80 officers working between 5:30 p.m. and 3:30 a.m. was added across all precincts to perform high-visibility patrols to deter and detect violent crimes. See Aaron et al. (2019), Rau (2021), and McDonald (2021).

Table 3: Search and conditional hit rates by Z_i

		White		Minority	
Day	Shift	Search	Cond. Hit	Search	Cond. Hit
Weekday	Morning	0.0376	0.2617	0.0603	0.2265
Weekday	Evening	0.1268	0.1774	0.1528	0.1826
Weekday	Night	0.2711	0.1080	0.2381	0.1645
Weekend	Morning	0.0372	0.2656	0.1091	0.1272
Weekend	Evening	0.1349	0.1044	0.1259	0.1597
Weekend	Night	0.2753	0.0562	0.2334	0.1064
Mean		0.1158	0.1958	0.1387	0.1868

Notes: Search and conditional hit rates account for the control variables. The mean rates for the observed data are calculated by weighting each setting by the proportion of stops in the data made in each setting, and taking a weighted average of the rates across the settings.

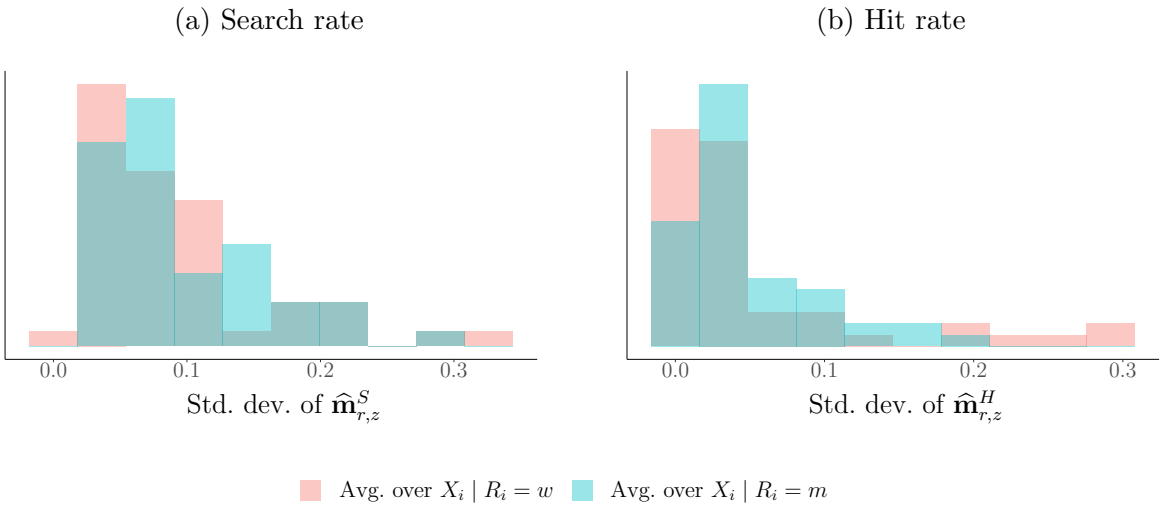
I choose $\mathbf{g}$ to be finer at lower levels of risk since Table 1 shows that the average conditional hit rates are between 21% and 24%, which suggests most drivers searched are relatively low-risk. Table 3 presents the conditional hit rates for each setting after accounting for controls. The average conditional hit rates remain low, ranging from 5% to 27%. The model also implies that drivers who are searched represent the riskiest subset of drivers who are stopped. In conjunction with the low conditional hit rates, this further suggests that most drivers stopped are low-risk. I incorporate this into the model by imposing the monotonicity restriction in (13), requiring that $\mathbf{p}_{r,z}$ is decreasing as risk increases for all $(r, z) \in \{w, m\} \times \mathcal{Z}$.

I do not impose any restrictions on $\boldsymbol{\sigma}$ except that it is non-decreasing in risk (as implied by Assumption 1) and lies in the unit interval.

The sample moments $\widehat{\mathbf{m}}^S_{r,z}$, $\widehat{\mathbf{m}}^H_{r,z}$ are obtained using the predicted probabilities from logistic regressions. Since traffic searches and hits can be rare events for some officers, I use Firth's logistic regression with intercept-correction to obtain unbiased estimates of the search and hit rates (Puhr et al., 2017).[26] To construct $\widehat{\mathbf{m}}^S_{r,z}$, I first regress $Search_i$ on setting Z_i and controls X_i conditional on race $R_i = r$. This provides an estimate of $\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z, X_i = x\}$. I then set $\widehat{\mathbf{m}}^S_{r,z}$ equal to the predicted probabilities

[26] Firth's logistic regression reduces the bias in coefficient estimates in small samples. However, it biases predicted probabilities towards 0.5. In a simulation study, Puhr et al. (2017) show that the bias in the predicted probabilities can be corrected by adjusting the intercept term. This adjustment also debiases predicted probabilities for rare events, and outperforms other methods seeking to debias logistic regressions in rare events data, including King and Zeng (2001).

Figure 5: Variation in search and hit rates across Z_i



Note: The left (right) panel shows the distribution of the standard deviation of search (hit) rates across R_i and Z_i. The standard deviation of the search (hit) rates across R_i and Z_i is calculated for each officer, and the histograms show the distribution of those standard deviations. The histograms in red (blue) correspond to the case where $\widehat{\mathbf{m}}^S_{r,z}$ and $\widehat{\mathbf{m}}^H_{r,z}$ are obtained by averaging the fitted search and hit rates from logistic regressions over the distribution of $X_i \mid R_i = w$ ($X_i \mid R_i = m$).

averaged over the sample distribution of X_i for either race of drivers, i.e.,

$$\widehat{\mathbf{m}}^S_{r,z} = \widehat{\mathbb{E}}\left[\widehat{\mathbb{P}}\{Search_i \mid R_i = r, Z_i = z, X_i\} \;\middle|\; R_i = r'\right] \text{ for } r' \in \{w, m\}. \tag{18}$$

