



Detecting Bias in Traffic Searches: Examining False Searches of Innocent Drivers

Margaret A. Meyer¹ · Richard Gonzalez¹

Accepted: 18 February 2024
© The Author(s) 2024

Abstract

Purpose We test current models of racial bias in policing, identify limitations, and propose a test of racial bias, that does not depend on unknown population contraband rate.

Methods We conceptualize police officer search decisions as a 2 (search/no search) by 2 (contraband present/absent) table, with missing data (if the police did not search, the presence of contraband is unknown). We constrain the feasible problem space using properties of a 2 x 2 contingency table. Then we examine all possible feasible 2 x 2 tables to identify instances of racial differences in police officer hit and false alarm rates. To do this, we develop a new test of racial bias, the Overlapping Condition Test. We analyze state and county data across 25 United States police departments.

Results These departments have an observable racial difference in false alarm rate regardless of the true value of missing data (under every feasible 2 x 2 table there is a racial difference). This effect is found in 10 out of 14 state police departments and 9 out of 11 local departments across the United States. That is, for every feasible real world scenario police officers have lower false alarm rates for White drivers than Black drivers.

Conclusion We interpret this difference in false alarm rate as a threshold bias. That is, officers use different criteria for searching Black drivers than White drivers and this conclusion is not qualified by the unknown contraband rate. Future directions should explore how police officers make the decision to search drivers and develop interventions to address the racial bias in search rate.

Keywords Racial bias · Policing · Mathematical modeling · Law · Signal detection theory

How Can We Assess Bias in Traffic Searches?

A body of research aims to determine if police officers are biased, without capturing the performance of police officers who search drivers. First, we explore the utility of the hit and false alarm rates from the 2 x 2 table in the analysis of bias in traffic searches. Then,

✉ Margaret A. Meyer
margem@umich.edu

Richard Gonzalez
gonzo@umich.edu

¹ Psychology, University of Michigan, 530 Church St., Ann Arbor, MI 48109, USA

we explain the difficulty in modeling police officer searches without utilizing performance measures such as hit and false alarm rates. Finally, we explore definitions of bias in police search and explain some of the key debates in the literature using a common framework.

Data Reporting and the Police

In order to understand or model police officer bias from the available data we must understand the nature of reporting policing data. Issues with policing data stem from the way in which these data are recorded. Police officers are not subject to the same mandatory data reporting practices that exist in other institutions (Goff and Kahn 2012; Groves and Cork 2009). Beyond data-reporting practices, police officers can engage in a wide variety of behaviors before their conduct needs to be formally recorded as a “stop” or a “search.” Police officers can engage in “pre-selection” tactics to improve the likelihood of recovering contraband if they engage in a search. They can ask questions, poke their head inside the car, use a flashlight to look in the windows, and even engage in behaviors we would consider to be searches, but that do not rise to the occasion of a search under the United States Constitution’s Fourth Amendment (see Gross and Barnes 2002, for a full review). Finally, police officers often self report the race of the driver. This is problematic as a body of research suggests that police officers are mislabeling minority searches as White (Luh 2022; Kocieniewski 1999; Gross and Barnes 2002). This distinction in what constitutes a search and how it is reported becomes even more convoluted at the level of pedestrian stops. Thus, we restrict our discussion to police officer searches of driver vehicles during traffic stops.

2 x 2 Contingency Table

We explore a new model of police search behavior that relies upon a 2 x 2 table of officer decisions and real world outcomes. Prior to interacting with the police, a driver makes the decision to either carry contraband, or not. When patrol officers encounter drivers in routine stops, they must determine whether they believe the driver has contraband and thus the officer can choose to either search or not search the car. As a result, there are four potential outcomes of this situation:

1. The officer is correct, the driver they searched had contraband (hit)
2. The officer is wrong, the driver they searched did not have contraband (false alarm)
3. The officer is correct, the driver they did not search did not have contraband (correct rejection)
4. The officer is wrong, the driver they did not search had contraband (miss).

We can use a 2 x 2 contingency table to model police officer search decisions and outcomes (see Table 1), where the columns represent the real world (contraband present v. absent) and the rows represent officer decision (search v. no search). In this table, the row sums correspond to the total proportion of drivers searched and not searched, and the column sums correspond to the proportion of drivers who had or did not have contraband. A police officer might have engaged in reporting bias, or even engaged in tactics prior to the

Table 1 Probabilities for the 2 x 2 contingency table

	Contraband	No contraband	Total
Search	$p(S \cap C)$	$p(S \cap \bar{C})$	$p(S)$
No search	$p(\bar{S} \cap C)$	$p(\bar{S} \cap \bar{C})$	$p(\bar{S}) = 1 - p(S)$
Total	$p(C)$	$p(\bar{C}) = 1 - p(C)$	1

The bold values represents the cells containing data from publicly available policing databases. S refers to search; C refers to contraband; bar above S or C denotes negation

search, to increase their likelihood of recovering contraband, and this behavior, while noteworthy, cannot be observed or studied with the available data.¹

The four cells in the 2 x 2 table can also be expressed as joint probabilities (Table 1). These probabilities provide the structures necessary for understanding police officer performance on the task of searching drivers and for deriving appropriate measures of bias.

Hit and False Alarm Rate

A common method of analyzing 2 x 2 tables is to consider the hit and false alarm rate. An important distinction must be drawn. The hit rate discussed in the context of the 2 x 2 table is not the same as the “hit rate” as discussed in the policing literature. In particular, the 2 x 2 definition of hit rate is a measure of officer accuracy that is conditional upon the amount of contraband found. However, the “hit rate” discussed in the policing literature is the likelihood the officer will find contraband conditional upon their decision to search the driver, which we refer to as the positive predictive value (PPV).²

To calculate hit rate (a conditional probability based on the information in Table 1), we divide the joint probability that the driver possesses contraband and the officer searches by the probability that the driver possesses contraband

$$HR = \frac{p(S \cap C)}{p(C)}. \tag{1}$$

We calculate a police officer’s PPV, a different conditional probability using the information in Table 1, by dividing the joint probability that the driver possesses contraband and the officer searches by the probability the officer searches

$$PPV = \frac{p(S \cap C)}{p(S)}. \tag{2}$$

The numerators of these two equations are identical (the joint probability the driver possesses contraband and the officer searches). The distinction between these definitions is

¹ In this paper we do not consider situations where traffic stops were not made.

² This distinction can be further illustrated through an example. An officer with a 70% PPV will find contraband 70% of the time they search, while an officer with a 70% hit rate will have found contraband 70% of the time it was present. Some authors give a more careful definition of policing “hit rate” referring to it as a conditional hit rate (Shea 2022; Lundberg 2022) points out the incorrect use of chi-squared tests when comparing PPV for different groups.

in the denominator (the probability an officer conducts a search or the probability a driver possesses contraband).

The other metric of analyzing the 2 x 2 table we will discuss is the false alarm rate, yet a different conditional probability. To calculate false alarm rate, we divide the probability of false alarms by the probability the driver does not possess contraband

$$FA = \frac{p(S \cap \bar{C})}{p(\bar{C})}. \quad (3)$$

What do these metrics mean? Police officer hit-rate tells us how good an officer is at detecting contraband, when it exists. For instance, an officer with 90% hit rate will correctly identify 90% of drivers with contraband. False alarm rate represents how good an officer is at determining when contraband is not present, when it does not exist. An officer with 90% false alarm rate will incorrectly search 90% of drivers without contraband. In the context of policing, this represents the proportion of times an officer searched an innocent driver. It is important to recognize that a 90% hit rate does not automatically imply a 10% false alarm rate, because the two probabilities are logically independent. As a boundary condition, a perfect officer would have a hit rate of 100% and a false alarm rate of 0%. While these metrics are commonly used in other domains, such as the medical literature,³ they have yet to be considered in the policing literature.

