Testing for Racial Bias in Police Traffic Searches

Joshua Shea*

May 11, 2024

Abstract

I develop a framework to detect and measure racial bias in police traffic searches. Officers are evaluated individually, permitting unrestricted officer heterogeneity and nonrandom assignment of drivers to officers. By using a threshold model with random thresholds, the direction and intensity of bias can vary with the probability that a driver carries contraband. Sharp bounds on the intensity of bias are derived using bilinear programs. I evaluate 50 officers from the Metropolitan Nashville Police Department and find 6 officers to be biased. Estimates suggest that the intensity of bias varies with the probability that a driver carries contraband.

JEL codes: C21, C36, J15, K14, K42

Keywords: Racial bias, police traffic search, bilinear program, partial identification, instrumental variable, non-convex optimization.

*Email: jkcshea@illinois.edu. Department of Economics, University of Illinois Urbana Champaign. I am deeply grateful to Alexander Torgovitsky, Stéphane Bonhomme, and Peter Hull, who have provided invaluable 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, and participants at the Becker Applied Economics Workshop at the University of Chicago. A special thank you to Laura Sale, Francisco del Villar, Dan Kashner, Eyo Herstad, Nadav Kunievsky, Jiarui Liu, Jonas Lieber, and Myungkou Shin. 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, measuring the extent to which bias contributes to the disparities, if at all, is often difficult due to data limitations and the challenges they impose on the methodology.

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. Similar to earlier papers, the officer is modeled to search drivers only if their probability of carrying contraband ("risk") exceeds a threshold. This threshold represents the officer's preference for searching drivers, as well as other factors that influence the search decision. Whereas recent papers have required or assumed fixed thresholds for each race of drivers, I allow the thresholds to be random. This permits a richer form of bias where the direction and intensity of bias may depend on the risk of the driver. Biased officers are not restricted to searching all drivers of one race with a given level of risk, while searching none of the equally risky drivers of another race, as implied by a fixed threshold. Instead, biased officers can search both groups of drivers at different intensities, e.g., whites with 10% risk are searched 20% of the time, and equally risky minorities are searched 40% of the time. Officers can also 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. I show how the dependence between bias and risk may be partially identified despite how risk is unobserved and cannot be credibly estimated.

Testing for bias entails testing whether the sharp identified set for the distributions of officer thresholds (i.e., the smallest set of distributions consistent with the model and data) includes an equivalent pair of distributions for white and minority drivers. If not, then the officer's thresholds must differ by race, implying he is biased. The intensity of bias may be inferred from how dissimilar the distributions of thresholds are across race. This new approach utilizes bilinear programs (BPs), which resemble linear programs but also include bilinear terms (i.e., the product of two distinct variables in the BP). The bilinear terms imply a non-convex optimization problem, yet BPs can be solved to provable global optimality. Restrictions on the model can be layered in a transparent manner as constraints to the program. BPs are not only novel in the context of discrimination, but also in the

context of partial identification and econometrics in general.

Identifying the distributions of officer thresholds is aided by an instrumental variable (IV). The intuition for identification 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. In my setting, the instrument shifts the distribution of risk among drivers stopped without shifting the officer's threshold. For each race of drivers, this generates a sequence of data points that must be consistent with a single distribution of thresholds, thereby constraining what the distribution can be. Since the risk of drivers stopped can be varied for each officer separately, the proposed methods can be applied to each officer separately. This permits unrestricted heterogeneity across officers in their distributions of thresholds and risk. The proposed methods may also be applied without an instrument, although this implies a weaker test for bias and wider bounds on the intensity of bias.

My identification strategy differs from the approach popularized by Arnold et al. (2018), which relies on instruments that shift the thresholds of decision makers while ensuring all decision makers face the same distribution of risk (e.g., random assignment of judges to defendants). In the context of police traffic searches, such instruments are difficult to find because officers select their own distributions of risk by selecting whom to stop. Changing an officer's thresholds—whether the threshold pertains to stopping or searching drivers—may affect who is stopped, which in turn affects the distribution of risk the officer faces. Moreover, different officers patrol different precincts and shifts, which naturally leads to differences in the types of drivers officers interact with.

The proposed methods are not specific to police traffic searches, and extend to settings where Becker's (1957; 1993) outcome test may be applied. More generally, the methods provide a way to partially identify treatment effects on latent thresholds in binary choice models when there is sample selection. The treatment can be discrete, and parametric assumptions on the random threshold are not required.

I apply the proposed methods on a panel data set tracking officers in the Metropolitan Nashville Police Department (MNPD) between 2010 and 2019. Due to the computational demands of the methods, I restrict my attention to the 50 officers with the most number of searches. On average, these officers have made over 2,100 stops and 250 searches for each group of drivers, and account for one third of all searches in the data. I use the time and day of the traffic stop as the instrument, controlling for driver and neighborhood characteristics. Across two sets of estimates, six officers fail the test at the 5% significance level. For each of these officers, I estimate bounds on 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, searched, and killed by police officers compared to white civilians (Gelman et al., 2007; Pierson et al., 2020).[1] However, it is challenging to determine the extent to which these disparities stem from racial bias. This is because researchers have limited information on civilians, officers, and their interactions. In this section, I summarize earlier approaches to detecting racial bias in spite of these limitations.

Knowles et al. (2001) lay the foundation for detecting racial bias in traffic searches by operationalizing the outcome test proposed by Becker (1957, 1993). Officers are modeled as being homogeneous and only search drivers whose risk of carrying contraband exceeds a threshold.[2] Bias is defined as a racial disparity in thresholds. The researcher's objective is thus to recover the thresholds for black and white drivers.

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, risk is unobserved. To recover the threshold, Knowles et al. (2001) use an equilibrium model where heterogeneous drivers decide whether to carry contraband in response to the probability they are searched. In equilibrium, drivers carry contraband with probability equal to a fixed threshold they face given their race.[3] This results in a straightforward test for bias: if officers have different success ("hit") rates conditional on searching white and minority drivers, then officers are biased. In addition to providing testable implications for racial bias, equilibrium models allow officers to coordinate and are able to incorporate constraints officers may face in their frequencies of traffic stops and searches.[4]

Anwar and Fang (2006) propose an alternative test that allows for heterogeneity in officer

[1] See also the Fatal Force database by the Washington Post.

[2] Persico and Todd (2006) generalize the model to allow heterogeneity across officers.

[3] 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.

[4] See Persico (2002) for an equilibrium model where officers are constrained in the volume of traffic searches they can conduct. See Persico and Todd (2006) for a discussion of how unbiased officers may adjust their search decisions to compensate for the actions of biased officers.

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 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.

More 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 restrictions formalized by Canay et al. (2023),[5] the thresholds of all decision makers can be point identified using the marginal treatment effect framework of Heckman and Vytlacil (2005). These restrictions 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 determine 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.[6] The implications are not satisfied and the author points to different distributions of risk across officers as a 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.[7] In the structural component of their paper, 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.[8]

Other papers have used statistical approaches to test whether civilian race has an effect

[5]See Canay et al. (2020a,b), Arnold et al. (2020), and Hull (2021) for a discussion on the restrictions.

[6]Frandsen et al. (2023) propose a test for the exclusion and monotonicity assumptions when using random assignment of judges as instruments. However, their setting is such that outcomes are not censored, whereas outcomes are censored in Arnold et al. (2018) and in police traffic searches (defendants who are not released cannot commit pretrial misconduct, and drivers who are not searched cannot be reported as possessing contraband).

[7]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.

[8]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.

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; Gaebler et al., 2022). 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) due to sample selection.[9]

3 Model

In this section I model the search decision of a single officer (he) for drivers who are stopped (she).[10] Since officers can be analyzed individually, I suppress the officer index 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, which allows different officers to form different assessments of the driver's risk. The econometrician only observes R_i but not V_i; any other characteristics of the driver and the stop observed by the econometrician are implicitly conditioned on throughout.

The driver may or may not 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 there was a "hit," i.e., contraband was found,

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

[9] The authors show that, under a principal strata framework, identifying the effect of race on post-stop decisions is only possible in the knife-edge scenario where the biases from sample selection and omitted variables cancel each other out. See Knox et al. (2020b) and Gaebler et al. (2020, 2022) for further discussion.

[10] If officers work in pairs, then the search decision corresponds to a pair of officers, and pairs of officers are assumed to be fixed across stops.

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).

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 is 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.[11] The setting is observed by both the officer and econometrician and will play the role of an instrument.