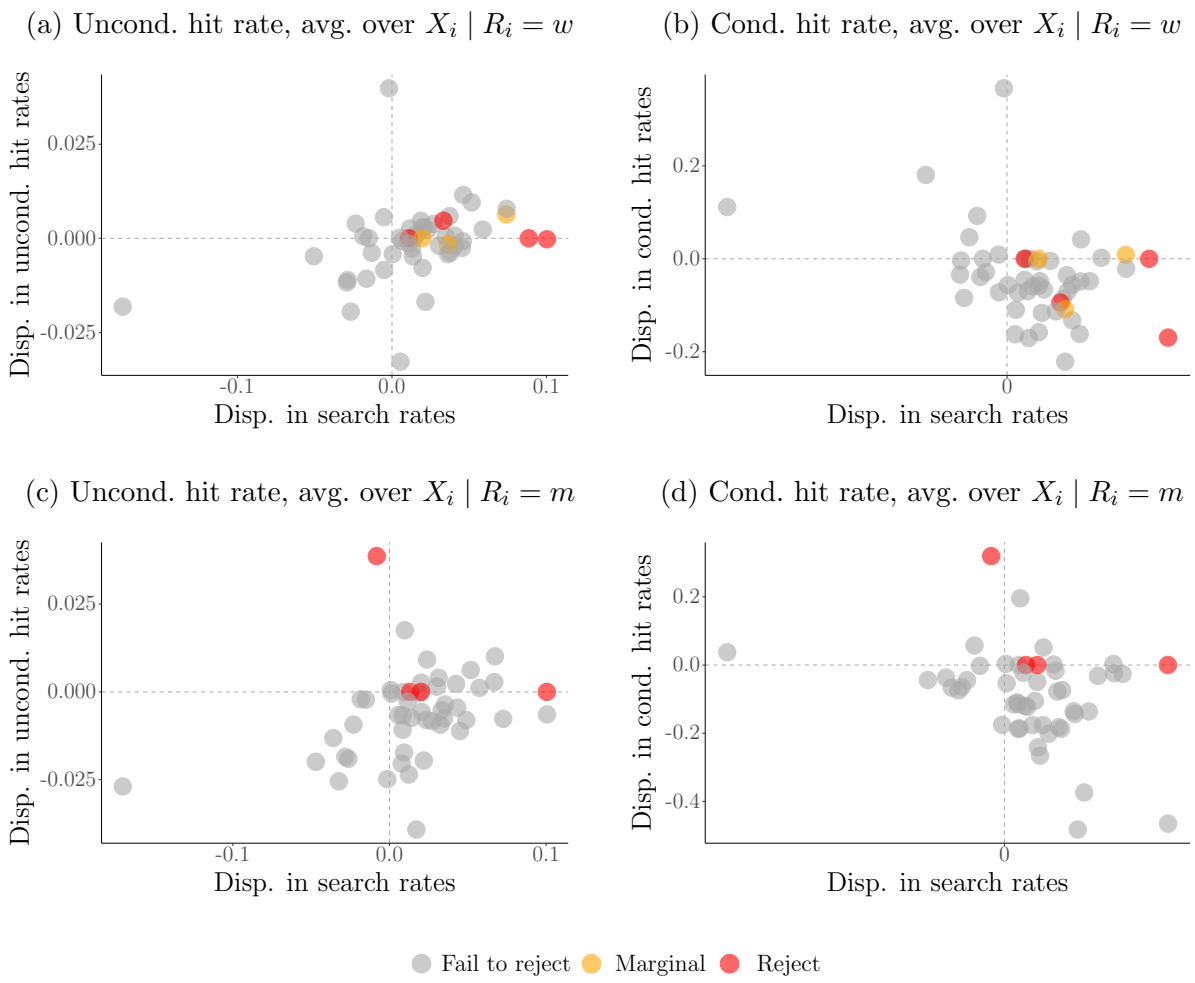
In Section 5.5, I present results for both $r' = w$ and $r' = m$. This approach allows me to control for X_i such that the estimates are representative of each race.

The hit rates $\widehat{\mathbf{m}}^H_{r,z}$ are estimated as in (18), except I regress Hit_i on Z_i and X_i conditional on each race.

Figure 5 summarizes the variation in search and hit rates generated by Z_i within officer. The figure is obtained by calculating the standard deviations of $\{\widehat{\mathbf{m}}^S_{r,z}\}$ and $\{\widehat{\mathbf{m}}^H_{r,z}\}$ across R_i and Z_i for each officer. The histograms of these standard deviations are presented in Figure 5. If the officer is biased, then greater variation in search and hit rates increases the power of the test by making it more difficult to find a single preference generating the data for both groups of drivers. Figure 5 shows that Z_i generates greater variation in search rates compared to hit rates, suggesting that the power of the test in my sample stems primarily from the variation in search rates.[27]

[27]There are three officers with no variation in hit rates as they have never found contraband despite having searched many drivers. For these officers, the criterion and test exclusively depend on (7). To see why, note that if $\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} > 0$ but $\mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} = 0$ for some r and

Figure 6: Racial disparities in search and hit rates by officer



Note: Each point corresponds to an individual officer. Search and hit rates of each officer are averaged across the different settings, controlling for observed characteristics of the driver. Positive disparities indicate that minority drivers have higher rates compared to white drivers. Red points indicate officers for whom the null hypothesis of being unbiased is rejected at the 5% significance level. Orange points indicate officers for whom the null hypothesis is close to being rejected (rejection at the 6% significance level). Grey points indicate the remaining officers for whom the null hypothesis is not rejected.

5.5 Results

When averaging the search and hit rates over $X_i \mid R_i = w$, I reject the null hypothesis that the officer is unbiased at the 5% significance level for four of the 50 officers. In addition, three officers are at the margin of failing the test, i.e., the null hypothesis may only be rejected for them at the 6% significance level.

When averaging the search and hit rates over $X_i \mid R_i = m$, I again reject the null hypothesis for four officers. Compared to the previous case where search and hit rates were averaged over $X_i \mid R_i = w$, two of these officers also failed the test, and one of these officers was at the margin of failing.

