Testing for Racial Bias in Police Traffic Searches

Joshua Shea*

December 9, 2021

[Link to most recent draft]

Abstract

I construct a flexible test for racial bias in police traffic searches that is valid amid sample selection and statistical discrimination. The test uses instrumental variables that shift the distribution of drivers stopped without shifting the officer's search preferences. These instruments enable the test to be performed separately for each officer, thus permitting unrestricted heterogeneity in their preferences and beliefs. By modeling search decisions stochastically, I allow the direction and intensity of bias to depend on the officer's beliefs, and I derive sharp bounds on various measures of intensity. I apply the test to 50 officers in the Metropolitan Nashville Police Department, and find evidence suggesting 8 to 14 officers are biased, with 7 being biased against minorities. I also find evidence suggesting the intensity of bias decreases for riskier drivers.

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

1 Introduction

Since the police killings of Eric Garner, Michael Brown, and Tamir Rice in 2014, the potentially fatal cost of encountering police for Black Americans has become a central theme in public and political discourse. The cost of these events is not only the welfare and lives of Blacks, but also the credibility and authority of the police (Manski and Nagin, 2017; Owens, 2020).[1] Over the last two decades, confidence in the police among African Americans has fallen from 37% to 19%, with 84% believing they are treated unfairly by the police (Jones, 2021; Horowitz et al., 2019).[2] This decline in confidence has led to growing demand for police accountability, a lack of which has historically been the norm.[3] In the cases of Eric Garner, Michael Brown, and Tamir Rice, none of the officers responsible for the deaths were indicted by a grand jury, an outcome that sparked protests across the country against racial bias in policing.

But holding officers accountable for their actions by establishing misconduct is difficult for several reasons.[4] In particular, we do not know the thoughts of an officer, making it difficult to determine if he has abused his discretion. This is exemplified by a comment from a juror involved in the case against officer Jeronimo Yanez, who

[1] Trinkner et al. (2019) find when officers are perceived as being racist, they feel their authority is lessened. This correlates with their condoning of greater use of force to establish control.

[2] Confidence in the police among whites has fallen from 63% to 56% in the last two decades, and 63% of whites believe Blacks are treated unfairly by the police (Jones, 2021; Horowitz et al., 2019).

[3] See Morin and Stepler (2016) and https://www.bloomberg.com/news/articles/2020-06-09/a-history-of-protests-against-police-brutality

[4] For example, officers refuse to testify against their colleagues ('Blue Wall of Silence'), and prosecutors typically refuse to go after officers in fear of jeopardizing their working relationship with the police department. See https://fivethirtyeight.com/features/why-its-still-so-rare-for-police-officers-to-face-legal-consequences-for-misconduct/ and https://www.vox.com/21497089/derek-chauvin-george-floyd-trial-police-prosecutions-black-lives-matter.

was acquitted for shooting and killing Philando Castile in front of his girlfriend and her four-year-old daughter during a traffic stop in 2016:[5]

> "It just came down to us not being able to see what was going on in the car. Some of us were saying that there was some recklessness there, but that didn't stick because we didn't know what escalated the situation: was [Yanez] really seeing a gun? We felt [Yanez] was an honest guy... and in the end, we had to go on his word, and that's what it came down to."

In this paper, I develop a test for racial bias that recognizes the disparity in the information available to the officer versus the researcher. Specifically, I use a partial identification approach to test whether individual officers have different preferences for searching white and minority drivers who are stopped. These preferences govern an economic choice model for searching drivers that depends on the officer's beliefs about how likely a driver carries contraband (e.g., drugs or weapons), which are unobserved by the researcher. Using this approach, I am able to make sharp inferences on how the officer's decisions depend on his beliefs. Restrictions on the officer's preferences and the probability that drivers carry contraband may be layered in a flexible and transparent manner. The test does not require officers to be randomly assigned to drivers and may be performed for each officer separately.

The test for bias checks whether the sharp identified set for the officer's search preferences includes an equivalent pair of preferences for white and minority drivers. If not, then the officer's preferences must differ by race, implying he is biased. The partial identification approach permits the test to be valid even when officers have dissimilar beliefs for white and minority drivers, which can occur for several reasons. One possibility is statistical discrimination, where the quality of the signals used by

[5] https://www.mprnews.org/story/2017/06/23/74-seconds-yanez-juror

officers to form their beliefs differ for white and minority drivers (Aigner and Cain, 1977). Another possibility is sample selection, which occurs because officers choose to stop different types of drivers depending on their race (e.g., bias in traffic stops). A third possibility is that white and minority drivers in population are simply different. Implementing the test entails solving a bilinear program, a type of non-convex problem that can be solved to global optimality. Bilinear programs are not only novel in the context of discrimination, but also in the context of partial identification.

A distinguishing feature of my test is how I model an officer's search decision. Similar to earlier papers, the officer is modeled to search drivers only if their probability of carrying contraband ('risk') exceeds a threshold, where the threshold represents the officer's search preference. However, whereas recent papers have required or assumed fixed thresholds (see Canay et al. (2020) and Hull (2021) for a discussion on this restriction), I use a random threshold. Consequently, there is no longer a single driver at the margin of search, but a 'marginal driver' at every level of risk. This means a biased officer is not restricted to searching all drivers of one race with a given level of risk, while searching none of the equally risky drivers of the other race, as implied by a fixed threshold. Instead, the officer can search both groups of drivers at different intensities, e.g., whites with 10% risk are searched 10% of the time, whereas equally risky minorities are searched 20% of the time. Officers can even change direction of bias depending on the level of risk, e.g., whites with 10% risk are half as likely to be searched compared to equally risky minorities, but whites with 20% risk are twice as likely to be searched compared to equally risky minorities. Using a bilinear program, it is feasible to estimate sharp bounds on these differences conditional on the risk of drivers. The random threshold therefore permits a richer and more refined analysis of racial bias, where the direction and intensity of bias may depend on unobserved (to the researcher) characteristics of the driver.

Identification stems from an instrumental variable (IV) that shifts the distribution of risk among drivers stopped without shifting the officer's preferences. This variation allows the researcher to 'trace' out the preferences of the officer, similar to how an IV is used to overcome simultaneity when estimating demand curves. More intuitively, the identification argument is simply that an officer's preferences may be learned by observing him make search decisions for many types of drivers. Since it is possible to vary the risk of drivers stopped for each officer separately, it is possible to apply the proposed methods on each officer separately.

I apply the test using data collected by the Metropolitan Nashville Police Department (MNPD) between 2010 and 2019. As with all police traffic data, the data are limited to drivers whom officers have chosen to stop. Due to the scarcity of traffic searches, I restrict my attention to the 50 officers with the most number of searches.[6] I find that 8 (14) officers fail the test at the 5% (10%) significance level. Among these officers, minority drivers are at least 8.3 percentage points more likely to be searched on average compared to equally risky white drivers, which is large considering these officers only search white drivers 6.7% of the time. The variation in the bounds on these disparities suggest that the intensity of bias diminishes 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 presents the data; Section 6 discusses the estimation procedure; Section 7 presents the estimates; and Section 8 concludes.

[6]The original data contains over 2,200 officers. These 50 officers are responsible for a third of all searches in the data.

2 Literature review

It is well documented that Black civilians are disproportionately impacted by policing across the US. Compared to their white counterparts, Black pedestrians are stopped up to six times as often (Gelman et al., 2007); Black motorists who are stopped are searched twice as often (Pierson et al., 2020); and Black civilians are almost seven times more likely to be killed by an officer.[7] Whether these disparities are the outcome of racial bias is a difficult question because officers are given a lot of discretion in making their decisions, but the researcher does not know what the officer is thinking at the time of the decision. In this section, I summarize earlier approaches to answering this question.

The seminal paper by Knowles et al. (2001) lays the foundation for detecting racial bias in traffic searches through its use of the outcome test proposed by Becker (1957, 1993). In Knowles et al. (2001), officers are homogeneous and only search drivers whose perceived risk of carrying contraband exceeds a fixed threshold. Officers are assumed to have accurate beliefs, meaning that the risk they perceive is equal to the true risk. If the thresholds differ for white and minority drivers, then officers are racially biased, so the researcher's objective is to recover these thresholds. If risk is observed by the researcher and continuously distributed over the unit interval, then the thresholds are identified from the risk of the white and minority drivers at the margin of search.

[7] Source: Fatal Force, Washington Post. Since 2015, the Washington Post has recorded every fatal shooting by on-duty officers in the US. This was a response to their investigation in the previous year revealing that the FBI undercounted fatal police shootings since police departments are not required to report such incidents and only do so voluntarily. Similarly, Collaborators et al. (2021) found over half of all deaths in the US due to police violence between 1980 and 2018 were unreported in the National Vital Statistics System.

However, because risk is not actually observed, the researcher must use a different strategy. Developing an econometric solution was considered to be difficult and has only recently been done (Ayres, 2002; Arnold et al., 2018). So Knowles et al. (2001) instead form a game-theoretic argument that drivers of the same race have the same risk in equilibrium, placing every driver at the margin of search.[8] In addition, the authors show that if officers are unbiased, then all white and minority drivers carry contraband with equal probability. This results in a straightforward test for bias: if officers have different success ('hit') rates when searching white and minority drivers, then officers are biased. However, the model's assumption of homogeneous officers, as well as its implication of homogeneous drivers (within race), may both be rejected using officer-level data. For instance, Ba et al. (2021) find that the rate at which officers stop, arrest, and use force against civilians varies with the race and sex of the officer.[9] Also, the variation across MNPD officers in the success rate of a search reveals that drivers are not homogeneous in risk.

Anwar and Fang (2006) propose an alternative test that allows for heterogeneity in officer decisions and driver risk.[10] Using a similar model for officers as Knowles et al. (2001) (i.e., officers only search drivers whose risk exceed a fixed threshold), but allowing different officers to have different thresholds, Anwar and Fang (2006)

[8]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. So in equilibrium, all drivers carry contraband with equal probability and officers search at random.

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

[10]Antonovics and Knight (2009) independently developed the same test as 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 same regardless of the race of the driver. While this approach can detect bias, it cannot determine which group is biased, nor which group of drivers is being discriminated against.

Recently, Arnold et al. (2018) made an important contribution to the literature by using random assignment of decision makers as an instrument to detect bias. However, their method is not directly applicable to the setting of this paper since officers select who to stop and are therefore non-randomly assigned to drivers. Nevertheless, Arnold et al. (2018) provide an empirical strategy to recover the thresholds of decision makers without trivializing the distribution of risk for each race (as in Knowles et al., 2001). The authors use the marginal treatment effect framework of Heckman and Vytlacil (2005) to compare the decisions of a continuum of decision makers, and show that this point-identifies the thresholds of all decision makers. This method is referred to as the marginal outcome test, and is valid as long as all decision makers have common distributions of risk (hence the importance of random assignment) and fixed thresholds (Canay et al., 2020). To see whether this approach extends to police traffic searches, Gelbach (2021) tests three implications of the model on police traffic data from Florida and Texas. These implications are not satisfied in the data, and the author points to different distributions of risk across officers as the potential reason. Papers using the marginal outcome test to study bias in policing therefore require restrictions on the distributions of risk. For example, Marx (2021) requires the distributions to either be common across officers, or that the distribution for one officer second-order stochastically dominates that of another. Feigenberg and Miller (2021) do not restrict how the distributions of risk differ across officers, but assume that they are independent of the officer's threshold and rule out sample selection

on unobservables.[11] While Arnold et al. (2020) extend the method of Arnold et al. (2018) to allow decision makers to face different distributions of risk, it comes at the expense of parametric assumptions on the joint distribution of thresholds and risk.[12]

Other papers have used statistical approaches to test whether civilian race has an effect on police decisions, including stop-and-frisk and use of force (Ridgeway, 2006; Grogger and Ridgeway, 2006; Gelman et al., 2007; Ridgeway and MacDonald, 2009; Goel et al., 2016a,b; Fryer Jr, 2019; MacDonald and Fagan, 2019; Knox et al., 2020; Gaebler et al., 2020). 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. (2020) is noteworthy for emphasizing how difficult it is to identify the effect of race on post-stop decisions alone (e.g., use of force, traffic searches) because of sample selection. The authors show that, under a principal strata framework, this is only possible in the knife-edge scenario where the biases from sample selection and omitted variables cancel each other out. Consequently, in their paper the effect of race on post-stop decisions includes the effect of race on stop decisions as well. The reason I am able to measure bias in the search decision alone in spite of sample selection is because I impose an economic choice model.

An area in the literature that has received increasing attention is inaccurate beliefs (see Bohren et al., 2019, 2020; Bordalo et al., 2016). In my setting, this corresponds to an officer incorrectly assessing the driver's risk, which can generate patterns of searches that resemble bias even when the officer is unbiased. Hence, the concern is that tests for bias conflate inaccurate beliefs with racial bias. Given the diffi-

[11] 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.

[12] See also Simoiu et al. (2017), Pierson et al. (2018), Pierson et al. (2020), and Chan et al. (2019), who impose similar parametric restrictions to identify thresholds of decision makers.

culty of distinguishing between the two without knowing the decision maker's beliefs, researchers have begun using experiments to elicit beliefs of decision makers when studying discrimination (Bohren et al., 2020). In the case of observational data, only Hull (2021) provides sufficient conditions to reject the hypothesis of accurate beliefs when using the marginal outcome test, although the framework in this paper permits a similar test for inaccurate beliefs.

3 Model

In this section, I model the search decision of a single officer (he) for drivers who are stopped (she). The analysis allows for heterogeneity across officers, but I suppress the officer subscripts for brevity. I also suppress the notation indicating the analysis is conditional on drivers who are stopped. This conditioning is important to keep in mind, though, as it affects the interpretation of the model's assumptions.

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. Some of the components in V_i may be observed by the officer prior to the stop; some components may also be observable to one officer but not another, so that officers vary in their perceptiveness and form different beliefs about the driver's risk. The researcher only observes R_i but not V_i; any other observed characteristics are implicitly conditioned on throughout.

The driver may carry contraband (e.g., drugs, weapons), denoted by $Guilty_i \in \{0, 1\}$, but the officer does not know this unless he searches the driver, denoted by

$Search_i \in \{0, 1\}$. At the end of each traffic stop, the officer reports whether or not contraband was found, referred to as a 'hit',

$$Hit_i \equiv Search_i \times Guilty_i,$$

which is observed by the researcher. It is therefore assumed that the officer finds contraband if and only if he searches a guilty driver.

Finally, drivers are assumed to be drawn from a distribution that depends on the setting of the stop, $Z_i \in \mathcal{Z}$. For example, Z_i may consist of the hour and day of the stop, and the interpretation 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.[13] The setting is observed by both the officer and researcher, and will play the role of an instrument.

When deciding whether to search, the officer considers the four possible outcomes of his decision: (i) searching an innocent driver; (ii) searching a guilty driver; (iii) not searching an innocent driver; and (iv) not searching a guilty driver. Associated with each outcome is an *ex post* utility that the officer learns after interacting with the driver and observing all of her characteristics, but prior to making his search decision. Let $\mathcal{U}_i^s(g; r)$ denote this utility when $Search_i = s$ and $Guilty_i = g$ for drivers with race $R_i = r$. Note that these utilities are random and can vary across drivers who are observationally equivalent to the officer. The distributions of these utilities represent the officer's search preferences, and the objective of the test is to

[13] If there are variables that inform the officer's stop decision and are visible for one value of Z_i but not another (e.g., race is visible during the day before stopping a driver, but is not visible at night), then the distribution of drivers stopped will vary with Z_i even if the composition of drivers on the road do not. This type of variation is used in the Veil of Darkness test by Grogger and Ridgeway (2006) to test whether race affects the stop decision.

detect whether race has a direct effect on these distributions. To do this, I make the following assumption about the utilities $\{\mathcal{U}_i^0(g;r), \mathcal{U}_i^1(g;r)\}_{(g,r)\in\{0,1\}\times\{w,m\}}$, which I denote by $\{\mathcal{U}_i^s\}$ for brevity.

Assumption 1.

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

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

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

Assumption 1(i) states that the officer prefers to make the correct decision by searching guilty drivers and not searching innocent drivers. This implies that officers are more likely to search drivers who are more likely to carry contraband. This would be violated if the officer instead prefers to make the wrong decision by searching innocent drivers and releasing guilty drivers.

Assumption 1(ii) states that the utilities are drawn from a common distribution, which allows me to pool the drivers of the same race together to infer the officer's preferences. Conditioning the analysis on variables that affect the distribution of utilities (e.g., age and sex of the driver) helps to satisfy this assumption. If instead $\{\mathcal{U}_i^s\}$ and $\{\mathcal{U}_{i'}^s\}$ were drawn from different distributions for every $i \neq i'$, then the researcher must infer the search preference faced by each driver using one observation only.

Assumption 1(iii) is the key assumption of the model and determines how racial bias is defined and how it may be detected, so I spend more time discussing its parts. The independence between the utilities $\{\mathcal{U}_i^s\}$ and the setting Z_i is the exogeneity assumption for the instrument, whose role is to shift the distribution of risk (by changing the drivers stopped) without shifting the officer's preferences. This variation

is what allows me to learn about the officer's preference. Moreover, this instrument enables the test to be applied to each officer separately since the variation in risk may be generated within officer. This instrument separates my approach from those using random assignment of decision makers as the instrument (Arnold et al., 2018, 2020), which instead aim to shift the officer's preferences without shifting the distribution of risk. Such an instrument is harder to justify in the setting of police searches since officers choose their distribution of risk by choosing who to stop. Randomly assigning officers to roads does not resolve this problem, as it does not guarantee that officers stop the same drivers.

The independence between the utilities and whether the driver is guilty means that the officer can only infer how likely a driver carries contraband using the characteristics of the driver and setting of the stop, but not his utilities. This rules out clairvoyance, where the officer infers the driver's guilt using information beyond what is provided by the driver.

Finally, the independence between the utilities and the unobserved characteristics of the driver allows the researcher to make a direct link between officer preferences and race. That is, any dependence between the officer's preferences and the driver is through the race of the driver. This part of Assumption 1(iii) is crucial to any test of racial bias and has generated discussion among researchers. I elaborate on this point after defining the officer's search decision rule.

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

Definition 1. *The officer is* racially biased in traffic searches *if* $(\mathcal{U}_i^0(0;w), \mathcal{U}_i^0(1;w), \mathcal{U}_i^1(0;w), \mathcal{U}_i^1(1;w))$ *and* $(\mathcal{U}_i^0(0;m), \mathcal{U}_i^0(1;m), \mathcal{U}_i^1(0;m), \mathcal{U}_i^1(1;m))$

do not share the same distribution.

The objective of the test is thus to determine whether the distribution of utilities depends on race.

3.2 Search decision

To map the preferences of the officer to the data, I assume the officer chooses the search decision that maximizes his utility. Since the driver's guilt is not known to him, he chooses the decision that maximizes his expected utility,

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

where

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

is the officer's belief of how likely the driver caries contraband, which I refer to as the 'risk' of the driver. His search decision may then be written as

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

where

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

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

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

Corollary 1.

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

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

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

Proof. See Appendix A. ■

So instead of comparing distribution of $\{\mathcal{U}_i^s\}$ across races to detect bias, it suffices to compare T_i across races.[14]

This approach of comparing thresholds to detect bias is standard in the literature. Many papers model the choice of the decision maker as in (1), except their thresholds are deterministic functions of race (Knowles et al., 2001; Anwar and Fang, 2006; Arnold et al., 2018), which corresponds to the special case where $T_i \mid R_i = r$ is degenerate for $r \in \{w, m\}$.[15] But in all the papers, the threshold is independent of the unobserved characteristics of the driver after conditioning on race. This independence

[14]Types of biases that generate identical thresholds will be impossible to detect. For instance, let the constant $\overline{u}$ denote the bias, and suppose $\mathcal{U}_i^s(g; w) = \mathcal{U}_i^s(g; m) + \overline{u}$ for $(s, g) \in \{0, 1\}^2$. Then $\overline{u}$ drops out of T_i, and the officer has the same decision rule for both groups of drivers. I ignore these cases since the bias neither affects the search decisions nor the impact on drivers.

[15]A threshold that is a deterministic function of race can be obtained by assuming $\{\mathcal{U}_i^s\}$ are degenerate random variables. Another way is to begin with decision rule (1) and then assume that

property is crucial to any test for racial bias, as it allows the researcher to make a direct link between an officer's preferences and a driver's race. Without it, any dependence between the utilities and race may be argued to be a result of confounding variables in V_i (Canay et al., 2020).

To provide some justification for this independence assumption, I refer to the MNPD manual, which states that "individuals shall only be subjected to stops, seizures, or detentions upon reasonable suspicions" based on "specific and articulable facts, which taken together with rational inferences from those facts, reasonably warrant an officer to believe that criminal activity is afoot."[16] So conditional on the risk of the driver, the officer should be impartial to drivers with different characteristics. However, the researcher may expect this mandate to be violated by officers letting their decisions be influenced by their preferences toward certain characteristics. For example, an officer may feel differently about searching a young male driver versus an elderly female driver. But for condition (ii) in Corollary 1 to hold, an officer may *only* be partial to the race of the driver. The analysis must therefore be conditional on all non-race characteristics of the driver that the officer is partial to.

The unobserved characteristics V_i then only affect the search decision through the risk of the driver. This in turn implies that omitted variables, sample selection, and statistical discrimination—the usual confounders of bias—only affect the search decision through $G(R_i, Z_i, V_i)$. The econometric challenge of detecting bias is thus to separately infer the distributions of T_i and $G(R_i, Z_i, V_i)$; I defer this discussion to Section 4.

the threshold is a deterministic function of race. Hull (2021) provides conditions where such a decision rule is a normalization of a more general decision rule ranking drivers in the order they will be searched.

[16]Section 4.40 of Metropolitan Nashville Police Department Manual (2018).

To elaborate on how the three confounders affect the distribution of risk, consider an example where V_i is the color of the car, and red cars are more likely to contain contraband. Omitted variable bias pertains to differences in the distribution of V_i across races in population, e.g., whites are twice as likely to drive red cars than minorities in population. So even if the officer stopped drivers at random, the distribution of risk may differ across race since the underlying determinant V_i differs across race. Sample selection pertains to differences in the distribution of V_i across races for drivers who are stopped, e.g., the officer may prefer to stop minority drivers in red cars, so conditional on being stopped, whites are only half as likely to be in red cars than minorities, despite how whites are twice as likely to drive red cars in population. Finally, statistical discrimination (in the sense of Aigner and Cain (1977)) pertains to how V_i maps to risk differently for white and minority drivers, e.g., red cars are correlated with possessing contraband for whites, but not for minorities. This notion of statistical discrimination also extends to other officers, where different officers observe different components of V_i (Arnold et al., 2020; Hull, 2021). For example, an experienced officer may know to consider the direction of travel along a highway when assessing the driver's risk (Barnes, 2004), whereas an inexperienced officer may not. Since the test may be applied to each officer separately, I place no restrictions on how different officers infer the risk of the driver.

The instrument Z_i also affects the search decision exclusively through the risk of the driver. But unlike the confounders of bias, the variation in risk generated by Z_i is helpful in partially identifying the distribution of T_i. The intuition for this is that, by seeing how an officer makes his search decisions in a variety of settings, I can build a profile for the types of drivers he likes to search, and then compare the profiles for white and minority drivers. In Section 4, I show it is possible to detect bias without Z_i by only using the variation in search decisions generated by R_i. However, such a

test will be weak.

Notice that, conditional on race, Z_i may shift $G(R_i, Z_i, V_i)$ in two ways. The first is through shifting the distribution of V_i, e.g., $G(R_i, Z_i, V_i) = G(R_i, V_i)$ but $Z_i \not\perp\!\!\!\perp V_i \mid R_i$. An example of this is if the time of the traffic stop contains no information on whether the driver is guilty, but criminals and drug dealers tend to drive at night. The second way Z_i may shift risk is to have a direct effect on $G(R_i, Z_i, V_i)$, i.e., $G(R_i, z_1, V_i) \neq G(R_i, z_2, V_i)$ for $z_1 \neq z_2$. This reflects how the same signals can be interpreted differently depending on the setting of the stop (Engel and Johnson, 2006; Novak and Chamlin, 2012). For example, stopping a white driver in a predominantly white suburb may not arouse much suspicion, whereas stopping the same driver in a predominantly Black neighborhood may lead to more questions. Similarly, stopping a high school student in the afternoon shortly after school has ended is less suspicious than stopping the same student late into the night.

Under Assumption 1 and decision rule (1), the probability that a driver is searched only depends on the race and the risk of the driver,

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

where the third equality follows from Assumption 1(iii), and F_X denotes the CDF of random variable X. For each level of risk $g \in [0, 1]$, the officer's bias may be measured

by

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

with $\beta(g) > 0$ indicating bias against minority drivers with risk g. Since $\beta(g)$ can vary with g and even change sign, the intensity and direction of bias can vary with the unobserved (to the researcher) risk of the driver. This feature of the model arises from the random threshold and distinguishes it from earlier models, under which the officer searches all drivers with a given level of risk or none at all (Knowles et al., 2001; Anwar and Fang, 2006; Arnold et al., 2018; Hull, 2021). A random threshold thus extends the notion of the marginal driver to every level of risk and permits a more nuanced analysis of bias.[17] I show in Section 4 how sharp bounds on $\beta(\cdot)$ may be derived.

A concern with all tests of racial bias, including this one, is the accuracy of the decision maker's beliefs and whether it is possible to distinguish between inaccurate beliefs and racial bias. To illustrate the problem, suppose an unbiased officer incorrectly believes minority drivers are twice as risky as they truly are. His search decision may then be written as

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

where $\widetilde{T}_i \equiv T_i/(1 + \mathbb{1}\{R_i = m\})$. So in this example, the effect of inaccurate beliefs is observationally equivalent to the officer drawing thresholds that are half as large

[17] But if the researcher wishes to maintain a fixed threshold, the methods I propose can also accommodate this.

for minorities compared to whites. The test may then incorrectly detect bias since $\widetilde{T}_i \not\perp R_i$, despite how T_i is the true object of interest. Nevertheless, these tests for bias are still valuable since the effects of inaccurate beliefs and bias are the same for drivers. The test may serve as a preliminary check to determine which officers ought to be reviewed, and further investigation may reveal whether differences in search behavior stem from bias or inaccurate beliefs. In the next section, I also show how certain cases of inaccurate beliefs can be detected using the proposed methods.