The officer's search decision may be written as

$$Search_i \equiv \mathbb{1}\left\{G(R_i, Z_i, V_i) \geq T_i\right\}, \tag{1}$$

where

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

is the probability that the driver caries contraband, which I refer to as the "risk" of the driver; and T_i is a random threshold that may be interpreted as the officer's perceived cost of searching a driver. As discussed below, this model is a relaxation of the Extended Roy Model proposed by Canay et al. (2023) as I allow T_i to be random even after conditioning on all characteristics of the driver observable to the officer (see Appendix A.1 for the derivation of the model).[12] This allows the model to reflect how an officer's threshold may vary for reasons other than the driver (e.g., the officer may receive idiosyncratic shocks to his motivation, risk aversion, and perceptiveness across traffic stops). This also permits a richer notion of bias where the intensity and direction of bias can vary with the risk of the driver.

To derive a logically valid test for bias, I assume the following.

Assumption 1.

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

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

(iii) $Z_i \not\perp\!\!\!\perp V_i \mid R_i = r$ *for* $r \in \{w, m\}$.

[11] If there are variables that inform the officer's stop decision and are visible for some values of Z_i but not others, 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.

[12] T_i can be written as $T_i = t(R_i) + \varepsilon_i$, where $t(\cdot)$ is a deterministic function and ε_i is a shock that may depend on race. The standard model with constant thresholds corresponds to the case where $\varepsilon_i = 0$.

Assumption 1(i) allows me to pool observations within driver race to infer an officer's thresholds. Assumption 1(ii) states that, conditional on driver race R_i, the officer's threshold is jointly independent of the guilt of the driver $Guilty_i$, the setting Z_i, and the unobserved (to the econometrician) driver characteristics, V_i. Assumption 1(iii) states that, conditional on R_i, the instrument Z_i may be used to shift V_i. Assumptions 1(ii)–(iii) are particularly important for the methodology and are discussed in greater detail below.

Under (1) and Assumption 1, the probability that a driver is searched is

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

where the third equality follows from Assumption 1(ii), and $F_{T|R}$ denotes the CDF of T_i conditional on R_i. The probability a driver is searched is equal to the probability the officer's threshold falls below the driver's risk.[13] This leads to the following definition of bias.

Definition 1.

(i) *The officer is racially biased if* $F_{T|R}(\cdot \mid w) \neq F_{T|R}(\cdot \mid m)$.

(ii) *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 $\beta(g)$ *measues the intensity of bias at risk* g. *If* $\beta(g) > 0$ $(\beta(g) < 0)$, *then the officer is biased against minority (white) drivers with risk* g.

Definition 1 is similar to the definition of racial τ-bias proposed by Canay et al. (2023). However, since $\beta(g)$ can vary with g and change sign, the intensity and direction of bias can vary with the risk of the driver, allowing for a more nuanced analysis of bias. This feature of the model arises from the random threshold and distinguishes my model from earlier models

[13] An "essentialist" perspective of race views race as a fixed set of characteristics determined by ancestry. A "constructivist" perspective of race views race as a social categorization, and the race perceived by the officer is a function of the physical and contextual features of the driver (Rose, 2023). This poses new challenges to measuring racial bias as a change in race necessarily coincides with changes in other driver characteristics. Conditioning the analyses on these race-related characteristics may therefore mask the effect of race. Similar to Arnold et al. (2018), the methods I present are valid even under a constructivist framework since traffic searches are only warranted based on the risk of the driver ("unobserved dimension reduction").

where, conditional on race and risk, an officer searches all drivers or none at all.[14] I show in Section 4 how sharp bounds on $\beta(\cdot)$ may be derived.

3.2 Discussion

In the remainder of this section, I discuss Assumptions 1(ii)-1(iii) in greater detail, and the potential challenges to satisfying these two assumptions. I also discuss how to select an instrument, and how the methods extend to other settings.

3.2.1 Exogeneity assumption

There are three independence conditions in Assumption 1(ii): (i) $T_i \perp\!\!\!\perp \mathit{Guilty}_i \mid R_i$; (ii) $T_i \perp\!\!\!\perp Z_i \mid R_i$; (iii) $T_i \perp\!\!\!\perp V_i \mid R_i$. The first condition simply restricts the officer to infer the probability a driver carries contraband using only details he observes during the traffic stop—R_i, V_i, and Z_i—and not from his thresholds. The second condition relates to the instrument, so I postpone its discussion to the next section dedicated to the instrument.

I focus my discussion here on the third independence condition, which ensures that V_i influences the search decision exclusively through the risk of the driver and not through the threshold. This is similar to how an Extended Roy Model permits only one side of the model to depend on variables unobserved by the econometrician. As discussed by Canay et al. (2023), without such an assumption, it is impossible to distinguish between differences in V_i across races from differences in the distribution of T_i across races, with the latter difference being racial bias.[15]

How well $T_i \perp\!\!\!\perp V_i \mid R_i$ is supported depends on the interpretation of T_i and the richness of the data. As shown in Appendix A.1, the threshold can be decomposed as

$$T_i = B_i + C_i + M_i,$$

where B_i reflects the officer's taste for searching drivers, C_i reflects other considerations made by the officer (e.g., safety),[16] and M_i reflects the error in assessing a driver's risk. Supporting the condition $T_i \perp\!\!\!\perp V_i \mid R_i$ amounts to supporting the condition $(B_i, C_i, M_i) \perp\!\!\!\perp V_i \mid R_i$.[17] One approach to achieving this is to simplify the interpretation of T_i by assuming away some of

[14] Fixed thresholds can be implemented in my framework by imposing integrality constraints on the distribution of officer thresholds.

[15] Canay et al. (2023) primarily focus their discussion on the marginal outcome test of Arnold et al. (2018). The random threshold precludes the use of the marginal outcome test as a means to test for bias since there is no longer a single value of risk associated with a driver at the margin of search for each race.

[16] Kleinberg et al. (2018) refer to this as omitted payoff bias.

[17] It is possible that $(B_i, C_i, M_i) \not\perp\!\!\!\perp V_i \mid R_i$, yet $T_i \perp\!\!\!\perp V_i \mid R_i$. I ignore such edge cases.

its components. For example, Knowles et al. (2001), Anwar and Fang (2006), and Persico (2009) assume $M_i = 0$, although such an assumption may be difficult to justify. Another approach is to restrict the channels through which V_i and T_i depend on each other, and then condition the analysis on an appropriate set of covariates to break the dependence. The success of this approach depends on the credibility of the restrictions and the richness of the data. I employ the second approach in the application.

For instance, suppose officers observe a driver's criminal history and derive greater utility from searching drivers with criminal histories compared to drivers without one. This will be reflected in different distributions of B_i for drivers with and without criminal histories. If a driver's criminal history is not observed by the econometrician, then it becomes a component of V_i and Assumption 1(ii) is violated. This invalidates the test since racial disparities in the distribution of T_i may potentially be explained by racial disparities in criminal histories rather than bias. This issue may be addressed by conditioning the analyses on drivers' criminal histories, as in Feigenberg and Miller (2022), as well as other factors that may influence an officer's search preference (e.g., non-race driver demographics).

Officers may also consider factors besides their taste for searching vehicles. These factors are captured by C_i. For example, officers may consider the opportunity cost to searching a vehicle, such as the time spent on other police calls. A potential concern is that minority drivers are more likely to be stopped in areas with greater need of policing compared to white drivers. This leads to a higher opportunity cost of searching minority drivers, creating racial disparities in C_i and T_i. The disparity in T_i therefore cannot be interpreted as bias, as the disparity may instead be driven by differences in demands for policing. To avoid this issue, the analysis should condition on variables correlated with R_i and C_i, such as the volume of calls for police services. Otherwise, these variables enter into V_i, violating Assumption 1(ii) and invalidating the test.

Finally, officers may make errors when assessing the risk of the driver. These errors are represented by M_i. Separating M_i from the other components in T_i is difficult and may require direct measures of officer beliefs (Bohren et al., 2019, 2023).[18] If measurement error exists but is assumed to be idiosyncratic, then the proposed methods remain valid, although any bias detected cannot be interpreted as being taste-based. But if officers are assumed to be better at inferring the risk of drivers who are nervous as opposed to calm, and the demeanor of the driver is contained in V_i, then $M_i \not\perp\!\!\!\perp V_i \mid R_i$, again violating Assumption 1(ii) and invalidating the test.

Satisfying $T_i \perp\!\!\!\perp V_i \mid R_i$ can therefore be a challenge. These challenges are not specific to

[18] Alternatively, one can assume that B_i and C_i are fixed conditional on race, and $\mathbb{E}[M_i \mid R_i = r] = 0$ for $r \in \{w, m\}$. All variation in T_i therefore stems from M_i.

my methodology or setting, but extend to earlier methods and other settings where there may be multiple determinants to the decision maker's threshold. Supporting Assumption 1 thus requires careful economic reasoning and sufficiently rich data.