Models of Search Behavior are Limited by the Data Available

While the hit and false alarm metrics assess officer bias, policing data sets have limited data. Policing data sets record if an officer made the decision to search and the outcome of those searches (see Table 1 where observed cells are denoted in yellow). If a police officer did not search the driver, we do not have data on whether or not the driver was/was not carrying contraband. This lack of data for non-searches has been referred to as the denominator problem (Knox and Mummolo 2020b; Walker 2001), and it limits current approaches to studying bias in police officer searches (see Goff and Kahn 2012, for a full review). For example, the hit and false alarm rates cannot be computed without knowing the base rate of contraband in the population. Knox and Mummolo have gone into great detail explaining limitations of studying racial bias without knowing the population base rate in the context of the causal inference literature (Knox et al. 2020; Knox and Mummolo 2020b, 2020). Our approach focuses on all feasible 2 x 2 tables, but the limitations are analogous.

We Can Still Use the 2 x 2 Table

Despite these limitations, we recognize that we only need to know one additional value to completely solve the 2 x 2 table. We denote the number of misses (i.e., times the police officer did not search but missed contraband) as x . Then, using the constraints imposed by the table: all entries sum to the total number of traffic stops and each entry in a row/column

³ Hit and false alarm rate are also known in the medical testing literature as sensitivity and specificity, respectively (Lalkhen and McCluskey 2008). Sensitivity refers to how “good” a diagnostic test is at identifying an unhealthy person, while specificity refers to how “good” a diagnostic test is at identifying a healthy person (the complement of the false alarm rate). In the context of policing that would be equivalent to how well an officer can detect the presence (sensitivity) or absence (specificity) of contraband.

Table 2 Convenient labels for the cells of the 2 x 2 table

	Contraband	No Contraband	Total
Search	a	b	S
No search	x	$\bar{S} - x$	\bar{S}
Total	C	\bar{C}	T

sums to the row/column total, we define all feasible values of x . This approach of considering all such scenarios is in the spirit of the causal inference approach using potential outcomes taken by Knox et al. (2020).

While we are not able to estimate the x -value in actual data sets without making additional assumptions or using additional data to enable estimation, its range can be constrained with only one assumption. The number of times an officer missed contraband when it was present (the x -value) can only be so large without the officer’s false alarm rate exceeding their hit rate. This would imply that the officer performs worse than chance, which we assume is not the case.

In particular, we assert that police officers’ hit rates should be at least as good as their false alarm rates⁴ (e.g., we assume $HR = p(S|C) \geq p(S|\bar{C}) = FA$). Surely, the police are searching more cars with contraband than cars without contraband. Psychologically, this assumption is supported by the idea that at the point officers make the decision to search (and it is formally recorded), police officers are hoping to find contraband, so it seems implausible that they would systematically perform worse than chance. Data reporting bias will likely be in favor of omitting false alarms (thus decreasing the false alarm rate). Moreover, we are currently discussing behavior in the aggregate, and individual officer might perform worse than chance, but police officers as a whole are performing at least as good as chance. With this assumption and the properties of the 2 x 2 table, we observe an upper bound on x .

To derive the upper bound on x , we return to the definitions of hit and false alarm rate (see Eqs. 1 and 3). We express the joint probability in each cell in Table 2 in terms of numerical values. The equation for hit rate is

$$HR = \frac{a}{a + x}. \tag{4}$$

The equation for false alarm rate is

$$FA = \frac{b}{b + (\bar{S} - x)}. \tag{5}$$

We then constrain the hit rate to be greater than or equal to the false alarm rate and some simple algebra yields

$$x \leq \bar{S} \frac{a}{a + b}. \tag{6}$$

The rightmost term is the PPV (Eq. 2), so the upper bound is the product of the complement of the search rate and the PPV.

⁴ This assumption is analogous to setting the sensitivity parameter in Signal Detection Theory to 0.

Table 3 Estimating the 2 x 2 table in Durham NC

	Contraband	No contraband	Total
Search	5632	16074	21706
No Search	x	$304318 - x$	304318
Total	$5632 + x$	$16074 + (304318 - x)$	326024

We can identify all feasible 2 x 2 tables using the upper bound on x . This approach can be applied to policing data directly. For example, we use data from the county of Durham, North Carolina from December 2001–2015 (a total of 326,024 traffic stops). These data were collected by the Durham police department and made accessible by the Stanford Open Policing Consortium (Pierson et al. 2020).⁵ In Table 3, we see the current known values for these data.

Using these known values, we calculate this upper bound on the possible x -values. In particular, the x -value is constrained by Eq. 6:

$$x \leq 304318 * \frac{5632}{16074 + 5632} \quad (7)$$

$$x \leq 78960.$$

Originally, the unconstrained x ranged from $x \in [0, 304318]$, but now $x \in [0, 78960]$; this constraint on x reduced the possible values to 25% of the original possibilities.⁶

Defining Racial Bias

A 2 x 2 Definition of Bias

To understand racial differences in officer performance, we need to examine multiple 2 x 2 contingency tables like the one shown in Table 1. For example, considering only two races, White and Black/African American, we would have one 2 x 2 for the White drivers and one 2 x 2 for the Black drivers. Racial bias would then be defined by measures that examine differences across two or more 2 x 2 contingency tables. This definition of racial bias is distinct from the bias metrics that can be computed within a single 2 x 2. Thus, we define “racial bias” as racial difference between metrics of performance. The two metrics we will compare across 2 x 2 tables are hit and false alarm rates.

To compare these hit and false alarm rates across race, we need to know the missing x -values in each 2 x 2 table, yet this is not possible with currently available data. We propose an alternative. We examine all feasible x -values for both White and Black drivers and compute the resulting differences between the metrics across the two sets of feasible 2 x 2 tables. Continuing with the example of Durham, North Carolina, we observe two 2 x 2 tables, one for White and one for Black drivers (see Table 4 below).⁷

⁵ The interested reader can access these open source data at <https://openpolicing.stanford.edu/data/>.

⁶ Because a partial stop is not possible, we round all decimals down to the nearest integer (i.e., apply a floor function) for purposes of reporting values for the 2 x 2 table in this manuscript.

⁷ The total Durham data contains drivers of all races. We dropped drivers whose race was categorized as Asian/Pacific Islander, Hispanic, or other due to low sample size.

Table 4 Durham NC 2 x 2 table by race

Black drivers			
	Contraband	No contraband	Total
Search	4203	11179	15382
No Search	x	$165469 - x$	165469
Total	$4203 + x$	$176648 - x$	180851
White drivers			
	Contraband	No contraband	Total
Search	957	2565	3522
No Search	x	$93893 - x$	93893
Total	$957 + x$	$96458 - x$	97415

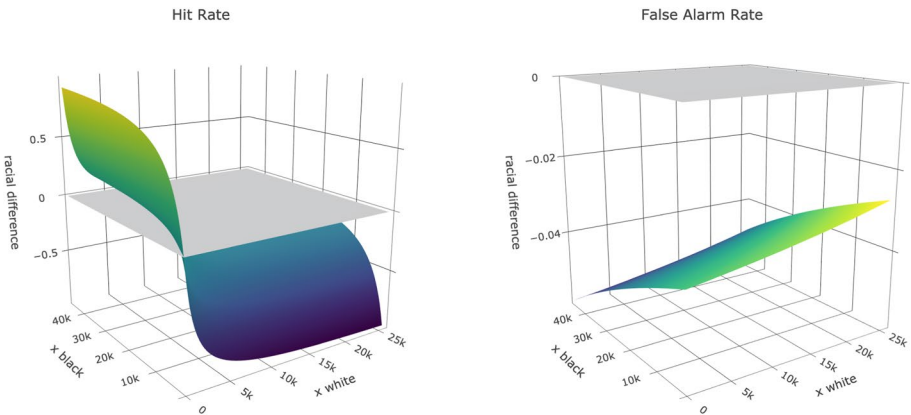


Fig. 1 Durham NC Data (Table 4): Racial difference in hit rate (left) and false alarm rate (right) for all possible combinations of x -values