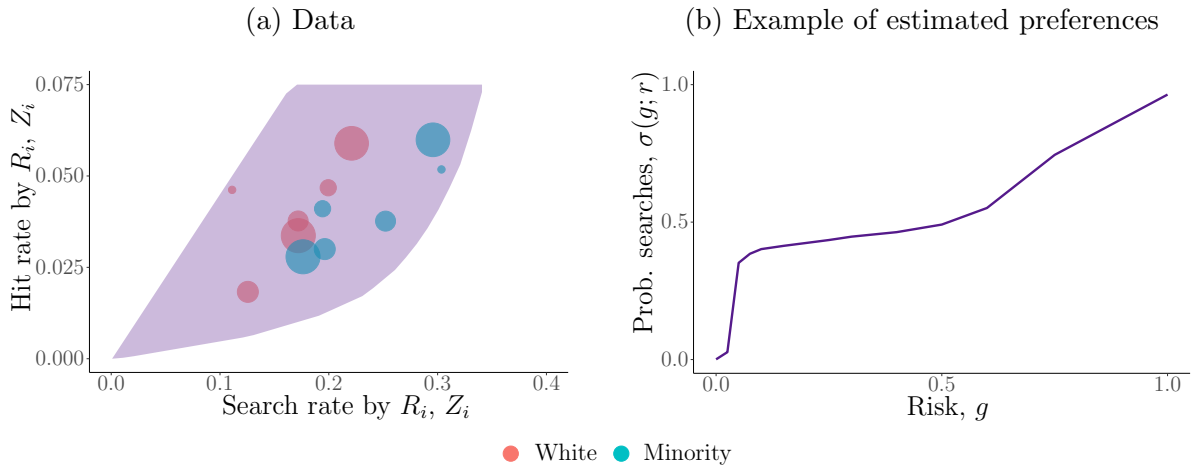
The change in the set of officers who fail the test suggests that bias may depend on observable characteristics of the driver and traffic stop, X_i. Table 2 compares the distribution of X_i for white and minority drivers, and shows that minority drivers are on average younger, and are stopped in different precincts in areas with higher crime rates, lower income, and lower proportion of white residents. Some of these differences in X_i may generate or remove biases. Note, however, that these differences in X_i are balanced across both groups of drivers when testing for bias, so the test is not conflating differences in X_i across race with differences in search preferences across race.

Figure 6 shows the relationship between the racial disparities in search and hit rates, and whether an officer is flagged as being racially biased. Positive disparities in search (hit) rates indicate that minority drivers have higher search (hit) rates compared to whites. The top two panels correspond to the case where search and hit rates are averaged over the distribution of $X_i \mid R_i = w$, and the bottom two panels correspond to the case where the rates are averaged over the distribution of $X_i \mid R_i = m$. Not surprisingly, officers with large racial disparities in search or conditional hit rates are flagged as racially biased. However, the test is also able to detect bias when such disparities are small. This suggests the proposed methodology is able to pick up subtleties in the data that may escape earlier tests only comparing search and hit rates across groups of drivers.

Figure 7 presents the data for an officer who passes the test, i.e., bias is not detected. The circles in the left panel represent the search and hit rates by race and setting, and the size of the circles indicate the number of stops associated with the setting. The purple region shows the set of data points that are consistent with the preference in the right panel. Since all the data points lie inside the purple region, it is possible for the observed data for white and minority drivers to be generated by a common preference. Applying the test from Section 4, I cannot reject the null hypothesis that the officer is unbiased at the 5% significant level.

z, then it must that $\mathbb{P}\{G_i = 0 \mid R_i = r, Z_i = z\} = 1$. This distribution of risk ensures the hit rate is 0. The test thus only involves finding preferences $\{\boldsymbol{\sigma}_r\}$ to match the search rates $\{\mathbf{m}^S_{r,z}\}$.