3.2.2 Choosing instrumental variables

In an ideal experimental setting, the distribution of an officer's threshold can be identified by exposing him to drivers with various levels of risk, and then measuring the probability of searches conditional on race and risk.[19] Bias may then be tested for and measured by comparing the distributions of thresholds across race. The IV attempts to replicate this experiment by varying the risk of the drivers without varying the thresholds. This requires the instrument satisfy a relevance condition, $Z_i \not\perp\!\!\!\perp V_i \mid R_i$, and an exogeneity condition, $T_i \perp\!\!\!\perp Z_i \mid R_i$. Since risk is never observed by the econometrician and the guilt status of drivers is only revealed for those who are searched, the distribution of thresholds can only be partially identified. In Section 4, I show an example where the proposed methods are able to detect bias even without an instrument, although the methods are strengthened by having one.

The intuition behind the partial identification result is similar to that of using an IV to identify a demand curve, where the instrument exclusively shifts the supply curve to trace out the demand curve. In my setting, Z_i exclusively shifts the distribution of risk through shifting V_i, generating a sequence of search and hit rates that constrain what the distribution of T_i can be for each race. This identification argument neither restricts how V_i varies across race, nor how $G(R_i, Z_i, V_i)$ depends on V_i for either race. Identification is then attainable regardless of how differently $G(R_i, V_i, Z_i)$ is distributed for each race of drivers, and is robust to sample selection in traffic stops and statistical discrimination.[20] The identification argument also does not restrict how T_i or V_i varies across officers. This is because the variation in $G(R_i, Z_i, V_i)$ through Z_i *within* officer provides information on the individual officer's thresholds. This paper thus offers a new test for bias in settings where decision makers are either exogenously or endogenously exposed to different distributions of individuals, and does so by providing a method to analyze each decision maker separately.

There are two approaches to selecting an instrument. The first is to consider variables

[19] Assuming that $T_i \perp\!\!\!\perp (V_i, \mathit{Guilty}_i) \mid R_i$, as in Assumption 1(ii).

[20] Regressing outcomes on race dummies and covariates is not a valid approach to test for racial bias. This is because such tests implicitly require the econometrician to balance G_i across races, which may not be possible. For instance, if V_i is distributed differently across races, then G_i is distributed differently across races and the regression suffers from omitted variable bias. This can arise from sample selection, where officers stop different types of white and minority drivers. If G_i depends on V_i differently across races, then G_i is again distributed differently across races and the regression conflates taste-based and statistical discrimination (in the sense of Aigner and Cain (1977)).

that are related to risk but independent of the threshold. An example of such a variable is traffic diversions, which disrupt the usual flow of traffic and force drivers to take routes they usually would not. This can change the composition of drivers in a police precinct, thereby changing the distribution of risk faced by officers of that precinct. If the traffic diversion stems from a road closure or traffic accident, then it may be reasonable to assume that T_i and its components are unaffected by the diversion. While data on certain forms of traffic diversions are available (e.g., road closures), the variation in risk they generate may be too small to effectively detect racial bias in police traffic searches. For this reason, I do not use traffic diversions as the instrument in the application.

The second approach to finding an instrument—which I employ in the application—is to consider variables that are related to risk but only indirectly related to the threshold. For instance, suppose T_i depends on Z_i only through a random vector W_i. If W_i is observed by the econometrician, then the dependence between T_i and Z_i can be broken by conditioning on W_i. In the application, I model W_i to be factors related to safety and choose Z_i to be the time and day of the stop. I assume that an officer's thresholds vary with Z_i to the extent that his safety varies with the time and day of a stop. Then conditional on these factors, the officer is equally willing to search drivers across different times and days.

As with conventional instruments, the primary challenge of choosing an IV is arguing that Z_i satisfies the exogeneity condition, that being $Z_i \perp\!\!\!\perp T_i \mid R_i$. For instance, in the first example above where traffic diversions are used as an instrument, not all traffic diversions will be suitable instruments. Traffic diversions caused by a police investigation or a crime likely encourage officers to search more frequently. Such kinds of traffic diversions should then be excluded as instruments.

The second approach to choosing an instrument faces the same challenge. For instance, let Z_i again be the time and day of stop. There may be a concern that Z_i is endogenous since officers may prefer working morning shifts on weekdays over night shifts on weekends, or certain officers are assigned to certain shifts based on their willingness to search. While this may induce a correlation between Z_i and T_i *across* officers, the proposed methods remain valid as long as an *individual* officer's threshold is independent of Z_i since officers are evaluated separately. I assume this to be the case in the application. However, if individual officers prefer to search more during the night than during the day, and white and minority motorists drive at different times, then the test may conflate racial bias with the changes in thresholds driven by Z_i.

Another challenge with this second approach to choosing an IV is that W_i must include all factors underlying the correlation between Z_i and T_i, and W_i must be observed. For instance, if W_i includes the fatigue of the officer, a variable that may vary by time and day

but is never reported in the data, then it will be impossible to condition on W_i. In such cases, the test may conflate changes in thresholds associated with W_i with racial bias.

3.2.3 Other applications

The proposed methods are not specific to police traffic searches and extend to other settings where an outcome test is appropriate. This may include parole release (Mechoulan and Sahuguet, 2015), pretrial detention (Arnold et al., 2018, 2022), mortgage approvals (Dobbie et al., 2021), and the labor market (Becker, 1957). More generally, the methodology is a way to nonparametrically identify treatment effects on the thresholds in models similar to (1), with the treatment being the driver's race in my application. Treatments are not restricted to be binary and can be discrete. The latent index compared against the threshold need not be a probability and can be generalized to an expectation, as in Dobbie et al. (2021).[21]

4 Detecting and measuring racial bias

In line with Becker (1957, 1993), the test I propose tests whether an officer's search decisions are consistent with him being unbiased. If they are not, then the officer is deemed biased.

4.1 Defining the test

For each traffic stop, I observe the driver's race, R_i; the setting of the stop, Z_i (e.g., time and day); the search decision, $Search_i$; and whether contraband is found, Hit_i. From these variables, I construct the officer's search and hit rates for race $r \in \{w, m\}$ and setting $z \in \mathcal{Z}$,

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

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

[21] MacLeod et al. (2017) study how college reputation affects labor market outcomes. The proposed methods may be applied to answer this question. For example, consider a firm that hires workers whose expected productivity exceeds a threshold, but may be biased towards applicants from high-ranked universities. That is, the threshold for applicants may vary depending on the ranking of their alma mater. The proposed methodology may be used to separate the value of a university's ranking from the university's contribution to its students' productivity and test whether productivity thresholds depend on the ranking of an applicant's university. A potential instrument for such an application are the demographics of an applicant's parents, which are known to affect children's adult outcomes (Heckman and Mosso, 2014). Such information likely correlates with an applicant's productivity, but may be unknown to the firm and therefore may not affect a firm's thresholds.

These equations follow from the law of iterated expectations and Assumption 1.[22] The conditional hit rate is the probability that contraband is found conditional on a traffic search and is equal to the ratio of (3) and (2),

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

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 $(V_i, T_i, Guilty_i) \mid R_i, Z_i$ 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 distribution of the officer's threshold 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 distribution of the officer's

[22] See Appendix A.2 for the full derivation.

threshold is

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

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

Corollary 1. *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 1 follows immediately from Definition 1. ■

A test built around Corollary 1 will be conservative since $\Sigma^{\star}$ may be non-empty even when the officer is biased. Nevertheless, since $\Sigma^{\dagger}$ is sharp, Corollary 1 is the strongest testable implication of the model for unbiasedness.

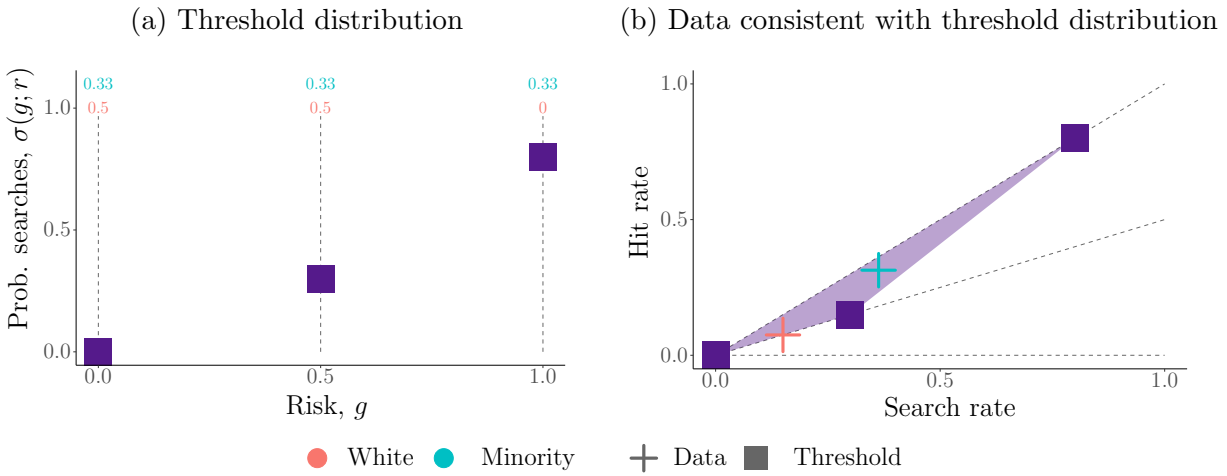
4.2 Intuition

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 distribution of thresholds, 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 the random threshold. 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 the same distribution of thresholds 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.