Using Eq. 6 we know the feasible $x_{black} \in [0, 45212]$ and the feasible $x_{white} \in [0, 25512]$. We can compute the difference in hit rate for Black and White drivers for each combination of x -values. A useful tool for visualizing how the hit rate differs for White and Black drivers is a three dimensional surface resulting from placing the feasible x -values for White drivers on one axis, the feasible x -values for Black drivers on another axis, and the difference between the hit rates corresponding to those two x -values on the z -axis. This creates a three-dimensional surface with points above the plane $z = 0$ representing a higher hit rate for White drivers and points below the plane $z = 0$ representing a hit rate that is higher for Black drivers. As seen in the left panel of Fig. 1,⁸ there are combinations of feasible x -values that show higher hit rates for White drivers and higher hit rates for Black drivers. Given that the feasible tables can produce differences in hit rates between the two races on

⁸ For this plot and future plots, two of the axes are the feasible x -values for White and Black drivers and they each have different ranges.

either side of 0, we cannot make claims about the pattern in the real world without more data or making more assumptions. But what about officer false alarm rates?

We can also use the feasible x -value to examine racial differences in false alarm rates. Using the same constraint on the x -values, we plotted the racial difference in false alarm rate as a three-dimensional surface (the right panel in Fig. 1). Surprisingly, for any combination of x -values we observe the false alarm rate for Black drivers is greater than the false alarm rate for White drivers (i.e., the feasible surface is below the plane $z = 0$). This means that for all feasible real-world scenarios, the false alarm rate for Black drivers (FA_B) is higher than that for White drivers (FA_W). While we do not know the actual values of x for White and Black drivers, we can see that for all feasible values the racial difference is always in the same direction. This racial difference is staggering and requires further investigation.

The Overlapping Condition Test

The three-dimensional surfaces (Fig. 1) demonstrate that for any combination of feasible x -values, $FA_B > FA_W$. However, producing three-dimensional plots is time consuming and can be difficult to interpret for those with limited mathematical knowledge. We have developed a simple method to test for racial bias in false alarm rate by checking one point: the case when the FA_B is minimized and the FA_W is maximized. This is a test of the first possible point of overlap. If $\min(FA_B) \leq \max(FA_W)$, then there exists at least one pair of x -values such that racial bias is not present.

The minimum FA_B occurs when $x_B = 0$. Substituting $x_B = 0$ into Eq 5. we obtain the minimum false alarm rate

$$\min(FA_B) = \frac{b_B}{\bar{S}_B + b_B}. \tag{8}$$

The maximum hit rate for White drivers serves as an upper bound on FA_W .

$$\max(FA_W) = \frac{b_W}{S_W - x_W + b_W} = \frac{a_W}{a_W + x_W}. \tag{9}$$

We substitute the upper bound for x_W from Eq. 6 in for x_W and simple algebra yields,

$$\max(FA_W) = \frac{S_W}{S_W + \bar{S}_W}. \tag{10}$$

That is, the highest false alarm rate will occur when $FA_W = p(S_W)$.

Therefore, the test of the overlapping condition is a police department has some combination of x -values that yield no racial bias when,

$$\max(FA_W) = \frac{S_W}{S_W + \bar{S}_W} \geq \frac{b_B}{\bar{S}_B + b_B} = \min(FA_B). \tag{11}$$

The overlapping condition test states: If $\max(FA_W) \geq \min(FA_B)$ then there exists some combination of x_W and x_B such that $FA_B \leq FA_W$. That is, there exists at least one pair of x -values where there is no racial bias. A department passes the overlapping condition test when there is some combination of x -values that overlap and they fail the test when there

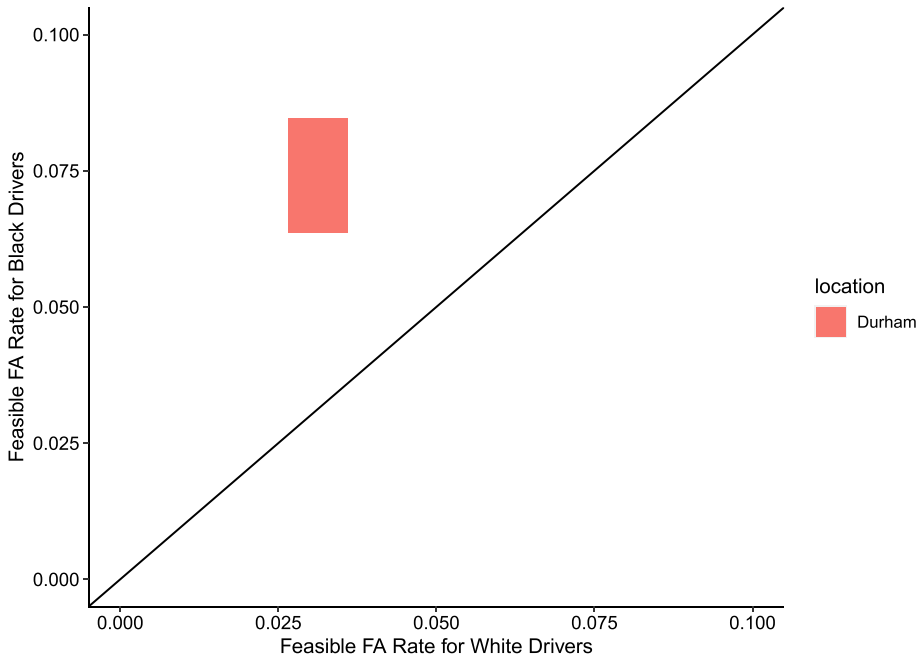


Fig. 2 Durham, NC data (Table 4): graphical representation of the overlapping condition Test

is no overlap (i.e., are racially biased for all feasible 2×2 tables).⁹ In other words, this test adjudicates whether or not a given department's data gives rise to racial bias under all rates of contraband.

We represent the overlapping condition test graphically (see Fig. 2). We plot feasible false alarm rates for Black drivers on the vertical axis and feasible false alarm rates for and White drivers on the horizontal axis. When these values are equal, the feasible space will overlap the identity line. Points above the identity line indicate higher false alarm rates for Black drivers and points below the identity line, higher false alarm rates for White drivers. The further the rectangle is from the identity line the greater the difference in false alarm rate by race. The point of the rectangle plot defined by the overlapping condition test on FA_W and FA_B is the bottom right corner (closest to the identity line). This corner will be the first to cross the identity line when there are combinations of feasible x -values that yield no-bias. Figure 2 provides a quick visual representation of the overlapping condition test.

As shown in Fig. 2, the bottom right corner does not overlap the identity line: Durham, NC fails the overlapping condition test. That is, for combinations of feasible values of x for White and Black drivers there is no overlap in the false alarm rate for White and Black

⁹ In a sense this “worst case scenario” is like a null hypothesis test, where in states that fail the overlapping condition test we reject the null hypothesis that there is no bias, but in states that pass the overlapping condition test we do not have enough evidence to reject the null that there is no bias. We are not performing a test of significance, but we include this intuition to help the reader.