Figure 7: Example where bias is not detected

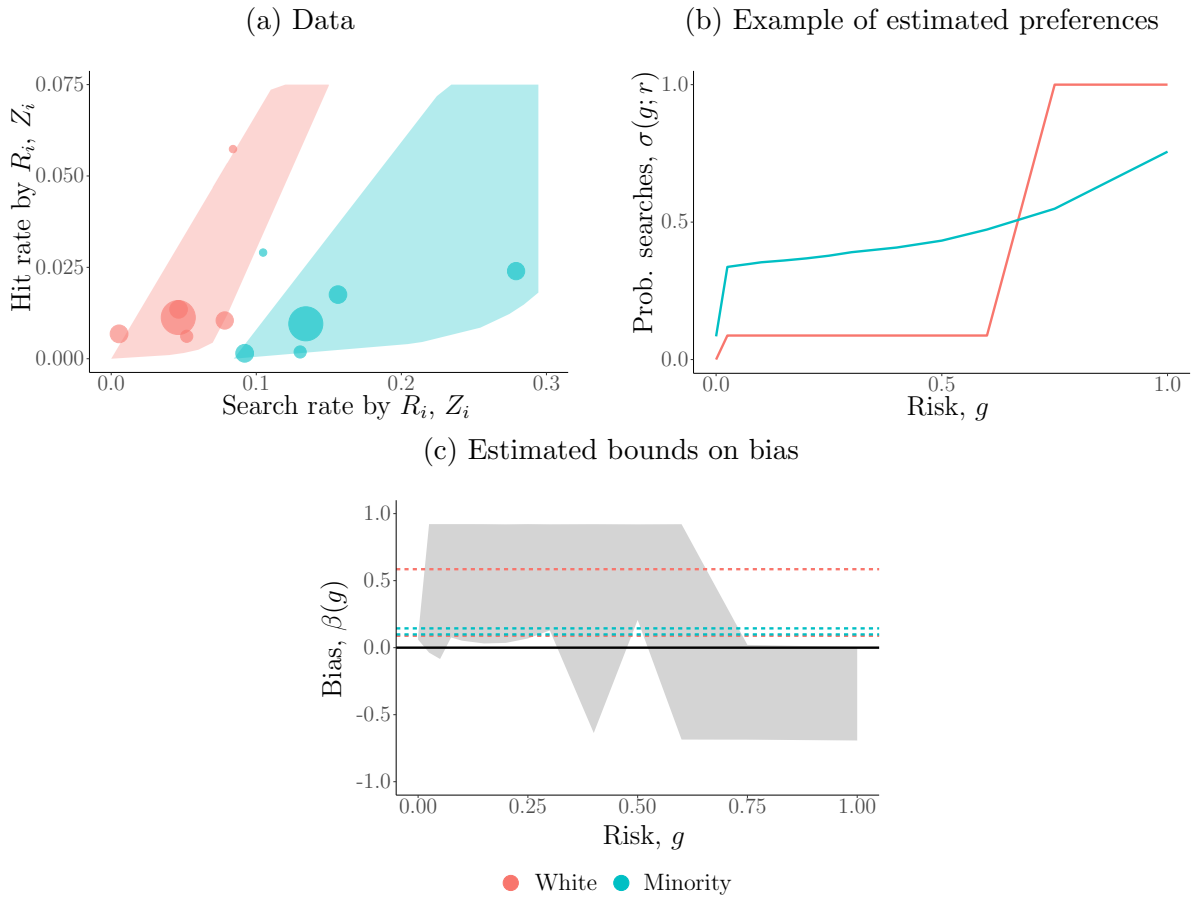


Note: Each dot in the left panel corresponds to the search and hit rates for a particular race and setting. The size of the dots represents the number of stops the data are associated with. Search and hit rates are averaged over the distribution of $X_i \mid R_i = w$. The purple polygon in the left panel represents the data that can be generated by the preference shown in the right panel.

Figure 8 presents the data for an officer who fails the test, i.e., bias is detected. The top right panel presents one possible set of estimates of the officer's search preferences. The bottom panel presents the estimated bounds on the bias, with the gray band showing the bounds conditional on risk, and the dashed lines showing the bounds on the average bias. The red (blue) dashed lines indicate the average bias when the distribution of risk is consistent with that of white (minority) drivers in the data. The estimated bounds on the bias conditional on risk suggest that this officer searches minority drivers more than equally risky white drivers when risk falls below 0.3. However, as risk increases, the bounds on bias decrease and become non-positive, suggesting the officer changes direction of bias when the risk becomes sufficiently large. This change in the direction of bias is also seen in the top right panel, where the two curves intersect. On average, minority drivers are estimated to be searched at least 8.8 percentage points less if they were treated as white drivers, holding their distribution of risk constant. In contrast, white drivers are estimated to be searched at least 9.9 percentage points more on average if they were treated as minority drivers, holding their distribution of risk constant. These estimates are large in magnitude considering this officer searches 4.5% of white drivers and 14.6% of minority drivers.

Figure 9 presents the data for another officer who fails the test. This officer searches 4.0% of white drivers and 12.8% of minority drivers, and his search rates are relatively stable across settings. However, this officer has never found contraband on either group of drivers, suggesting the officer only interacts with with zero-risk drivers. The racial disparity in search

Figure 8: Example where bias is detected



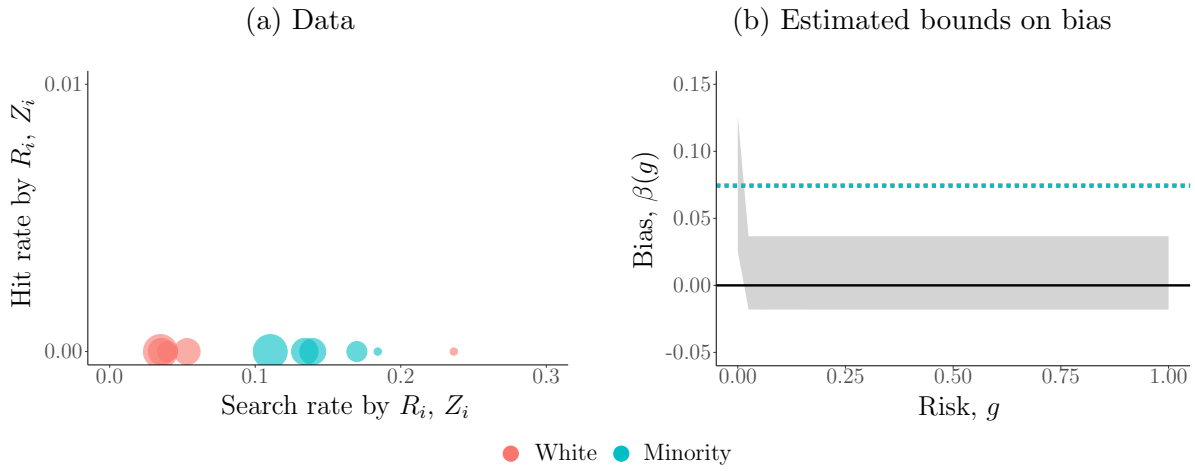
Note: Search and hit rates are averaged over the distribution of $X_i \mid R_i = w$. The red (blue) dashed lines in the bottom panel indicate the bounds on the average bias when the distribution of risk is consistent with that of white (minority) drivers.

rates is therefore not justified and bias is detected. Moreover, if the officer is searching only a fraction of zero-risk drivers from each race, then the officer is effectively searching at random. Feigenberg and Miller (2022) also find evidence to suggest officers search at random. Such behavior is consistent with the model I propose in (1) and necessitates a random threshold. In contrast, random searches contradict earlier models with fixed thresholds, which imply officers search all drivers with a given level of risk or none at all.

See Online Appendix C for the estimated bias for all the officers who fail the test.

Figure 10 presents the estimated bounds on the average bias for the officers who fail the test. The left panel corresponds to the estimates where the search and hit rates are averaged over $X_i \mid R_i = w$, and the right panel corresponds to the estimates where the search and hit rates are averaged over $X \mid R_i = m$. The red bounds correspond to the bias being averaged over the distribution of risk of white drivers and indicate how much more white drivers

Figure 9: Example of random thresholds



Note: Search and hit rates are averaged over the distribution of $X_i \mid R_i = w$. If white drivers were treated as minority drivers, holding their risk constant, they would be searched approximately 7.5 percentage points more on average. If minority drivers were treated as white drivers, holding their risk constant, they would be searched between 7.4 and 7.5 percentage points less on average.