4 Testing for racial bias

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

4.1 Defining the test

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

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

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

See Appendix A for the derivations. The instrument Z_i varies the search and hit rates by varying the distributions of risk. From the ratio of these rates, I also obtain the conditional hit rate, which is the probability that contraband is found conditional on a traffic search,

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

To define the identified set of the model, let $\mathcal{F}$ denote the space of distributions of $(R_i, Z_i, V_i, T_i, Guilty_i)$ satisfying Assumption 1. The sharp identified set is then

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

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 instead consider a projection of the identified set when testing for bias. To define this projection, I introduce the following notation,

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

where G_i denotes the risk in stop i; $\sigma(\cdot; r)$ denotes the probability a driver with of race r is searched conditional on her risk, and represents the officer's search preference for race r. The parameters of interest are then $\sigma(\cdot; w)$ and $\sigma(\cdot; m)$, and the distribution of risk conditional on race is

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

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 parameters of interest is

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

A test for racial bias immediately follows from (6).

Corollary 2. *Under (1) and Assumption 1, if there does not exist $\sigma^\star \in \Sigma$ such that $(\sigma^\star, \sigma^\star) \in \Sigma^\star$, then the officer is biased.*

Proof. By the definition of an unbiased officer (see Definition 1 and property (iii) of Corollary 1), the contrapositive of the corollary is true. ■

Since $\Sigma^\star$ is sharp, Corollary 2 is the strongest testable implication of the model for unbiasedness. As discussed below, having variation in the search and hit rates across Z_i strengthens the test by reducing the size of $\Sigma^\star$.

4.2 Building intuition for the test

4.2.1 When $|\text{supp}\,(G)\,| = 2$

In this section, I show how the variation in search and hit rates can reveal whether an officer is biased. But instead of inferring whether $\sigma(\cdot; r)$ is the same for $r \in \{w, m\}$, it is easier and equivalent to infer whether

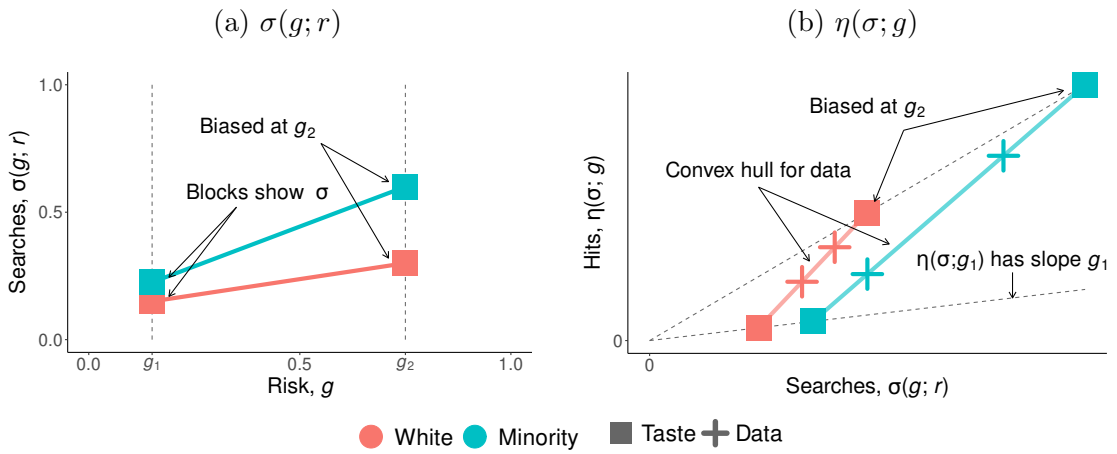
$$\eta(\sigma(\cdot; r); g) \equiv g\ \sigma(g; r)$$

depends on r, where η maps the probability a driver is searched to the probability of a hit, conditional on the risk of the driver. More intuitively, η maps the volume of traffic searches to the volume of hits for drivers with a certain level of risk, and is akin to a constant-returns-to-scale production function, e.g., searching drivers with risk g twice as often will double the amount of contraband recovered from them. The advantage of comparing η across races instead of σ is that η has a convenient geometric relationship with the data that lessens the computational burden of the test. I show in Appendix A that $\sigma(\cdot; r)$ depends on r if and only if $\eta(\sigma(\cdot, r); \cdot)$ depends on r.

To provide a simple illustration of how Z_i can reveal the officer's preferences, suppose that drivers stopped are either low- or high-risk so that $\text{supp}\,(G) = \{g_1, g_2\}$ with $g_1 < g_2$. The left panel of Figure 1 displays $\sigma(\cdot; w)$ and $\sigma(\cdot; m)$ for an officer, and the blocks indicate how often the officer searches a driver conditional on her race and risk. Since the configurations of the blocks differ for white and minority drivers, it follows that $\sigma(\cdot; w) \neq \sigma(\cdot; m)$ and the officer is biased.

The right panel displays $\eta(\sigma(\cdot; w); \cdot)$ and $\eta(\sigma(\cdot; m); \cdot)$. The dashed lines depict $\eta(\sigma(\cdot; w); g)$ for $g \in \{g_1, g_2\}$, and show how much contraband an officer will recover if he exclusively searches drivers with a given level of risk. Both of the dashed lines

Figure 1: Testing for bias when $|\text{supp}\,(G)\,| = 2$



intersect the origin because no contraband will be found if no searches occur; and both lines are linear since every additional search is equally likely to uncover contraband, with the slopes being equal to the risk of the driver. How far the blocks lie along the dashed lines indicate how often the low- and high-risk drivers are searched. If the positions of the blocks differ by race in the left panel, then they will also differ in the right panel, which is why bias may be detected by comparing $\eta(\sigma(\cdot;r);\cdot)$ across races.

To see how Z_i is informative of preferences, consider the data points

$$\mathcal{D}(r) \equiv \{(\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\}, \mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\})\}_{z \in \mathcal{Z}}$$

for $r \in \{w, m\}$, depicted by the crosses in the right panel of Figure 1. As implied by (4)–(5), for each race, the data points are convex combinations of the blocks and must therefore lie inside the convex hull of the blocks. Since there are only two levels of risk/blocks for each race, the convex hulls are simply the colored lines in the right panel. The higher up a data point lies along the line, the greater the proportion of high-risk drivers stopped at that setting.

If the officer is unbiased, then the position of the blocks in both panels of Figure 1

should be the same across races. This means that $\mathcal{D}(w)$ and $\mathcal{D}(m)$ should lie inside the same convex hull/along the same line. To show this, let

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

denote the distribution of risk conditional on the race of the driver and setting of the stop. Then the search and hit rates for an unbiased officer are

$$\begin{aligned}
\mathbb{P}\{\mathit{Search}_i \mid R_i = r, Z_i = z\} &= \sigma^\star(g_1)\ p_{r,z}(g_1) + \sigma^\star(g_2)\ (1 - p_{r,z}(g_1)), \\
\mathbb{P}\{\mathit{Hit}_i \mid R_i = r, Z_i = z\} &= g_1\ \sigma^\star(g_1)\ p_{r,z}(g_1) + g_2\ \sigma^\star(g_2)\ (1 - p_{r,z}(g_1)),
\end{aligned}$$

for some race-neutral $\sigma^\star \in \Sigma$. The linear relationship that the search and hit rates have with $p_{r,z}(g_1)$ indicate that the data lie on a line, and that variation in the data only stems from differences in the composition of drivers. Since the data in Figure 1 do not lie on a line, the officer is revealed to be biased.

In this special case where $|\text{supp}\,(G)\,| = 2$, the officer's preferences are summarized by the colored lines containing the data, and the IV enables us to 'trace' out these preferences, similar to how an IV may be used to trace out a demand curve. This result can also be viewed as a control function approach, since the instrument provides a way to condition on the unobserved risk of the drivers. Regardless, when $|\text{supp}\,(G)\,| = 2$, a simple IV regression is able to detect bias.

Proposition 1. *Suppose* $|\text{supp}\,(G)\,| = 2$ *and* $Var[\mathit{Search}_i \mid R_i = r, Z_i] > 0$ *for* $r \in \{w, m\}$. *Then*

$$\mathbb{E}[\mathit{Hit}_i \mid R_i = r, Z_i = z] = \alpha_0(r) + \alpha_1(r)\ \mathbb{E}[\mathit{Search}_i \mid R_i = r, Z_i = z],$$

where $\alpha_0(r) \leq 0$ *and* $\alpha_1(r) > 0$ *for* $r \in \{w, m\}$. *The coefficients* $\alpha_0(r)$, $\alpha_1(r)$ *are identified by an IV regression of* Hit_i *on* $Search_i$, *using* Z_i *as an instrument.*

(i) *If* $\alpha_0(w) \neq \alpha_0(m)$ *or* $\alpha_1(w) \neq \alpha_1(m)$, *then the officer is biased.*

(ii) *If* $\alpha_0(r_1) > \alpha_0(r_2)$ *and* $\alpha_1(r_1) > \alpha_1(r_2)$, *then the officer is biased against race* r_2.

Proof. See Appendix A. ■

Notice how condition (ii) in Proposition 1 is required to infer the direction of bias, since bias may now vary with unobserved risk. I show later what may be learned about the direction and intensity of bias at each level of risk.

If the researcher does not have an instrument, Proposition 1 still offers a way to test for bias.

Corollary 3. *Suppose* $|\text{supp}\,(G)\,| = 2$ *and* $Var[Search_i \mid R_i] > 0$ *for* $r \in \{w, m\}$. *Consider the regression of* $\mathbb{E}[Hit_i \mid R_i]$ *on* $\mathbb{E}[Search_i \mid R_i]$ *and an intercept so that*

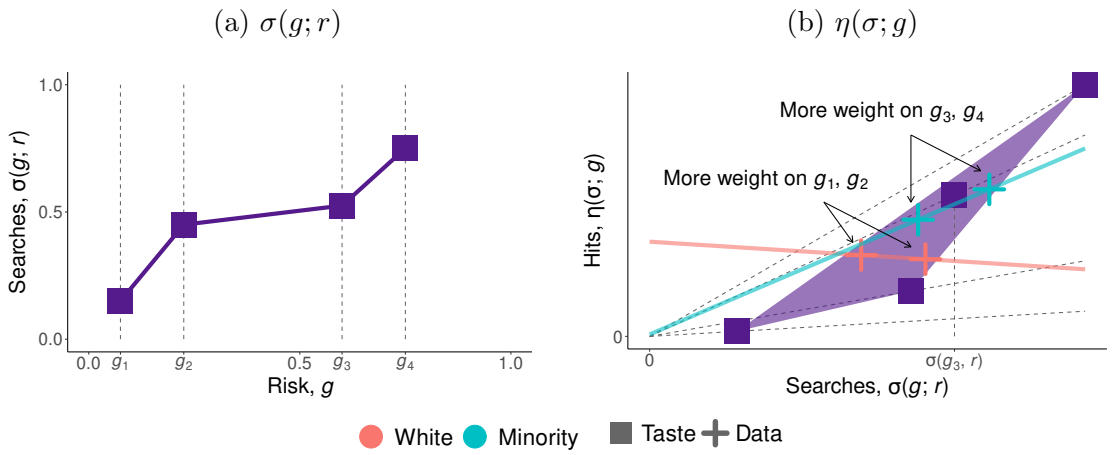
$$\mathbb{E}[Hit_i \mid R_i = r] = \alpha_0 + \alpha_1 \mathbb{E}[Search_i \mid R_i = r].$$

If $\alpha_0 > 0$ *or* $\alpha_1 \leq 0$, *then the officer is biased.*

Proof. See Appendix A. ■

The intuition behind Corollary 3 is that an unbiased officer should search minorities more often than whites only if minorities have more high-risk drivers. If this is indeed the case, then the officer should experience more hits with minorities, and there should be a positive relationship between the search and hit rates. If instead the relationship is weak or negative, then the higher search rates cannot be due to there being more high-risk drivers and must be due to bias.

Figure 2: An unbiased officer when $|\text{supp}\,(C)\,| > 2$



4.2.2 When $|\text{supp}\,(G)\,| > 2$

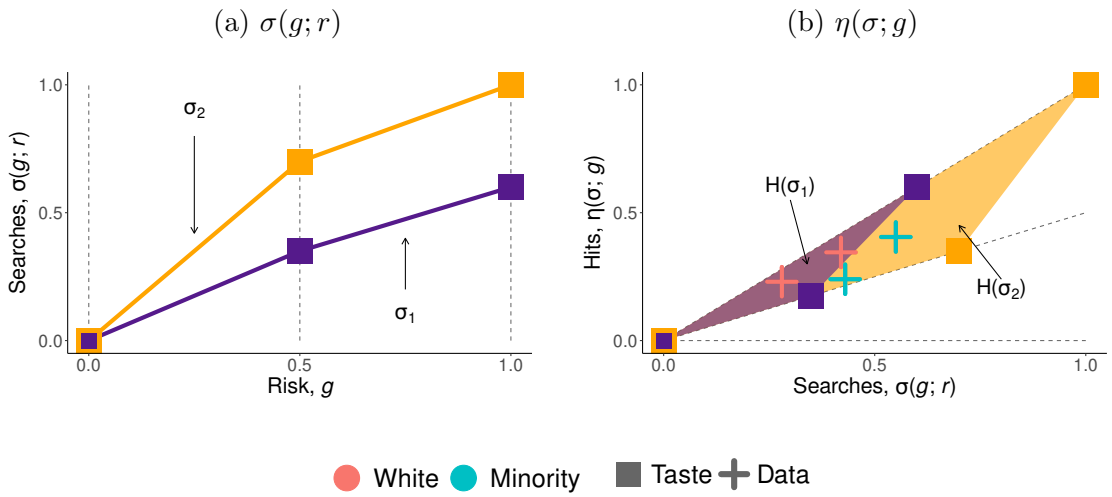
Testing for bias when $|\text{supp}\,(G)\,| > 2$ is much more difficult. The IV result in Proposition 1 does not extend to this setting because the convex hulls containing the data need not be lines, so $\mathcal{D}(w)$ and $\mathcal{D}(m)$ need not lie on the same line for an unbiased officer. An example of this is provided in Figure 2, which displays the preferences and data for unbiased officer, where the purple triangle in the right panel represents the convex hull of the blocks containing the data.

To highlight another challenge, as well as illustrate how officer preferences may only be partially identified, notice that the convex hull in Figure 2 contains the block corresponding to $\sigma(g_3; r)$. While the vertices of the convex hull may be identified at infinity, the block inside of the convex hull cannot. This means that the officer's preference for searching drivers with $G_i = g_3$ may never be recovered, even with infinite values of Z_i. Bias among drivers with $G_i = g_3$ may therefore remain undetected.

To show how Corollary 2 may be used to detect bias in this general setting, define

$$\mathcal{H}(\sigma(\cdot; r)) \equiv \text{conv}\left(\{(\sigma(g; r), g\ \sigma(g; r))\}_{g \in \text{supp}(G)}\right)$$

Figure 3: Applying Corollary 2 when $|\text{supp}\,(C)\,| = \{0, 0.5, 1\}$



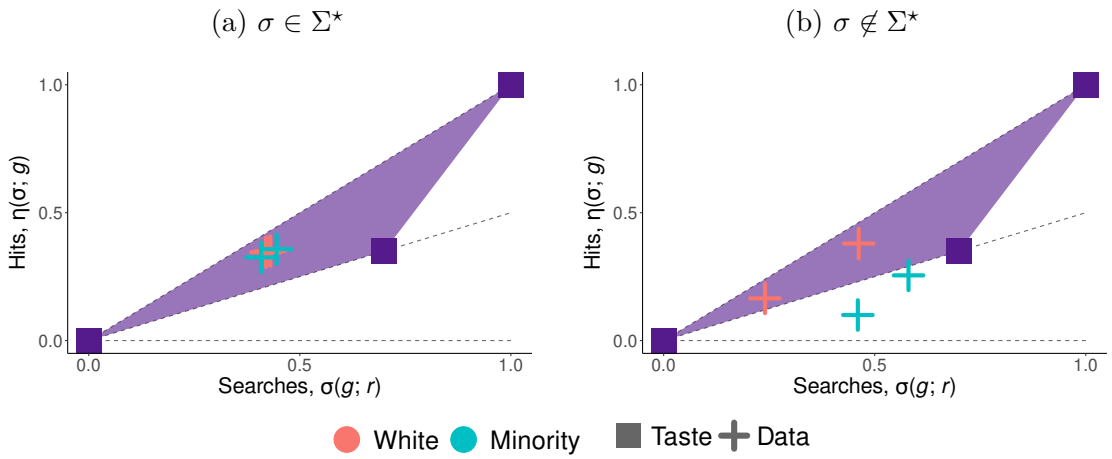
to be the convex hull generated by preference $\sigma(\cdot; r)$ that is supposed to contain the data $\mathcal{D}(r)$ for $r \in \{w, m\}$. In Figure 2, $\mathcal{H}(\sigma(\cdot; r))$ is depicted by the purple triangle. Corollary 2 asks whether there exists a race-neutral $\sigma^\star \in \Sigma$ such that

$$\mathcal{D}(w), \mathcal{D}(m) \subseteq \mathcal{H}(\sigma^\star). \tag{7}$$

If so, the officer may be unbiased; and if not, the officer must be biased. Figure 3 provides an example of how this test may be performed by iteratively checking whether $\sigma \in \Sigma$ satisfy (7). In practice, iterating over the elements in Σ is an inefficient way to test for bias, but it is helpful at conveying the idea of the test; I show below how Corollary 2 may be implemented by solving a bilinear programming (BP) problem instead. Figure 3 shows that $\sigma_1 \in \Sigma$ (purple triangle) is not a suitable candidate for $\sigma^\star$ since $\mathcal{D}(m) \not\subseteq \mathcal{H}(\sigma_1)$. In contrast, $\sigma_2 \in \Sigma$ (orange triangle) is a possible candidate for $\sigma^\star$ since $\mathcal{D}(w), \mathcal{D}(m) \subseteq \mathcal{H}(\sigma_2)$. The officer in this example thus passes the test, and it is feasible for him to be unbiased.

Figure 4 demonstrates how the strength of the test depends on the variation in

Figure 4: How variation in $\mathcal{D}(w)$, $\mathcal{D}(m)$ affects the strength of the test



$\mathcal{D}(w)$ and $\mathcal{D}(m)$. If there is little variation and the data are clustered around a point, then it is easy to find a $\sigma^\star \in \Sigma$ such that $\mathcal{D}(w), \mathcal{D}(m) \subseteq \mathcal{H}(\sigma^\star)$ and hard to detect bias. If instead there is a lot of variation in the data, then it becomes more difficult to find such a $\sigma^\star \in \Sigma$ and therefore easier to detect bias.

4.3 Applying Corollary 2 using a bilinear program

In this section, I show how Corollary 2 may be implemented as a bilinear program. For the problem to be computationally feasible, I continue to discretize G_i so that $\text{supp}\,(G) = \{g_0, \ldots, g_K\}$ for $K < \infty$ and

$$\mathbb{E}[Search_i \mid R_i = r, Z_i = z] = \sum_{k=0}^{K} \sigma(g_k; r)\ p_{r,z}(g_k),$$

$$\mathbb{E}[Hit_i \mid R_i = r, Z_i = z] = \sum_{k=0}^{K} g_k\ \sigma(g_k; r)\ p_{r,z}(g_k).$$

However, under appropriate restrictions on the model, the BP problem is feasible even when G_i is continuous; I discuss this in Section 4.4.

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

$$\begin{aligned}
\mathbf{m}^S_{r,z} &\equiv \mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z\} \\
\mathbf{m}^H_{r,z} &\equiv \mathbb{P}\{Hit_i = 1 \mid R_i = r, Z_i = z\} \\
\mathbf{g} &\equiv \{g_0, \ldots, g_K\} \\
\varsigma &\equiv (\sigma^\star(g_0), \ldots, \sigma^\star(g_K)) \\
\mathbf{p}_{r,z} &\equiv (p_{r,z}(g_0), \ldots, p_{r,z}(g_K))
\end{aligned}$$

The objects $\mathbf{m}^S_{r,z}$, $\mathbf{m}^H_{r,z}$ are the search and hit rates for each race r and setting z; and the vector $\mathbf{g}$ is the support of G_i. I assume these three objects are known to the researcher. The unknown parameters of the BP problem are ς, which are the values of $\sigma^\star(\cdot)$ 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 notational brevity, I refer to the distributions of risk by $\{\mathbf{p}_{r,z}\}$. The objective of the problem is to optimize over these parameters to match the conditional moments $\mathbf{m}^S_{r,z}$, $\mathbf{m}^H_{r,z}$.

To ensure that the parameters of the model are consistent with their definitions, I impose two baseline constraints. The first is that

$$0 \leq \varsigma_k \leq \varsigma_{k+1} \leq 1 \text{ for } k = 0, \ldots, K-1,$$

where ς_k denotes the k^{th} component of ς. This ensures $\sigma^\star \in \Sigma$, as required by Corollary 2. The second is that

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

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 all $(r,z) \in$

$\{w, m\} \times \mathcal{Z}$, as required by the definition of $\Sigma^\star$. These restrictions are linear and may be written as

$$\mathbf{A} \begin{bmatrix} \varsigma \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

where matrix $\mathbf{A}$ and vector $\mathbf{b}$ characterize the restrictions (see Appendix B for more details). 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.

Corollary 2 may then be implemented as follows.

Proposition 2. *Define the criterion $Q^\star$ as the solution to the following BP program,*

$$Q^\star \equiv \min_{\varsigma, \{\mathbf{p}_{r,z}\}} \sum_{r,z} \left|\varsigma' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z}\right| + \sum_{r,z} \left|(\mathbf{g} \odot \varsigma)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z}\right| \tag{8}$$

$$\text{s.t.} \quad \mathbf{A} \begin{bmatrix} \varsigma \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

where $\odot$ denotes the Hadamard (element-wise) product. The officer is biased if $Q^\star > 0$.

Proof. If $Q^\star > 0$, then there exists an $(r, z) \in \{w, m\} \times \mathcal{Z}$ such that $|\varsigma' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z}| > 0$ or $|(\mathbf{g} \odot \varsigma)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z}| > 0$. Then there does not exist a $\sigma^\star \in \Sigma$ such that (4)–(5) are satisfied for all $(r, z) \in \{w, m\} \times \mathcal{Z}$. Then by Corollary 2, the officer is biased. ■

The criterion $Q^\star$ in Proposition 2 is the minimum ℓ_1-norm between the moments

of the model and the moments of the data. In theory, other norms may be used, but I choose the ℓ_1-norm because it leaves (8) quadratic, which is easier to solve.[18]

Note that any of the constraints in (8) may be tested. To do this, define

$$\varsigma_r \equiv (\sigma(g_0; r), \cdots, \sigma(g_K; r))$$

to be the officer's preference for race r, and consider the BP problem of matching only the moments for race r when the constraints are defined by $(\mathbf{A}, \mathbf{b})$,

$$Q^\star_{C,r}(\mathbf{A}, \mathbf{b}) \equiv \min_{\varsigma_r, \{\mathbf{p}_{r,z}\}} \sum_z \left|\varsigma_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z}\right| + \sum_z \left|(\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z}\right| \tag{9}$$

$$\text{s.t.} \quad \mathbf{A} \begin{bmatrix} \varsigma_r \\ \mathbf{p}_{r,1} \\ \vdots \\ \mathbf{p}_{r,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}.$$

Suppose that the set of constraints $(\mathbf{A}_1, \mathbf{b}_1)$ is a strict subset of $(\mathbf{A}_2, \mathbf{b}_2)$, i.e., the rows of $\mathbf{A}_1$ are a strict subset of those of $\mathbf{A}_2$, and likewise for $\mathbf{b}_1$ and $\mathbf{b}_2$. Then if $Q^\star_{C,r}(\mathbf{A}_1, \mathbf{b}_1) = 0 < Q^\star_{C,r}(\mathbf{A}_2, \mathbf{b}_2)$, the additional constraints in $(\mathbf{A}_2, \mathbf{b}_2)$ may be rejected.

For example, suppose $(\mathbf{A}_1, \mathbf{b}_1)$ and $(\mathbf{A}_2, \mathbf{b}_2)$ contain all the restrictions on ς_r and $(\mathbf{p}_{r,z})$ as $(\mathbf{A}, \mathbf{b})$ in (8), except $(\mathbf{A}_1, \mathbf{b}_2)$ excludes the monotonicity restriction on ς. Then if $Q^\star_{C,r}(\mathbf{A}_1, \mathbf{b}_1) = 0 < Q^\star_{C,r}(\mathbf{A}_2, \mathbf{b}_2)$, it means the monotonicity restriction cannot be satisfied and may be rejected. That is, the data reveals the officer is searching certain drivers with greater probability compared to other drivers with lower risk. There are two interpretations of this result. The first is that Assumption 1 is violated

[18] If I instead use the ℓ_2-norm, then (8) becomes quartic.

and the model is misspecified. The second is that Assumption 1 holds but the officer has inaccurate beliefs, which is similar to the interpretation used by Hull (2021) to detect inaccurate beliefs using the marginal outcome test.

4.4 Adding restrictions

The framework allows the researcher to strengthen the test by adding restrictions to Σ and $\mathcal{F}_G$ in a transparent, modular fashion. All additional restrictions may also be easily tested, as just described.