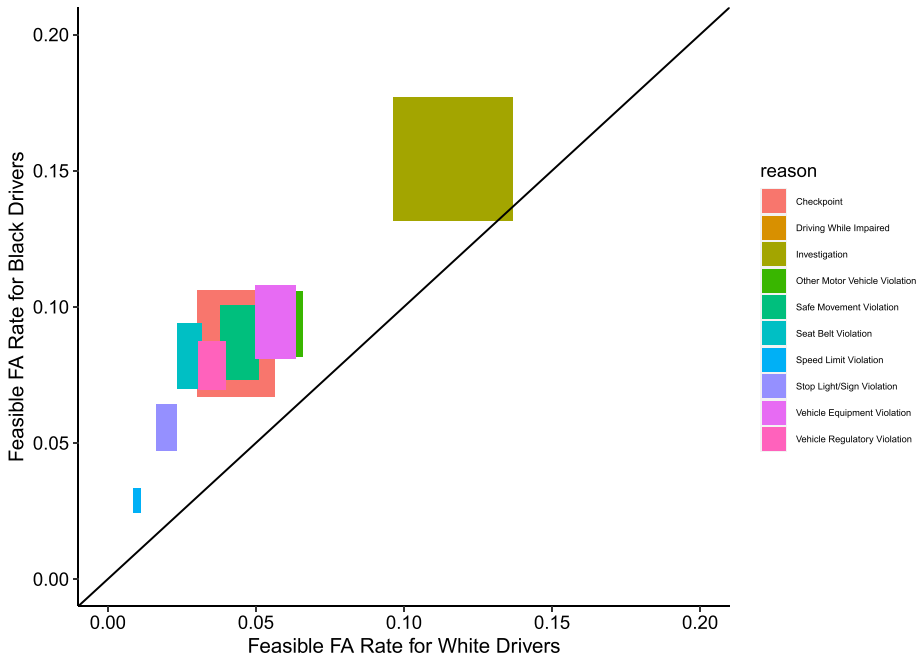


Fig. 3 Durham, NC data: overlapping condition test for stop reason

drivers. We will utilize this analytical tool to further understand the nature of racial bias in Durham.

Examining if Reasons for Stop and Search Moderates the False Alarm Finding

Police officers in Durham, NC fail the overlapping condition test. However, it is possible that officers may exhibit different false alarm rates under different reasons for a stop or a search due to a difference in the reason they conducted a search. Work by Baumgartner et al. (2018) suggests that investigatory stops show greater discrimination because officers have more power to discriminate with their greater discretion (this was found using the *PPV* metric of analyzing racial bias). Thus, in the following analysis we utilize the overlapping condition test to examine if the reason for a stop or search moderates the racial difference in false alarm rate.

Reason for Stop

In Durham, police officers recorded 10 different reasons why they stopped drivers, from legitimate traffic violations such as stop light/stop sign violations to pretextual reasons such as “other movement violation.” We disaggregated the data by stop reason, creating a separate 2 x 2 table for each reason, and observed that Durham fails the overlapping

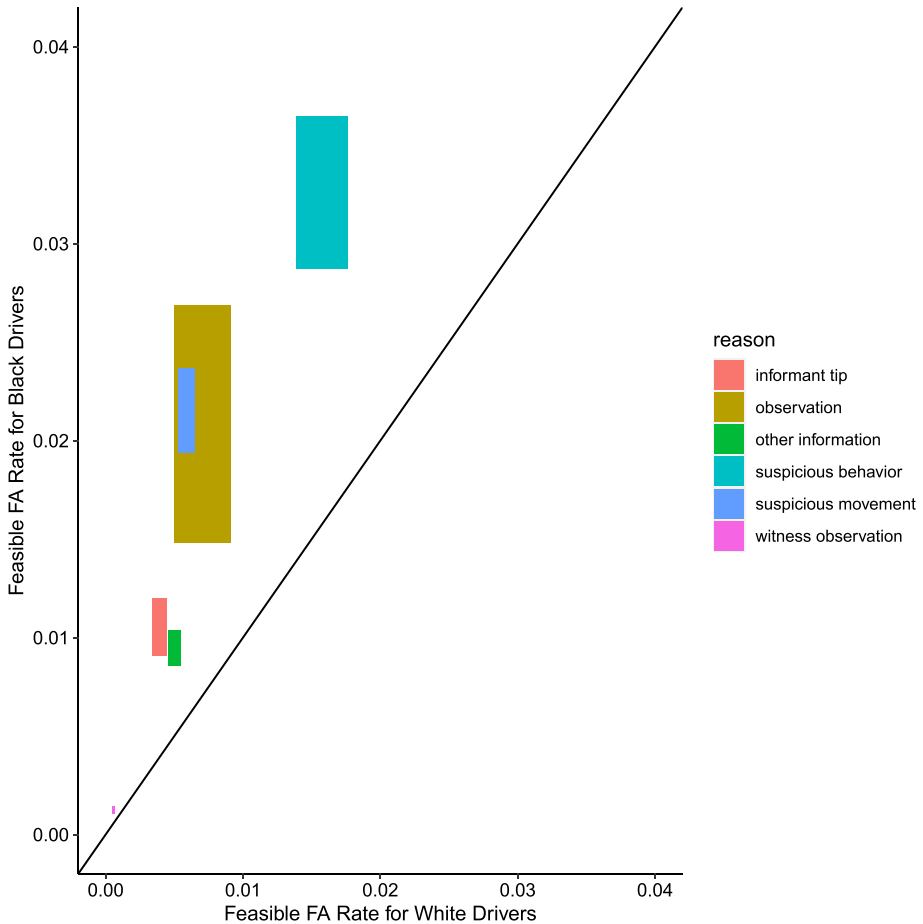


Fig. 4 Durham, NC data: overlapping condition test for search reason

condition test for all stop reasons except investigatory stops. Figure 3 represents this finding graphically.

As shown in Fig. 3, only the investigatory stop rectangle overlaps the identity line. This means that the reason for stop does not attenuate the racial bias in false alarm rate except for stops for “investigations.” Investigatory stops pass the overlapping condition test because some combinations of feasible x -values yields no bias. This suggests that without further information we cannot make strong claims about a lack of racial bias in investigatory stops because we do not know the true false alarm rate value.¹⁰

¹⁰ As mentioned elsewhere in this paper, these results are contingent upon stops that did occur and does not address the space of stops that could have happened but did not occur.

Reason for Search

We also use the overlapping condition test to examine if the reason for search moderates the racial difference in false alarm rate in Durham. We are unable to make separate 2 x 2 tables in the same way we have previously because there is no “no search” for a particular search reason. We do not have data on the number of times an officer could have, but decided not to, search for a specific reason. To work around this, we assume that every person could have been searched for every reason. We set the total amount of no searches for a given reason equal to the total amount of no-searches. The upper bound (Eq. 6) is the number of no searches multiplied by the proportion of searches that yielded a hit. By assuming the number of no-searches as the largest possible value we will create the largest possible feasible region for x , a conservative test of our condition. As shown in Fig. 4, according to the overlapping condition test there is racial bias in false alarm rate for all such reasons.¹¹ Overall, these results suggest that whatever is driving racial bias in police officer searches it is pervasive across different reasons for stop and search.

Analysis of 98 Million Traffic Stops Across 14 States

To examine the generalizability of these results, we analyzed data from 14 state police departments from the years 1999–2017, which amounted to approximately 98 million traffic stops.¹² We included all state patrols with search data available through the Stanford Open Policing database (Pierson et al. 2020) in our analyses. The results are presented below (see Supplemental Materials for full extent of our analyses).