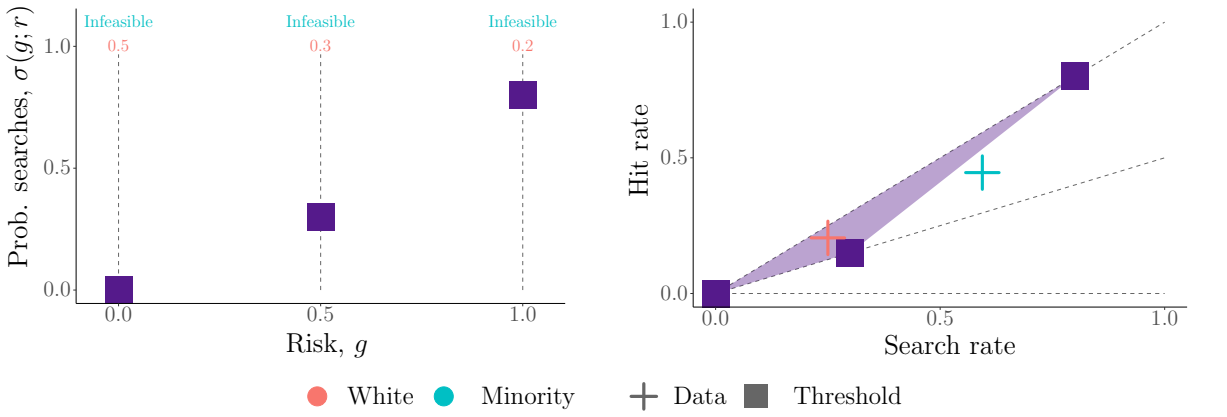
Figure 2 presents the case where only the red cross lies in the convex hull generated by the distribution in the left panel. This implies that the blue cross is generated by a different distribution. Corollary 1 states that if the officer is unbiased, then there must exist a distribution of thresholds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

Figure 1: How search and hit rates are generated



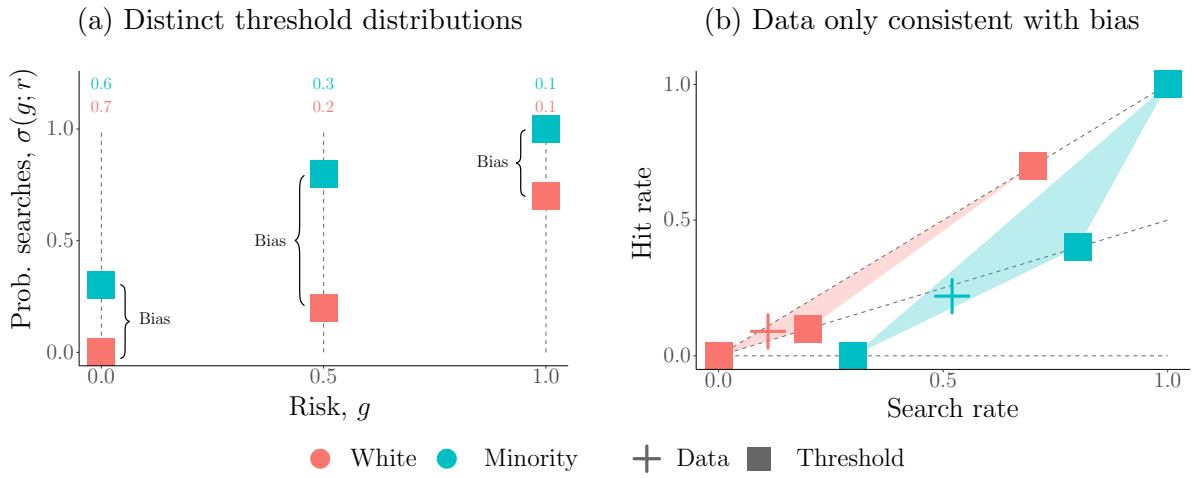
Note: The squares in each figure represent the officer's random threshold. Data that are consistent with the officer's threshold must lie inside the purple region in the right panel. The colored crosses in the right panel represent the observed search and hit rates. Since the data points lie inside the purple region, it is possible that they are generated by the distribution of thresholds 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 thresholds



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

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



Note: Since it is impossible to find a single distribution of thresholds generating the data for both white and minority drivers, the officer must be biased. Any pair of thresholds 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 distributions of thresholds consistent with the data, I am able to derive bounds on various measures of bias (e.g., bias conditional on risk, bias averaged over risk).

Figure 3 presents the case where no single distribution of thresholds can generate the data for both groups of drivers. Racial bias is therefore detected. Note that bias is detected even though there is only one data point for each race of drivers, which corresponds to the case of having no instrument. If an instrument were available, there would be multiple data points for each race of drivers. This strengthens the test by making it more difficult to find a single distribution of thresholds capable of generating all the 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 distributions of thresholds 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 distributions of thresholds and risks that are consistent with the data, I derive bounds on various measures of bias.

4.3 Implementation

Corollary 1 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.[23] For simplicity, suppose that G_i is discrete and $\text{supp}\,(G_i) = \{g_1, \ldots, g_K\}$ for finite

[23]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).

K. Then (4)–(5) become

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

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

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 the distributions of thresholds and risk if G_i is continuously distributed.

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 probabilities $\mathbf{m}^S_{r,z}$, $\mathbf{m}^H_{r,z}$ 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 $\sigma(\cdot; r) \in \Sigma$ 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 distributions of thresholds 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 of which are linear in the parameters of the model. The first set of constraints 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 $\sigma(\cdot; r) \in \Sigma$.

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 $(r,z) \in \{w,m\} \times \mathcal{Z}$. 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

$$Q(\{\boldsymbol{\sigma}_r\},\{\mathbf{p}_{r,z}\}) \equiv \sum_{r,z} \Big| \overbrace{\boldsymbol{\sigma}_r' \mathbf{p}_{r,z}}^{\text{Bilinear terms}} - \mathbf{m}_{r,z}^S \Big| + \sum_{r,z} \Big| \overbrace{(\mathbf{g} \odot \boldsymbol{\sigma}_r)' \mathbf{p}_{r,z}}^{\text{Bilinear terms}} - \mathbf{m}_{r,z}^H \Big|,$$

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 1.

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

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

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

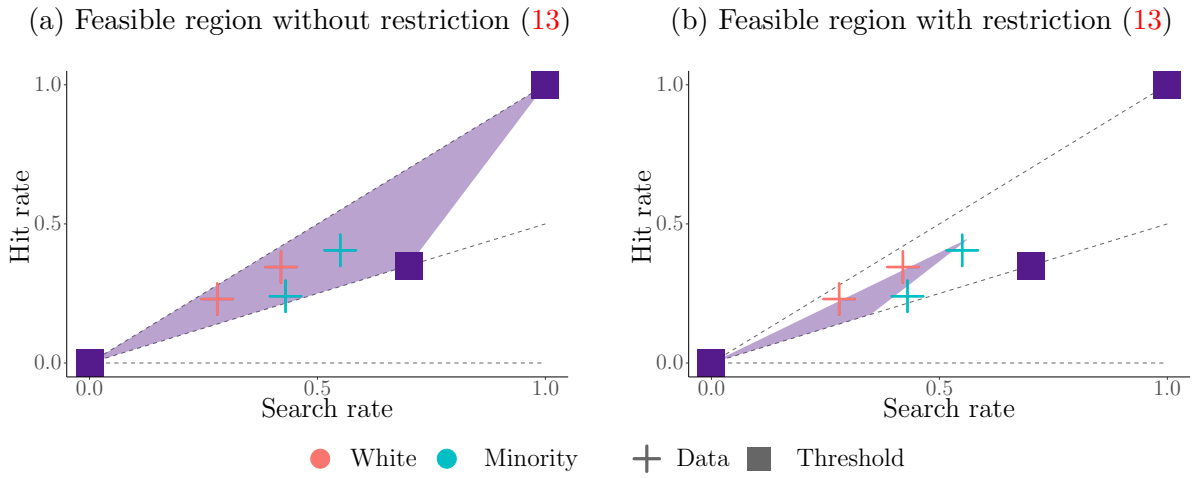
Proof. The constraint $\boldsymbol{\sigma}_w = \boldsymbol{\sigma}_m$ restricts the officer to be unbiased. If $Q^\star > 0$, then (7) or (8) is violated for some $(r,z) \in \{w,m\} \times \mathcal{Z}$. Then by Corollary 1, the officer is biased. ■

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

The test can be strengthened by restricting Σ and $\mathcal{F}_G$. This is straightforward to do if the restrictions can be written as linear, bilinear, or quadratic constraints to the BP problem.