For example, if the researcher believes $\sigma^\star(\cdot)$ is smooth, then it can be modeled as a Bernstein polynomial, which has several convenient properties. First, it is highly flexible. Second, it is linear in its parameters, so the test remains as a BP problem. Third, restrictions on its range and derivatives take the form of linear constraints and are easy to impose.[19]

The framework also nests the earlier models in the literature where $Search_i = \mathbb{1}\{G_i \geq t(R_i)\}$ for some deterministic function t. These models effectively impose an

[19]The Bernstein basis of degree L is defined by

$$\mathbf{b}_l^L(g) \equiv \binom{L}{l}(1-g)^{L-l}g^l$$

for $l = 0, \ldots L$ and $g \in [0,1]$. So $\sigma^\star$ can be modeled as

$$\sigma^\star(g) = \sum_{l=0}^{L} \theta_l \mathbf{b}_l^L(g)$$

for some $\theta \equiv (\theta_0, \ldots, \theta_L)$. See Appendix B for a summary of Bernstein polynomials and their properties. See Farouki (2012) for a more detailed summary.

integrality constraint on ς so that

$$\sigma^{\star}(g_k) \in \{0,1\} \text{ for } k = 0, \dots K. \tag{10}$$

The researcher may also impose restrictions on the distributions of risk, $\{\mathbf{p}_{r,z}\}$. For example, restrictions such as (ranking across settings)

$$\mathbb{E}[G_i \mid R_i = r, Z_i = z_1] \leq \cdots \leq \mathbb{E}[G_i \mid R_i = r, Z_i = z_{|\mathcal{Z}|}].$$

and (ranking across races)

$$\mathbb{E}[G_i \mid R_i = w, Z_i = z] \lesseqgtr \mathbb{E}[G_i \mid R_i = m, Z_i = z].$$

are linear constraints and easy to impose.[20] It is also possible to model $\{\mathbf{p}_{r,z}\}$ as Bernstein polynomials.[21] If both ς and $\{\mathbf{p}_{r,z}\}$ are modeled as Bernstein polynomials, then the BP problem is feasible even when G_i is continuous—I will consider such a specification in future drafts.[22] But when making restrictions on $\{\mathbf{p}_{r,z}\}$, the researcher

[20] The inequality constraint $\mathbb{E}[G_i \mid R_i = r_1, Z_i = z_1] \leq \mathbb{E}[G_i \mid R_i = r_2, Z_i = z_2]$ may be written as

$$\mathbf{g}'\mathbf{p}_{r_1,z_1} \leq \mathbf{g}'\mathbf{p}_{r_2,z_2}.$$

[21] This is not an unreasonable assumption since the Beta density function, which is used to model random probabilities, is itself a Bernstein polynomial. This restricts the density function of risk to be smooth and allows the researcher to impose various shape restrictions.

[22] If ς and $\mathbf{p}_{r,z}$ are Bernstein polynomials, then their product—which is used to construct $\varsigma'\mathbf{p}_{r,z}$ in the objective function in the BP problem—is also a Bernstein polynomial. The unknown coefficients of this polynomial are bilinear functions of the unknown coefficients of ς and $\mathbf{p}_{r,z}$. Since the integral of any Bernstein basis polynomial of degree L over the unit interval is $(1+L)^{-1}$, Proposition 2 may

must bear in mind that the restrictions pertain to the risk of drivers stopped. So these restrictions should be justified by evidence in the data, or reasonable priors on the distribution of risk in the population and how sample selection interacts with this distribution.

To provide a visual example of how these restrictions strengthen the test, consider restricting the mass of drivers to be decreasing in risk,

$$\mathbf{p}_{r,z,k} \geq \mathbf{p}_{r,z,k+1} \text{ for } k = 0, \dots K-1 \text{ and } (r,z) \in \{w,m\} \times \mathcal{Z}. \tag{11}$$

This assumption is plausible as long as the mass of low-risk drivers in population is sufficiently large. See Appendix B for a simulated example of this restriction holding true even when the officer is much more likely to stop high-risk drivers.[23] Figure 5 shows how this restriction shrinks $\mathcal{H}(\sigma_2)$ from the example in Figure 3. So while it was previously feasible for the officer to be unbiased and have search preference σ_2, this is no longer true after imposing (11). In fact, there do not exist any race-neutral $\sigma^\star \in \Sigma$ capable of generating the data for both races while satisfying (11). The test thus detects bias under this restriction.

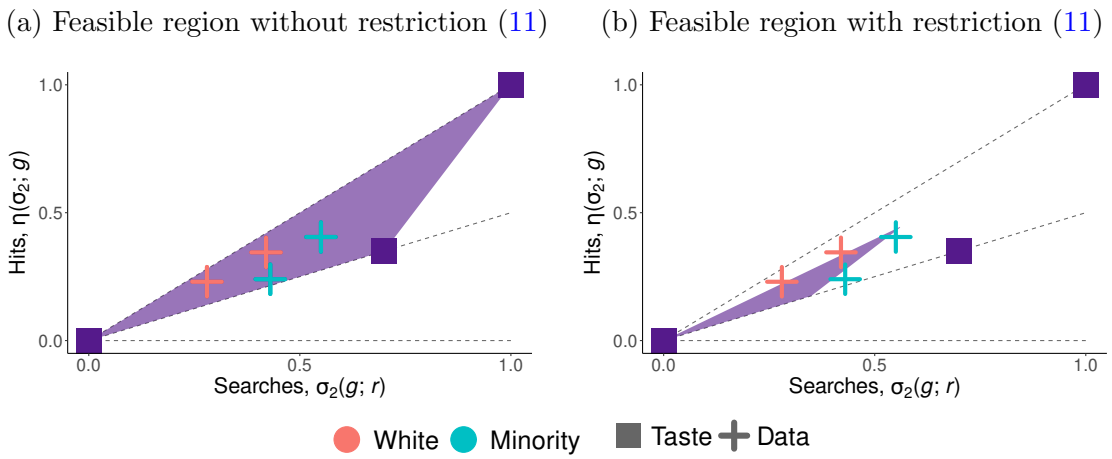
4.5 Determining the direction and intensity of bias

If the test detects bias, the next step is to determine *how* the officer is biased. This may be done in two ways. The first is to derive bounds on $\beta(g_k) \equiv \sigma(g_k;m) - \sigma(g_k;w)$

still be applied by solving a BP problem that is feasible even when G_i is continuous.

[23] It is not unreasonable to assume that most drivers in the population are low-risk. If the mass of low-risk drivers in population is sufficiently large compared to high-risk drivers, then the officer may still primarily stop low-risk drivers, even if he prefers to stop high-risk drivers.

Figure 5: Strengthening the test by restricting $\{\mathbf{p}_{r,z}\}$



for $k = 0, \ldots, K$, whose sharp identified set is

$$\mathcal{B}_k \equiv \{b \in \mathbb{R} : b = \sigma(g_k; m) - \sigma(g_k; w) \text{ for } (\sigma(\cdot; w), \sigma(\cdot; m)) \in \Sigma^\star\}.$$

The sharp bounds on $\beta(g_k)$ may then be obtained as follows.

Proposition 3. *For $g_k \in \text{supp}\,(G)$, the sharp bounds on $\beta(g_k)$ are obtained by solving the following BP problem,*

$$\begin{aligned} \beta_{\text{lb}}(g_k), \beta_{\text{ub}}(g_k) \equiv & \min/\max_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{w,z}\}, \{\mathbf{p}_{m,z}\}} \varsigma_{m,k} - \varsigma_{w,k} \\ \text{s.t.} \quad & \sum_{r,z} \left| \varsigma_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z} \right| + \sum_{r,z} \left| (\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z} \right| = 0 \\ & \mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}, \end{aligned} \tag{12}$$

where $\varsigma_{r,k}$ *is the* k^{th} *element of* ς_r.

Proof. The constraints in (12) characterize the sharp identified set $\Sigma^\star$, so the bounds are sharp by definition. ■

The bounds in Proposition 3 are sharp in the sense that they equal the smallest and largest values in the identified set $\mathcal{B}_k$. However, because bilinear programs are non-convex, $\mathcal{B}_k$ need not be the full interval $[\beta_{\text{lb}}(g_k), \beta_{\text{ub}}(g_k)]$, and may instead be a union of disjoint intervals contained in $[\beta_{\text{lb}}(g_k), \beta_{\text{ub}}(g_k)]$. I focus the discussion on the simpler bounds in Proposition 3, although $\mathcal{B}_k$ may be identified by 'inverting' (12), i.e., minimizing the criterion subject to the constraint that $\beta(g_k) = b$ for some $b \in \mathbb{R}$, where $b \in \mathcal{B}_k$ if and only if the criterion is zero. This procedure is similar to inverting a statistical test to construct a confidence interval. See Appendix B for more details.

If $\beta_{\text{lb}}(g_k) > 0$, then the officer is biased against minorities with risk g_k; and if $\beta_{\text{ub}}(g_k) < 0$, then the officer is biased against whites. So it is possible that the officer is biased against one race when risk equals g_1, but biased against the other when risk equals g_2. It is also possible for the officer to fail the test in Proposition 2, but have $\beta_{\text{lb}}(g) < 0 < \beta_{\text{ub}}(g)$ for all $g \in \text{supp}\,(G)$. For such an officer, I can detect that he is biased, but cannot determine the direction of bias. If the researcher has a strong prior on the direction of bias, then this prior may be imposed via a sign restriction on β, e.g., $\beta(g_k) \geq 0$ for all $k = 0, \ldots, K$ so that bias is always against minorities.[24]

The second way to measure the direction and intensity of bias is to average $\beta(g)$ across $g \in \text{supp}\,(G)$. To do this, the researcher must first choose a weight $\omega = (\omega_0, \ldots, \omega_K)$ such that $\omega_k \in [0, 1]$ for $k = 0, \ldots, K$ and $\sum_{k=0}^{K} \omega_k = 1$. The average

[24] The direction of bias can be fixed for a subset of values of risk instead.

bias is then defined as

$$\mathbb{E}[\beta(G_i);\omega] \equiv \sum_{k=0}^{K} \omega_k \ \beta(g_k),$$

and its sharp identified set is

$$\mathcal{E} \equiv \left\{ b \in \mathbb{R} : b = \sum_{k=0}^{K} \omega_k \ (\sigma(g_k; m) - \sigma(g_k; w) \text{ for } (\sigma(\cdot; w), \sigma(\cdot; m)) \in \Sigma^\star \right\}.$$

What $\mathbb{E}[\beta(G_i);\omega]$ measures is the average difference in the probability that equally risky white and minority drivers are searched, under the counterfactual where $\mathbb{P}\{G_i = g_k \mid R_i = r\} = \omega_k$ for $r \in \{w, m\}$.

Proposition 4. *For a choice of ω, the sharp bounds on $\mathbb{E}[\beta(G_i);\omega]$ are obtained by solving the following BP problem,*

$$\begin{aligned} \mathbb{E}[\beta(G_i);\omega]_{\text{lb}}, \mathbb{E}[\beta(G_i);\omega]_{\text{ub}} \equiv & \min/\max_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{w,z}\}, \{\mathbf{p}_{m,z}\}} \ \omega' (\varsigma_m - \varsigma_w) \\ \text{s.t.} \quad & \sum_{r,z} \left| \varsigma_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z} \right| + \sum_{r,z} \left| (\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z} \right| = 0 \\ & \mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}. \end{aligned} \tag{13}$$

Proof. The constraints in (13) characterize the sharp identified set $\Sigma^\star$, so the bounds are sharp by definition. ■

Similar to the bounds in Proposition 3, the bounds in Proposition 4 are sharp in

the sense that they equal the smallest and largest values in the identified set $\mathcal{E}$. See Appendix B for how $\mathcal{E}$ maybe fully recovered by inverting (13). These bounds may also be seen as a non-parametric, partial identification approach to the decomposition methods of Oaxaca (1973), Blinder (1973), and DiNardo et al. (1996).[25]

For many interesting counterfactuals, the weights ω may be functions of $\{\mathbf{p}_{r,z}\}$ and therefore be unknown. For example, the researcher may wish to know what the average bias is if minority drivers are equally risky as white drivers in the data. This corresponds to choosing

$$\begin{aligned}\omega_k &= \mathbb{P}\{G_i = g_k \mid R_i = w\} \\ &= \sum_{z \in \mathcal{Z}} \mathbb{P}\{G_i = g_k \mid R_i = w, Z_i = z\}\ \mathbb{P}\{Z_i = z \mid R_i = w\} \\ &= \sum_{z \in \mathcal{Z}} p_{w,z}(g_k)\ q_{w,z}, \end{aligned} \tag{14}$$

for $k = 0, \ldots, K$, where the second equality follows from law of iterated expectations, and $q_{w,z} = \mathbb{P}\{Z_i = z \mid R_i = w\}$ is observed in the data. More generally, the researcher can choose ω to be

$$\omega = \mathbf{P}_w \mathbf{q}_w + \mathbf{P}_m \mathbf{q}_m,$$

where $\mathbf{P}_r$ is an unknown $K \times |\mathcal{Z}|$ matrix whose l^{th} column is $\mathbf{p}_{r,z_l}$; and $\mathbf{q}_r$ is a known vector whose purpose is to weight the different settings Z_i for race r. Proposition 4

[25] These methods decompose average outcomes into structural and composition effects. By reweighting the structural effects, the authors are able to form counterfactuals. This is similar to how I decompose the search and hit rates into σ and $p_{r,z}$, and reweight σ to construct counterfactuals. See Fortin et al. (2011) for a summary of decomposition methods in economics.

continues to hold under this choice of ω since the objective function

$$\begin{aligned}\omega'(\varsigma_m - \varsigma_w) &= (\mathbf{P}_w\mathbf{q}_w + \mathbf{P}_m\mathbf{q}_m)' \ (\varsigma_m - \varsigma_w) \\ &= \sum_{r\in\{w,m\}} \sum_{l=1}^{|\mathcal{Z}|} \mathbf{q}_{r,l} \sum_{k=0}^{K} \underbrace{\mathbf{p}_{r,z,k} \ (\varsigma_{m,k} - \varsigma_{w,k})}_{\text{Bilinear terms}}\end{aligned}$$

is still bilinear.

5 Data

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 MNPD traffic stop data

Each observation in the data is a traffic stop made by an officer. The researcher observes the driver's race, age, sex, and state of registration, but does not observe sensitive information such as her license number and vehicle tag number. All information on the officer is hidden, other than an anonymized identifier. Logistic details of the traffic stop are observed and include the date, time, address, and geocoordinates of the stop. The researcher observes the reason for the traffic stop, whether a search occurred, why the search occurred, whether any contraband was found, and the outcome of the stop (i.e., arrest, citation, warning). However, there are no detailed descriptions of the interaction between the officer and driver.[26]

[26] For a small number of stops, a short note written by the officer summarizing the stop is available.

Although though the data categorize contraband into weapons and drugs, I treat all forms of contraband as being the same. This is because traffic searches are infrequent and typically unsuccessful, so there are relatively few traffic stops that uncover contraband. This makes it infeasible to evaluate the officer's search decision for weapons and drugs separately. It is also unknown whether the officer was searching for weapons or drugs to begin with.

I supplement the MNPD police traffic data with additional MNPD data on criminal incidents and calls for services.[27] The purpose of these data are to control for environmental variables that may correlate with the setting of the stop but also affect the officer's preference toward searching a driver. For the same reasons, I also include local measures of racial composition and median household income from the American Community Survey of the US Census Bureau.

5.2 Restricting the sample

To study bias in traffic searches, the search decision must be discretionary. So traffic searches motivated by rules or mandates are excluded from the study. This includes searches that are incidental to an arrest, inventory searches, and searches based on warrants.[28] In total, 28% of the traffic searches in the data must be discarded for potentially being non-discretionary.

I restrict my attention to the 50 officers with the highest number of traffic searches.

[27] I restrict both 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.

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

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

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 make 2,180 stops and 250 searches for white drivers, and 2,268 stops and 314 searches for minority drivers. Remarkably, this small fraction of officers make up a third of all searches in the data.

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

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

5.3 Context variable Z_i

The officer's search preferences are realized after he observes driver characteristics R_i, V_i and before he searches the driver. The options for Z_i depend on the determinants of the officer's preferences, which I assume include basic demographic variables (e.g., race, age, sex of the driver), the reason for the stop, the interaction with the driver,

and the surrounding environment. For example, the officer may feel less comfortable searching female drivers than male drivers, suggesting that female drivers face larger draws of T_i. A motorist stopped for reckless driving may earn the ire of an officer and thereby face lower draws of T_i. A charismatic driver may be able to dissuade the officer from searching, or at least discourage him, implying larger draws of T_i. Finally, patrolling a dangerous neighborhood may put the officer on edge, resulting in lower draws of T_i. For Z_i to satisfy the independence conditions in Assumption 1, it is necessary for me to control for the determinants of officer preferences that may be correlated with Z_i, as they may induce a correlation between Z_i and T_i.

In the application, 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 (7am–3pm), evening (3pm–11pm), or night (11pm–7am), giving me up to six values of Z_i for each officer. Below, I discuss the controls I use to support this assumption. Table 2 provides summary statistics for these controls, and Tables 3–4 show how they vary with Z_i.

The first set of controls are the observable (to the researcher) characteristics of the driver, i.e., race, age, sex, state of registration. To see why this is necessary, imagine that officers prefer not to inconvenience elderly female drivers by searching them, but do not feel such reservations toward college-age male drivers. Tables 3–4 show that drivers who are stopped late at night are younger and more likely to be male compared to drivers stopped earlier in the day. So if elderly females primarily drive in the mornings, and college-age males primarily drive late at night when there is no school, then Z_i may violate the independence assumption.

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. I categorize the reason for stop into three groups: driving-related reasons, non-driving related

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.

Table 3: Controls by Z_i, white drivers

	Weekday			Weekend		
	Morning	Evening	Night	Morning	Evening	Night
Driver characteristics						
Male	0.5792	0.6072	0.6521	0.5912	0.6230	0.6286
Age	39.46	36.37	34.43	40.98	35.89	32.01
Out of state	0.0684	0.0533	0.0615	0.0738	0.0654	0.0805
Reason for stop						
Driving	0.8686	0.8502	0.9413	0.8920	0.8948	0.9376
Non-driving	0.1212	0.1403	0.0395	0.0999	0.0862	0.0353
Investigation	0.0103	0.0094	0.0192	0.0081	0.0190	0.0271
Reason for search						
Plain view	0.1252	0.3814	0.6302	0.0781	0.5954	0.7899
Consent	0.7318	0.5345	0.3297	0.8438	0.3505	0.1759
Probable Cause	0.1430	0.0842	0.0401	0.0781	0.0541	0.0342
Location						
Highway	0.1445	0.0877	0.1279	0.0854	0.1049	0.1313
Precinct 1	0.0503	0.0568	0.1505	0.0517	0.1205	0.1449
Precinct 2	0.1429	0.1324	0.0550	0.0505	0.1179	0.0548
Precinct 3	0.0846	0.1011	0.1226	0.2909	0.1631	0.1197
Precinct 4	0.0300	0.0312	0.0527	0.0331	0.0478	0.1170
Precinct 5	0.4180	0.3815	0.2656	0.1545	0.2913	0.1969
Precinct 6	0.0572	0.0331	0.0149	0.0314	0.0289	0.0156
Precinct 7	0.1219	0.1736	0.1527	0.0935	0.0918	0.0846
Precinct 8	0.0949	0.0904	0.1859	0.2944	0.1388	0.2664
Census tract demographics						
Percent white	0.6044	0.5610	0.5986	0.5733	0.5689	0.6174
Median household income	51915	45154	48417	49006	45648	49590
Crime incident rate	0.0193	0.0434	0.0208	0.0049	0.0128	0.0205
Calls for MNPD services	0.0232	0.0251	0.0096	0.0087	0.0131	0.0197

Notes: Crime and call rates are per capita and are restricted to those pertaining to violent crimes, theft, or drugs. Rates for reasons for search are calculated using only stops involving searches. All other rates are estimated using all stops in the data.

Table 4: Controls by Z_i, minority drivers

	Weekday			Weekend		
	Morning	Evening	Night	Morning	Evening	Night
Driver characteristics						
Male	0.5606	0.5963	0.6663	0.6127	0.6111	0.6540
Age	36.12	34.56	32.89	37.84	34.06	31.45
Out of state	0.0350	0.0272	0.0368	0.0421	0.0323	0.0480
Reason for stop						
Driving	0.8630	0.8547	0.9305	0.8755	0.8984	0.9375
Non-driving	0.1229	0.1323	0.0472	0.1067	0.0874	0.0348
Investigation	0.0141	0.0130	0.0224	0.0177	0.0142	0.0276
Reason for search						
Plain view	0.1126	0.2034	0.3451	0.0439	0.2728	0.5067
Consent	0.6723	0.6319	0.5501	0.8772	0.5754	0.4083
Probable Cause	0.2151	0.1647	0.1048	0.0789	0.1518	0.0850
Location						
Highway	0.0780	0.0388	0.0930	0.0570	0.0483	0.0987
Precinct 1	0.0348	0.0272	0.1215	0.0306	0.0456	0.1123
Precinct 2	0.1623	0.2362	0.1008	0.0598	0.1764	0.1047
Precinct 3	0.1186	0.1216	0.1734	0.3652	0.2204	0.1963
Precinct 4	0.0227	0.0169	0.0271	0.0124	0.0265	0.0864
Precinct 5	0.2948	0.2509	0.2398	0.1063	0.2505	0.1842
Precinct 6	0.1508	0.1120	0.0443	0.1364	0.1046	0.0490
Precinct 7	0.1480	0.1642	0.1909	0.0857	0.0865	0.1012
Precinct 8	0.0679	0.0711	0.1023	0.2034	0.0895	0.1659
Census tract demographics						
Percent white	0.4814	0.4078	0.4925	0.4424	0.4147	0.5200
Median household income	46113	37043	42050	42478	36647	43331
Crime incident rate	0.0237	0.0619	0.0232	0.0060	0.0176	0.0187
Calls for MNPD services	0.0221	0.0293	0.0090	0.0089	0.0139	0.0164

Notes: Crime and call rates are per capita and are restricted to those pertaining to violent crimes, theft, or drugs. Rates for reasons for search are calculated using only stops involving searches. All other rates are estimated using all stops in the data.

reasons, and investigative reasons.[29] If officers have different search preferences for drivers depending on the (reported) reason they are stopped, and the (reported) reason for stops varies with the day or shift, then it becomes necessary to condition on why the driver is stopped.[30] For example, Makofske (2020) shows that officers in Louisville arrest 40% of drivers stopped for failing to signal, compared to 1% of drivers stopped for any other reason. This suggests that certain stops are pretextual, and that the reason for stop may indicate the officer's search preference. In addition, the data show a 10% increase in the proportion of stops being attributed to driving-related reasons between the evening and night shifts.[31] If the reason for stop is indeed correlated with both search preferences and setting, then it is necessary to condition on it to satisfy Assumption 1.

Reasons for traffic searches include driver consent, probable cause, and plain view of contraband, and provide some insight into the interaction between the officer and driver. To see why it is necessary to condition on the reason for search, consider a traffic stop where the driver behaves belligerently. Not only may her behavior raise the officer's suspicion that she is hiding contraband and there is probable cause to search her, but it may also frustrate the officer and result in a lower draw of T_i. In contrast, a respectful driver may be disarming, and the officer may instead ask

[29] Driving-related reasons correspond to how the driver maneuvers her vehicle and how she interacts with other drivers on the road. This 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.

[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.

[31] In a study on endogenous driving behavior, Kalinowski et al. (2020) find that minority drivers adjust their driving behavior during the day, when their race is more visible to the officer.

for consent to search, or forgo the search entirely. Consequently, the search basis and officer preference may be correlated. If this type of behavior is also correlated with Z_i—e.g., belligerent drivers are more common during the weekend night shifts—then it becomes necessary to condition on the search basis. Tables 3–4 show strong correlation between the setting Z_i and reason for searches.

The final set of controls pertains to the environment where the stop takes place. This includes whether the stop is on a highway, the police precinct, the racial composition and income level of the census tract, the crime rate of the census tract, and the frequency of calls for MNPD services from the census tract. The concern is that an officer may be more cautious and mindful of his safety in some neighborhoods, but more carefree in others, and his mindset may influence his search preferences. For example, 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. If officers patrol different locations depending on the time and day, then this may induce a correlation between the officer's preferences and the setting of the stop, making these controls necessary.

6 Estimation

Estimation is performed in several steps and done separately for each officer. First, I estimate an individual officer's search and hit rates. Then I test for bias using a bilinear program. If the officer is biased, I then construct bounds on $\beta(g)$ for each g,

as well as $\mathbb{E}[\beta(g); \omega]$ for some ω.

6.1 Estimating search and hit rates

Let X_i denote the vector of control variables. I condition on $X_i = \overline{x}$ throughout, where $\overline{x}$ is defined as follows. For continuous controls, $\overline{x}_j = (1/n) \sum_i X_{i,j}$, i.e., the sample average, where j indexes components in the vector of controls. For categorical controls, $\overline{x}_j = \arg\max_{x_j} \widehat{\mathbb{P}}\{X_{i,j} = x_j\}$, i.e., the sample mode, where $\widehat{\mathbb{P}}$ denotes the empirical distribution.

To construct the search rate $\widehat{\mathbb{E}}[Search_i \mid R_i, Z_i, X_i]$, I use a logistic regression. To construct the hit rate $\widehat{\mathbb{E}}[Hit_i \mid R_i, Z_i, X_i]$, I use the relation

$$\mathbb{E}[Hit_i \mid R_i, Z_i, X_i] = \mathbb{E}[Hit_i \mid Search_i = 1, R_i, Z_i, X_i] \, \mathbb{E}[Search_i \mid R_i, Z_i, X_i].$$

So I first estimate the conditional hit rate $\widehat{\mathbb{E}}[Hit_i \mid Search_i = 1, R_i, Z_i, X_i]$ using a logistic regression on the subsample of drivers who are searched. I then scale these estimates by the estimated search rates, $\widehat{\mathbb{E}}[Search_i \mid R_i, Z_i, X_i]$.

An alternative approach to estimating the hit rate is to simply regress Hit_i on R_i, Z_i, and X_i. However, for some officers, this results in nonsensical estimates of the hit rate that exceed their search rate. This would imply a conditional hit rate greater than 1, which is not possible.