We observe that in the aggregate these police departments fail the overlapping condition test. False alarm rates for White drivers do not overlap false alarm rates for Black drivers $\max(FA_W) = .027 < \min(FA_B) = .034$. However, as shown in Fig. 5, there is heterogeneity in the degree of racial bias observed in different states. We observe that 10 of our 14 states fail the overlapping condition test (these rectangles do not intersect the identity line), while four states (California, Texas, South Carolina, and Massachusetts) pass the overlapping condition test.¹³

We expect variation in the degree of racial bias between different states. Different departments have different policies for searching drivers, resulting in different patterns of the hit and false alarm rates. Moreover, different departments have different data reporting standards, further influencing the hit and false alarm rates (Department of Justice 2009).

¹¹ One might assume that police officers engaging in plain view searches (coded observation here) would not show bias in false alarm rate because the contraband is visible. Instead, we observe a racial bias in false alarm rate for this search reason. These feasible false alarm rates are admittedly low, however they are even lower for White drivers than Black drivers. We speculate this is related to the effect observed in Eberhardt et al. (2004), whereby officers are more likely to misidentify an object as contraband when interacting with a Black driver. This would explain the higher false alarm rate for Black drivers in this context.

¹² Selection criteria for our analyses was State Patrol Data that included information about the search (and its respective outcome). There are other states that have data from different agencies (e.g., NYPD versus New York State Police). To keep the type of department consistent, we focused on State Police data. Moreover, not all departments had data available during the entirety of this time period.

¹³ Additional research can (and should) estimate whether the actual x -value falls in that corner where the box overlaps. This is especially important because the region of overlap is relatively small compared to the area of feasible values (see Texas for example).

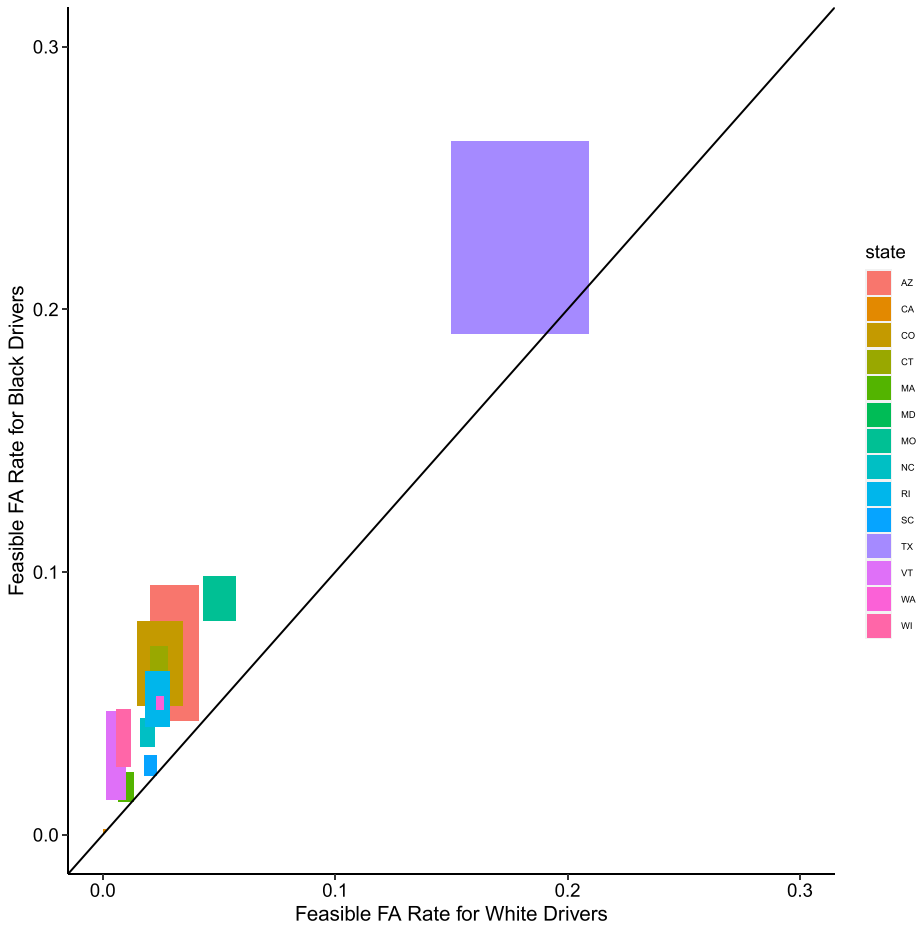


Fig. 5 US state patrol data: overlapping condition test for 14 US states

Future work should examine the differences in these department policies to determine what separates states with little racial bias from states with great racial bias in false alarm rates. For states that have further local municipal department data available, we conducted the same analysis and the same pattern of results held. That is, 9 out of 11 municipalities failed the Overlapping Condition test (see Supplemental materials).

A Psychological Analysis of Trade-offs

Now that we have described the general pattern of the data at the state and local level, we consider potential explanations for this behavior. In an ideal world, officers would maximize their hit rate and minimize their false alarm rate simultaneously. However, it is not possible to satisfy a perfect hit and false alarm rate simultaneously; officers must either pick a metric (i.e., hit or false alarm rate) to optimize, or engage in a trade-off between these metrics (see, e.g., Green and Swets 1966). A police officer motivated to find contraband

every time it is present (i.e., maximize hits) should search every single person. An officer motivated to never search an innocent person (i.e., minimize false alarms), should never search. We do not observe either of these patterns in our data. This suggests that police officers are optimizing a different objective function, possibly one that relies on a trade-off between hit and false alarm rates.

Signal detection theory (Green and Swets 1966) proposes one set of assumptions that leads to a trade-off model that imposes a decision threshold on the standard likelihood ratio. This decision threshold can be related to the probability of contraband, as well as utility terms corresponding to each of the four cells of the 2 x 2 table: For a rigorous construction of this model see Green and Swets (1966), Coombs et al. (1970) and Macmillan and Creelman (2005).

Police departments that fail the overlapping condition test have a false alarm rate bias. According to signal detection theory, we can describe this bias through a threshold parameter β . Specifically, a failure of the overlapping test can be modeled through a difference in the value the likelihood ratio must exceed before an officer searches a driver, such that it is easier for a police officer's likelihood ratio to exceed the threshold β for a Black driver than the corresponding β for a White driver, for any value of x . This means that a police officer needs less evidence to search a Black driver than a White driver.

We are not the first to consider a threshold measure as a way to model biased policing. Notably, Simoiu et al. (2017) developed a threshold test, later improved in Pierson et al. (2018), to examine police officer performance independent of the problems posed by infra-marginality. Our model differs from their approach. Their approach requires a statistical model with parameter estimates defined over the 2 x 2 with an unobserved row, whereas our approach examines all feasible scenarios for the unobserved data. We are using mathematical constraints to study the feasible space over the unobserved data, but our model is currently limited in its ability to capture heterogeneity across officers and departments in a statistical framework. Simoiu et al. (2017) and subsequent models of police bias have implemented heterogeneity in officers and precincts through a random effects model in a Bayesian context.

Current Models of Racial Bias in the Literature

Considering all feasible real world scenarios leads to new approaches to analyze racial bias in policing such as the overlapping condition test. We can also use the approach of considering all feasible real world scenarios to highlight debates surrounding current models of racial bias in policing. Current measures of bias in the policing literature focus on the observed row of the 2 x 2 table involving searches. This is both a consequence of the data available to researchers (Knox and Mummolo 2020b; Goff and Kahn 2012) and Bayes Theorem. We first briefly review the prominent methods and then consider how the analysis of all feasible 2 x 2 tables provides new insight into their limitations.