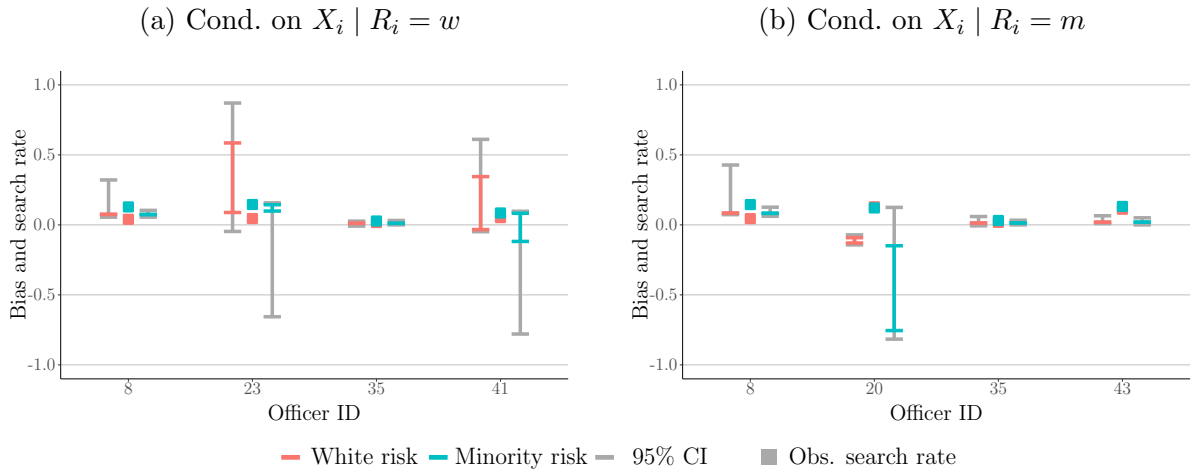
would be searched if they were treated as minorities. The blue bounds correspond to the bias being averaged over the distribution of risk for minority drivers and indicate how much less minority drivers would be searched if they were treated as whites. The gray bounds correspond to the 95% confidence interval. The squares indicate the search rate observed in the data.

When averaging search and hit rates over $X_i \mid R_i = w$, the estimates suggest white drivers would be searched at least 2.4 percentage points more on average (83.1% more relative to the observed search rate of 2.9%) by these biased officers if the drivers were treated as minorities, holding their risk constant. In contrast, minority drivers would be searched at least 1.2 percentage points less on average (17.3% less relative to the observed search rate of 6.8%) if they were treated as whites, holding their risk constant.[28]

When averaging search and hit rates over $X_i \mid R_i = m$, the estimates suggest white drivers would be searched at least 1.2 percentage points more on average (22.3% more relative to the observed search rate of 5.2%) by these biased officers if the drivers were treated as minorities, holding their risk constant. Minority drivers would be searched at most 8.7 percentage points more on average (99% more relative to the observed search rate of 8.8%) if they were treated as whites, holding their risk constant. The possible increase in search rates is because the estimated bounds on the average bias for one officer are $[-0.755, -0.149]$, suggesting he is

[28]These estimates are obtained by taking a weighted average of the red or blue lower bounds, with the weights being equal to the proportion of stops (conditional on race) made by each officer.

Figure 10: Bounds on average bias $\mathbb{E}[\beta(G_i)]$ for biased officers



Note: The left (right) panel shows the estimated bounds when search and hit rates are averaged over the distribution of $X_i \mid R_i = w$ ($X_i \mid R_i = m$). Positive average bias indicates minority drivers are searched more often than equally risky white drivers on average. Red (blue) bounds indicate the bias averaged over the distribution of risk for white (minority) drivers in the data. Gray bounds indicate the 95% confidence interval. The colored squares indicate the search rates observed in the data for each officer.

biased against white drivers on average.[29]

6 Conclusion

In this paper, I provide a flexible approach to detecting and measuring racial bias in police traffic searches. The partial identification framework enables the test to be applied even amid sample selection on unobservables and statistical discrimination. In addition, by using an IV to vary the risk among drivers stopped, the methods I propose may be applied to individual officers, allowing for unrestricted heterogeneity in preferences and beliefs across officers.

This paper also contributes to the literature from a modeling standpoint, as earlier papers studying racial bias have either assumed or required choice models with deterministic thresholds, whereas I allow the threshold to be random. This relaxation permits a richer notion of bias, where the direction and intensity of bias may depend on the unobserved (to the researcher) risk of the driver. Sharp bounds on these measures immediately follow from the econometric model. Additional restrictions to tighten these bounds or strengthen the test may be imposed in a transparent and modular fashion. Implementing these methods

[29] This officer accounts for 15.7% of searches among the biased officers.

involves solving bilinear programs, which are novel in the literature on discrimination and econometrics in general.

I apply the proposed methods on police traffic data from the Metropolitan Nashville Police Department and find evidence to suggest 6 of the 50 officers analyzed are biased. For each of these officers, I am able to estimate the fraction of searches stemming from bias. The estimates suggest that the presence and intensity of bias for some officers vary with the observable characteristics and unobserved risk of the driver.

A natural extension of the paper is to apply these methods to other data sets. These methods can be applied to standard police traffic data, and the assumptions of the model can be supported by incorporating local demographic data that are typically public or available upon request, such as household incomes and crime rates. These methods can also be applied to study discrimination in different settings along different dimensions, such as testing for and measuring racial bias in healthcare or gender bias in labor markets.

References

Aaron, D., K. Mumford, and B. Reese (2019). New Initiative to Further Enhance Public Safety in Nashveill's Entertainment District. *Metropolitan Nashville Police Department Media Release*. https://www.nashville.gov/departments/police/news/new-initiative-further-enhance-public-safety-nashvilles-entertainment (accessed 6/19/2023).

Agan, A. and S. Starr (2018). Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. *The Quarterly Journal of Economics 133*(1), 191–235.