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 random threshold 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 random threshold, 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 distributions from $\Sigma^\star$.

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

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

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.[24]

Figure 4 demonstrates how this restriction strengthens the test. The same distribution of thresholds is depicted in the left and right panel. However, the range of data that can be generated by the threshold is reduced when (13) is imposed. In fact, while there exist random thresholds 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.

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

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

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

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

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 intensity of bias at each level of risk and $\boldsymbol{\omega}$ is a counterfactual distribution of risk.[25] 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 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 prior on the direction of bias, then a sign restriction on the elements

[25] 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.

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, $\beta(\cdot)$. 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 $\boldsymbol{\omega}$ equal to that distribution. For example, the average bias under the counterfactual where risk is uniformly distributed 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 averaged across settings for white drivers,

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

where the second equality follows by the law of iterated expectations, and $\mathbb{P}\{Z_i = z \mid R_i = w\}$ 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, I define the following terms. 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. 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.[26]

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

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

and

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

[26] $\widehat{Q}$ 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 be 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).

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

$$\begin{aligned} \theta_{\text{lb}}, \theta_{\text{ub}} \equiv & \min_{\boldsymbol{\omega}, \{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}} / \max \ \boldsymbol{\omega}' \left(\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_w\right) \\ \text{s.t.} \quad & \widehat{Q}(\{\boldsymbol{\sigma}_r\}, \{\mathbf{p}_{r,z}\}) \leq \widehat{Q}^{\star}_{\text{Biased}}, \ (9), \ (10), \ (11). \end{aligned}$$

I construct the confidence interval for the intensity of bias by inverting the test for bias. That is, 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

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.

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 the stop into three groups: driving-related reasons, non-driving reasons, and investigative reasons.[27] 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,[28] 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 thresholds.

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

[27]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.

[28]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.

warrants.[29] In total, 72% of the traffic searches in the data are retained.

I restrict my attention to the 50 officers with the largest 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. Surprisingly, this small fraction of officers make up one third of all the searches in the data.

Finally, I focus on comparing the officer's thresholds 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.

5.3 Control and instrumental variables

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. To support the independence condition in Assumption 1, I control for variables that may be correlated with both Z_i and officer thresholds. 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.

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.[30] 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 thresholds 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%

[29] 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.

[30] 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.

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.

of drivers stopped for any other reason. This suggests that certain stops in Louisville are pretextual and the reason for stopping a driver can affect an officer's thresholds. 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,[31] 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 tract; and the frequency of calls for MNPD services originating from the census tract. This accounts for the possible correlation between an officer's thresholds and Z_i induced by his surroundings. As discussed in Section 3.2.2, such a correlation may arise if officers are more likely to be in dangerous or high-crime neighborhoods at certain times (e.g., night shifts on weekends), and officers are more concerned about their safety or face higher opportunity costs to searching vehicles when in these neighborhoods.[32]

Some potential concerns with the data are endogenous shift assignments, ticket quotas, or officers being instructed to search more aggressively during certain times. Regarding endogenous shift assignments, see the discussion in Section 3.2.2. Regarding the ticket quotas, Tennessee has explicit laws banning quotas on traffic citations. Although this has not stopped departments from implementing such quotas, ticket quotas pertain to stop decisions of officers instead of search decisions.[33] As long as ticket quotas do not affect search thresholds, then the quotas only impact the search decisions through changing the distribution of risk via sample selection. 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.[34]

[31] In a study on endogenous driving behavior, Kalinowski et al. (2023) find that minority drivers adjust their driving behavior during the day, when their race is more visible to the officer.

[32] 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.

[33] The mayor of Ridgetop, TN attempted 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.

[34] In July of 2019, the MNPD introduced the Entertainment District Initiative, which assigned 17 addi-

A limitation of the data worth mentioning is the absence of criminal records for drivers. Controlling for this information is important if officer thresholds are believed to depend on past offenses of drivers. Unfortunately, police traffic stop data do not contain such information. Identifying information of drivers is also typically hidden, making it impossible to merge in criminal records for drivers. The data set constructed by Feigenberg and Miller (2022) is unique in its inclusion of driver criminal histories.

5.4 Setting up the BP problem

I discretize the support of risk to be

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

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 probabilities $\widehat{\mathbf{m}}^S_{r,z}$, $\widehat{\mathbf{m}}^H_{r,z}$ are estimated using 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).[35] 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

tional 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).

[35] 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).

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.

set $\widehat{\mathbf{m}}^S_{r,z}$ equal to the predicted probabilities 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.

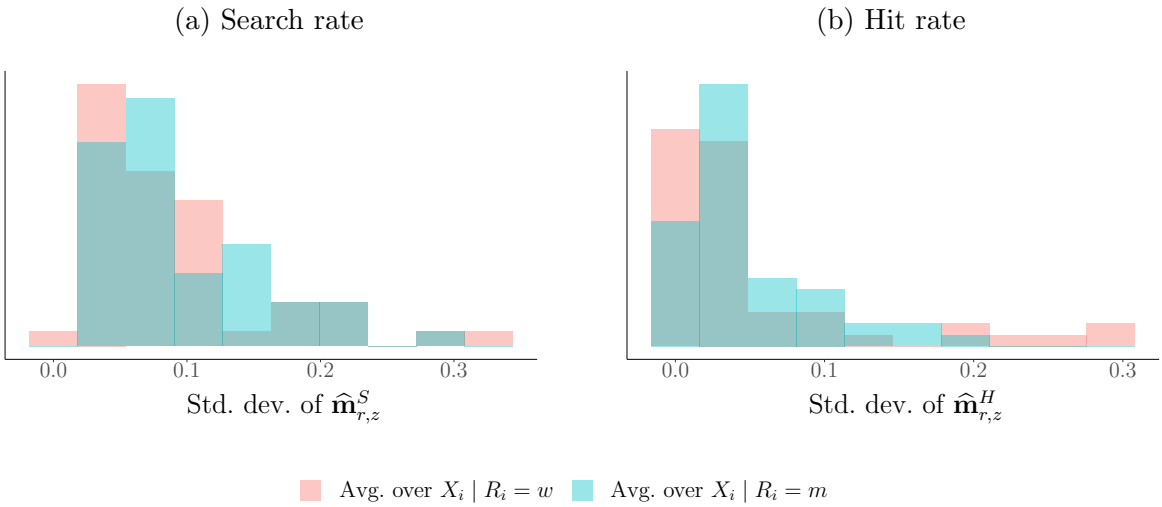
Figure 5 summarizes the variation in search and hit rates generated by Z_i within officer. The figure is obtained by calculating the standard deviations of $\{\widehat{\mathbf{m}}^S_{r,z}\}$ and $\{\widehat{\mathbf{m}}^H_{r,z}\}$ across R_i and Z_i for each officer, and then presenting the histogram of these standard deviations. Greater variation in search and hit rates increases the power of the test by making it more difficult to find a single distribution of thresholds generating the data for both groups of drivers.[36]

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

[36]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).

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 $X_i \mid R_i = w$ ($X_i \mid R_i = m$).

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.

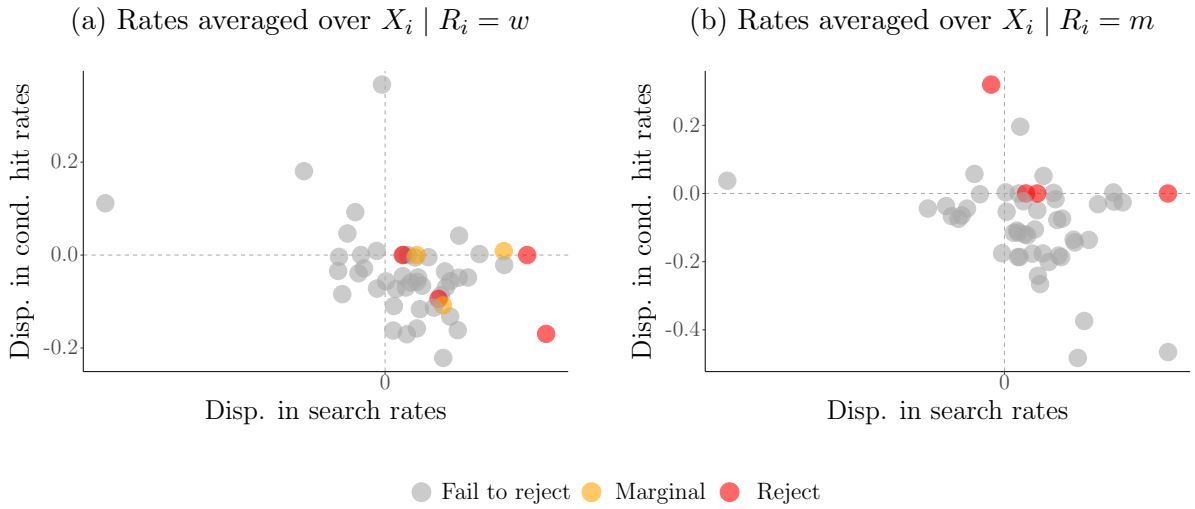
When averaging the search and hit rates over $X_i \mid R_i = m$, I again reject the null hypothesis for four officers. Compared to the previous case where search and hit rates were averaged over $X_i \mid R_i = w$, two of these officers also failed the test, and one of these officers was at the margin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

Correcting for multiple hypothesis testing using the Holm-Bonferroni method, I reject the null hypothesis for two officers whether averaging search and hit rates over $X_i \mid R_i = w$ or $X_i \mid R_i = m$. Given the conservative nature of the test and the conservative nature of the Holm-Bonferroni method, I focus my discussion on the estimates obtained without correcting for multiple hypothesis testing.