The Benchmark Test

The benchmark test uses the racial difference in search rates to assess discrimination. Scholars using the benchmark test define discrimination as police officers searching Black and White drivers at different rates (i.e., $p_B(S) \neq p_W(S)$), and by implication, no racial bias exists if the two population search rates are equal. However, equal search rates might miss

a serious racial bias caused by an underlying difference in base rates of contraband, a phenomenon referred to as statistical discrimination (Knowles et al. 2001; Becker 1971).

There have been notable advancements in the Benchmark test that include examining the probability of being searched as a function of the driver's race and other characteristics observable by the police officer in a logistic regression framework where these characteristics could be treated as control variables, or correlates. When race has no explanatory power, then no discrimination is present (Smith et al. 2004; Fagan and Davies 2001; Gelman et al. 2007; Engel and Calnon 2004). Due to the difficulty in identifying all observable characteristics this model has been adapted to consider propensity scores to match Black and White drivers on key observable characteristics, such as location (Ridgeway 2006). Another special case of the benchmark test is the "veil of darkness test" (Grogger and Ridgeway 2006; Pierson et al. 2020), which assumes that if police officers are not relying on race for their search decisions, then the proportion of drivers stopped that are Black should not change between light and dark hours (relatively easy vs. hard conditions to determine race). Ultimately, the Benchmark test can serve as a first approximation of racial bias, but in order to understand the full scope of police officer search decisions, researchers must also consider metrics that incorporate officer accuracy.

The Outcome Test

Some scholars recognized the limitations of the benchmark test and designed a different test that would begin to address concerns about police officer accuracy, the outcome test (Knowles et al. 2001; Ayres 2002; Baumgartner et al. 2018). In their paper, Knowles et al. (2001) developed a model of racial bias that differentiates between statistical discrimination (Arrow 1971) and a taste for discrimination (Becker 1971). In the KPT framework, officers have a taste for discrimination when the utility for searching a Black driver differs from that of searching a White driver. According to Knowles et al. (2001), police officers demonstrate statistical discrimination when Black and White drivers are searched at different rates and the observable characteristics of the drivers are identical. The KPT model recognizes the need for equal rates of contraband, but instead hinges on the key assumption that police officers and criminals will reach an equilibrium whereby the rates of contraband in the population should be equal for White and Black drivers. Other scholars have drawn attention to issues of monolithic police behavior and rational decision making with this model that question the possibility of an equilibrium (Anwar and Fang 2006; Engel 2008; Engel and Calnon 2004; Persico and Todd 2006, 2008; Engel and Tillyer 2008; Manski 2006).

The outcome test defines discrimination as police officers having different "hit rates" for White and Black drivers. Recall the "hit rate" they are describing is the PPV (Eq. 2) value on the 2×2 table. This formulation of the outcome test is related to the overlapping condition test, except instead of examining false alarm rates, the metric of interest is the PPV. The outcome test defines discrimination as when

$$\frac{p_B(S \cap C)}{p_B(S)} \neq \frac{p_W(S \cap C)}{p_W(S)}. \quad (12)$$

The outcome test relies on the conditional probability $p(C|S)$ computed with the entries from the single observed row of the 2×2 table. An equivalent computational approach for this conditional probability uses Bayes theorem and requires the complete 2×2 table. The posterior probability $p(C|S)$ can be computed through the standard odds form of Bayesian

Table 5 Example data: Benchmark test

Black Drivers			
	Contraband	No contraband	Total
Search	40	5	45
No search	x	$55 - x$	55
Total	$40 + x$	$60 - x$	100
White drivers			
	Contraband	No contraband	Total
Search	4	23	27
No search	x	$73 - x$	73
Total	$4 + x$	$96 - x$	100

updating, where the likelihood ratio (i.e., hit rate divided by false alarm rate) is multiplied by the contraband odds in the population. Thus, the outcome test is equivalent to the difference between the posterior probability for White and Black drivers. Any critique (or positive attribute) of the outcome test also applies to this particular Bayesian formulation of posterior probability.

The Problem of Inframarginality

The outcome test and its derivative methods neglect any confounding reason why a police officer would be more accurate when searching one race than others. Most importantly, these tests neglect the base rates of contraband that exist between Black and White drivers that might drive differences in accuracy. This problem is commonly known as the problem of inframarginality (Ayres 2002). This limitation has been discussed at length in the literature (Ayres 2002; Anwar and Fang 2006; Pierson et al. 2018; Knox and Mummolo 2020b).

Another way to think about this limitation is that it underscores the need to consider all four cells of the 2 x 2 table. We need to examine the instance where officers did not search because those who were guilty and not searched (i.e., misses) represent an important component of the situation. It is not sufficient to consider the accuracy of searches that occurred if we want to know how accurate an officer is when they search drivers in general. We also need to know how well an officer performs when they conclude a person does not have contraband so does not initiate a search. The underlying rates of contraband are directly relevant to understanding the true nature of police officer bias, and claims of causality cannot be made without the full information in the 2 x 2 table (see, Knox and Mummolo 2020, 2020b; Knox et al. 2020, for a review of these limitations).

Insights from the Full 2 x 2 Table

We combine some of these common debates in the literature, such as the inframarginality problem and the influence of base contraband rates, under one common framework if we consider all feasible scenarios. Using three-dimensional graphs we demonstrate that different values of contraband lead to different conclusions of bias. Consider the following hypothetical data set:

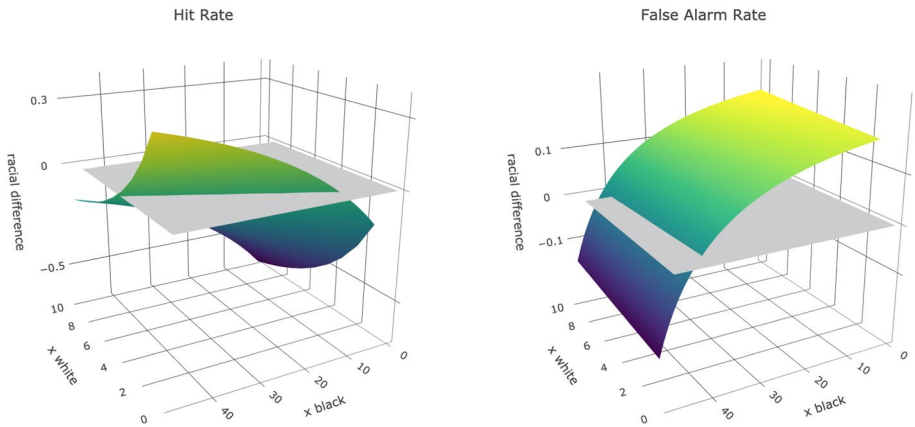


Fig. 6 Hypothetical data in Table 5: Racial difference in hit rate (left) and false alarm rate (right) for all possible combinations of x -values

For these data, the benchmark test would state that we are observing racial bias, $p_B(S) = .45 \neq .27 = p_W(S)$, and that police officers are searching more Black drivers than White drivers. On the other hand, the outcome test (Eq. 12) suggests an anti-white bias. That is, police are more accurate when searching Black drivers $PPV_B = .89 > .15 = PPV_W$.

So which test is correct? Should we say that police officers are searching more Black drivers than White drivers, so they are racially biased? Or, are officers unbiased because they find more contraband in Black drivers who were searched than White drivers who were searched?