Aigner, D. J. and G. G. Cain (1977). Statistical Theories of Discrimination in Labor Markets. *Indutrial and Labor Relations Review 30*(2), 175–187.

Andrews, D. W. and P. Guggenberger (2009). Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities. *Econometric Theory 25*(3), 669–709.

Andrews, D. W. and G. Soares (2010). Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica 78*(1), 119–157.

Anwar, S. and H. Fang (2006). An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence. *American Economic Review 96*(1), 127–151.

Arnold, D., W. Dobbie, and P. Hull (2022). Measuring Racial Discrimination in Bail Decisions. *American Economic Review 112*(9), 2992–3038.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial Bias in Bail Decisions. *The Quarterly Journal of Economics 133*(4), 1885–1932.

Ba, B. A., D. Knox, J. Mummolo, and R. Rivera (2021). The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago. *Science 371*(6530), 696–702.

Barnes, K. Y. (2004). Assessing the Counterfactual: The Efficacy of Drug Interdiction Absent Racial Profiling. *Duke Law Journal 54*, 1089.

Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2022). Consumer-lending Discrimination in the FinTech Era. *Journal of Financial Economics 143*(1), 30–56.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Becker, G. S. (1993). Nobel lecture: The Economic Way of Looking at Behavior. *Journal of Political Economy 101*(3), 385–409.

Bhutta, N. and A. Hizmo (2021). Do Minorities Pay More for Mortgages? *The Review of Financial Studies 34*(2), 763–789.

Blinder, A. S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources*, 436–455.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2022). Inaccurate Statistical Discrimination: An Identification Problem. *Working Paper*.

Bohren, J. A., A. Imas, and M. Rosenberg (2019). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review 109*(10), 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Bugni, F. A., I. A. Canay, and X. Shi (2015). Specification Tests for Partially Identified Models Defined by Moment Inequalities. *Journal of Econometrics 185*(1), 259–282.

Bugni, F. A., I. A. Canay, and X. Shi (2017). Inference for Subvectors and Other Functions of Partially Identified Parameters in Moment Inequality Models. *Quantitative Economics 8*(1), 1–38.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020a). On the Use of Outcome Tests for Detecting Bias in Decision Making. *National Bureau of Economic Research Working Paper No. w28789*.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020b). Reply to the Comment of Arnold, Dobbie, Yang (2020).

Canay, I. A., M. Mogstad, and J. Mountjoy (2022). On the Use of Outcome Tests for Detecting Bias in Decision Making. *National Bureau of Economic Research Working Paper No. w28789*.

Canay, I. A., A. Santos, and A. M. Shaikh (2013). On the Testability of Identification in Some Nonparametric Models with Endogeneity. *Econometrica 81*(6), 2535–2559.

Card, D., A. R. Cardoso, and P. Kline (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly journal of economics 131*(2), 633–686.

Chan, D. C., M. Gentzkow, and C. Yu (2022). Selection with Variation in Diagnostic Skill: Evidence from Radiologists. *The Quarterly Journal of Economics 137*(2), 729–783.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica: Journal of the Econometric Society*, 1001–1044.

Ductor, L., S. Goyal, and A. Prummer (2021). Gender and Collaboration. *The Review of Economics and Statistics*, 1–40.

Durlauf, S. N. and J. J. Heckman (2020). An Empirical Analysis of Racial Differences in Police Use of Force: A Comment. *Journal of Political Economy 128*(10), 3998–4002.

Engel, R. S. and R. Johnson (2006). Toward a Better Understanding of Racial and Ethnic Disparities in Search and Seizure Rates. *Journal of Criminal Justice 34*(6), 605–617.

Feigenberg, B. and C. Miller (2022). Would Eliminating Racial Disparities in Motor Vehicle Searches Have Efficiency Costs? *The Quarterly Journal of Economics 137*(1), 49–113.

Ferrier, D. (2019). Ferrier Files: Ridgetop Disbands Police Department After Illegal Ticket Quotas Exposed. *Fox 17 WZTV*. https://fox17.com/news/local/ferrier-files-ridgetop-disbands-police-department-after-illegal-ticket-quotas-exposed (accessed 6/19/2023).

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition Methods in Economics. In *Handbook of Labor Economics*, Volume 4, pp. 1–102. Elsevier.

Frandsen, B., L. Lefgren, and E. Leslie (2023). Judging Judge Fixed Effects. *American Economic Review 113*(1), 253–77.

Fryer Jr, R. G. (2019). An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy 127*(3), 1210–1261.

Gaebler, J., W. Cai, G. Basse, R. Shroff, S. Goel, and J. Hill (2020). Deconstructing Claims of Post-treatment Bias in Observational Studies of Discrimination. *arXiv preprint arXiv:2006.12460*.

Gelbach, J. B. (2021). Testing Economic Models of Discrimination in Criminal Justice. *Social Science Research Network No. 3784953*.

Gelman, A., J. Fagan, and A. Kiss (2007). An Analysis of the New York City Police Department's "Stop-and-frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association 102*(479), 813–823.

Goel, S., J. M. Rao, and R. Shroff (2016a). Personalized Risk Assessments in the Criminal Justice System. *American Economic Review 106*(5), 119–23.

Goel, S., J. M. Rao, and R. Shroff (2016b). Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *The Annals of Applied Statistics 10*(1), 365–394.

Goncalves, F. and S. Mello (2021). A Few Bad Apples? Racial Bias in Policing. *American Economic Review 111*(5), 1406–1441.

Grogger, J. and G. Ridgeway (2006). Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness. *Journal of the American Statistical Association 101*(475), 878–887.

Gurobi Optimization, Inc. (2021). Gurobi Optimizer Reference Manual.

Heckman, J. J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica 73*(3), 669–738.

Hernández-Murillo, R. and J. Knowles (2004). Racial Profiling or Racist Policing? Bounds Tests in Aggregate Data. *International Economic Review 45*(3), 959–989.

Hull, P. (2021). What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making. *National Bureau of Economic Research Working Paper No. w28503*.