To summarize the variation in search and hit rates generated by Z_i, Tables 5–6 present logistic regressions of the search and hit indicators on Z_i, conditional on race, controls, and officer fixed effects. For ease of interpretation, the estimates presented are the exponentiated logit coefficients and they reflect the multiplicative impact that a change of setting has on the odds of being searched or finding contraband. The estimates suggest that, relative to stops during weekday evenings, the average odds

of being searched can fall by 36% and rise by 68% across the settings. Conditional on being searched, the average odds of finding contraband can fall by 20% and increase by up to 168% across settings.

Figure 6 provides examples of the search and hit rates for specific officers after conditioning on controls. For officer 49, the data are fairly similar across races, which may be consistent with the officer being unbiased. But the same cannot be said of the other officers. For instance, officer 6 is an example where bias is certainly going to be detected, since the absence of any hits for both white and minority drivers suggests both groups of drivers stopped are low-risk, yet minorities are searched three times more often than white drivers. Also, to see that the conditional hit rate varies with setting at the officer level, simply note that the data for each race do not lie along a ray extending from the origin.[32]

6.2 Restrictions $\text{supp}\,(G)$ and $\{\mathbf{p}_{r,z}\}$

In order for the test to be computationally feasible, I discretize the support of risk to be

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

The grid $\mathbf{g}$ is deliberately chosen to be finer at lower levels of risk since the average conditional hit rate across officers is not particularly high, between 20% and 25% (see Table 1). This suggests most drivers searched are relatively low risk. In order to distinguish between these drivers, I allocate more points in the grid to lower-levels of

[32]The conditional hit rate for each data point is equal to the ratio of the y-coordinate and x-coordinate of the point. If the conditional hit rates are the same across all settings, then the data points will lie along a line extending from the origin.

Table 5: Pooled logistic regression of $Search_i$ on Z_i, X_i

Contraband found	Estimate	Std. Err.	p-value	C.I. lower	C.I. upper
White					
Weekend	**0.7855**	0.0498	0.0001	0.6937	0.8895
Morning	**0.6356**	0.0271	0	0.5846	0.6909
Night	**1.4779**	0.0666	0	1.353	1.6144
Weekend × Morning	0.9252	0.1382	0.6029	0.6903	1.24
Weekend × Night	**0.7845**	0.0638	0.0028	0.6689	0.9201
N	109,023				
Minority					
Weekend	**0.8027**	0.0389	0	0.7299	0.8828
Morning	**0.5737**	0.0203	0	0.5353	0.6148
Night	**1.4698**	0.0528	0	1.3698	1.5771
Weekend × Morning	**1.6803**	0.1525	0	1.4065	2.0074
Weekend × Night	**0.8504**	0.0535	0.01	0.7517	0.9621
N	113,405				

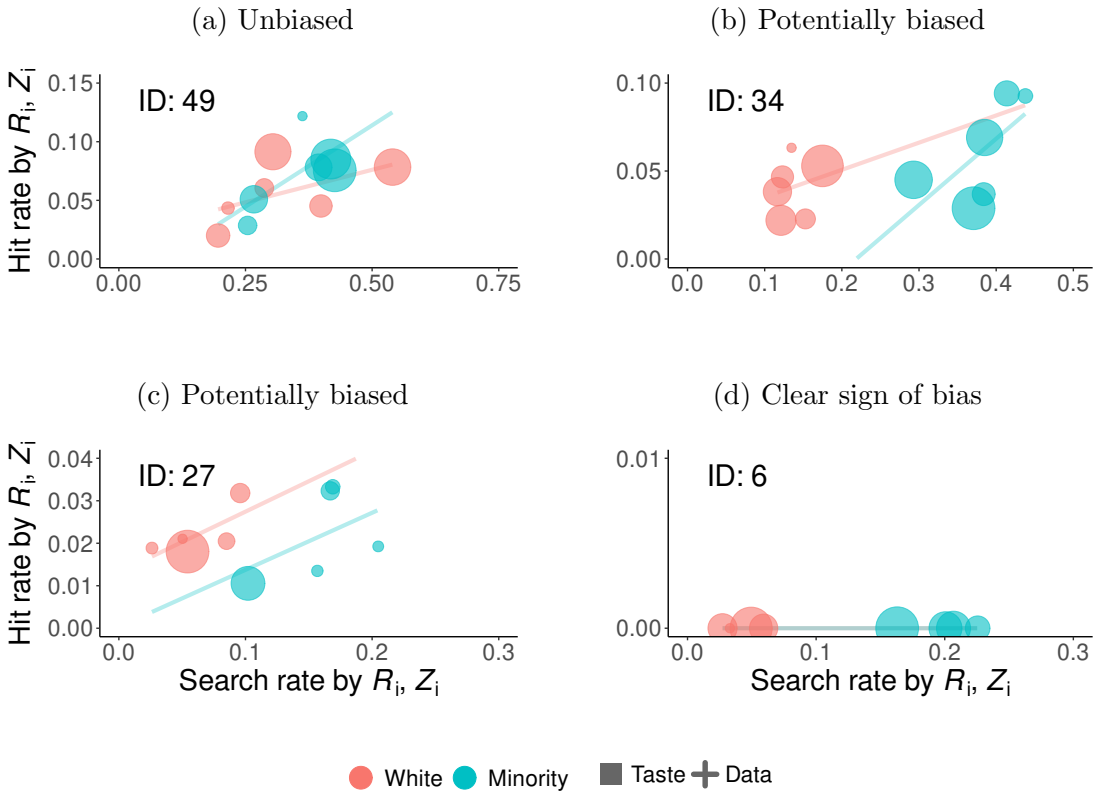
Notes: Estimates presented are the multiplicative impact a change of setting has on the odds ratio, and are relative to weekday evening stops. Estimates condition on officer fixed effects. Confidence intervals are at the 95% confidence level.

Table 6: Pooled logistic regression of Hit_i on Z_i, X_i, conditional on being searched

Contraband found	Estimate	Std. Err.	p-value	C.I. lower	C.I. upper
White					
Weekend	0.7987	0.1382	0.1938	0.569	1.1211
Morning	0.9345	0.0952	0.5063	0.7653	1.1411
Night	1.1581	0.122	0.1633	0.9421	1.4237
Weekend × Morning	**2.3021**	0.9085	0.0346	1.0622	4.9894
Weekend × Night	1.0303	0.2224	0.8899	0.6749	1.573
N	12,622				
Minority					
Weekend	0.8047	0.0968	0.0708	0.6357	1.0186
Morning	1.0007	0.0872	0.9938	0.8436	1.1871
Night	**1.4871**	0.1269	0	1.258	1.7578
Weekend × Morning	1.6823	0.4608	0.0575	0.9835	2.8777
Weekend × Night	0.9221	0.1445	0.6049	0.6783	1.2537
N	15,732				

Notes: Estimates presented are the multiplicative impact a change of setting has on the odds ratio, and are relative to weekday evening stops. Estimates condition on officer fixed effects. Confidence intervals are at the 95% confidence level.

Figure 6: Example of officer-level data



Note: Size of points correspond to number of traffic stops in setting Z_i. Search and hit rates are conditional on X_i.

risk.

Furthermore, since the drivers searched represent a riskier subset of the driver stopped, the low conditional hit rates suggest that most drivers stopped are relatively low risk as well. So I also impose the monotonicity restriction in (11), requiring that the PMF $\mathbf{p}_{r,z}$ is decreasing in risk for all $(r,z) \in \{w,m\} \times \mathcal{Z}$

I do not impose any restrictions on σ other than that it is non-decreasing in risk and bounded in the unit interval. But with more computational time, I can impose additional restrictions (e.g., model σ using Bernstein polynomials) in future versions of the draft.

6.3 Testing for bias

To test for bias, I solve the empirical counterpart to (8) in Proposition 2. I reweight the moments to improve efficiency, although an optimal weighting scheme has yet to be determined. The problem I solve is then

$$\widehat{Q}^{\star} \equiv \min_{\varsigma, \{\mathbf{p}_{r,z}\}} \sum_{r,z} \widehat{\mathbf{w}}^{S}_{r,z} \left| \varsigma' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{S}_{r,z} \right| + \sum_{r,z} \widehat{\mathbf{w}}^{H}_{r,z} \left| (\mathbf{g} \odot \varsigma)' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{H}_{r,z} \right| \tag{15}$$

$$\text{s.t.} \quad \mathbf{A} \begin{bmatrix} \varsigma \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

where

$$\widehat{\mathbf{m}}^{S}_{r,z} \equiv \widehat{\mathbb{P}}\{Search_i = 1 \mid R_i = r, Z_i = z, X_i = \overline{x}\}$$

$$\widehat{\mathbf{m}}^{H}_{r,z} \equiv \widehat{\mathbb{P}}\{Hit_i = 1 \mid R_i = r, Z_i = z, X_i = \overline{x}\}$$

and the weights are[33]

$$\widehat{\mathbf{w}}^S_{r,z} \equiv \frac{\sqrt{\sum_{i:R_i=r} \mathbb{1}\{R_i = r, Z_i = z\}}}{\widehat{\text{s.e.}}(\widehat{\mathbb{P}}\{Search_i = 1 \mid R_i = r, Z_i = z, X_i = \overline{x}\})},$$

$$\widehat{\mathbf{w}}^H_{r,z} \equiv \min\left\{ \frac{\sqrt{\sum_{i:R_i=r} \mathbb{1}\{R_i = r, Z_i = z\}}}{\widehat{\text{s.e.}}(\widehat{\mathbb{P}}\{Hit_i = 1 \mid R_i = r, Z_i = z, X_i = \overline{x}\})},\ 10\ \widehat{\mathbf{w}}^S_{r,z} \right\}.$$

The standard errors in the denominators of the weights adjust for how well the search and hit rates are estimated. These standard errors are estimated using a stratified bootstrap, where the number of stops drawn for a given R_i and Z_i is equal to that of the original sample. The numerator in the weights is the square-root of the number of traffic stops for a given race and setting. Its purpose is to account for how the standard errors may be artificially low for settings where the officer has made only a few stops. For example, if an officer makes five stops for $(R_i, Z_i) = (w, z)$ and happens to search the driver every time, then $\widehat{\text{s.e.}}(\{\widehat{\mathbb{P}}\{Search_i = 1 \mid R_i = w, Z_i = z, X_i = \overline{x}\})$ will be small. As a result, these search and hit rates will be weighted too heavily, and the test will primarily target these moments despite how they make up a small fraction of the officer's stops.

In addition, the weight assigned to the hit rate is limited to ten times that of the search rate. This is to prevent excessive weight being placed on the hit rates for officers who almost never find contraband, for whom $\widehat{\text{s.e.}}(\{\widehat{\mathbb{P}}\{Hit_i = 1 \mid R_i = w, Z_i = z, X_i = \overline{x}\})$ is small.

To conduct statistical inference, I make a modification to the test described in Proposition 2. In particular, I do not use $\widehat{Q}^\star$ as the test statistic to detect bias. Doing so not only tests whether the officer is biased, but also tests whether the model

[33] In practice, I also scale the weights by a constant to improve numerical stability when optimizing.

in general is misspecified. As a result, the distribution of $\widehat{Q}^\star$ may be overly dispersed and the test becomes very weak. Instead, I solve a second BP problem where the officer is allowed to have different preferences ς_w, ς_m for each race of drivers,

$$\widehat{Q}_B^\star \equiv \min_{\varsigma_w,\varsigma_m,\{\mathbf{p}_{r,z}\}} \sum_{r,z} \widehat{\mathbf{w}}_{r,z}^S \left|\varsigma_r' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}_{r,z}^S\right| + \sum_{r,z} \widehat{\mathbf{w}}_{r,z}^H \left|(\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}_{r,z}^H\right| \tag{16}$$

$$\text{s.t.} \quad \mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}.$$

I then construct the test statistic as

$$\widehat{\tau} \equiv \frac{\widehat{Q}^\star - \widehat{Q}_B^\star}{\widehat{Q}_B^\star},$$

which compares the fit of the model when the officer is restricted to being unbiased against the fit without the restriction. For example, if $\widehat{\tau} = 0.05$, that means the fit of the model under the unbiasedness restriction is 5% worse relative to the fit without the restriction. The test therefore only tests the unbiasedness restriction.[34]

To obtain the distribution of $\widehat{\tau}$, I use a stratified bootstrap, where the number of stops drawn for a given R_i and Z_i is equal to that of the original sample. I then reject that an officer is unbiased if the estimated α-quantile of τ exceeds threshold $\overline{\tau}$, for some choice of α and $\overline{\tau}$. This heuristic approach is not guaranteed to control the size of the test; a more formal approach is under development.

[34] See Bugni et al. (2015) and Chernozhukov et al. (2020) for a discussion on inference for partially identified models defined by moment restrictions.

Due to the computational demands of (15)–(16), inference is performed using 200 bootstrap samples and BP programs are terminated after five minutes. I construct $\widehat{\tau}$ conservatively whenever the BP program is terminated before being solved to global optimality. Specifically, when solving (15), a lower and upper bound on $\widehat{Q}^{\star}$ is obtained, and an optimal solution is reached when the lower and upper bounds coincide.[35] If (15) is terminated before this occurs, then the lower bound of $\widehat{Q}^{\star}$ is used to construct $\widehat{\tau}$. In contrast, if (16) is terminated before optimality is achieved, then the upper bound of $\widehat{Q}^{\star}_{B}$ is used to construct $\widehat{\tau}$. Together, this minimizes $\widehat{\tau}$, resulting in a more conservative test.

6.4 Bounding intensity of bias

If bias is detected, then the direction and intensity of the bias may be estimated by solving the empirical counterparts to (12)–(13) in Propositions 3–4. The bounds on

[35] In practice, optimality is achieved once the difference between the lower and upper bounds fall below some tolerance. In the application, I set the tolerance to be 1%.

$\beta(g_k)$ for $k = 0, \ldots, K$ are estimated by

$$\widehat{\beta}_{\text{lb}}(g_k), \widehat{\beta}_{\text{ub}}(g_k) \equiv \min/\max_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{w,z}\}, \{\mathbf{p}_{m,z}\}} \varsigma_{m,k} - \varsigma_{w,k} \tag{17}$$

$$\text{s.t.} \quad \sum_{w,z} \widehat{\mathbf{w}}^S_{w,z} \left| \varsigma_w' \mathbf{p}_{w,z} - \widehat{\mathbf{m}}^S_{w,z} \right| + \sum_{w,z} \widehat{\mathbf{w}}^H_{w,z} \left| (\mathbf{g} \odot \varsigma_w)' \mathbf{p}_{w,z} - \widehat{\mathbf{m}}^H_{w,z} \right| \leq \widehat{Q}^\star_{B,w}(1+\kappa)$$

$$\sum_{m,z} \widehat{\mathbf{w}}^S_{m,z} \left| \varsigma_m' \mathbf{p}_{m,z} - \widehat{\mathbf{m}}^S_{m,z} \right| + \sum_{m,z} \widehat{\mathbf{w}}^H_{m,z} \left| (\mathbf{g} \odot \varsigma_m)' \mathbf{p}_{m,z} - \widehat{\mathbf{m}}^H_{m,z} \right| \leq \widehat{Q}^\star_{B,m}(1+\kappa)$$

$$\mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

where the moments from the data have been replaced by their sample counterparts; $\widehat{Q}^\star_{B,r}$ is the minimized criterion for race r obtained in (16), i.e.,

$$\widehat{Q}^\star_{B,r} \equiv \sum_z \widehat{\mathbf{w}}^S_{r,z} \left| \widehat{\varsigma}_r' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^S_{r,z} \right| + \sum_z \widehat{\mathbf{w}}^H_{r,z} \left| (\mathbf{g} \odot \widehat{\varsigma}_r)' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^H_{r,z} \right|,$$

where $\widehat{\varsigma}_r$ is part of the solution to (16); and $\kappa \geq 0$ is a tuning parameter controlling the slackness in the moment matching criterion, where the slackness ensures the optimization problem is always feasible.[36] In the application, I set $\kappa = 0.001$. The bounds I present in Section 7 may be tightened by choosing a smaller value of κ.

To estimate the average bias, I define ω as in (14) and set $q_{w,z} = \widehat{\mathbb{P}}\{Z_i = z \mid R_i = w\}$. That is, the average bias corresponds to the average difference in the probability of being searched when minority drivers have the same distribution of risk as white

[36] The tuning parameter κ converges to zero as the number of traffic stops grows. See Mogstad et al. (2018) for another example of such a tuning parameter.

drivers in the data. The bounds on $\mathbb{E}[\beta(G_i);\omega]$ are estimated by

$$\widehat{\mathbb{E}}[\beta(G_i);\omega]_{\text{lb}}, \widehat{\mathbb{E}}[\beta(G_i);\omega]_{\text{ub}} \equiv \min/\max_{\varsigma_w,\varsigma_m,\{\mathbf{p}_{w,z}\},\{\mathbf{p}_{m,z}\}} \mathbf{q}_w' \mathbf{P}_w (\varsigma_m - \varsigma_w) \tag{18}$$

$$\text{s.t.} \quad \sum_z \widehat{\mathbf{w}}^S_{w,z} \left|\varsigma_w' \mathbf{p}_{w,z} - \widehat{\mathbf{m}}^S_{w,z}\right| + \sum_z \widehat{\mathbf{w}}^H_{w,z} \left|(\mathbf{g} \odot \varsigma_w)' \mathbf{p}_{w,z} - \widehat{\mathbf{m}}^H_{w,z}\right| \leq \widehat{Q}^\star_{B,w}(1+\kappa)$$

$$\sum_z \widehat{\mathbf{w}}^S_{m,z} \left|\varsigma_m' \mathbf{p}_{m,z} - \widehat{\mathbf{m}}^S_{m,z}\right| + \sum_z \widehat{\mathbf{w}}^H_{m,z} \left|(\mathbf{g} \odot \varsigma_m)' \mathbf{p}_{m,z} - \widehat{\mathbf{m}}^H_{m,z}\right| \leq \widehat{Q}^\star_{B,m}(1+\kappa)$$

$$\mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}.$$

The BP problems (17)–(18) are terminated after five minutes. If optimality is not yet reached, then the lower bound of the BP objective value is used to estimate the lower bound on the measure of bias, and the upper bound of the BP objective value is used to estimate the upper bound on the measure of bias.

To conduct inference, the researcher may construct the confidence interval for each measure of bias by inverting the test in Proposition 2. Specifically, for $b \in [-1, 1]$ and $k = 1, \ldots, K$, the researcher can test the restriction $\beta(g_k) = b$. If the test does not reject the restriction, then b is contained in the confidence interval for $\beta(g_k)$. The confidence interval for $\mathbb{E}[\beta(G_i);\omega]$ may be constructed in the same way. See Appendix B.4 for a full description of this procedure. However, this approach is computationally demanding since it involves iterating over a grid of values for $b \in [-1, 1]$, where each iteration entails bootstrapping a BP problem.

So in the application, I instead construct the 'backward' confidence interval described in Andrews and Han (2009). That is, the lower bound of the confidence

interval is equal to the $\alpha/2$-quantile of the bootstrap distribution of $\widehat{\mathbb{E}}[\beta(G_i);\omega]_{\text{lb}}$, and the upper bound is equal to the $(1-\alpha/2)$-quantile of the bootstrap distribution of $\widehat{\mathbb{E}}[\beta(G_i);\omega]_{\text{ub}}$. While Andrews and Han (2009) point out the coverage probability of the backward confidence interval may be incorrect, this heuristic approach may still be informative and is a stand-in until a more formal/feasible method of inference is developed.

7 Results

Table 7 displays the number of officers who fail the test under various specifications. Each column indicates the α-quantile of the empirical distribution of the test statistic $\widehat{\tau}$ used to detect bias, and each row indicates the threshold $\overline{\tau}$ for how much the model fit must worsen under the unbiasedness restriction before the officer is flagged as biased. Each entry indicates the number of officers who fail the test for a choice of α and $\overline{\tau}$. The counts are not adjusted for multiple hypothesis testing.

The size of the test increases as α increases and $\overline{\tau}$ decreases. For instance, consider the entry corresponding to $\alpha = \overline{\tau} = 0.05$ (third row, third column). The entry indicates that, for 8 officers, the unbiasedness restriction worsens the fit of the model by at least 5% (relative to the fit without the restriction) in 95% of the bootstrap samples. The entry to the right shows that, for 14 officers, the restriction worsens the fit by at least 5% in 90% of the bootstrap samples; and the entry above shows that, for 13 officers, the restriction worsens the fit by at least 2.5% for 95% of the bootstrap samples. I focus on the test results for $\alpha \in \{0.05, 0.1\}$ and $\tau = 0.05$, which flags 8 to 14 officers as biased.

Figure 7 presents the estimated bounds on the average bias for all officers, who are ordered to be decreasing in their estimated lower bound. The black bars indicate

Table 7: Number of biased officers

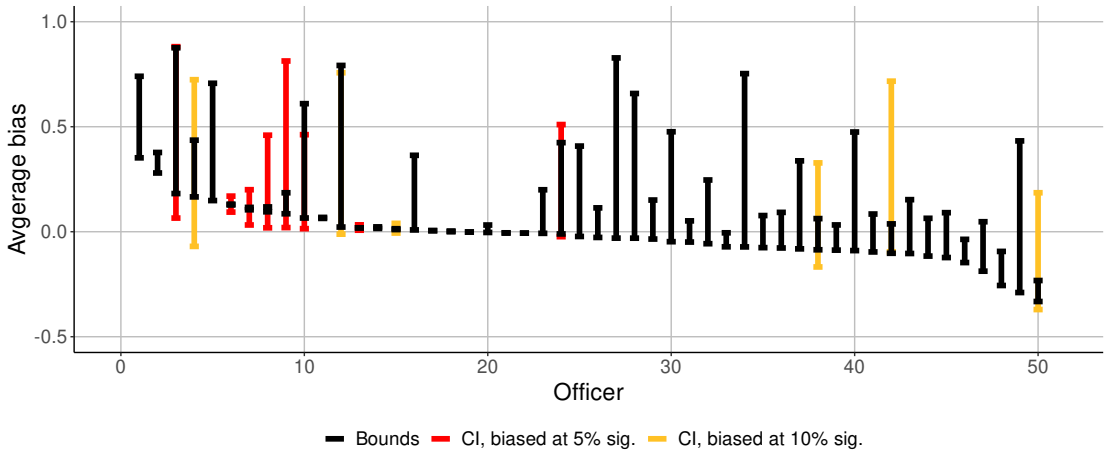
	α-percentile		
$\overline{\tau}$	0.01	0.05	0.10
0.000	16	31	42
0.025	7	13	24
0.050	6	8	14
0.100	5	6	9
0.200	4	6	6

Notes: Estimates are based on 200 bootstrap samples. An officer is biased if the α-percentile of the test statistic $\widehat{\tau}$ strictly exceeds threshold $\overline{\tau}$. The counts above do not adjust for multiple hypothesis testing.

the estimated bounds, and the colored bars indicate the 95% confidence interval for officers who fail the test. Notice how tight the bounds are for many of the officers. As discussed below, this follows from the data being highly informative of what the distributions of risk are for these officers. Also notice that the estimated bounds extend beyond the confidence intervals for officers 10 and 12. While this raises concerns over the validity of the inference procedure for the bounds on $\mathbb{E}[\beta(G_i); \omega]$, there are currently no formal methods for inference when solving bilinear programs. Developing a valid method is a possible area for future research.

For officers whose bounds on the average bias contain zero, it is possible for them to be unbiased or biased. For example, officers 25 through 32 pass the test and their bounds on the average bias all contain zero. In contrast, officer 24 fails the test when $\alpha = 0.05$ and officer 38 fails the test when $\alpha = 0.10$, but their bounds also contain zero. This can happen if the officer is indeed biased, but the direction of bias changes with the level of risk so that the positive and negative biases cancel each other out

Figure 7: Bounds on average bias $\mathbb{E}[\beta(G_i); \omega]$



Note: Positive average bias indicates that the officer searches minority drivers more often than equally risk white drivers on average. The 95% confidence intervals on the bounds are only shown for officers who fail the test at either the 5% or 10% significance level when $\overline{\tau} = 0.05$.

on average.[37] This corresponds to the case where $0 \in \mathcal{E}$.

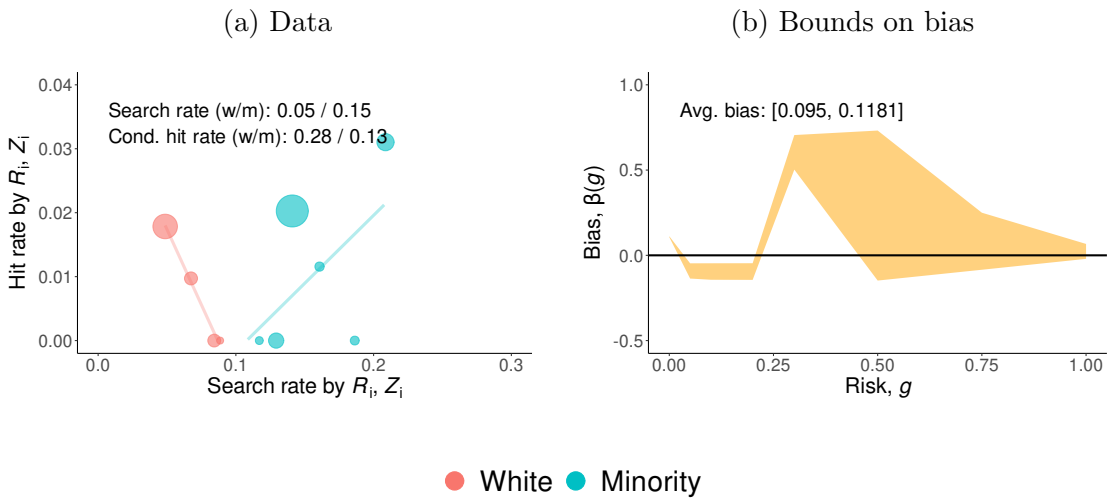
However, recall that $\mathcal{E} \subseteq [\mathbb{E}[\beta(G_i); \omega]_{\text{lb}}, \mathbb{E}[\beta(G_i); \omega]_{\text{ub}}]$. So even though the bounds on the average bias contain zero, it is possible that $0 \notin \mathcal{E}$. This corresponds to the case where the identified set of preferences is not consistent with an average bias of zero—therefore indicating the officer is biased—but is consistent with both positive and negative average bias. Likewise, since $\mathcal{B}_k \subseteq [\beta_{\text{lb}}(g_k), \beta_{\text{ub}}(g_k)]$, it is possible that $\beta_{\text{lb}}(g_k) < 0 < \beta_{\text{ub}}(g_k)$ but $0 \notin \mathcal{B}_k$ for all $k = 0, \ldots, K$, and the officer still fails the test. This highlights how Proposition 2 tests the strongest implication of unbiasedness, and tests based on $\mathbb{E}[\beta(G_i); \omega]$ or $\beta(g_k)$ will not be as powerful.[38]

Figures 8–11 take a closer look at individual officers and their bias conditional

[37] This can also happen if the BP problem is terminated prematurely and the bounds are constructed conservatively. If given more time, the bounds may shrink and exclude zero.