If we examine the racial difference surface (Fig. 6) generated by considering all feasible x -values ($x_{black} \in [0, 48]$ and $x_{white} \in [0, 10]$), we observe a subset of stops where the hit rate for Black drivers is less than the hit rate for White drivers and a subset of stops where the false alarm rate for Black drivers exceeds that for White drivers. In other words, the outcome test (under its basic assumptions) cannot work around the problem of infra-marginality; while it gives a more nuanced definition of bias than the benchmark test, the outcome test fails to capture potential key patterns in the data across different contraband rates.¹⁴

Discussion

Economists, psychologists, sociologist, lawyers, and computer scientists have expressed growing interest and concern for racial bias in police searches. Researchers typically frame the issue of police search bias as a decision problem. The application of decision theory

¹⁴ If we apply the outcome test to the data from Durham, North Carolina we observe a $PPV_B = .273$ which is almost identical to the $PPV_W = .272$. The outcome test would still conclude a racial bias in the data, which is consistent with the racial bias we observe in the false alarm rate using the Overlapping Condition Test. Notably, the benchmark test would have also picked up the discrepancy of higher search rates for Black drivers ($p_B(S) = .085$) than White drivers ($p_W(S) = .038$).

takes different forms from game theoretic decisions (Knowles et al. 2001), to causal inferences (Knox and Mummolo 2020b), and even distributions of signal and noise akin to signal detection theory (Pierson et al. 2018). Despite these different applications of decision theory, researchers have argued for a common framework to study racial bias in traffic stop data (Knox and Mummolo 2020b; Goff and Kahn 2012). While there are different approaches to modeling bias from policing data, a growing body of work continues to find that police officers exhibit racial bias against Black drivers. Several models interpret this bias in terms of police officers setting different thresholds for search (Simoiu et al. 2017; Pierson et al. 2018). Yet we still do not understand the psychology of how police officers set their thresholds.

How do the Police Set a Threshold?

One method of exploring how police set thresholds is to consider an ideal (unbiased) decision maker and the potential limitations that prevent an officer from achieving such a feat. In order to search without bias, police officers must consider the base rates of contraband in the population of White and Black drivers and search accordingly. This is no easy task. A body of research has shown that people ignore, or under weight, base rates when other stereotype-consistent information is present (Tversky and Kahneman 1974). That is, when estimating the likelihood that a person has contraband, the stereotype of the black criminal (Eberhardt et al. 2004; Entman 1992) may overpower base rate information. This is especially likely for police officers who make these split-second decisions. Importantly, this psychological description of the challenges of searching does not excuse the racial discrepancies in police officer search behavior, but rather offers researchers an opportunity to provide new interventions that account for this psychology.

While an ideal decision maker provides a first approximation, we need to define the officer's objective function. Scholars have argued that the goal of an officer is to maximize hits (Knowles et al. 2001), or minimize crime (Dominitz and Knowles 2006), or do something else altogether (Durlauf 2006). Carefully modeling officer thresholds will not only advance conclusions derived from these modeling techniques, but also advance methods where the key debate centers around the police officer's objective function. Recent mathematical results (e.g., Kleinberg et al. 2017) suggest important constraints on permissible solutions to tradeoffs over multiple objectives (i.e., permissible in the sense of not creating inconsistencies with desired properties such as calibration of predicted hit and false alarm rates). This is a promising area of research that can be combined with the present model of examining the feasible space of 2 x 2 tables.

Beyond tackling the officer's objective function, we need more direct tests of the proposed threshold models. For example, the threshold model used by Simoiu et al. (2017) is related to a class of models (see, e.g., Coombs et al. 1970) that imposes a particular constraint on the ROC curve (the trade-off curve between the hit rate and the false alarm rate) that differs from the pattern implied by signal detection theory. Evaluation of ROC curves at different levels of analysis (e.g., officer-level, department-level, state-level) could provide insight into underlying mechanisms involving thresholds.

Suggestions for Intervention

We found that police officers in the aggregate have a lower threshold for searching Black drivers than White drivers for all feasible real world scenarios in 10 out of 14 US states,

and 9 out of 11 local departments. We interpret this as a threshold difference, where police officers are more eager to identify an object as contraband when they are interacting with a Black driver as compared to a White driver. This suggests that we should design interventions that move a police officer's threshold for search such that they react less to small indications of signal. These interventions need not focus on the ability of an officer to detect contraband as this may adversely change the hit rate.¹⁵ Again, understanding the officer's objective function is key, as an officer who never searches will effectively minimize their false alarm rates (reducing their racial bias to zero), but they may no longer meet other objectives if they do not engage in any searches. Future work must collaborate with police departments on their desired objective functions and how interventions might facilitate these objectives while simultaneously reducing bias.

The Importance of Studying False Alarm Rates

The current literature on racial differences in search rate (benchmark test) or PPV values (outcome test) has ignored an important aspect of traffic stops, the false alarm rate. The false alarm rate captures how frequently the police inconvenience an innocent person instead of targeting an appropriately guilty person (hit rate). The consequences of targeting innocent people have had serious lasting impacts on communities. The targeting of innocent Black people in particular has led to decreased community trust in policing as well as several negative health outcomes (Del Toro et al. 2019; Legewie and Fagan 2019).

Different burdens of proof (e.g., probable cause versus reasonable suspicion) could contribute to different false alarm rates and hit rates. Nevertheless, police officers should be equally wrong (or right) about searching citizens under the same burden of proof (i.e., reason for stop/search). Our results suggest that police officers under the same conditions for stop and search incorrectly target more innocent Black drivers than innocent White drivers.

Limitations

Revisiting the Problem of Inframarginality

The overlapping condition test and other methods of considering all feasible scenarios are not immune to problems of inframarginality. With large samples at the state and county level, the observed data provide sufficient constraints on the feasible x -values to be able to apply the overlapping condition test on the false alarm rates across race. However, interventions will typically be at the individual officer level, and with smaller sample sizes at that level, the problem of inframarginality returns because the feasible values for false alarm rates will span more of the $[0, 1]$ interval. Similar to hit rates in the present work, with smaller sample sizes at the officer level, one may not be able to make claims about bias (or lack thereof) without addressing the problem of inframarginality. Future work needs to develop estimation methods for the unknown parts of the 2×2 table in order to evaluate the success of officer-level interventions.

¹⁵ For such interventions, one might consider turning to the SDT literature for insights on how to change the threshold (false alarm rate) without affecting performance (hit rate).

Reporting Bias

A developing area of research suggests that police officers systematically misreport driver race. In the case of Texas State Patrol officers, troopers mislabel minority searches as White drivers (Luh 2022). It is for this reason we exclude Hispanic drivers from our analysis, as they are 2.3 times more likely to be misreported as White when the search result in a false alarm (Luh 2022). While this reporting problem is indicative of its own form of bias, it does not change our conclusions of anti-Black bias. The misreporting of false alarm stops increases the numerator of the White false alarm fraction, making it less likely that we could conclude bias in all feasible scenarios. Despite this, of the 25 police departments we analyzed, 19 of them demonstrated racial bias in police search decisions. The few exceptions occur in four of the 14 states and two local municipalities (see the Supplemental materials) for extreme x pairs that yield small overlap in false alarm rates. That is, feasible x values for Whites near the upper bound where hit rates are near chance performance paired with feasible x values at the other extreme for Blacks where the Black hit rate exceeds 90%. These pairs of x 's are unlikely to occur in real data, but are in the subset of feasible 2×2 tables we consider. While reporting bias does not change our conclusions, it has serious implications for researchers building better predictive models of bias. Our work, and subsequent conclusions, are limited by our ability to detect and model these forms of reporting bias.