The results suggest 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; are stopped in different precincts; and are stopped in areas with higher crime rates, lower income, and lower proportion of white residents. 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 thresholds across race.

To see whether the proposed methods imply a simpler approach to detecting bias, Figure 6

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



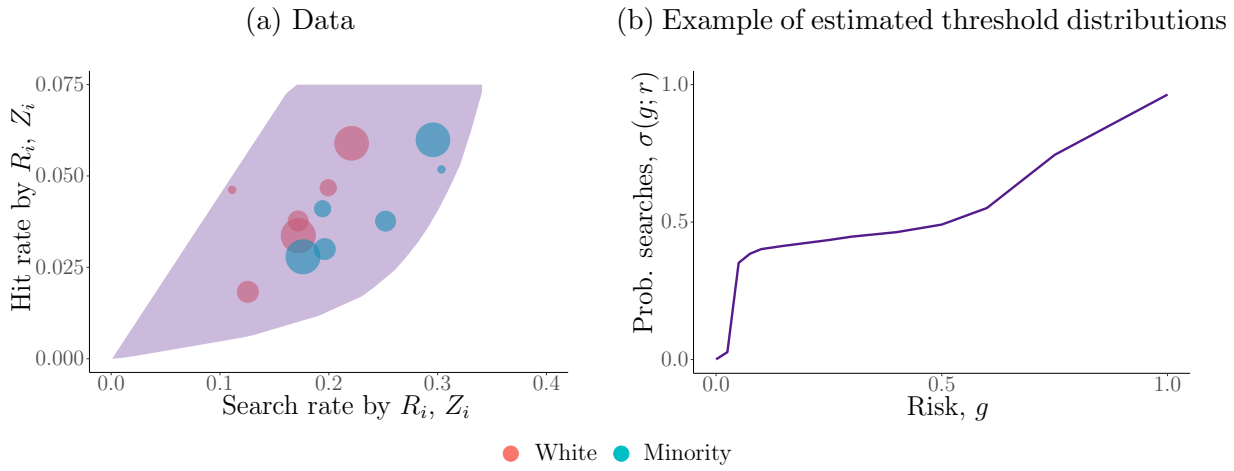
Note: Each point corresponds to an individual officer. Search and hit rates of each officer are averaged across the different settings, controlling for observed characteristics of the driver. Positive disparities indicate that minority drivers have higher rates compared to white drivers. Red points indicate 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.

shows the racial disparities in search and conditional hit rates for officers who are flagged as racially biased and those who are not. Positive disparities in search (hit) rates indicate that minority drivers have higher search (hit) rates compared to whites. The left panel averages search and hit rates over the distribution of $X_i \mid R_i = w$, and the right panel averages the rates over the distribution of $X_i \mid R_i = m$. There is no clear relationship between search and hit rate disparities and racial bias, as bias may or may not be detected regardless of the magnitude of the disparities.

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 threshold distribution 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 the same random threshold. Applying the test from Section 4, I cannot reject the null hypothesis that the officer is unbiased at the 5% significant level.

Figure 8 presents the data for an officer who fails the test, i.e., bias is detected. The top right panel presents one possible estimate of the officer's distribution of thresholds.

Figure 7: Example where bias is not detected

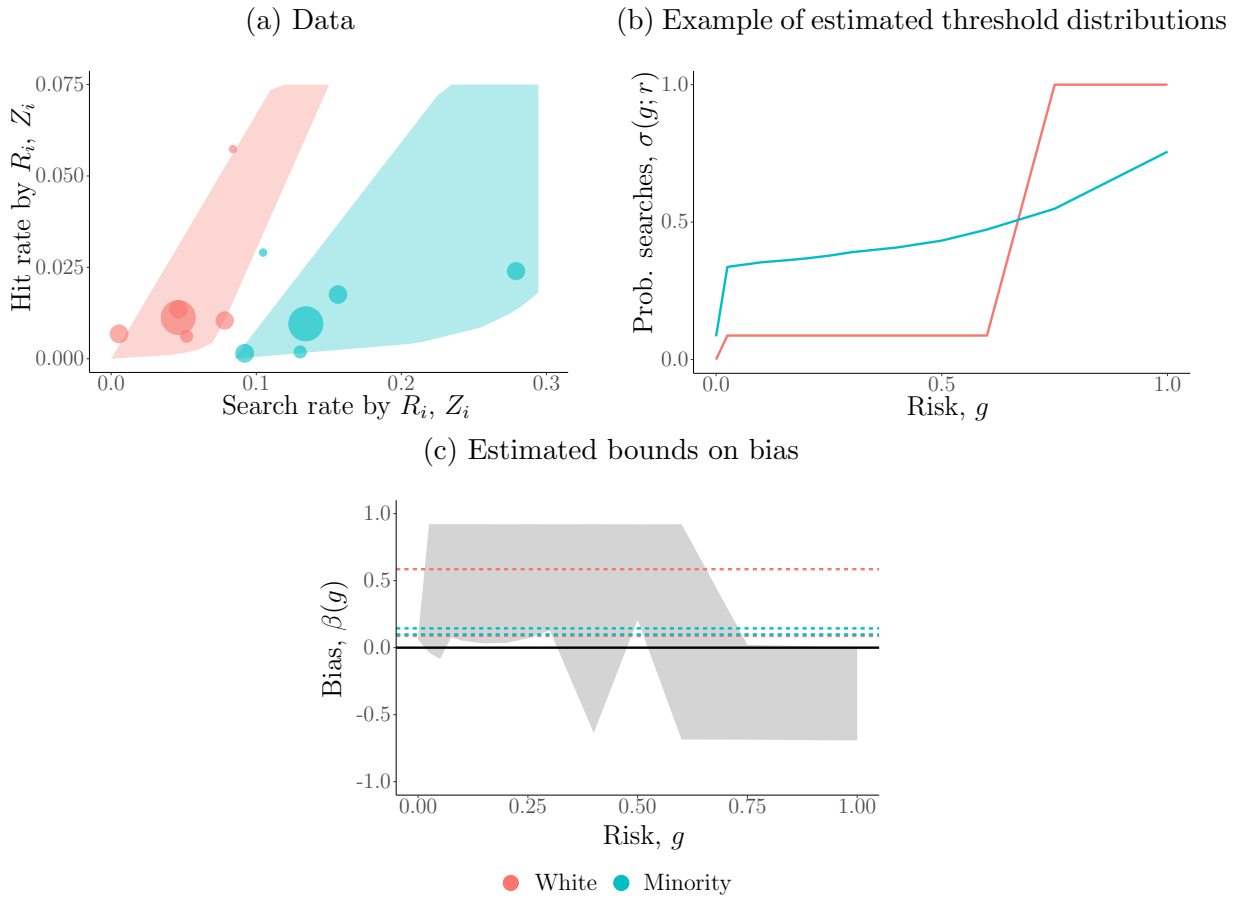


Note: Each dot in the left panel corresponds to the search and hit rates for a particular race and setting. The size of the dots represents the number of stops the data are associated with. Search and hit rates are averaged over the distribution of $X_i \mid R_i = w$. The purple polygon in the left panel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

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 to the extent that the direction of bias may switch once risk is sufficiently large. The top right panel shows two distributions of thresholds consistent with the data, and the change in direction of bias occurs where the two lines intersect. Minority drivers are estimated to be searched at least 8.8 percentage points less on average 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. This result may be interpreted in a few ways. The first is that the officer has almost

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.