[38] Proposition 2 may be thought of as jointly testing $\beta(g_k) = 0$ for *all* $k = 0, \ldots, K$.

Figure 8: Officer 8 switches direction of bias



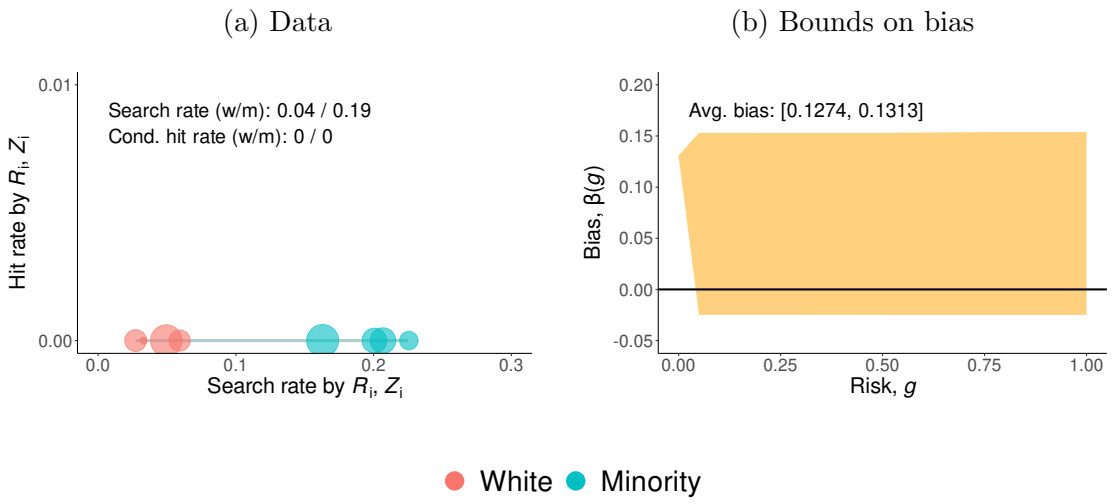
on risk. To reduce computational cost, I estimate the bounds on $\beta(g_k)$ for only a subset of $g_k \in \text{supp}\,(G)$.[39] Inference is still being computed for these bounds. In Appendix D, I show the estimated bounds for all 50 officers.

Figure 8 presents the data and bounds for officer 8, who fails the test when $\alpha = 0.05$ and is estimated to change his direction of bias. For zero-risk drivers, he is biased against minorities; as risk increases to 0.05 and 0.1, he becomes biased against white drivers; as risk increases to 0.3, he is again biased against minority drivers. Once the risk exceeds 0.5, the direction of bias is unknown, but the bounds on $\beta(\cdot)$ shrink towards zero, a common pattern in the estimates (see Appendix D). Intuitively, this makes sense, as it suggests that the officer's preferences have less of an impact on the search decision as it becomes increasingly apparent that the driver carries contraband.

Figure 9 presents the data and bounds for officer 6, who also fails the test when $\alpha = 0.05$. As his hit rates are approximately zero, it is implied that both groups of drivers stopped primarily have zero risk. Yet, the officer searches minority drivers

[39] I estimate the bounds on $\beta(g)$ for roughly every other point in $\mathbf{g}$, i.e., for g in $\{0, 0.05, 1, 0.2, 0.3, 0.5, 0.75, 1\}$.

Figure 9: Officer 6 is clearly biased against zero-risk drivers



15 percentage points more than white drivers, indicating that the officer is biased against zero-risk minority drivers. Moreover, since the data suggests both groups of drivers have similar distribution of risk, the differences in search rates must stem from differences in preferences, resulting in very tight bounds on the average bias.

Figure 10 shows an example where the officer appears to be biased at first glance. Specifically, the data alone shows that officer 34 is more than twice as likely to search minority drivers as white drivers, despite how he is half as likely to find contraband on minority drivers compared to white drivers. Nevertheless, he passes the test when $\alpha \in \{0.05, 0.1\}$.

Finally, Figure 11 presents an example where the data for white and minority drivers are similar. This suggests the identified sets of $\sigma(\cdot; w)$ and $\sigma(\cdot; m)$ are also similar, and the data may be generated using the same search preference for both groups of drivers. As expected, the officer passes the test.

Figure 10: Officer 34 passes the test

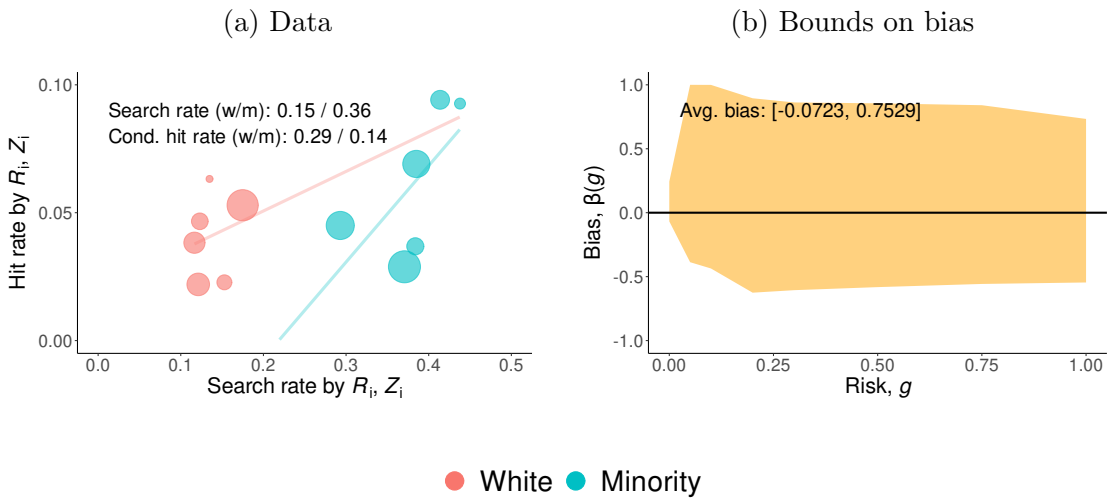
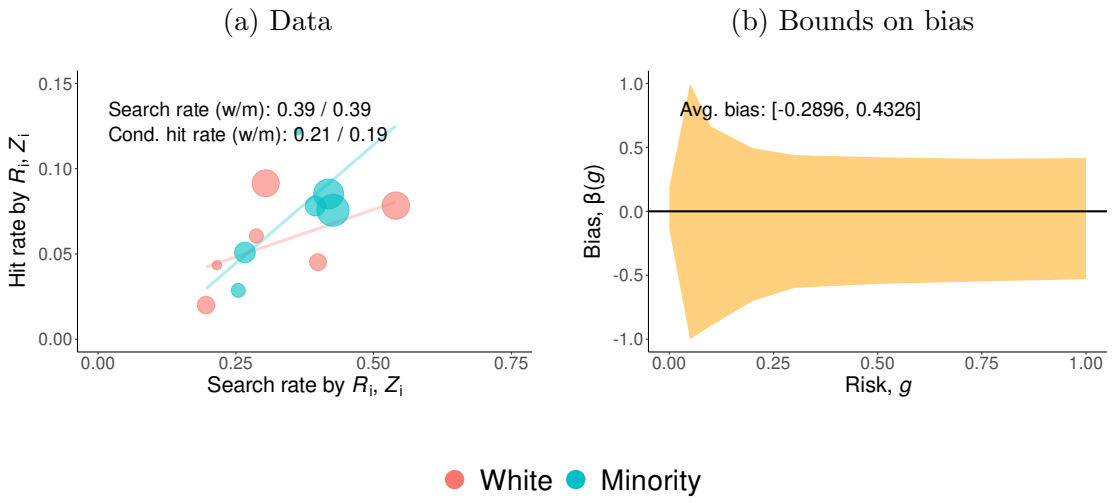


Figure 11: Officer 49 passes the test



8 Conclusion

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

in preferences and beliefs across officers.

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

Implementing these methods involves solving several bilinear programs, which is novel in the literature on discrimination. There is commercial software freely available to academic institutions capable of solving these problems to global optimality, making it feasible to estimate the sharp bounds discussed. Bilinear programs also have the potential to be used more generally to study mixture models, and is a possible area of future research. Another topic that requires further study is statistical inference for bilinear programs, for which there is currently no formal procedure.

I apply the proposed methods on police traffic data from the Metropolitan Nashville Police Department, and find evidence to suggest a large number of officers are biased. The estimates also suggest that officers are more likely to be biased against low-risk minority drivers, and the bias disappears as the risk of the driver increases. A convenient feature of these methods is that they may be performed using fairly standard police traffic data sets. The assumptions of the model are better satisfied when the police data are supplemented with local demographic data, such as household incomes and crime rates, and such data is often public or available upon request. So a natural extension of the paper is to apply these methods to other police data sets from across the US.

Another avenue for future research is to extend these methods to study bias in traffic stops. Although the framework in this paper was intended to circumvent the challenges that bias in traffic stops imposed on measuring bias in traffic searches, it would be interesting to decompose the *total* effect of bias on searches into the biases that occur before and after a driver is stopped. There are now commercial data sets on driver demographics collected by tracking smartphones in vehicles, and the availability of such data provides an opportunity to tackle this long-standing question in new ways.

References

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 S. Han (2009). Invalidity of the Bootstrap and the m out of n Bootstrap for Confidence Interval Endpoints Defined by Moment Inequalities. *The Econometrics Journal 12*(suppl_1), S172–S199.

Antonovics, K. and B. G. Knight (2009). A New Look at Racial Profiling: Evidence from the Boston Police Department. *The Review of Economics and Statistics 91*(1), 163–177.

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 (2020). Measuring Racial Discrimination in Bail Decisions. Technical report.

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

Ayres, I. (2002). Outcome Tests of Racial Disparities in Police Practices. *Justice Research and Policy 4*(1-2), 131–142.

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

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

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.

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 (2020). Inaccurate Statistical Discrimination: An Identification Problem. Technical report, National Bureau of Economic Research.

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

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

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

Canay, I. A., M. Mogstad, and J. Mountjoy (2020). On the Use of Outcome Tests for Detecing Bias in Decision Making. Technical report.

Chan, D. C., M. Gentzkow, and C. Yu (2019). Selection with Variation in Diagnostic Skill: Evidence from Radiologists. Technical report, National Bureau of Economic Research.

Chernozhukov, V., W. K. Newey, and A. Santos (2020). Constrained Conditional Moment Restriction Models. Technical report.

Collaborators, G. . P. V. U. S. et al. (2021). Fatal Police Violence by Race and State in the USA, 1980–2019: A Network Meta-Regression. *The Lancet 398*(10307), 1239–1255.

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.

Doha, E., A. Bhrawy, and M. Saker (2011). On the Derivatives of Bernstein Polynomials: An Application for the Solution of High Even-order Differential Equations. *Boundary Value Problems 2011*, 1–16.

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), 000–000.

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

Farouki, R. T. (2012). The Bernstein Polynomial Basis: A Centennial Retrospective. *Computer Aided Geometric Design 29*(6), 379–419.

Farouki, R. T. and V. Rajan (1988). Algorithms for polynomials in bernstein form. *Computer Aided Geometric Design 5*(1), 1–26.

Feigenberg, B. and C. Miller (2021). Would Eliminating Racial Disparities in Motor Vehicle Searches Impose Efficiency Costs? Technical report.

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

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

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

Gelbach, J. B. (2021). Testing Economic Models of Discrimination in Criminal Justice. Technical report.

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.

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.

Horowitz, J. M., A. Brown, and K. Cox (2019). Race in America 2019.

Hull, P. (2021). What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making. Technical report.

Jones, J. M. (2021). In U.S., Black Confidence in Police Recovers from 2020 Low.

Kalinowski, J., M. Ross, and S. L. Ross (2020). Endogenous Driving Behavior in Tests of Racial Profiling in Police Traffic Stops. Technical report, Working Paper.

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 (2020). Administrative Records Mask Racially Biased Policing. *American Political Science Review*, 1–19.

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

Makofske, M. (2020). Pretextual traffic stops and racial disparities in their use. Technical report.

Manski, C. F. and D. S. Nagin (2017). Assessing Benefits, Costs, and Disparate Racial Impacts of Confrontational Proactive Policing. *Proceedings of the National Academy of Sciences 114*(35), 9308–9313.

Marx, P. (2021). An absolute Test of Racial Prejudice. *The Journal of Law, Economics, and Organization*.

Metropolitan Nashville Police Department Manual (2018). https://www.nashville.gov/Police-Department/Department-Manual.aspx.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86*(5), 1589–1619.

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

Morin, R. and R. Stepler (2016). The Racial Confidence Gap in Police Performance. Technical report, Pew Research Center.

Novak, K. J. and M. B. Chamlin (2012). Racial threat, suspicion, and police behavior: The Impact of Race and Place in Traffic Enforcement. *Crime & Delinquency 58*(2), 275–300.

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

Owens, E. (2020). The Economics of Policing. *Handbook of Labor, Human Resources and Population Economics*, 1–30.

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*, 1–10.

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.

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.

Trinkner, R., E. M. Kerrison, and P. A. Goff (2019). The Force of Fear: Police Stereotype Threat, Self-Legitimacy, and Support for Excessive Force. *Law and human behavior 43*(5), 421.

A Proofs

A.1 Deriving the random threshold in (1)

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

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

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

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

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

A.2 Proof of Corollary 1

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

A.3 Deriving the search and hit rates

The search rate is derived as follows.

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

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

The hit rate is derived as follows.

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

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

written as

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

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

A.4 Testing for bias using η instead of σ

I show below that η depends on the race of the driver if an only if σ depends on the race of the driver. For all $g > 0$,

$$
\begin{aligned}
& \eta(\sigma(\cdot; w); g) = \eta(\sigma(\cdot; m); g) \\
\iff \quad & g\sigma(g; w) = g\sigma(g; m) \\
\iff \quad & \sigma(g; w) = \sigma(g; m).
\end{aligned}
$$

For $g = 0$, consider the graph of η,

$$\begin{aligned} & (\sigma(0,w), \eta(\sigma(\cdot;w);0)) = (\sigma(0,m), \eta(\sigma(\cdot;m);0)) \\ \iff & (\sigma(0,w), 0 \times \sigma(0;w)) = (\sigma(0,m), 0 \times \sigma(0;m)) \\ \iff & (\sigma(0,w), 0) = (\sigma(0,m), 0). \end{aligned}$$

So there exists a $g \in [0,1]$ such that $\eta(\sigma(\cdot;w);g) \neq \eta(\sigma(\cdot;m);g)$ if and only if $\sigma(g;w) \neq \sigma(g;m)$. Then $\eta(\sigma(\cdot;r);\cdot)$ depends on $r \in \{w,m\}$ if and only if $\sigma(\cdot;r)$ depends on $r \in \{w,m\}$.

A.5 Proof of Proposition 1

Proof. The proof proceeds in three steps. First, I show that there is a linear relationship between $\mathbb{E}[Search_i \mid R_i, Z_i]$ and $\mathbb{E}[Hit_i \mid R_i, Z_i]$. Second, I show that this linear relationship may be recovered by a linear IV regression. Third, I show that the officer is biased if the IV estimands differ by race.

The assumption that $Var[Search_i \mid R_i = r, Z_i] > 0$ is to rule out the cases where $\sigma(g_1;r) = \sigma(g_2;r) > 0$, or $\sigma(g_2;r) = 0$. In the first case, the observed search rates indicate the proportion of drivers the officer searches, regardless of their risk. If these rates differ across race, then the officer is immediately revealed to be biased. In the second case where $\sigma(g_2;r) = 0$, it must be that $\sigma(g_1;r) = 0$, since $\sigma(\cdot;r)$ is a non-decreasing function. It follows that no searches are observed at all, so $\alpha_0(r) = 0$ and α_1 is not well defined. But this is a trivial case that corresponds to an officer who never searches any driver, and the absence of any searches fully reveals the officer's preferences. So for the remainder of the proof, I assume $\sigma(g_2;r) > 0$ for $r \in \{w,m\}$, but allow $\sigma(g_1;r)$ to be 0.

To show the linear relationship between $\mathbb{E}[Hit_i \mid R_i, Z_i]$ and $\mathbb{E}[Search_i \mid R_i, Z_i]$, write the search and hit rates as

$$\mathbb{E}[Search_i \mid R_i = r, Z_i = z] = \sigma(g_1; r) + p_{r,z}(g_2)(\sigma(g_2; r) - \sigma(g_1; r)), \tag{A.5}$$

$$\mathbb{E}[Hit_i \mid R_i = r, Z_i = z] = g_1\ \sigma(g_1; r) + p_{r,z}(g_2)(g_2\ \sigma(g_2; r) - g_1\ \sigma(g_1; r)). \tag{A.6}$$

Solving for $p_{r,z}(g_2)$ in (A.5), I have

$$p_{r,z}(g_2) = \frac{\mathbb{E}[Search_i \mid R_i = r, Z_i = z] - \sigma(g_1; r)}{\sigma(g_2; r) - \sigma(g_1; r)}.$$

Substituting this expression for $p_{r,z}(g_2)$ into (A.6) and grouping terms, I have

$$\mathbb{E}[Hit_i \mid R_i = r, Z_i = z] = \alpha_0(r) + \alpha_1(r)\mathbb{E}[Search_i \mid R_i = r, Z_i = z], \tag{A.7}$$

where

$$\alpha_0(r) = -\frac{\sigma(g_1; r)\sigma(g_2; r)(g_2 - g_1)}{\sigma(g_2; r) - \sigma(g_1; r)} \leq 0,$$
$$\alpha_1(r) = \frac{g_2\ \sigma(g_2; r) - g_1\ \sigma(g_1; r)}{\sigma(g_2; r) - \sigma(g_1; r)} > 0.$$

This establishes the linear relationship between $\mathbb{E}[Search_i \mid R_i, Z_i]$ and $\mathbb{E}[Hit_i \mid R_i, Z_i]$, and that $\alpha_0(r) \leq 0$ and $\alpha_1(r) > 0$.

To show that α_0 and α_1 are identified by a linear IV regression, let

$$X_i' \equiv (1, Search_i)$$
$$W_i' \equiv (1, Z_i)$$
$$\alpha(r)' \equiv (\alpha_0(r), \alpha_1(r)).$$

To simplify the proof, suppose $\mathcal{Z} = \{0, 1\}$ so that $\alpha(r)$ is just identified. Then the IV estimand is

$$
\begin{aligned}
&\mathbb{E}[W_i X_i' \mid R_i = r]^{-1}\ \mathbb{E}[W_i\ Hit_i \mid R_i = r] \\
&= \mathbb{E}[\mathbb{E}[W_i X_i' \mid R_i = r, W_i] \mid R_i = r]^{-1} \mathbb{E}[\mathbb{E}[W_i\ Hit_i \mid R_i = r, W_i] \mid R_i = r] \\
&= \mathbb{E}[W_i\ \mathbb{E}[X_i' \mid R_i = r, W_i] \mid R_i = r]^{-1} \mathbb{E}[W_i\ \mathbb{E}[Hit_i \mid R_i = r, W_i] \mid R_i = r] \\
&= \mathbb{E}[W_i\ \mathbb{E}[X_i' \mid R_i = r, W_i] \mid R_i = r]^{-1} \mathbb{E}[W_i\ \mathbb{E}[X_i' \mid R_i = r, W_i] \mid R_i = r]\ \alpha(r) \\
&= \alpha(r),
\end{aligned}
$$

where the first equality if by law of iterated expectations; the second equality is by linearity of expecations; the third equality follows from (A.7) and linearity of expectations; and the final equality follows from matrix algebra.

From the definitions of $\alpha_0(r)$ and $\alpha_1(r)$, it follows that the officer is biased if $\alpha(w) \neq \alpha(m)$. To see why the converse does not hold, suppose $\sigma(g_1; w) = \sigma(g_1; m) = 0$, $\sigma(g_2; w) \neq \sigma(g_2; m)$, and $\sigma(g_2; w), \sigma(g_2; m) > 0$. Then $\alpha_0(w) = \alpha_0(m) = 0$ and $\alpha_1(w) = \alpha_1(m) = g_2$, even though the officer has different search preferences for white and minority drivers with risk g_2. ■

A.6 Proof of Corollary 3

Proof. To build off the proof of Proposition 1, suppose Z_i is unobserved by the researcher. Suppose also that the officer is unbiased so that $\alpha(0) = \alpha(1) = \alpha$. Then

the hit rate observed by the researcher for race r is

$$\begin{aligned}
\mathbb{E}[Hit_i \mid R_i = r] &= \mathbb{E}[\mathbb{E}[Hit_i \mid R_i = r = Z_i] \mid R_i = r] \\
&= \mathbb{E}[\alpha_0 + \alpha_1 \mathbb{E}[Search_i \mid R_i = r, Z_i] \mid R_i = r] \\
&= \alpha_0 + \alpha_1 \mathbb{E}[\mathbb{E}[Search_i \mid R_i = r, Z_i] \mid R_i = r] \\
&= \alpha_0 + \alpha_1 \mathbb{E}[Search_i \mid R_i = r],
\end{aligned}$$

where the first equality is by law of iterated expectations; the second equality follows from (A.7); the third equality follows from linearity of expectations; and the final equality follows from the law of iterated expectations.

Proposition 1 implies that $\alpha_0 \leq 0$ and $\alpha_1 > 0$ if the officer is unbiased. Then by contraposition, if $\alpha_0 > 0$ or $\alpha_1 \leq 0$, then the officer must be biased. ■

B Constraints in the bilinear programming problem

This section provides some examples of how to impose linear constraints in the bilinear program, as well as motivates monotonicity restriction (11) on the distributions of risk.

B.1 Imposing linear constraints

Consider the vector of variables $\mathbf{x}' = (x_0, \ldots, x_K)$. The monotonicity constraint

$$x_0 \leq x_1 \leq \cdots \leq x_K \tag{B.8}$$

may be written as

$$\begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \mathbf{x} \geq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

To reverse the direction of monotonicity, simply reverse the inequalities. Linear constraints of the form

$$\sum_{k=0}^{K} a_k x_k \lesseqgtr b \tag{B.9}$$

may be written as

$$\mathbf{a}'\mathbf{x} \lesseqgtr b,$$

where $\mathbf{a}' = (a_0, \ldots, a_K)$.

To ensure that the search probabilities $\varsigma_r = (\sigma(g_0; r), \ldots, \sigma(g_K; r))$ that are being optimized over are consistent with being a CDF of $T_i \mid R_i = r$ for $r \in \{w, m\}$, ς_r must be non-decreasing in index k, and each element must be in the unit interval. The non-decreasing property of ς_r takes the form of (B.8), and the bounds on each element of ς_r take the form of (B.9) (i.e., choose $\mathbf{a}$ to be a standard basis vector).

To ensure that the distribution of risk $\mathbf{p}_{r,z}$ is consistent with being a PMF, the elements of $\mathbf{p}_{r,z}$ must be in the unit interval and sum to 1. Both of these constraints take the form of (B.9). The researcher may also choose to impose monotonicity constraints on $\mathbf{p}_{r,z}$. These will take the form of (B.8).

The researcher may want to rank the average risk of drivers by race R_i, setting Z_i, or both. This constraint is straightforward to impose. To see how, write the average risk conditional on race and setting as

$$\begin{aligned}\mathbb{E}[Guilty_i \mid R_i = r, Z_i = z] &= \sum_{k=0}^{K} g_k \mathbf{p}_{r,z,k} \\ &= \mathbf{g}' \mathbf{p}_{r,z},\end{aligned}$$

where $\mathbf{g}' = (g_0, \ldots, g_K)$ is the vector of discretized risks. Then the ranking

$$\mathbb{E}[Guilty_i \mid R_i = r_1, Z_i = z_1] \leq \mathbb{E}[Guilty_i \mid R_i = r_2, Z_i = z_2]$$

takes the form

$$\begin{aligned} & \sum_{k=0}^{K} g_k \mathbf{p}_{r_1,z_2,k} \leq \sum_{k=0}^{K} g_k \mathbf{p}_{r_2,z_2,k} \\ \iff & \sum_{k=0}^{K} g_k \mathbf{p}_{r_1,z_1,k} - \sum_{k=0}^{K} g_k \mathbf{p}_{r_2,z_2,k} \leq 0 \\ \iff & \mathbf{g}'(\mathbf{p}_{r_1,z_1} - \mathbf{p}_{r_2,z_2}) \leq 0. \end{aligned}$$

This restriction has the same form as (B.9), with $\mathbf{a}' = (\mathbf{g}', -\mathbf{g}')$ and $\mathbf{x}' = (\mathbf{p}'_{r_1,z_1}, \mathbf{p}'_{r_2,z_2})$.

B.2 Motivating restrictions on the distribution of risk

To provide an example for how the PDF of risk for drivers stopped may be decreasing in risk, I consider the following model for traffic stops. Let $Stop_i \in \{0, 1\}$ denote the stop decision of an officer for driver i. Data is only available for drivers who are stopped, for whom $Stop_i = 1$. Let $\mathcal{U}^p_{P,i}(R_i)$ denote the random utility of stop decision p for driver i, and $\mathcal{U}^s_{S,i}(Guilty_i; R_i)$ denote the random utility of searching driver i. The search utilities $\{\mathcal{U}^s_{S,i}\}$ are as in the main paper, except I have included the additional 'S' subscript to distinguish it from the utilities from stopping a driver.