Kalinowski, J., M. B. Ross, and S. L. Ross (2021). Endogenous Driving Behavior in Tests of Racial Profiling. *National Bureau of Economic Research Working Paper No. w28789*.

King, G. and L. Zeng (2001). Logistic Regression in Rare Events Data. *Political analysis 9*(2), 137–163.

Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic Discrimination Among Large US Employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Knowles, J., N. Persico, and P. Todd (2001). Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy 109*(1), 203–229.

Knox, D., W. Lowe, and J. Mummolo (2020a). Administrative Records mask Racially Biased Policing. *American Political Science Review 114*(3), 619–637.

Knox, D., W. Lowe, and J. Mummolo (2020b). Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection? *Available at SSRN 3940802*.

MacDonald, J. M. and J. Fagan (2019). Using Shifts in Deployment and Operations to Test for Racial Bias in Police Stops. In *AEA Papers and Proceedings*, Volume 109, pp. 148–51.

Makofske, M. (2020). Pretextual Traffic Stops and Racial Disparities in their Use. *Working Paper*.

Marx, P. (2022). An Absolute Test of Racial Prejudice. *The Journal of Law, Economics, and Organization 38*(1), 42–91.

McCormick, G. P. (1976). Computability of Global Solutions to Factorable Nonconvex Programs: Part I—Convex Underestimating Problems. *Mathematical programming 10*(1), 147–175.

McDonald, H. (2021). Special Report: MNPD Puts 60 Extra Officers in Nashville's Entertainment District on Weekends. Why? *News Channel 5 Nashville*. https://www.newschannel5.com/news/special-report-mnpd-puts-60-extra-officers-in-nashvilles-entertainment-district-on-weekends-why (accessed 6/19/2023).

Mehlhorn, K., P. Sanders, and P. Sanders (2008). *Algorithms and Data Structures: The Basic Toolbox*, Volume 55. Springer.

Morin, R., K. Parker, R. Stepler, and A. Mercer (2017). Behind the Badge. *Pew Research Center*.

Novak, K. J. and M. B. Chamlin (2012). Racial Threat, Suspicion, and Police Behavior: The Impact of Race and Place in Traffic Enforcement. *Crime & Delinquency 58*(2), 275–300.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 693–709.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science 366*(6464), 447–453.

Onuchic, P. and D. Ray (2023). Signaling and discrimination in collaborative projects. *American Economic Review 113*(1), 210–52.

Pierson, E., S. Corbett-Davies, and S. Goel (2018). Fast Threshold Tests for Detecting Discrimination. In *International Conference on Artificial Intelligence and Statistics*, pp. 96–105.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, et al. (2020). A Large-scale Analysis of Racial Disparities in Police Stops Across the United States. *Nature Human Behaviour 4*(7), 736–745.

Puhr, R., G. Heinze, M. Nold, L. Lusa, and A. Geroldinger (2017). Firth's Logistic Regression with Rare Events: Accurate Effect Estimates and Predictions? *Statistics in medicine 36*(14), 2302–2317.

Rau, N. (2021). In Nashville, Mayor Cooper, Chief Drake Announce Policing Reforms to Address Murders, Gun Crimes. *Tennessee Lookout*. https://tennesseelookout.com/2021/02/01/in-nashville-mayor-cooper-chief-drake-announce-policing-reforms-to-address-murders-gun-crimes/ (accessed 6/19/2023).

Ridgeway, G. (2006). Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores. *Journal of Quantitative Criminology 22*(1), 1–29.

Ridgeway, G. and J. M. MacDonald (2009). Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops. *Journal of the American Statistical Association 104*(486), 661–668.

Roh, S. and M. Robinson (2009). A Geographic Approach to Racial Profiling: The Microanalysis and Macroanalysis of Racial Disparity in Traffic Stops. *Police Quarterly 12*(2), 137–169.

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021). Gender Differences in Recognition for Group Work. *Journal of Political Economy 129*(1), 101–147.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017). The Problem of Infra-Marginality in Outcome Tests for Discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Wasserman, M. (2023). Hours Constraints, Occupational Choice, and Gender: Evidence from Medical Residents. *The Review of Economic Studies 90*(3), 1535–1568.

Appendix

A Proofs

A.1 Deriving the random threshold in (1)

The officer wishes to maximize his expected utility. As shown in the main paper, the expected utility for decision $Search_i = s$ is

$$\begin{aligned}
&\mathbb{E}[\mathcal{U}_i^s(Guilty_i; R_i) \mid R_i = r, Z_i = z, V_i = v] \\
&= G(r,z,v)\, \mathcal{U}_i^s(1; R_i) \;\; + (1 - G(r,z,v))\, \mathcal{U}_i^s(0; R_i) \\
&= \mathcal{U}_i^s(0; R_i) + G(r,z,v)\, (\mathcal{U}_i^s(1; R_i) - \mathcal{U}_i^s(0; R_i))
\end{aligned}$$

So the officer chooses to search the driver if the expected utility from searching is at least as great as that of not searching, which is equivalent to

$$\begin{aligned}
&\mathbb{E}[\mathcal{U}_i^1(Guilty_i; R_i) \mid R_i = r, Z_i = z, V_i = v] \geq \mathbb{E}[\mathcal{U}_i^0(Guilty_i; R_i) \mid R_i = r, Z_i = z, V_i = v] \\
\iff &\mathcal{U}_i^1(0; R_i) + G(r,z,v)\, \big(\mathcal{U}_i^1(1; R_i) - \mathcal{U}_i^1(0; R_i)\big) \geq \mathcal{U}_i^0(0; R_i) + G(r,z,v)\, \big(\mathcal{U}_i^0(1; R_i) - \mathcal{U}_i^0(0; R_i)\big) \\
\iff &G(r,z,v) \begin{bmatrix} \big(\mathcal{U}_i^1(1; R_i) - \mathcal{U}_i^1(0; R_i)\big) \\ -\big(\mathcal{U}_i^0(1; R_i) - \mathcal{U}_i^0(0; R_i)\big) \end{bmatrix} \geq \mathcal{U}_i^0(0; R_i) - \mathcal{U}_i^1(0; R_i) \\
\iff &G(r,z,v) \geq \underbrace{\frac{\mathcal{U}_i^0(0; R_i) - \mathcal{U}_i^1(0; R_i)}{[\mathcal{U}_i^1(1; R_i) - \mathcal{U}_i^1(0; R_i)] - [\mathcal{U}_i^0(1; R_i) - \mathcal{U}_i^0(0; R_i)]}}_{\text{Random utility threshold } T_i}.
\end{aligned}$$

The final line follows from Assumption 1(i), which ensures the denominator in the expression for T_i is strictly positive.