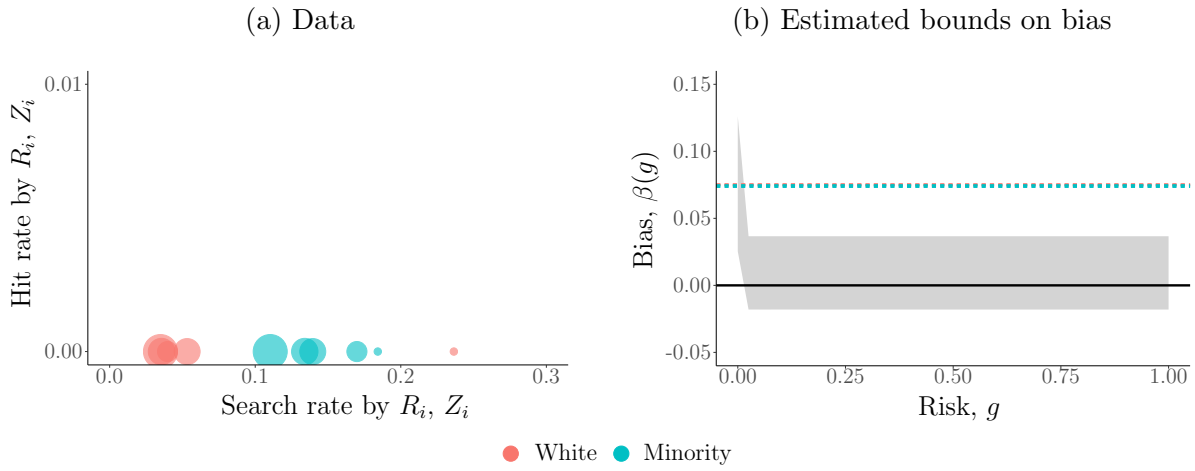
only stopped zero-risk drivers of either race. The racial disparity in search rates is therefore unjustified and bias is detected. Furthermore, if the officer is searching only a fraction of zero-risk drivers from each race, then there is some randomness to the search decision, as suggested by Feigenberg and Miller (2022). Such behavior is consistent with the random threshold proposed in (1), but not the fixed thresholds from the earlier literature.[37]

A second interpretation of this result is that the sample is too small to draw any conclusion. Although all drivers searched in the sample were not found to be guilty, the population risk may differ across the two groups of drivers, and this difference may justify the disparity in search rates. Nevertheless, this officer has stopped almost 3,000 white drivers and searched over 100 of them; and stopped over 3,600 minority drivers and searched over 500 of them.

A third and more pessimistic interpretation of this result is that the model is violated.

[37] A fixed threshold implies officers search all drivers with a given level of risk or none at all.

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.

For example, the officer may have been unable to find contraband in all the cases where it was present, which can be viewed as a violation of Assumption 1(ii).

Figure 10 presents the estimated bounds on the average bias for the officers who fail the test.[38] 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 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.

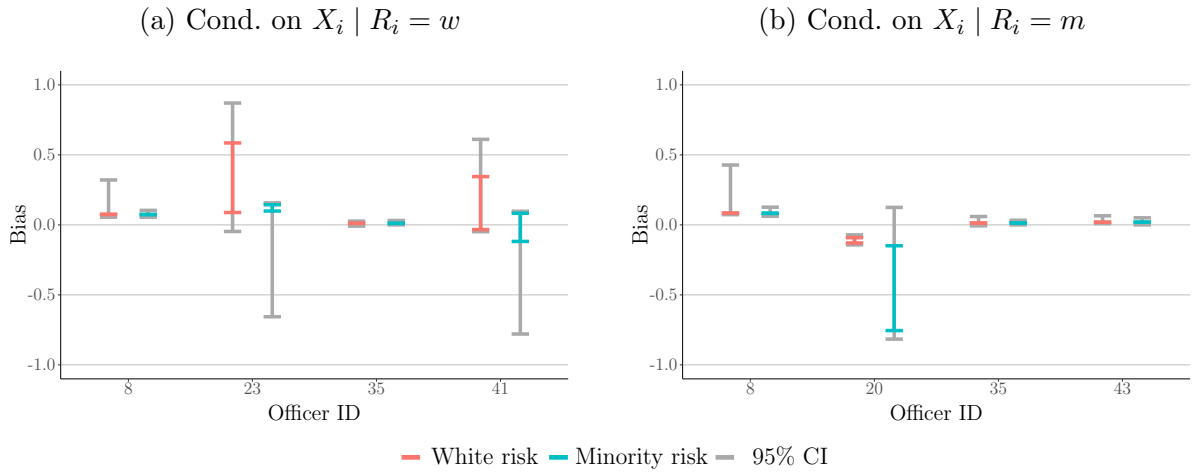
When averaging search and hit rates over $X_i \mid R_i = w$, the estimates suggest white drivers would be searched at least 2.4 percentage points more on average (83.1% more relative to the observed search rate of 2.9%) by the biased officers if the drivers were treated as minorities, holding their risk 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.[39]

When averaging search and hit rates over $X_i \mid R_i = m$, the estimates suggest white drivers

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

[39] 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.

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 the 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 biased against white drivers on average.[40]

For comparison, I also apply the test of Knowles et al. (2001) to these 50 officers. Table 4 presents the hit rates of white and minority drivers who are searched, as well as the p-values from Pearson χ^2 tests of equal hit rates. Under the model of Knowles et al. (2001), the null hypothesis of equal hit rates corresponds to officers being unbiased. Bold entries indicate the group of drivers that officers (as a collective) are biased against whenever bias is detected. Interestingly, officers appear to be biased against white drivers in most cases where bias is detected. This difference stems from how the test of Knowles et al. (2001) compares the police department-wide hit rate across white and minority drivers, whereas the method I propose compares both the search and hit rates in multiple settings across white and minority drivers for each officer. Because the test of Knowles et al. (2001) is derived from an equilibrium model, it is not intended to evaluate each officer separately and

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

cannot identify which officers are biased.[41]

6 Conclusion

In this paper, I provide a flexible approach to detect and measure racial bias in police traffic searches. The proposed methods are valid amid sample selection on unobservables and statistical discrimination. In addition, by using an IV to vary the risk among drivers stopped, the methods may be applied to individual officers, allowing for unrestricted heterogeneity across officers in the distributions of thresholds and risk.

This paper also contributes to the literature from a modeling standpoint, as earlier papers studying racial bias have either assumed or required models with deterministic thresholds once conditioning on race, 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 are obtained. Additional restrictions to tighten these bounds or strengthen the test are straightforward to impose. Implementing these methods 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 evaluated are biased. For each of these officers, I am able to estimate bounds on 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.

[41] A potential way to narrow down which officers are biased is to apply the test to different combinations of precinct, day, and shift, although there are still many officers operating within each combination.

Table 4: Knowles et al. (2001) test results

	Male and female drivers			Male drivers			Female drivers		
	White	Minority	p-value	White	Minority	p-value	White	Minority	p-value
All	**0.145**	0.174	0.000***	**0.151**	0.179	0.000***	**0.134**	0.158	0.005***
Precinct 1	0.186	0.192	0.779	0.187	0.172	0.505	**0.180**	0.287	0.039**
Precinct 2	0.206	0.217	0.495	0.202	0.215	0.5	0.214	0.224	0.74
Precinct 3	0.182	**0.144**	0.008***	0.171	0.150	0.204	0.210	**0.112**	0.001***
Precinct 4	**0.022**	0.083	0.000***	**0.022**	0.097	0.000***	**0.022**	0.051	0.068*
Precinct 5	0.203	**0.183**	0.054*	0.208	**0.183**	0.046**	0.193	0.183	0.618
Precinct 6	**0.167**	0.238	0.015**	**0.160**	0.240	0.033**	0.177	0.234	0.249
Precinct 7	0.183	**0.161**	0.073*	0.186	0.175	0.503	0.179	**0.127**	0.011**
Precinct 8	**0.053**	0.125	0.000***	**0.064**	0.138	0.000***	**0.035**	0.093	0.000***
Weekday morning	0.262	**0.227**	0.024**	0.267	**0.232**	0.057*	0.251	0.206	0.135
Weekday evening	0.177	0.183	0.55	0.184	0.185	0.928	0.164	0.174	0.528
Weekday night	**0.108**	0.165	0.000***	**0.114**	0.170	0.000***	**0.096**	0.146	0.000***
Weekend morning	0.266	**0.127**	0.014**	0.163	0.152	0.873	0.476	**0.023**	0.000***
Weekend evening	**0.104**	0.160	0.002***	**0.110**	0.162	0.014**	**0.091**	0.150	0.063*
Weekend night	**0.056**	0.106	0.000***	**0.055**	0.109	0.000***	**0.059**	0.100	0.017**

Notes: In each vertical panel, the first two columns present the hit rates for white and minority drivers who are searched. The third column presents the p-value from a Pearson χ^2 test for equal hit rates, conditional on certain characteristics of the driver and traffic stop. Bold hit rates indicate the group of drivers which officers are biased against. * 10%, ** 5%, *** 1% significance.