Conclusion

The present work establishes a method of examining threshold bias across feasible real world scenarios without additional estimation of unknown data and without further assumptions beyond the constraint on the relation between hit rate and false alarm rate. Using the Overlapping Condition Test we conclude that in the aggregate police officers demonstrate racial bias in their false alarm rates. This is indicated by higher false alarm rates for Black drivers across feasible values of the unobserved data. We interpret this false alarm difference as a difference in officer search criteria (i.e., threshold). These results are consistent with other models of racial bias in the literature that use different decision theory models to capture racial bias in policing. Consequently, interventions need to train officers on how to set thresholds for Black and White drivers to achieve equal false alarm rates. Importantly, for such interventions to be effective, future work should assess how officers set their search thresholds so that interventions designed to eliminate racial bias can target those processes.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10940-024-09585-4>.

Acknowledgements We thank Sam Gross, Phoebe Ellsworth and Amie Gordon for helpful comments on previous versions.

Author Contributions MAM contributed to the idea development, mathematical derivations, and manuscript writing. RG contributed to idea development, mathematical derivations, and provided feedback on manuscript drafts.

Funding Not applicable

Availability of data and materials We used data made publicly available from the Stanford Open Policing Project. The full database can be found at this link: <https://openpolicing.stanford.edu/data/>. Our cleaned data is uploaded to a git repository and will be made available upon acceptance.

Code availability Our code is available on a git repository: <https://github.com/margem/detecting-bias>.

Declarations

Conflicts of interest We declare no known conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anwar S, Fang H (2006) An alternative test of racial prejudice in motor vehicle searches: theory and evidence. *Am Econ Rev* 96(1):127–151
- Arrow K (1971) The theory of discrimination (working paper No. 403). Princeton University, Department of Economics, Industrial Relations Section, NJ, USA
- Ayres I (2002) Outcome Tests of Racial Disparities in Police Practices. *Justice Res Policy* 4(1–2):131–142
- Baumgartner F, Epp D, Shoub K (2018) Suspect citizens: What 20 million traffic stops tell us about policing and race. Cambridge University Press, Cambridge
- Becker GS (1971) The economics of discrimination, 2nd edn. University of Chicago Press, Chicago
- Coombs CH, Dawes RM, Tversky A (1970) Mathematical psychology: an elementary introduction. Prentice-Hall, Hoboken
- Del Toro J, Lloyd T, Buchanan KS, Robins SJ, Bencharit LZ, Smiedt MG, Goff PA (2019) The criminogenic and psychological effects of police stops on adolescent black and Latino boys. *Proc Natl Acad Sci* 116(17):8261–8268
- Department of Justice (2009). Ensuring the quality, credibility, and relevance of U.S. Justice statistics. National Academies Press, Washington, D.C. (P: 12671) <https://doi.org/10.17226/12671>
- Dominitz J, Knowles J (2006) Crime minimisation and racial bias: What can we learn from police search data? *Econ J* 116(515):F368–F384
- Durlauf SN (2006) Assessing racial profiling. *Econ J* 116(515):F402–F426
- Eberhardt JL, Goff PA, Purdie VJ, Davies PG (2004) Seeing black: race, crime, and visual processing. *J Pers Soc Psychol* 87:876–893
- Engel RS (2008) A critique of the outcome test in racial profiling research. *Justice Q* 25(1):1–36
- Engel RS, Calnon JM (2004) Examining the influence of drivers' characteristics during traffic stops with police: results from a national survey. *Justice Q* 21(1):49–90
- Engel RS, Tillyer R (2008) Searching for equilibrium: the tenuous nature of the outcome test. *Justice Q* 25(1):54–71
- Entman RM (1992) Blacks in the news: television, modern racism and cultural change. *J Q* 89(2):341–361
- Fagan J, Davies G (2001) Street stops and broken windows: terry, race and disorder in New York City, SSRN Electron J
- Gelman A, Fagan J, Kiss A (2007) An analysis of the New York city police department's "stop-and-frisk" policy in the context of claims of racial bias. *J Am Stat Assoc* 102(479):813–823

- Goff PA, Kahn KB (2012) Racial bias in policing: why we know less than we should: policing racism. *Soc Issues Policy Rev* 6(1):177–210
- Green DM, Swets JA (1966) *Signal detection theory and psychophysics*. Wiley, Hoboken
- Grogger J, Ridgeway G (2006) Testing for racial profiling in traffic stops from behind a veil of darkness. *J Am Stat Assoc* 101(475):878–887
- Gross SR, Barnes KY (2002) Road work: racial profiling and drug interdiction on the highway. *Michigan Law Review*, 101(3)
- Groves R, Cork D (Eds.) (2009) *Ensuring the quality, credibility, and relevance of U.S. justice statistics*. National Academies Press, Washington, D.C. (P 12671) <https://doi.org/10.17226/12671>
- Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. *Innov Theor Comput Sci Conf* 43(1–43):23
- Knowles J, Persico N, Todd P (2001) Racial bias in motor vehicle searches: theory and evidence. *J Polit Econ* 109(1):203–229
- Knox D, Lowe W, Mummolo J (2020) Administrative Records Mask Racially Biased Policing. *Am Polit Sci Rev* 114(3):619–637
- Knox D, Mummolo J (2020) Making inferences about racial disparities in police violence. *Proc Natl Acad Sci* 117(3):1261–1262
- Knox D, Mummolo J (2020b) Toward a general causal framework for the study of racial bias in policing. *J Polit Inst Polit Econ*, 1
- Kocieniewski D (1999) Trenton charges 2 troopers with falsifying drivers' race. *The New York Times*, New York
- Lalkhen AG, McCluskey A (2008) Clinical tests: sensitivity and specificity. *Continuing education in Anaesthesia critical care & pain* 8(6):221–223
- Legewie J, Fagan J (2019) Aggressive policing and the educational performance of minority youth. *Am Sociol Rev* 84(2):220–247
- Luh E (2022) Not so black and white: uncovering racial bias through systematically misreported trooper reports (working paper). University of Michigan. University of Michigan Criminal Justice Administrative Records System, Ann Arbor, MI.
- Lundberg A (2022) Statistical power and search intensity bias in hit rates tests of discrimination. *J Quant Criminol* 38(4):979–1002
- Macmillan NA, Creelman CD (2005) *Detection theory: a user's guide*, 2nd edn. Lawrence Erlbaum Associates Publishers, Mahwah
- Manski CF (2006) Profiling: introduction to the feature. *Econ J* 116(515):F347–F350
- Persico N, Todd P (2006) Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita. *Econ J* 116(515):F351–F367
- Persico N, Todd PE (2008) The Hit rates test for racial bias in motor-vehicle searches. *Justice Q* 25(1):37–53
- Pierson E, Corbett-Davies S, Goel S (2018) Fast threshold tests for detecting discrimination. *proceedings of the twenty-first international conference on artificial intelligence and statistics* (pp. 96–105). PMLR
- Pierson E, Simoiu C, Overgoor J, Corbett-Davies S, Jenson D, Shoemaker A, Goel S (2020) A large-scale analysis of racial disparities in police stops across the United States. *Nat Hum Behav* 4(7):736–745
- Ridgeway G (2006) Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J Quant Criminol* 22(1):1–29
- Shea J (2022) Testing for racial bias in police traffic searches. University of Illinois Champaign, Urbana
- Simoiu C, Corbett-Davies S, Goel S (2017) The problem of infra-marginality in outcome tests for discrimination. *Ann Appl Stat* 11(3):1193–1216
- Smith WR, Tomaskovic-Devey D, Zingraff MT, Mason HM, Warren PY, Wright CP (2004) *The North Carolina highway traffic study* (grant report No. 204021). North Carolina State University, Raleigh North Carolina, USA: National Institute of Justice US Department of Justice
- Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. *Science* 185(4157):1124–1131
- Walker S (2001) Searching for the denominator: problems with police traffic stop data and an early warning system solution. *Justice Res Policy* 3(1):63–95

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.