A.2 Deriving the search and hit rates

The search rate is derived as follows.

$$
\begin{aligned}
&\mathbb{E}[Search_i \mid R_i = r, Z_i = z] \\
&= \mathbb{E}[\mathbb{E}[Search_i \mid R_i = r, Z_i = z, V_i] \mid R_i = r, Z_i = z] && \text{(A.1)} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}\{G(R_i, Z_i, V_i) \geq T_i\} \mid R_i = r, Z_i = z, V_i] \mid R_i = r, Z_i = z] && \text{(A.2)} \\
&= \mathbb{E}[F_{T|R}(G(r, z, V_i) \mid r) \mid R_i = r, Z_i = z] && \text{(A.3)} \\
&= \int_{\mathcal{V}} F_{T|R}(G(r, z, v) \mid r)\ dF_{V|R,Z}(v \mid r, z),
\end{aligned}
$$

where the first equality is by law of iterated expectations; the second equality is by substituting the definition of $Search_i$; the third equality follows from $T_i \perp\!\!\!\perp (Z_i, V_i) \mid R_i$ imposed by property (ii) in Corollary 1; the final equality follows by definition of conditional expectations.

The hit rate is derived as follows.

$$
\begin{aligned}
&\mathbb{E}[Hit_i \mid R_i = r, Z_i = z] \\
&= \mathbb{E}[\mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i] \mid R_i = r, Z_i = z] \\
&= \int_{\mathcal{V}} \mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i = v]\ dF_{V|R,Z}(v \mid r, z), && \text{(A.4)}
\end{aligned}
$$

where the first equality is by law of iterated expectations; and the second equality is by definition of conditional expectations. The expectation in the integrand may be written as

$$
\begin{aligned}
&\mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i = v] \\
&= \mathbb{E}[Search_i \times Guilty_i \mid R_i = r, Z_i = z, V_i = v] \\
&= \mathbb{E}[Guilty_i \mid Search_i = 1, R_i = r, Z_i = z, V_i = v]\ \mathbb{E}[Search_i \mid R_i = r, Z_i = z, V_i = v] \\
&= \mathbb{E}[Guilty_i \mid G(r, z, v) > T_i, R_i = r, Z_i = z, V_i = v]\ \mathbb{E}[Search_i \mid R_i = r, Z_i = z, V_i = v] \\
&= \mathbb{E}[Guilty_i \mid R_i = r, Z_i = z, V_i = v]\ \mathbb{E}[Search_i \mid R_i = r, Z_i = z, V_i = v] \\
&= G(r, z, v)\ F_{T|R}(G(r, z, v) \mid r),
\end{aligned}
$$

where the first equality follows by definition of Hit_i; the second equality follows by law of iterated expectations, and that $Search_i \times Guilty_i = 0$ when $Search_i = 0$; the third equality follows from the definition of $Search_i$; the fourth equality follows from $T_i \perp\!\!\!\perp Guilty_i \mid R_i, Z_i, V_i$ from Corollary 1; and the final equality follows by definition of $G(\cdot, \cdot, \cdot)$, as well as from (A.1)–(A.3). Substituting this expression for $\mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i = v]$ into (A.4) completes

the derivation of the hit rate.

B Identifying and conducting inference on θ

B.1 Constructing Θ

The bounds in Proposition 2 are sharp in the sense that they are the smallest and largest values of Θ. However, because bilinear programs are non-convex, Θ need not be the full interval derived in Proposition 2. Θ may be recovered by solving the following BP problem for $t \in [-1, 1]$,

$$\begin{aligned} Q^\star_\theta(t) \equiv & \min_{\boldsymbol{\omega}, \{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}} Q(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) \\ \text{s.t.} \quad & \boldsymbol{\omega}'(\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w) = t, \ (9), \ (10), \ (11). \end{aligned}$$

The level of bias t is in Θ if and only if $Q^\star_\theta(t) = 0$.

B.2 Confidence intervals for Θ

The confidence interval for θ may be constructed by inverting the test for racial bias. To determine whether $t \in [-1, 1]$ is in the confidence interval, I first solve the following BP problem,

$$\begin{aligned} \widehat{Q}^\star_\theta(t) \equiv & \min_{\boldsymbol{\omega}, \{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}} \widehat{Q}(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) \\ \text{s.t.} \quad & \boldsymbol{\omega}'(\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w) = t, \ (9), \ (10), \ (11). \end{aligned}$$

I then construct the test statistic

$$\widehat{\tau}_\theta(t) = \widehat{Q}^\star_\theta(t) - \widehat{Q}^\star_{\text{Bias}}, \tag{B.5}$$

which compares the fit of the model when the officer is restricted have bias $\theta = t$ against the fit when the officer is allowed to have any level of bias.

To estimate the distribution of $\widehat{\tau}_\theta(t)$ under the null hypothesis that $t \in \Theta$, I resample the data B times. For each resampled dataset, indexed by $b = 1, \ldots, B$, I calculate (B.5) and denote it by $\widehat{\tau}_{\theta,b}(t)$. Define $\widehat{\tau}^{\text{Null}}_{\theta,b}(t) \equiv \widehat{\tau}_{\theta,b}(t) - \widehat{\tau}_\theta(t)$. Then t does not enter the $(1-\alpha)$ confidence interval for θ if $\widehat{\tau}$ exceeds the $1-\alpha$ quantile of $\{\widehat{\tau}^{\text{Null}}_{\theta,b}\}$.