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.

Arnold, D., W. Dobbie, and C. S. Yang (2020). Comment on Canay, Mogstad, Mountjoy (2020).

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 (2023). Inaccurate Statistical Discrimination: An Identification Problem. *Review of Economics and Statistics*, 1–45.

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

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 (2023). On the Use of Outcome Tests for Detecting Bias in Decision Making. *Review of Economic Studies*.

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.

De Boor, C. (2001). *A Practical Guide to Splines, Revised Edition*, Volume 27. Springer-Verlag New York.

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.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2021). Measuring Bias in Consumer Lending. *The Review of Economic Studies 88*(6), 2799–2832.

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.

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*.

Gaebler, J., W. Cai, G. Basse, R. Shroff, S. Goel, and J. Hill (2022). A Causal Framework for Observational Studies of Discrimination. *Statistics and public policy 9*(1), 26–48.

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 S. Mosso (2014). The Economics of Human Development and Social Mobility. *Annu. Rev. Econ. 6*(1), 689–733.

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. J., M. B. Ross, and S. L. Ross (2023). Endogenous Driving Behavior in Tests of Racial Profiling. *Journal of Human Resources*.

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

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

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.

MacLeod, W. B., E. Riehl, J. E. Saavedra, and M. Urquiola (2017). The Big Sort: College Reputation and Labor Market Outcomes. *American Economic Journal: Applied Economics 9*(3), 223–261.

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).

Mechoulan, S. and N. Sahuguet (2015). Assessing Racial Disparities in Parole Release. *The Journal of Legal Studies 44*(1), 39–74.

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

Mørken, K. (1991). Some Identities for Products and Degree Raising of Splines. *Constructive Approximation 7*, 195–208.

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.

Persico, N. (2002). Racial Profiling, Fairness, and Effectiveness of Policing. *American Economic Review 92*(5), 1472–1497.

Persico, N. (2009). Racial Profiling? Detecting Bias Using Statistical Evidence. *Annu. Rev. Econ. 1*(1), 229–254.

Persico, N. and P. Todd (2006). Generalising the Hit Rates Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita. *The Economic Journal 116*(515), F351–F367.

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.

Rose, E. K. (2023). A Constructivist Perspective on Empirical Discrimination Research. *Journal of Economic Literature 61*(3), 906–923.

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.

Shea, J. and A. Torgovitsky (2023). ivmte: An R Package for Extrapolating Instrumental Variable Estimates Away From Compliers. *Observational Studies 9*(2), 1–42.

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 Derivations

A.1 Deriving the random threshold in (1)

My model is an extension of the model by Canay et al. (2023). Let $Y_{si} \in \{0,1\}$ denote the potential outcome of whether contraband is found on driver i when $Search_i = s$.[42] I assume that contraband is found if and only if the driver is carrying contraband and searched, which implies $Guilty_i = Y_{1i} - Y_{0i}$. Let $R_i \in \{w, m\}$ denote the race of the driver, V_i denote the characteristics of the driver observed by the officer and not the econometrician, and Z_i denote the instrument. Variables observed by both the officer and econometrician are implicitly conditioned on and I suppress their notation for brevity.

Let $\widetilde{B}_i$ denote the officer's taste for searching driver i. Let $\widetilde{C}_i$ denote the cost of searching driver i, which represents other considerations besides bias. To allow for measurement error in the risk assessment, let $Y^\star_{si} \in \{0,1\}$ denote the potential outcome the officer considers when making his decision, which can differ from the true Y_{si}. The officer solves the following problem when deciding whether to search driver i,

$$\max_{s\in\{0,1\}} \mathcal{E}\left[Y^\star_{si} + s(\widetilde{B}_i - \widetilde{C}_i) \mid R_i = r, Z_i = z, V_i = v\right] + U_{si},$$

where $\mathcal{E}[\cdot \mid R_i = r, Z_i = z, V_i = v]$ denotes the officer's subjective conditional expectation; and $U_{si} = s(\check{B}_i - \check{C}_i) + \check{M}_{si}$ is a sum of random shocks to search tastes ($\check{B}_i$), search costs ($\check{C}_i$), and error in risk assessment ($\check{M}_{si}$). The solution to the above problem is

$$Search_i = \mathbb{1}\left\{\mathcal{E}[Y^\star_{1i} - Y^\star_{0i} \mid R_i = r, Z_i = z, V_i = v] \geq B_i(r,z,v) + C_i(r,z,v) - (\check{M}_{1i} - \check{M}_{0i})\right\},$$

where

$$\begin{aligned} B_i(r,z,v) &= -\left(\mathcal{E}[\widetilde{B}_i \mid R_i = r, Z_i = z, V_i = v] + \check{B}_i\right), \\ C_i(r,z,v) &= \mathcal{E}[\widetilde{C}_i \mid R_i = r, Z_i = z, V_i = v] + \check{C}_i \end{aligned}$$

To collect the effect of measurement error into a single term, define

$$m(r,z,v) = \mathbb{E}[Y_{1i} - Y_{0i} \mid R_i = r, Z_i = z, V_i = v] - \mathcal{E}[Y^\star_{1i} - Y^\star_{0i} \mid R_i = r, Z_i = z, V_i = v]$$

[42] The methods extend to the case where Y_{si} is continuous.

to be the deviation between the true risk of the driver and the officer's risk assessment, conditional on R_i, Z_i, and V_i. The officer's search decision may then be written as

$$Search_i = \mathbb{1}\Big\{ \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid R_i = r, Z_i = z, V_i = v]}_{\mathbb{P}\{Guilty_i | R_i=r, Z_i=z, V_i=v\}} \geq \underbrace{B_i(r,z,v) + C_i(r,z,v) + M_i(r,z,v)}_{T_i | R_i=r, Z_i=z, V_i=v} \Big\}, \tag{A.1}$$

where

$$M_i(r,z,v) = m(r,z,v) - (\check{M}_{1i} - \check{M}_{0i}).$$

As in Canay et al. (2023), the threshold in (A.1) includes terms representing search tastes, search costs, and measurement error in risk. However, by introducing shocks $\check{B}_i$, $\check{C}_i$, $\check{M}_i$, the threshold is random even after conditioning on R_i, Z_i, and V_i. This allows the direction and intensity of bias to vary with the risk of the driver.

Assumption 1(ii) is satisfied by imposing the following mean independence assumption that is analogous to the Extended Roy Model restriction in Canay et al. (2023),

$$\begin{aligned} \mathcal{E}[\widetilde{B}_i \mid R_i = r, Z_i = z, V_i = v] &= \mathcal{E}[\widetilde{B}_i \mid R_i = r], \\ \mathcal{E}[\widetilde{C}_i \mid R_i = r, Z_i = z, V_i = v] &= \mathcal{E}[\widetilde{C}_i \mid R_i = r], \\ m(r,z,v) &= m(r), \end{aligned}$$

and assuming the shocks to search tastes, search costs, and measurement error are effectively idiosyncratic,

$$(\check{B}_i, \check{C}_{si}, \check{M}_{0i}, \check{M}_{1i}) \perp\!\!\!\perp (Guilty_i, Z_i, V_i).$$

A.2 Deriving the search and hit rates

The search rate is derived as follows.

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

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 Assumption 1; and the final equality follows by definition of conditional expectations.

The hit rate is derived as follows.

$$\begin{aligned}
&\mathbb{E}[Hit_i \mid R_i = r, Z_i = z] \\
&= \mathbb{E}[\mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i] \mid R_i = r, Z_i = z] \\
&= \int_{\mathcal{V}} \mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i = v] \; dF_{V|R,Z}(v \mid r, z),
\end{aligned} \tag{A.3}$$

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 Assumption 1; and the final equality follows by definition of $G(\cdot, \cdot, \cdot)$, as well as from (A.2). Substituting this expression for $\mathbb{E}[Hit_i \mid R_i = r, Z_i = z, V_i = v]$ into (A.3) completes the derivation of the hit rate.

B Identifying and conducting inference on Θ

B.1 Identifying 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 $t \in [-1, 1]$,

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

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

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

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

I then construct the test statistic

$$\widehat{\tau}_\theta(t) = \widehat{Q}_\theta^\star(t) - \widehat{Q}_{\text{Bias}}^\star, \tag{B.4}$$

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.4) and denote it by $\widehat{\tau}_{\theta,b}(t)$. Define $\widehat{\tau}_{\theta,b}^{\text{Null}}(t) \equiv \widehat{\tau}_{\theta,b}(t) - \widehat{\tau}_\theta(t)$. Then t does not enter the $(1-\alpha)$ confidence interval for θ if $\widehat{\tau}$ exceeds the $1-\alpha$ quantile of $\{\widehat{\tau}_{\theta,b}^{\text{Null}}\}$.