Before stopping the driver, the officer observes R_i, Z_i, and V_i^{pre}, where V_i^{pre} is a subvector of $V_i' \equiv (V_i^{\text{pre}'}, V_i^{\text{post}'})$. So V_i^{pre} contains variables that the officer observes without having to make a stop, such as the make of the vehicle and the speed it was traveling at; and V_i^{post} includes variables that the officer only observes after stopping and interacting with the driver, such as the demeanor of the driver and the smell of the vehicle interior. As in the main paper, the researcher observes no components of V_i. The officer also knows the stop utilities $\{\mathcal{U}^p_{P,i}\}$ before stopping the driver, similar to how he knows $\{\mathcal{U}^s_{S,i}\}$ before searching the driver.

To make his stop decision, the officer considers the expected utility from stopping a driver and not stopping a driver. The reason why he maximizes the expected utility is because he does not know whether he will search the driver afterwards, and if he does, whether the driver will be guilty. So the officer's stop decision may be expressed as

$$\begin{aligned} Stop_i &\equiv \arg\max_{p\in\{0,1\}} \mathbb{1}\{p=1\}\left(\mathcal{U}^1_{P,i}(R_i) + \mathbb{E}[\mathcal{U}^{Search_i}_{S,i}(Guilty_i; R_i) \mid R_i, Z_i, V_i^{\text{pre}}]\right) \\ &\qquad + \mathbb{1}\{p=0\}\mathcal{U}^0_{P,i}(R_i) \\ &= \mathbb{1}\left\{\mathcal{U}^1_{P,i}(R_i) + \mathbb{E}[\mathcal{U}^{Search_i}_{S,i}(Guilty_i; R_i) \mid R_i, Z_i, V_i^{\text{pre}}] \geq \mathcal{U}^0_{P,i}(R_i)\right\} \\ &= \mathbb{1}\left\{\mathbb{E}[\mathcal{U}^{Search_i}_{S,i}(Guilty_i; R_i) \mid R_i, Z_i, V_i^{\text{pre}}] \geq T_i^{Stop}\right\}, \end{aligned}$$

where $T_i^{Stop} \equiv \mathcal{U}^0_{P,i}(R_i) - \mathcal{U}^1_{P,i}(R_i)$ is a random utility threshold. To distinguish between the thresholds for stop and search decisions, let T_i^{Search} denote the utility threshold for searches.

Assumption B1. $\{\mathcal{U}^p_{P,i}\} \perp\!\!\!\perp (\{\mathcal{U}^s_{S,i}\}, Z_i, V_i^{\text{pre}})$.

Corollary B1. $T_i^{Stop} \perp\!\!\!\perp (T_i^{Search}, Z_i, V_i^{\text{pre}})$.

The independence between $\{\mathcal{U}^p_{P,i}\}$ and $\{\mathcal{U}^s_{S,i}\}$ is imposed to ensure that Assumption 1(ii)–1(iii) in the main paper is satisfied. To see why, suppose the stop and search preferences are correlated and let V_i^{post} contain $\{\mathcal{U}^p_{P,i}\}$. Then Assumption 1(iii) is immediately violated. Assumption 1(ii) is also violated since the officer's draws of $\{\mathcal{U}^s_{S,i}\}$ may differ for drivers i and j of race r with $\mathcal{U}^1_{P,i}(r) \neq \mathcal{U}^1_{P,j}(r)$. The independence between $\{\mathcal{U}^p_{P,i}\}$ and (Z_i, V_i^{pre}) is not require and is imposed to simplify the model.

Note that Assumption B1 does not imply there is no relationship an officers stop and search preferences. That is, it does not preclude officers who are eager to stop minority drivers to also be eager to search minority drivers. Instead, it imposes

that the draws of the random utilities/thresholds in the stop and search decision are independent of each other. This is admittedly a strong assumption, but without it, other strong assumptions are required in order to detect bias in searches while explicitly modeling traffic stops.

The probability the officer stops a driver is then

$$\begin{aligned}
&\mathbb{P}\{Stop_i = 1 \mid R_i = r, Z_i = z, V_i^{pre} = v\} \\
&= \mathbb{P}\{\mathbb{E}[\mathcal{U}_{S,i}^{Search_i}(Guilty_i; R_i) \mid R_i, Z_i, V_i^{\text{pre}}] \geq T_i^{Stop} \mid R_i = r, Z_i = z, V_i^{pre} = v\} \\
&= F_{T^{Stop}|R}(\mathbb{E}[\mathcal{U}_{S,i}^{Search_i}(Guilty_i; R_i) \mid R_i = r, Z_i = z, V_i^{\text{pre}} = v] \mid r),
\end{aligned}$$

where the last equality follows from Corollary B1. To see that this probability depends on the risk of the driver, we can apply the law of iterated expectations to the expectation inside of the CDF,

$$\begin{aligned}
&\mathbb{E}[\mathcal{U}_{S,i}^{Search_i}(Guilty_i; R_i) \mid R_i = r, Z_i = z, V_i^{pre} = v] \\
&= \sum_{s=0}^{1} \mathbb{E}[\mathcal{U}_{S,i}^{s}(Guilty_i; R_i) \mid Search_i = s, R_i = r, Z_i = z, V_i^{pre} = v] \times \\
&\qquad\qquad \mathbb{P}\{Search_i = s \mid R_i = r, Z_i = z, V_i^{pre} = v\}.
\end{aligned}$$

Consider the terms in the summand when $s = 1$. Applying the law of iterated expectations again, I have

$$\begin{aligned}
&\mathbb{E}[\mathcal{U}_{S,i}^{1}(Guilty_i; R_i) \mid S_i = 1, R_i = r, Z_i = z, V_i^{pre} = v] \\
&= \mathbb{E}[\mathbb{E}[\mathcal{U}_{S,i}^{1}(Guilty_i; R_i) \mid S_i = 1, R_i = r, Z_i = z, V_i] \mid S_i = 1, R_i = r, Z_i = z, V_i^{pre} = v] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathcal{U}_{S,i}^{1}(Guilty_i; R_i) \mid \underbrace{G(r, z, V_i)}_{\text{Risk}} \geq T_i^{Search}, R_i = r, Z_i = z, V_i\right] \middle| \begin{array}{l} S_i = 1, R_i = r, \\ Z_i = z, V_i^{pre} = v \end{array}\right]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z, V_i^{pre} = v\} \\
&= \mathbb{E}[\mathbb{P}\{Search_i = 1 \mid R_i = r, Z_i = z, V_i\} \mid R_i = r, Z_i = z, V_i^{pre} = v] \\
&= \mathbb{E}[F_{T^{Search}|R}(\underbrace{G(r, z, V_i)}_{\text{Risk}} \mid r) \mid R_i = r, Z_i = z, V_i^{pre} = v],
\end{aligned}$$

where the last equality follows from the model in Section 3 of the main paper.

Suppose that the reduced form relationship between $\mathbb{P}\{Stop_i = 1 \mid R_i = r, Z_i = z\}$ and $G(R_i, Z_i, V_i)$ is as shown in the top panel of Figure 12, where the officer has a 1% probability of stopping a driver with zero risk, and a 50% probability of stopping a driver with unit risk. Denote this relationship by by $\pi_{r,z}$, i.e.,

$$\pi_{r,z}(g) \equiv \mathbb{P}\{Stop_i = 1 \mid R_i = r, Z_i = z, G_i = g\}.$$

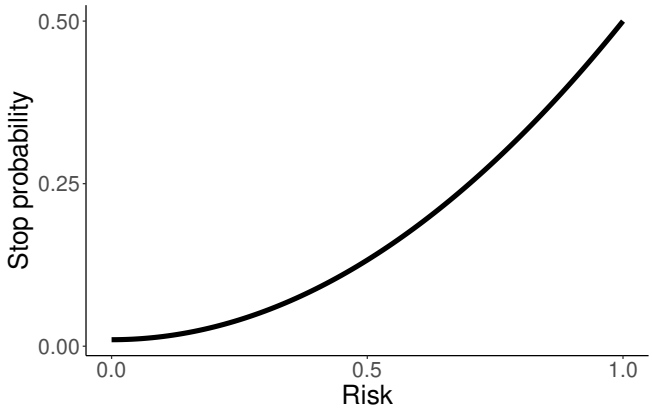
Suppose also that the population distribution of risk is as shown in the middle panel, and is equal to a beta distribution with shape parameters 1 and 9 and a mean of 0.1, i.e., 10% of drivers carry contraband. Denote the density by $f_{G|R,Z}(\cdot \mid r, z)$. Then conditional on being stopped, the distribution of risk is as shown in the bottom panel, and maybe written as

$$f_{G|Stop,R,Z}(g \mid 1, r, z) = \frac{\pi_{r,z}(g)\ f_{G|R,Z}(g \mid r, z)}{\int_0^1 \pi_{r,z}(g')\ f_{G|R,Z}(g' \mid r, z)\ dg'}.$$

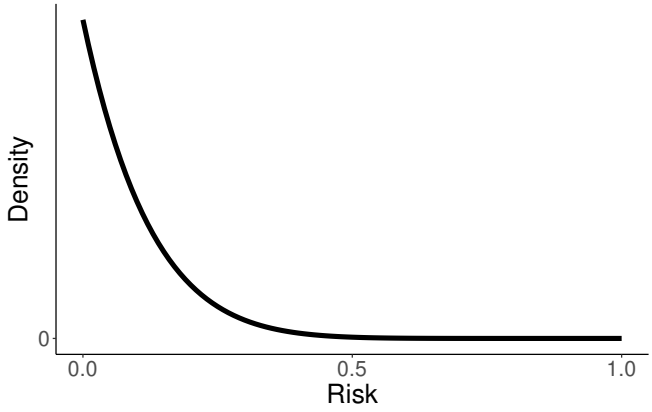
So in spite of the officer's preference for stopping high-risk drivers, the proportion of low-risk drivers is sufficiently large so that the density of risk post-stop is strictly decreasing.

Figure 12: Monotone-decreasing density for risk of drivers stopped

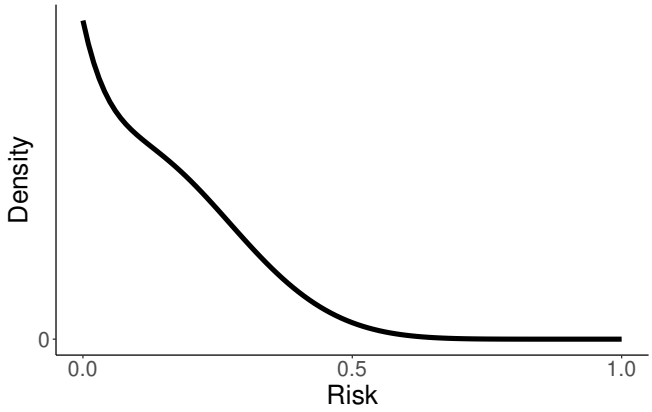
(a) Probability of stopping a driver



(b) Population distribution of risk



(c) Sample distribution of risk



B.3 Recovering the identified sets $\mathcal{B}_k$ and $\mathcal{E}$

The identified set for $\beta(g_k)$ may be recovered by solving

$$Q^{\star}_{\beta_k}(b) \equiv \min_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{w,z}\}, \{\mathbf{p}_{m,z}\}} \sum_{r,z} \left| \varsigma_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z} \right| + \sum_{r,z} \left| (\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z} \right|$$

$$\text{s.t.} \quad \varsigma_{m,k} - \varsigma_{w,k} = b$$

$$\mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}$$

for all $b \in [-1, 1]$. Then $b \in \mathcal{B}_k$ if and only if $Q^{\star}_{\beta_k}(b) = 0$.

Likewise, the identified set for $\mathbb{E}[\beta(G_i); \omega]$ may be recovered by solving

$$Q^{\star}_{\mathbb{E}[\beta(G);\omega]}(b) \equiv \min_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{w,z}\}, \{\mathbf{p}_{m,z}\}} \sum_{r,z} \left| \varsigma_r' \mathbf{p}_{r,z} - \mathbf{m}^S_{r,z} \right| + \sum_{r,z} \left| (\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \mathbf{m}^H_{r,z} \right|$$

$$\text{s.t.} \quad \omega'(\varsigma_m - \varsigma_w) = b$$

$$\mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b}$$

for all $b \in [-1, 1]$. Then $b \in \mathcal{E}$ if and only if $Q^{\star}_{\mathbb{E}[\beta(G);\omega]}(b) = 0$.

B.4 Constructing confidence intervals for $\beta(g_k)$ and $\mathbb{E}[\beta(G_i); \omega]$

The confidence intervals for $\beta(g_k)$ for $k = 1, \ldots, K$ may be constructed by inverting the test for racial bias. To determine whether $b \in [-1, 1]$ is in the confidence interval, the researcher must solve

$$\widehat{Q}^{\star}_{\beta(g_k)}(b) \equiv \min_{\varsigma_w, \varsigma_m, \{\mathbf{p}_{r,z}\}} \sum_{r,z} \widehat{\mathbf{w}}^{S}_{r,z} \left| \varsigma_r' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{S}_{r,z} \right| + \sum_{r,z} \widehat{\mathbf{w}}^{H}_{r,z} \left| (\mathbf{g} \odot \varsigma_r)' \mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{H}_{r,z} \right|$$

$$\text{s.t.} \quad \varsigma_{m,k} - \varsigma_{w,k} = b$$

$$\mathbf{A} \begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

which is the BP problem in (16) with the additional constraint that the intensity of bias at g_k is equal to b. The researcher can then construct the test statistic

$$\widehat{\tau}_{\beta(g_k)}(b) = \frac{\widehat{Q}^{\star}_{\beta(g_k)}(b) - \widehat{Q}^{\star}_{B}}{\widehat{Q}^{\star}_{B}},$$

which compares the fit of the model when the officer is restricted to have $\beta(g_k) = b$ against the fit without the restriction. The distribution of $\widehat{\tau}_{\beta(g_k)}(b)$ may be estimated using the bootstrap, and the hypothesis that $\beta(g_k) = b$ is rejected if the α-quantile of the bootstrap distribution is sufficiently large, for some value of $\alpha \in [0, 1]$. If the hypothesis is not rejected, then b enters into the $(1 - \alpha)$-confidence interval of $\beta(g_k)$. Again, this heuristic approach is not guaranteed to generate confidence intervals with the correct coverage probabilities, but may still be informative and is a stand-in until

a formal method for inference is developed.

The confidence intervals for $\mathbb{E}[\beta(G_i);\omega]$ maybe constructed in the same way. First solve

$$\widehat{Q}^{\star}_{\beta(g_k)}(b) \equiv \min_{\varsigma_w,\varsigma_m,\{\mathbf{p}_{r,z}\}} \sum_{r,z} \widehat{\mathbf{w}}^{S}_{r,z} \left|\varsigma_r'\mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{S}_{r,z}\right| + \sum_{r,z} \widehat{\mathbf{w}}^{H}_{r,z} \left|(\mathbf{g}\odot\varsigma_r)'\mathbf{p}_{r,z} - \widehat{\mathbf{m}}^{H}_{r,z}\right|$$

$$\text{s.t.} \quad \mathbf{q}_w'\mathbf{P}_w(\varsigma_m - \varsigma_w) = b$$

$$\mathbf{A}\begin{bmatrix} \varsigma_w \\ \varsigma_m \\ \mathbf{p}_{w,1} \\ \vdots \\ \mathbf{p}_{m,|\mathcal{Z}|} \end{bmatrix} \leq \mathbf{b},$$

which is the BP problem in (16) with the additional constraint that the average intensity of bias is equal to b. The following test statistic may then be constructed,

$$\widehat{\tau}_{\mathbb{E}[\beta(G_i);\omega]}(b) = \frac{\widehat{Q}^{\star}_{\mathbb{E}[\beta(G_i);\omega]}(b) - \widehat{Q}^{\star}_{B}}{\widehat{Q}^{\star}_{B}},$$

which compares the fit of the model when the officer is restricted to have $\mathbb{E}[\beta(G_i);\omega] = b$ against the fit without the restriction. The same bootstrap procedure above may then be applied to determine whether b is in the confidence interval of $\mathbb{E}[\beta(G_i);\omega]$.

C Bernstein polynomials

In this section, I briefly discuss some properties of Bernstein polynomials. See Farouki and Rajan (1988), Doha et al. (2011), and Farouki (2012) for more details.

The Bernstein basis of degree L is defined by

$$\mathbf{b}_l^L(g) \equiv \binom{L}{l}(1-g)^{L-l}g^l$$

for $l = 0, \ldots L$ and $g \in [0, 1]$. A Bernstein polynomial of degree L has the form

$$f(g) = \sum_{l=0}^{L} \theta_l \mathbf{b}_l^L(g)$$

for some $\theta \equiv (\theta_0, \ldots, \theta_L)$.

Suppose $\sigma^\star$ is modeled as a Bernstein polynomial, i.e., $\sigma^\star(g) = f(g)$. To see how this affects the bilinear program, consider the bilinear terms $\varsigma' \mathbf{p}_{r,z}$ from the objective function of the bilinear program in Proposition 2. These terms become

$$\varsigma' \mathbf{p}_{r,z} = \sum_{k=0}^{K} \sum_{l=0}^{L} \mathbf{b}_l^L(g_k) \underbrace{\theta_l \mathbf{p}_{r,z,k}}_{\text{Bilinear terms}},$$

where $\{\mathbf{b}_l^L(g_k)\}_{l=0,\ldots,L;k=0\ldots K}$ are known values. The BP program optimizes over θ and $\{\mathbf{p}_{r,z}\}$.

Imposing shape constraints on Bernstein polynomials is straightforward. The polynomial $f(g)$ satisfies

$$\min_l \theta_l \leq f(g) \leq \max_l \theta_l.$$

So a Bernstein polynomial may be bounded above or below simply by bounding its

coefficients. For example, f may be constrained to be in the unit interval by imposing the restriction $0 \leq \theta_l \leq 1$ for $l = 1, \ldots, L$.

To impose that f is monotonic increasing, add the restriction

$$\theta_0 \leq \theta_1 \leq \cdots \leq \theta_L,$$

which has the same form as (B.8). To impose that f is monotonic decreasing, simply reverse the inequalities.

The derivative of a Bernstein polynomial is also a Bernstein polynomial. So the shape constraints above may be used to constrain the derivatives of $f(g)$ as well. Doha et al. (2011) show that the q^{th} derivative of $f(g)$ is

$$f^{(q)}(g) = \sum_{l=0}^{L} \sum_{i=-q}^{q} \theta_{l-i}\ C_i(l, L, q)\ \mathbf{b}_l^L(g),$$

where

$$C_i(l, L, q) = q! \sum_{j=0}^{q} (-1)^{j+q} \binom{q}{j} \binom{l}{j+i} \binom{L-l}{q-j-i}.$$

So $f^{(q)}(g)$ is a Bernstein polynomial of degree L with coefficients $\left\{\sum_{i=-q}^{q} \theta_{l-i}\ C_i(l, L, q)\right\}_{l=0}^{L}$, where $C_i(l, L, q)$ are known constants. The derivatives may then also be restricted to fall within some interval and be monotonic.

A product of Bernstein polynomials is also a Bernstein polynomial. For instance, let

$$h(g) = \sum_{n=0}^{N} \pi_n \mathbf{b}_n^N(g)$$

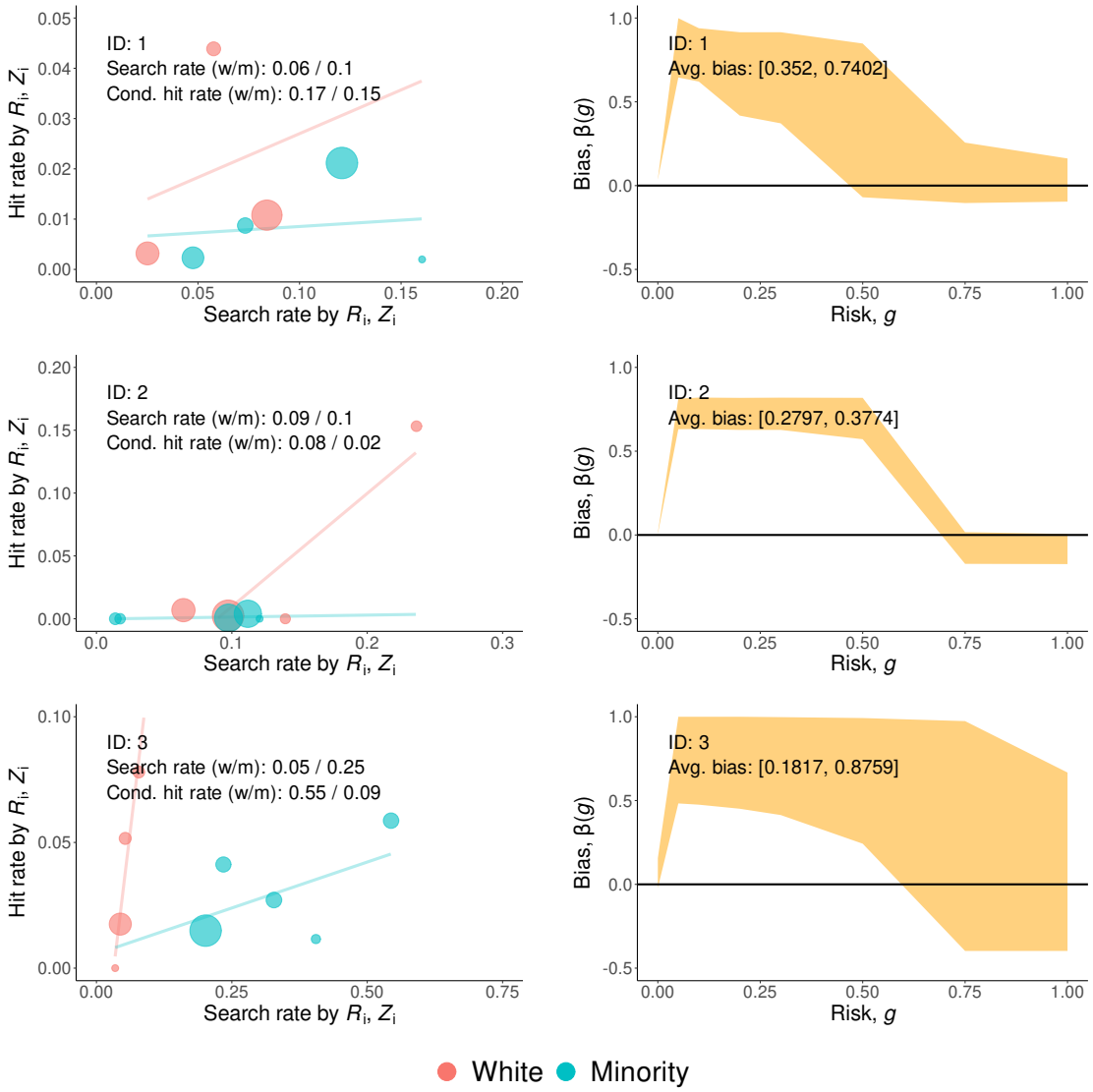
for some $\pi \equiv (\pi_0, \ldots, \pi_N)$. Then

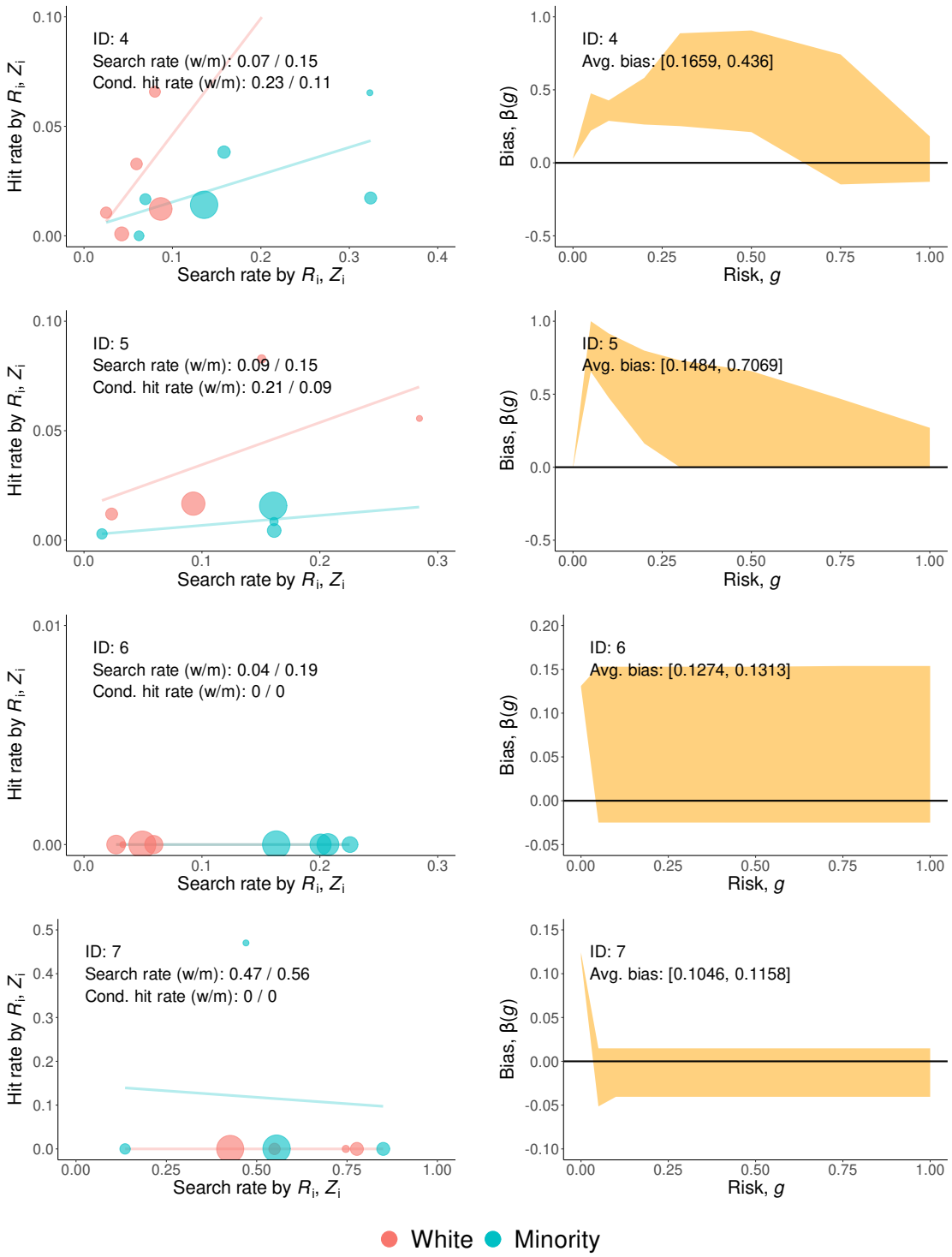
$$f(g)\ h(g) = \sum_{i=0}^{L+N} \left[\sum_{j=\max\{0,i-N\}}^{\min\{L,i\}} \frac{\binom{L}{j}\binom{N}{i-j}}{\binom{L+N}{i}} \underbrace{\theta_j\ \pi_{i-j}}_{\text{Bilinear terms}} \right] \mathbf{b}_i^{L+N}(g). \tag{C.10}$$

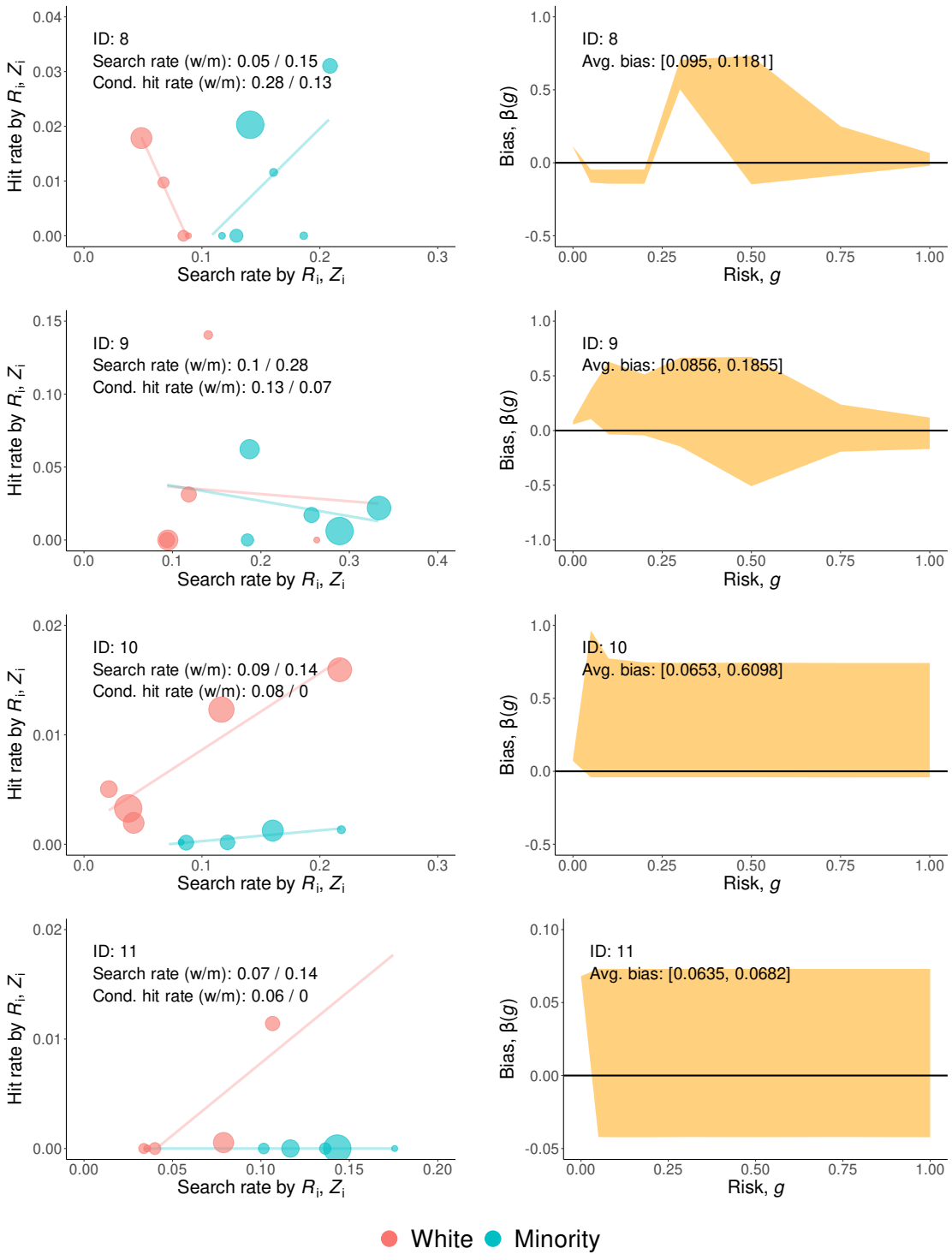
This means it is possible to model both σ and $\{\mathbf{p}_{r,z}\}$ as Bernstein polynomials. The bilinear program optimizes over θ and π.

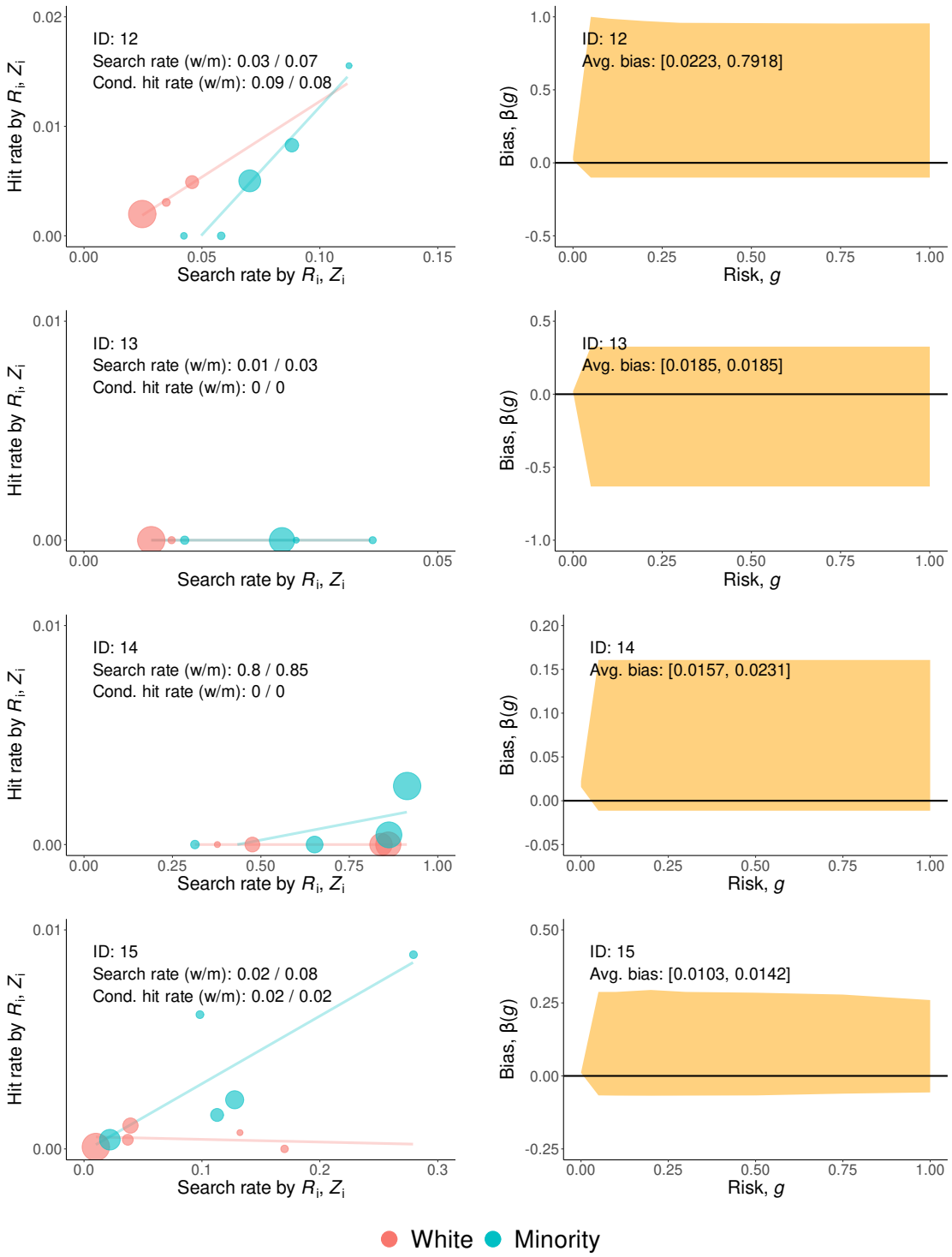
D Full set of estimates

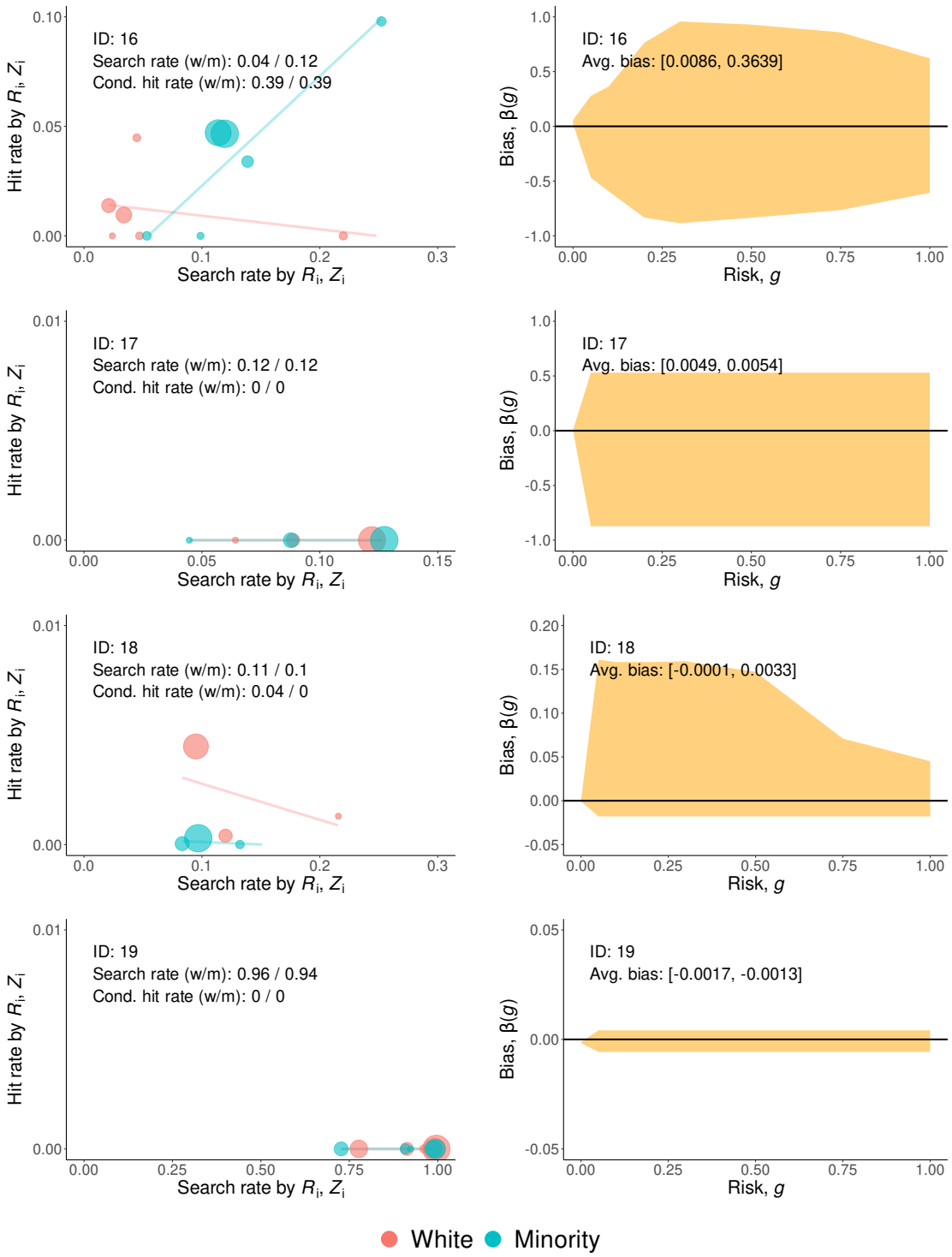
The weights ω are chosen so that $\mathbb{E}[\beta(G_i); \omega]$ measures the average difference in the probability that equally risky white and minority drivers are searched, under the counterfactual where the distribution of risk for minority drivers is equal to that of white drivers in the data.

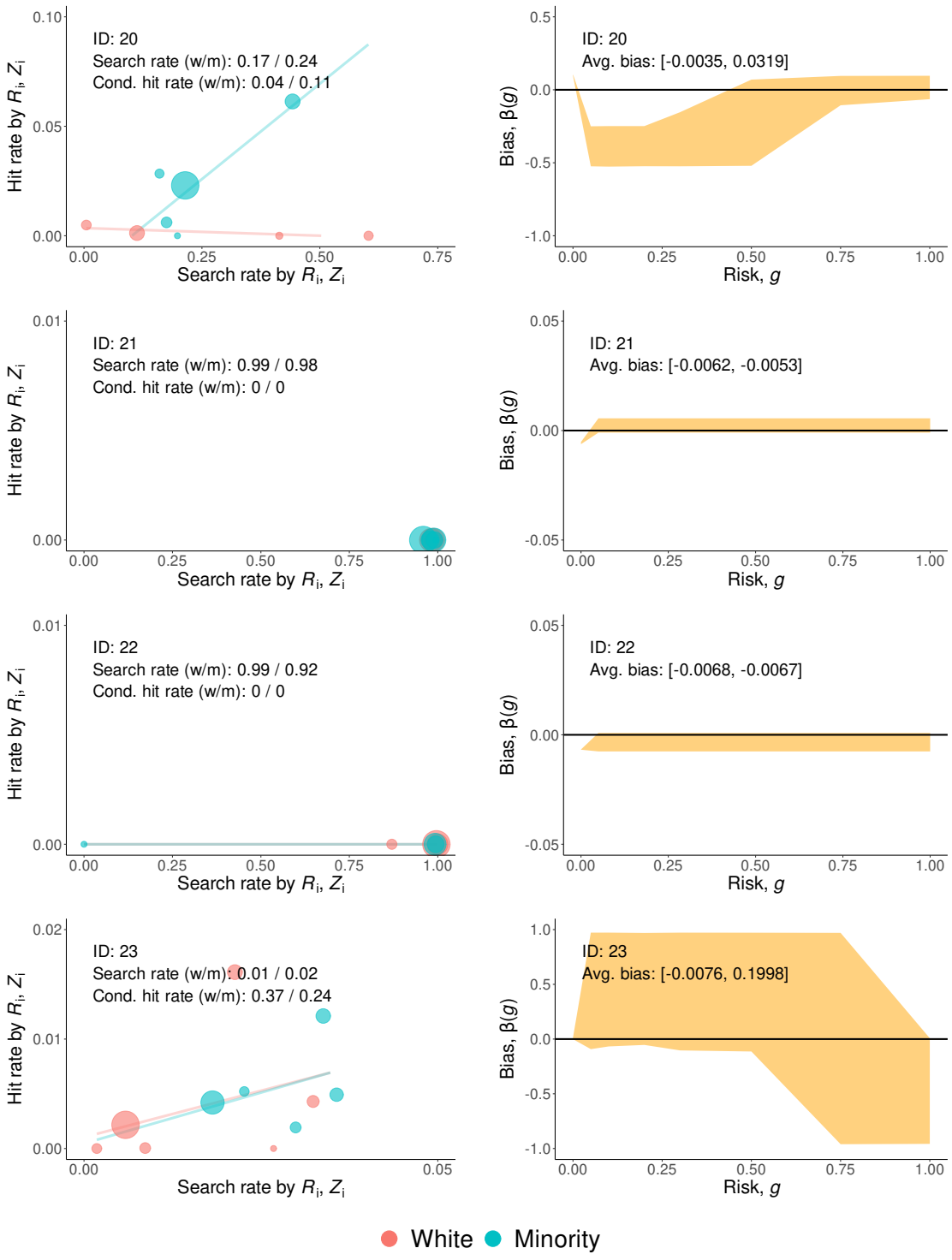


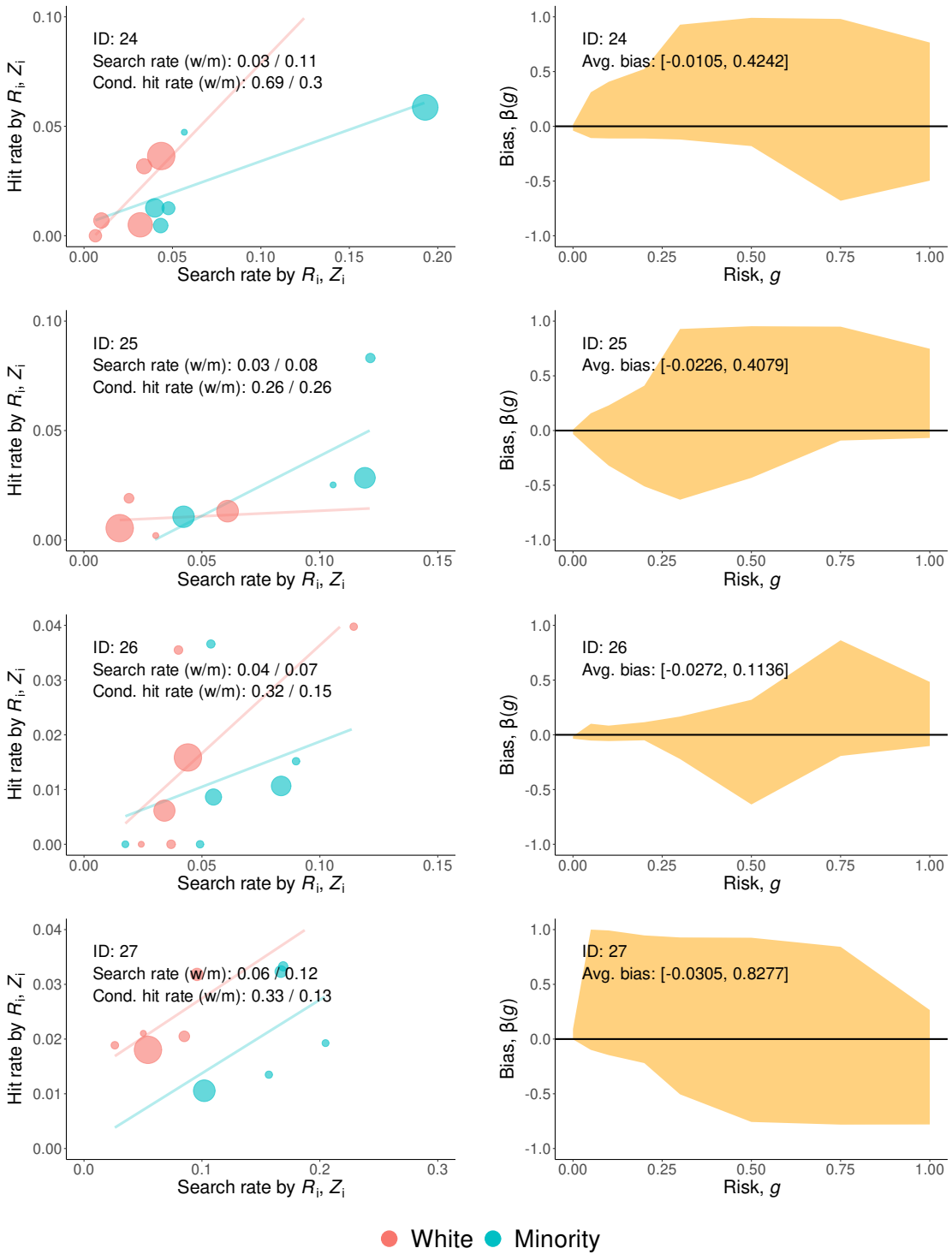

ID: 4
Search rate (w/m): 0.07 / 0.15
Cond. hit rate (w/m): 0.23 / 0.11
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 4
Avg. bias: [0.1659, 0.436]
Bias, β(g)
Risk, g
ID: 5
Search rate (w/m): 0.09 / 0.15
Cond. hit rate (w/m): 0.21 / 0.09
ID: 5
Avg. bias: [0.1484, 0.7069]
ID: 6
Search rate (w/m): 0.04 / 0.19
Cond. hit rate (w/m): 0 / 0
ID: 6
Avg. bias: [0.1274, 0.1313]
ID: 7
Search rate (w/m): 0.47 / 0.56
Cond. hit rate (w/m): 0 / 0
ID: 7
Avg. bias: [0.1046, 0.1158]
White
Minority

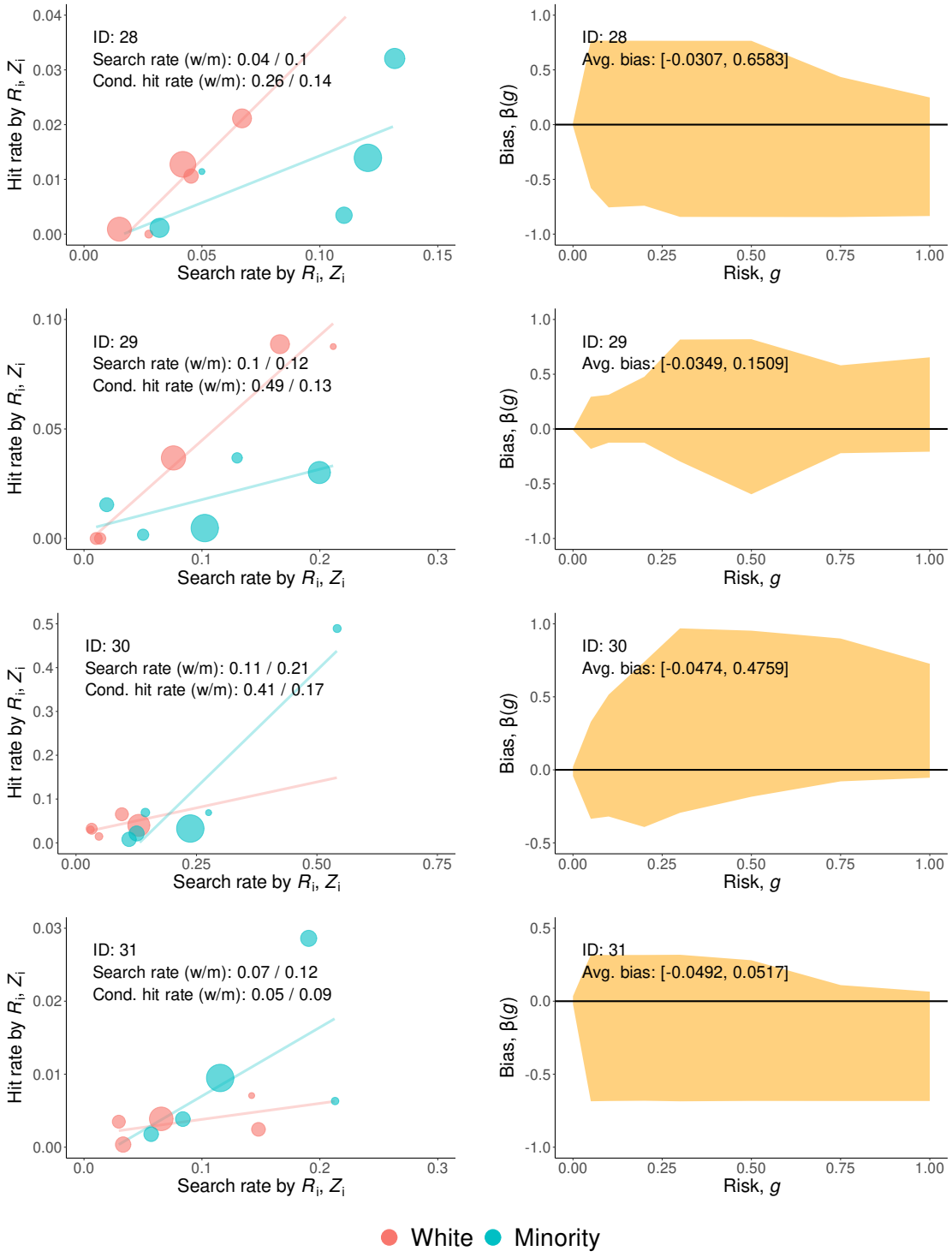

ID: 8
Search rate (w/m): 0.05 / 0.15
Cond. hit rate (w/m): 0.28 / 0.13
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 8
Avg. bias: [0.095, 0.1181]
Bias, β(g)
Risk, g
ID: 9
Search rate (w/m): 0.1 / 0.28
Cond. hit rate (w/m): 0.13 / 0.07
ID: 9
Avg. bias: [0.0856, 0.1855]
ID: 10
Search rate (w/m): 0.09 / 0.14
Cond. hit rate (w/m): 0.08 / 0
ID: 10
Avg. bias: [0.0653, 0.6098]
ID: 11
Search rate (w/m): 0.07 / 0.14
Cond. hit rate (w/m): 0.06 / 0
ID: 11
Avg. bias: [0.0635, 0.0682]
White
Minority

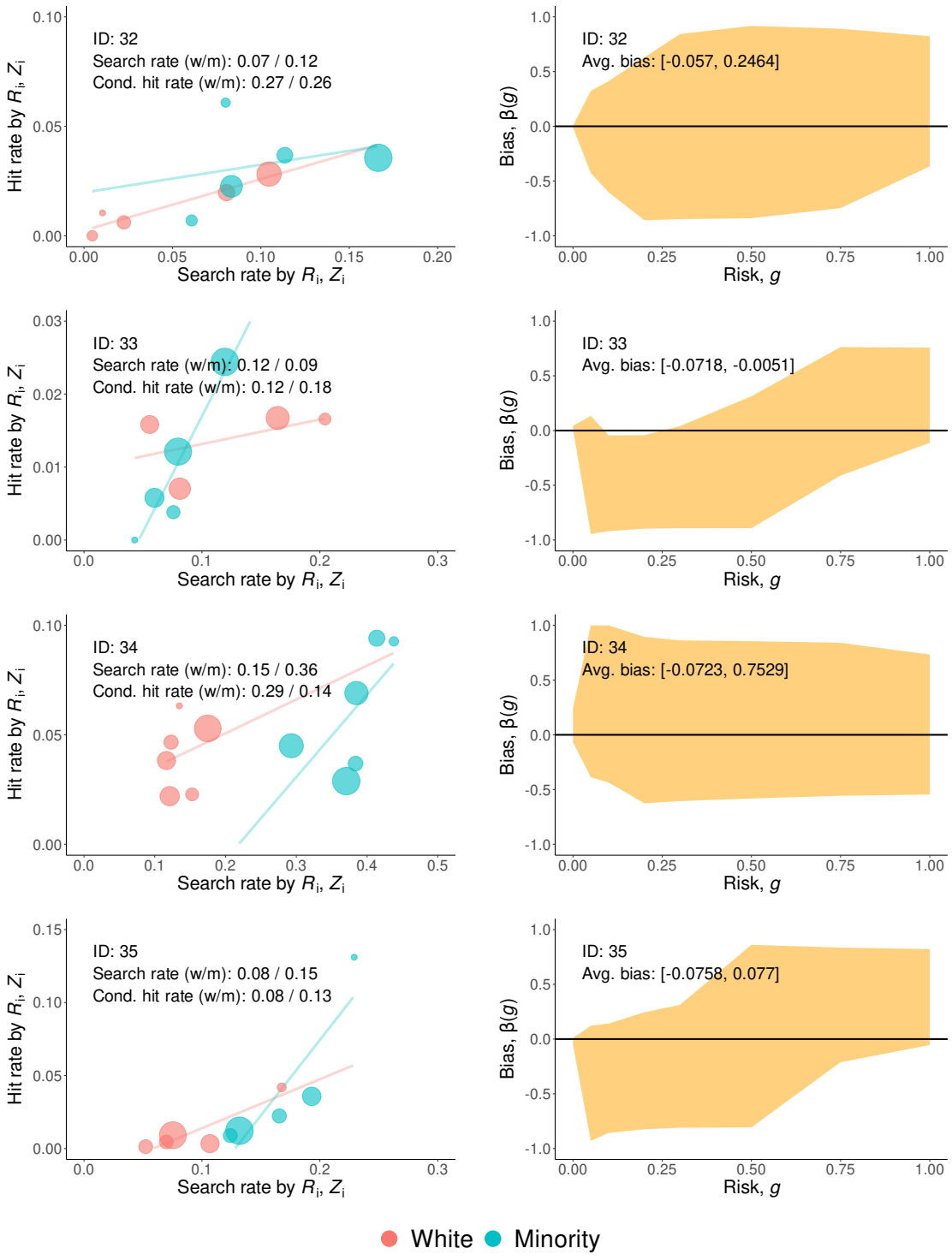

ID: 12
Search rate (w/m): 0.03 / 0.07
Cond. hit rate (w/m): 0.09 / 0.08
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 12
Avg. bias: [0.0223, 0.7918]
Bias, β(g)
Risk, g
ID: 13
Search rate (w/m): 0.01 / 0.03
Cond. hit rate (w/m): 0 / 0
ID: 13
Avg. bias: [0.0185, 0.0185]
ID: 14
Search rate (w/m): 0.8 / 0.85
Cond. hit rate (w/m): 0 / 0
ID: 14
Avg. bias: [0.0157, 0.0231]
ID: 15
Search rate (w/m): 0.02 / 0.08
Cond. hit rate (w/m): 0.02 / 0.02
ID: 15
Avg. bias: [0.0103, 0.0142]
White
Minority

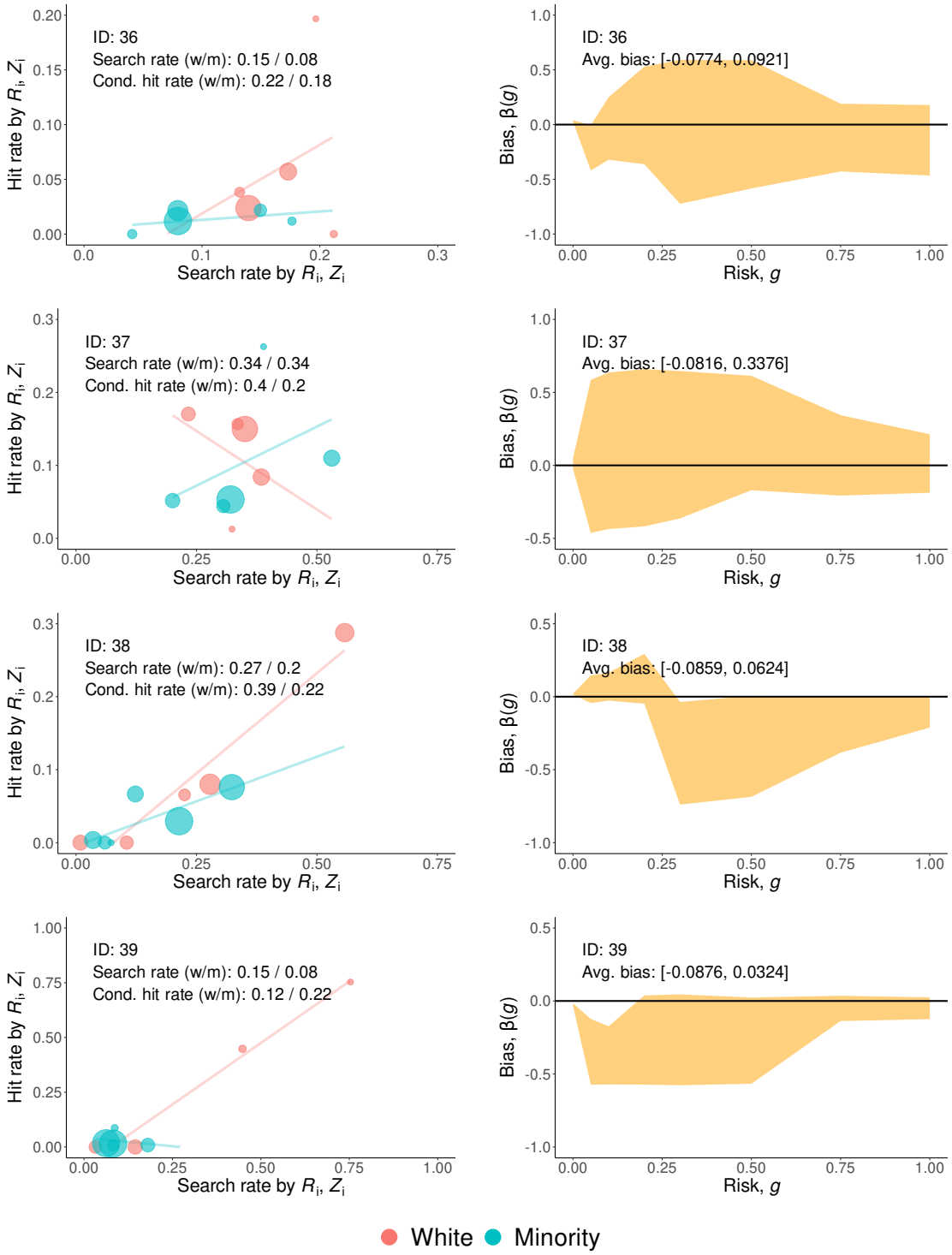

ID: 16
Search rate (w/m): 0.04 / 0.12
Cond. hit rate (w/m): 0.39 / 0.39
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 16
Avg. bias: [0.0086, 0.3639]
Bias, β(g)
Risk, g
ID: 17
Search rate (w/m): 0.12 / 0.12
Cond. hit rate (w/m): 0 / 0
ID: 17
Avg. bias: [0.0049, 0.0054]
ID: 18
Search rate (w/m): 0.11 / 0.1
Cond. hit rate (w/m): 0.04 / 0
ID: 18
Avg. bias: [-0.0001, 0.0033]
ID: 19
Search rate (w/m): 0.96 / 0.94
Cond. hit rate (w/m): 0 / 0
ID: 19
Avg. bias: [-0.0017, -0.0013]
White
Minority

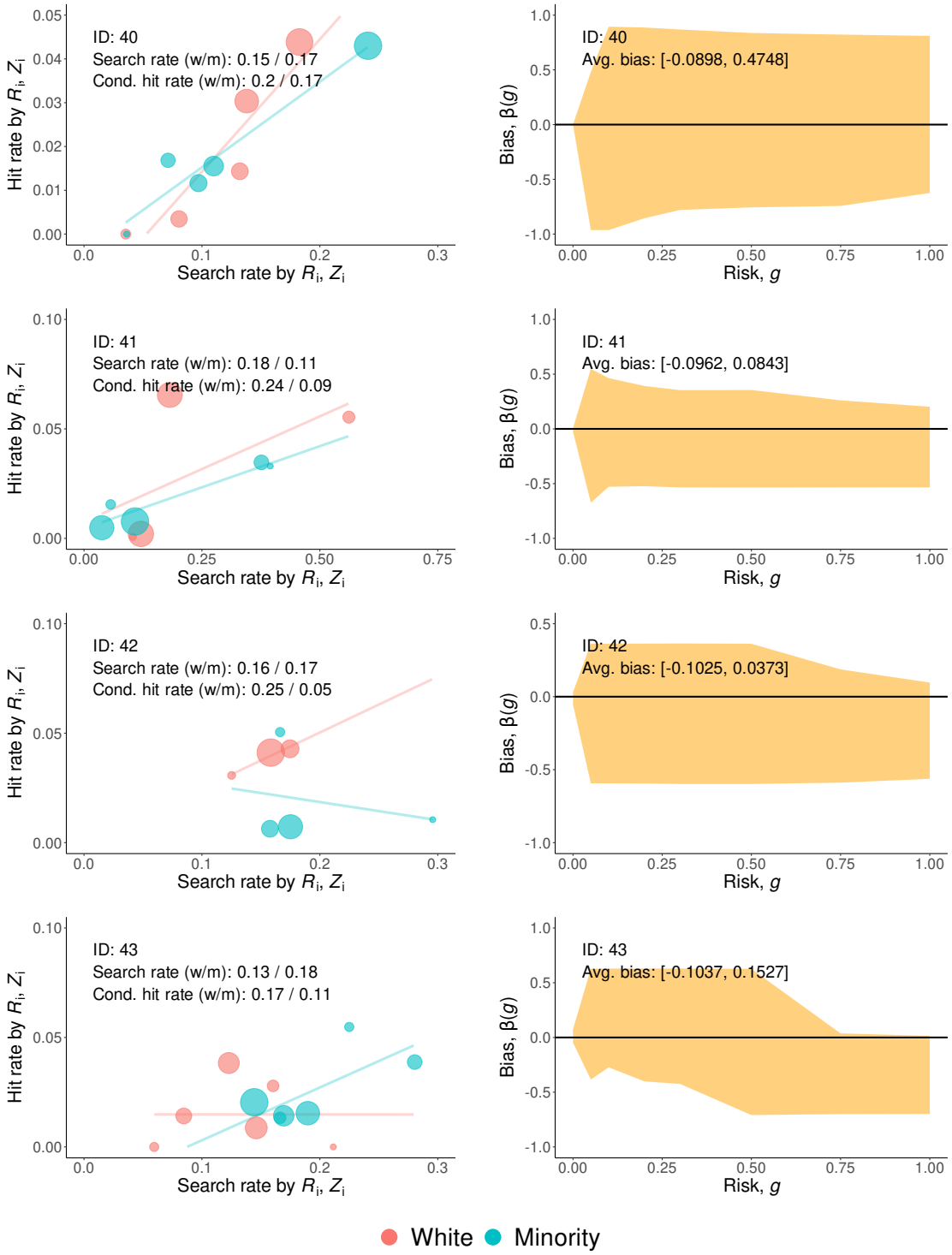

ID: 20
Search rate (w/m): 0.17 / 0.24
Cond. hit rate (w/m): 0.04 / 0.11
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 20
Avg. bias: [-0.0035, 0.0319]
Bias, β(g)
Risk, g
ID: 21
Search rate (w/m): 0.99 / 0.98
Cond. hit rate (w/m): 0 / 0
ID: 21
Avg. bias: [-0.0062, -0.0053]
ID: 22
Search rate (w/m): 0.99 / 0.92
Cond. hit rate (w/m): 0 / 0
ID: 22
Avg. bias: [-0.0068, -0.0067]
ID: 23
Search rate (w/m): 0.01 / 0.02
Cond. hit rate (w/m): 0.37 / 0.24
ID: 23
Avg. bias: [-0.0076, 0.1998]
White
Minority

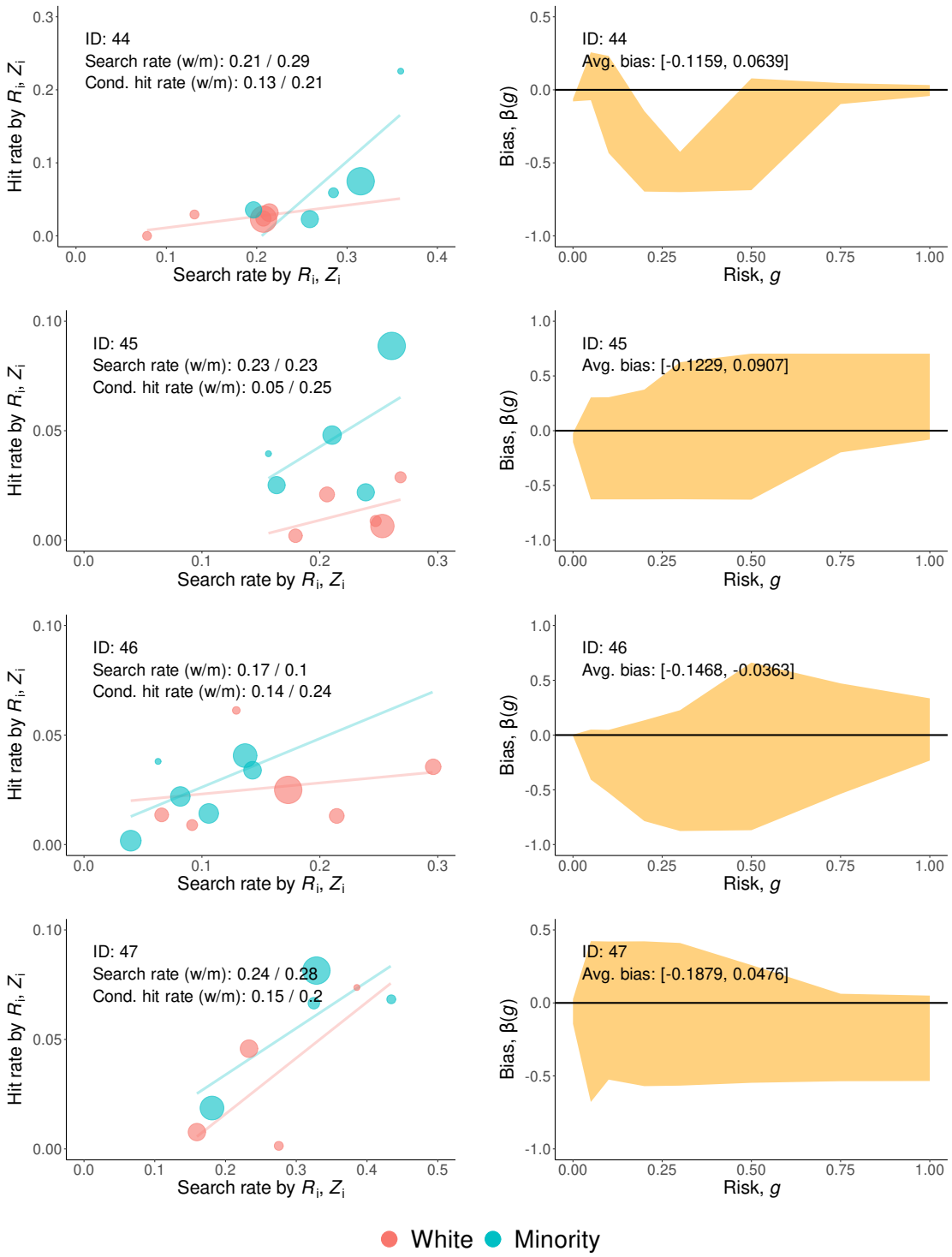

ID: 24
Search rate (w/m): 0.03 / 0.11
Cond. hit rate (w/m): 0.69 / 0.3
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 24
Avg. bias: [-0.0105, 0.4242]
Bias, β(g)
Risk, g
ID: 25
Search rate (w/m): 0.03 / 0.08
Cond. hit rate (w/m): 0.26 / 0.26
ID: 25
Avg. bias: [-0.0226, 0.4079]
ID: 26
Search rate (w/m): 0.04 / 0.07
Cond. hit rate (w/m): 0.32 / 0.15
ID: 26
Avg. bias: [-0.0272, 0.1136]
ID: 27
Search rate (w/m): 0.06 / 0.12
Cond. hit rate (w/m): 0.33 / 0.13
ID: 27
Avg. bias: [-0.0305, 0.8277]
White
Minority

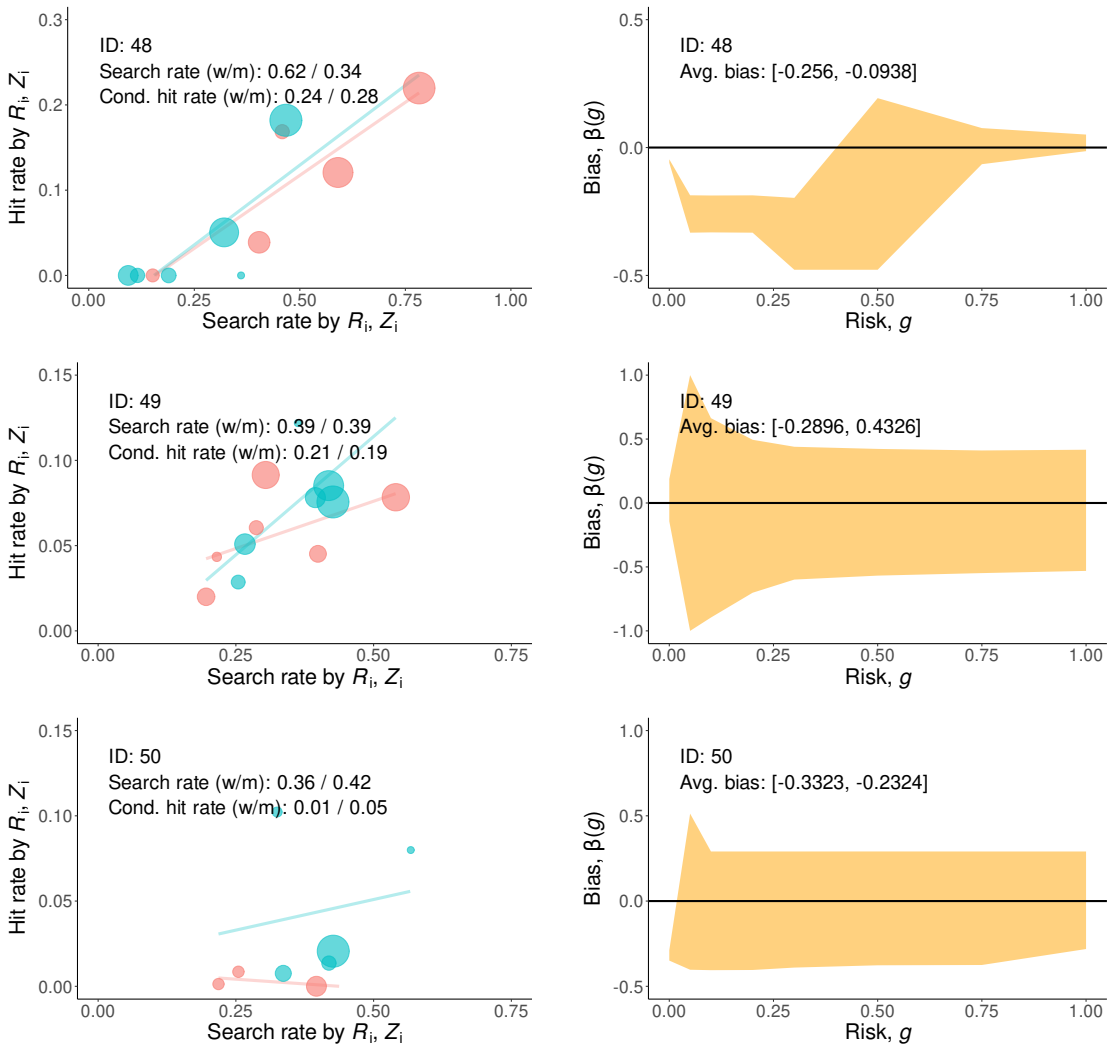

ID: 28
Search rate (w/m): 0.04 / 0.1
Cond. hit rate (w/m): 0.26 / 0.14
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 28
Avg. bias: [-0.0307, 0.6583]
Bias, β(g)
Risk, g
ID: 29
Search rate (w/m): 0.1 / 0.12
Cond. hit rate (w/m): 0.49 / 0.13
ID: 29
Avg. bias: [-0.0349, 0.1509]
ID: 30
Search rate (w/m): 0.11 / 0.21
Cond. hit rate (w/m): 0.41 / 0.17
ID: 30
Avg. bias: [-0.0474, 0.4759]
ID: 31
Search rate (w/m): 0.07 / 0.12
Cond. hit rate (w/m): 0.05 / 0.09
ID: 31
Avg. bias: [-0.0492, 0.0517]
White
Minority


ID: 32
Search rate (w/m): 0.07 / 0.12
Cond. hit rate (w/m): 0.27 / 0.26
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 32
Avg. bias: [-0.057, 0.2464]
Bias, β(g)
Risk, g
ID: 33
Search rate (w/m): 0.12 / 0.09
Cond. hit rate (w/m): 0.12 / 0.18
ID: 33
Avg. bias: [-0.0718, -0.0051]
ID: 34
Search rate (w/m): 0.15 / 0.36
Cond. hit rate (w/m): 0.29 / 0.14
ID: 34
Avg. bias: [-0.0723, 0.7529]
ID: 35
Search rate (w/m): 0.08 / 0.15
Cond. hit rate (w/m): 0.08 / 0.13
ID: 35
Avg. bias: [-0.0758, 0.077]
White
Minority


ID: 36
Search rate (w/m): 0.15 / 0.08
Cond. hit rate (w/m): 0.22 / 0.18
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 36
Avg. bias: [-0.0774, 0.0921]
Bias, β(g)
Risk, g
ID: 37
Search rate (w/m): 0.34 / 0.34
Cond. hit rate (w/m): 0.4 / 0.2
ID: 37
Avg. bias: [-0.0816, 0.3376]
ID: 38
Search rate (w/m): 0.27 / 0.2
Cond. hit rate (w/m): 0.39 / 0.22
ID: 38
Avg. bias: [-0.0859, 0.0624]
ID: 39
Search rate (w/m): 0.15 / 0.08
Cond. hit rate (w/m): 0.12 / 0.22
ID: 39
Avg. bias: [-0.0876, 0.0324]
White
Minority


ID: 40
Search rate (w/m): 0.15 / 0.17
Cond. hit rate (w/m): 0.2 / 0.17
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 40
Avg. bias: [-0.0898, 0.4748]
Bias, β(g)
Risk, g
ID: 41
Search rate (w/m): 0.18 / 0.11
Cond. hit rate (w/m): 0.24 / 0.09
ID: 41
Avg. bias: [-0.0962, 0.0843]
ID: 42
Search rate (w/m): 0.16 / 0.17
Cond. hit rate (w/m): 0.25 / 0.05
ID: 42
Avg. bias: [-0.1025, 0.0373]
ID: 43
Search rate (w/m): 0.13 / 0.18
Cond. hit rate (w/m): 0.17 / 0.11
ID: 43
Avg. bias: [-0.1037, 0.1527]
White
Minority


ID: 44
Search rate (w/m): 0.21 / 0.29
Cond. hit rate (w/m): 0.13 / 0.21
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 44
Avg. bias: [-0.1159, 0.0639]
Bias, β(g)
Risk, g
ID: 45
Search rate (w/m): 0.23 / 0.23
Cond. hit rate (w/m): 0.05 / 0.25
ID: 45
Avg. bias: [-0.1229, 0.0907]
ID: 46
Search rate (w/m): 0.17 / 0.1
Cond. hit rate (w/m): 0.14 / 0.24
ID: 46
Avg. bias: [-0.1468, -0.0363]
ID: 47
Search rate (w/m): 0.24 / 0.28
Cond. hit rate (w/m): 0.15 / 0.2
ID: 47
Avg. bias: [-0.1879, 0.0476]
White
Minority


ID: 48
Search rate (w/m): 0.62 / 0.34
Cond. hit rate (w/m): 0.24 / 0.28
Hit rate by R_i, Z_i
Search rate by R_i, Z_i
ID: 48
Avg. bias: [-0.256, -0.0938]
Bias, β(g)
Risk, g
ID: 49
Search rate (w/m): 0.39 / 0.39
Cond. hit rate (w/m): 0.21 / 0.19
ID: 49
Avg. bias: [-0.2896, 0.4326]
ID: 50
Search rate (w/m): 0.36 / 0.42
Cond. hit rate (w/m): 0.01 / 0.05
ID: 50
Avg. bias: [-0.3323, -0.2324]
White
Minority